Development and application of bioinformatic tools for the representation and analysis of genetic diversity

THE COVER

The cover illustrates artistically the kind of data and methods used in this thesis, as well as displays screen shots of the PDA and the DPDB websites that have been created.

Kindly designed by Miquel Ràmia.

Development and application of bioinformatic tools for the representation and analysis of genetic diversity

Desenvolupament i aplicació d'eines bioinformàtiques per a la representació i anàlisi de la diversitat genètica

Desarrollo y aplicación de herramientas bioinformáticas para la representación y análisis de la diversidad genética

DOCTORAL THESIS

Sònia Casillas Viladerrams

Universitat Autònoma de Barcelona · Facultat de Biociències · Departament de Genètica i de Microbiologia

Bellaterra, 2007

Memòria presentada per la Llicenciada en Biologia Sònia Casillas Viladerrams per a optar al grau de Doctora en Ciències Biològiques.

Bellaterra, a 26 d'octubre del 2007

El Doctor Antonio Barbadilla Prados, Professor titular del Departament de Genètica i de Microbiologia de la Facultat de Biociències de la Universitat Autònoma de Barcelona, i el Doctor Alfredo Ruiz Panadero, Catedràtic del mateix departament,

CERTIFIQUEN: que la Sònia Casillas Viladerrams ha dut a terme sota la seva direcció el treball de recerca realitzat en el Departament de Genètica i de Microbiologia de la Facultat de Biociències de la Universitat Autònoma de Barcelona que ha portat a l'elaboració d'aquesta Tesi Doctoral, titulada *"Desenvolupament i aplicació d'eines bioinformàtiques per a la representació i anàlisi de la diversitat genètica"*.

I perquè consti als efectes oportuns, signen el present certificat a Bellaterra, a 26 d'octubre del 2007.

Dr. Antonio Barbadilla Prados

Dr. Alfredo Ruiz Panadero

— Al Jordi, pel teu amor i suport incondicional

— Als pares, per haver-me guiat i aconsellat sempre

— A la iaia, pel teu inesgotable optimisme

— A la Georgina, la Berta i tota la resta de família i amics, per les estones tant agradables compartides i perquè el temps que he dedicat a aquesta tesi no l'he pogut dedicar a vosaltres

Table of Contents

Preface

Genetic variation is the cornerstone of biological evolution. The description and explanation of the forces controlling genetic variation within and between populations is the main goal of population genetics (LEWONTIN 2002). The deciphering of an explosive number of nucleotide sequences in different genes and species has changed radically the scope of population genetics, transforming it from an empirically insufficient science into a powerfully explanatory interdisciplinary endeavor, where high-throughput instruments generating new sequence data are integrated with bioinformatic tools for data mining and management, and advanced theoretical and statistical models for data interpretation.

This thesis is an integrative and comprehensive bioinformatics and population genetics project whose central topic is the genetic diversity of populations. It is accomplished in three sequential steps: (i) the development of tools for data mining, processing, filtering and quality checking of raw data, (ii) the generation of databases of knowledge from refined data obtained in the first step, and (iii) the testing of hypotheses that require multi-species and/or multi-locus data. In the first part of the thesis, we have developed PDA —Pipeline Diversity Analysis—, an open-source, web-based tool that allows the exploration of polymorphism in large datasets of heterogeneous DNA sequences. This tool feeds from the millions of haplotypic sequences from individual studies that are stored in the main molecular biology databases, and generates high-quality, population genetics data that can be used to describe patterns of nucleotide variation in any species or gene. All the extracted and analyzed data resulting from the first part of this thesis is used in the second step to create a comprehensive on-line resource that provides searchable collections of polymorphic sequences with their associated diversity measures in the genus *Drosophila* (DPDB —Drosophila Polymorphism Database—). This resource means an ambitious pledge to test the efficiency of the system created in the first part.

Finally, two different studies that make use of the modules of data mining and analysis developed are shown. First, we study patterns of sequence variation to infer constraint and adaptation in *Drosophila* conserved noncoding sequences (CNSs). For this study we have used population genetics re-sequencing data from *D. melanogaster* together with comparative genomic data from other *Drosophila* species. We show that patterns of nucleotide sequence evolution in *Drosophila* CNSs are incompatible with the notion that mutational cold-spots explain these conserved blocks. Rather, the results support the hypothesis that CNSs are maintained by the action of purifying selection. The second study focuses on the coding evolution of *Hox* genes, a class of essential transcription factors expressed early in development that are involved in the specification of regional identities along the anteroposterior body axis. We have measured the rates of nucleotide divergence and fixation of insertions and deletions of three *Hox* genes, and compared them with those of three *Hox*-derived genes and a set non-*Hox* genes to test the hypothesis that *Hox* genes evolve slowly. Our results show that both the number of nonsynonymous substitutions and the degree of functional constraint are not significantly different between *Hox* and non-*Hox* genes, and that *Hox* and *Hox*-derived genes contain significantly more insertions and deletions than non-*Hox* genes in their coding sequences. Thus, *Hox* genes evolve faster than other essential genes expressed early in development, with complex expression patterns or with long introns rich in *cis*-regulatory elements.

As a whole, the works presented in this thesis round a complete bioinformatics project off, including all the necessary steps from mining the data to generating new scientific knowledge. More interestingly, the outcome of each step is the seed of multiple possible studies in the next step, and thus this thesis has many applications for the scientific community.

ACKNOWLEDGEMENTS

Quisiera agradecer muy sinceramente al Dr. Antonio Barbadilla y al Dr. Alfredo Ruiz la posibilidad que me han dado de llevar a cabo este trabajo en su grupo de investigación. A vosotros os debo los buenos consejos, muchísimas correcciones y todo lo que me habéis enseñado para crecer tanto a nivel científico como personal. Muchas gracias por la confianza que siempre habéis depositado en mí, por estar siempre accesibles cuando os he necesitado y por todo vuestro esfuerzo y dedicación para que este proyecto saliera adelante.

I also thank Dr. Casey M. Bergman for giving me the opportunity to stay in his lab and supervise one of the works presented in this thesis. Thanks a lot for your follow-up guidance and support, for being always ready to help with anything, and… also for not throwing me out after 'destroying' my computer the first day in your lab :-)

També voldria adreçar un molt sincer agraïment a tots els meus companys, sobretot a la Raquel, la Bárbara, l'Emilio, la Marta, l'Alejandra, la Natalia i la Cristina. Moltes gràcies per les estones compartides, els cafès, el vostre interès i en molts casos el vostre ajut en aquest projecte, i per la vostra comprensió i ànims en les temporades més difícils.

I also thank my colleagues in Manchester —Francesco, Winfried and Nora— for very nice days together, making the 3-months stay in Manchester much more easy with amusing table tennis matches and a pleasant 'Cappuccino & Sudoku time' every day.

També voldria agrair al Miquel el disseny i realització de la portada d'aquesta tesi; m'has estalviat molta feina!

Agrair també a la Montse Sales, la Montse Danés, la Julia, la Maite i la Conchi el seu ajut logístic i administratiu.

I finalment, a les diferents fonts de finançament que han fet possible la realització d'aquesta tesis doctoral: e-Biointel, pel finançament durant els primers mesos, el Ministerio de Ciencia y Tecnología per quatre anys de beca predoctoral FPI (BES-2003-0416) i la Comissió Europea per una beca Marie Curie de tres mesos (HPMT-GH-01-00285-13).

Pròleg

La variació genètica és la pedra angular de l'evolució biològica. La descripció i explicació de les forces que controlen la variació genètica dins i entre poblacions és el principal objectiu de la genètica de poblacions (LEWONTIN 2002). L'obtenció d'un número explosiu de seqüències nucleotídiques a diferents gens i espècies ha canviat radicalment les perspectives de la genètica de poblacions, transformant-la des d'una ciència empírica insuficient fins a un esforç interdisciplinari de gran abast, on els aparells de generació de noves seqüències a gran escala s'integren amb eines bioinformàtiques per a l'extracció i gestió de dades, juntament amb avançats models teòrics i estadístics per a la seva interpretació.

Aquesta tesi és un projecte de bioinformàtica i genètica de poblacions complet, l'objectiu principal del qual és l'estudi de la diversitat genètica a les poblacions. S'ha dut a terme en tres passos seqüencials: (i) el desenvolupament d'eines per a l'extracció, processat, filtrat i control de qualitat de seqüències nucleotídiques, (ii) la generació de bases de dades de coneixement a partir de les dades obtingudes a la primera part i (iii) la prova d'hipòtesis que requereixen de dades de varies espècies i loci. A la primera part de la tesi hem desenvolupat PDA (Pipeline Diversity Analysis), una aplicació Web de codi obert que permet l'exploració del polimorfisme a grans conjunts de seqüències de DNA heterogènies. Aquesta eina es nodreix dels milions de seqüències haplotípiques d'estudis individuals que hi ha emmagatzemades a les principals bases de dades moleculars i genera dades de genètica de poblacions que poden ser utilitzades per a descriure patrons de variació nucleotídica a qualsevol espècie o gen. Totes les dades extretes i analitzades a la primera part de la tesi són utilitzades a la segona part per a crear un recurs via Web complet que proporciona col·leccions de seqüències polimòrfiques amb les seves mesures de diversitat associades en el gènere *Drosophila* (DPDB, Drosophila Polymorphism Database). Aquest recurs ha significat un repte ambiciós que ha permès posar a prova l'eficiència del sistema creat a la primera part.

Finalment, s'inclouen dos estudis que utilitzen els mòduls d'extracció i anàlisi de dades desenvolupats a la primera part. En el primer, hem estudiat patrons de variació genètica per a inferir selecció negativa i positiva a seqüències conservades no codificadores a *Drosophila*. Per a aquest estudi hem utilitzat dades de re-seqüenciació a *D. melanogaster* junt amb dades genòmiques comparatives a d'altres espècies de *Drosophila* per a demostrar que les regions fredes de mutació no poden explicar aquests blocs conservats. Els resultats mostren que les seqüències conservades no codificadores són mantingudes per l'acció de la selecció purificadora. El segon estudi es centra en l'evolució codificant dels gens *Hox*, una classe de factors de transcripció essencials en el desenvolupament primerenc que estan involucrats en l'especificació de les regions al llarg de l'eix anteroposterior del cos. Hem mesurat les taxes de divergència nucleotídica i de fixació d'insercions i delecions a tres gens *Hox*, i les hem comparat amb les de tres gens derivats de *Hox* i un conjunt de gens no *Hox* per a provar la hipòtesi que els gens *Hox* evolucionen lentament. Els resultats mostren que tant el número de substitucions no sinònimes com el grau de constrenyiment funcional no són significativament diferents entre els gens *Hox* i els no *Hox*, i que els gens *Hox* i els derivats de *Hox* contenen significativament més insercions i delecions que els gens no *Hox* a les seves seqüències codificants. Per tant, els gens *Hox* evolucionen més ràpidament que altres gens essencials expressats al desenvolupament primerenc, amb patrons d'expressió complexos o amb introns llargs rics en elements *cis*-reguladors.

Resumint, els treballs presentats a aquesta tesi tanquen un cicle complet de projecte bioinformàtic, incloent tots els passos necessaris des de l'extracció de dades fins a la generació de nou coneixement científic. És més, el resultat de cada pas és la llavor per a múltiples possibles estudis en el següent pas, i per tant aquesta tesi té moltes aplicacions per a la comunitat científica.

Prólogo

La variación genética es la piedra angular de la evolución biológica. La descripción y explicación de las fuerzas que controlan la variación genética dentro y entre poblaciones es el principal objetivo de la genética de poblaciones (LEWONTIN 2002). La obtención de un número explosivo de secuencias nucleotídicas en distintos genes y especies ha cambiado radicalmente las perspectivas de la genética de poblaciones, transformándola desde una ciencia empírica insuficiente a un esfuerzo interdisciplinario de un gran alcance, donde los aparatos de generación de nuevas secuencias a gran escala se integran con herramientas bioinformáticas para la extracción y gestión de datos, junto a avanzados modelos teóricos y estadísticos para su interpretación.

Esta tesis es un proyecto de bioinformática y genética de poblaciones completo, cuyo objetivo es el estudio de la diversidad genética en las poblaciones, que se ha llevado a cabo en tres pasos secuenciales: (i) el desarrollo de herramientas para la extracción, procesado, filtrado y control de calidad de secuencias nucleotídicas, (ii) la generación de bases de datos de conocimiento a partir de los datos obtenidos en la primera parte y (iii) la puesta a prueba de hipótesis que requieren de datos de varias especies y loci. En la primera parte de la tesis hemos desarrollado PDA (Pipeline Diversity Analysis), una aplicación Web de código abierto que permite la exploración del polimorfismo en grandes conjuntos de secuencias de DNA heterogéneas. Esta herramienta se alimenta de los millones de secuencias haplotípicas de estudios individuales que hay almacenados en las principales bases de datos moleculares y genera datos de genética de poblaciones que pueden ser utilizados para describir patrones de variación nucleotídica en cualquier especie o gen. Todos los datos extraídos y analizados en la primera parte de la tesis son utilizados en la segunda parte para crear un recurso vía Web completo que proporciona colecciones de secuencias polimórficas con sus medidas de diversidad asociadas en el género *Drosophila* (DPDB, Drosophila Polymorphism Database). Este recurso ha significado un reto ambicioso que ha permitido poner a prueba la eficiencia del sistema creado en la primera parte.

Finalmente, se incluyen dos estudios que utilizan los módulos de extracción y análisis de datos desarrollados en la primera parte. En el primero, hemos estudiado los patrones de variación genética en secuencias conservadas no codificadoras para inferir selección negativa y positiva en *Drosophila*. En este estudio hemos utilizado datos de re-secuenciación en *D. melanogaster* junto con datos genómicos comparativos en otras especies de *Drosophila* para demostrar que las regiones frías de mutación no pueden explicar estos bloques conservados. Los resultados muestran que las secuencias conservadas no codificadoras se mantienen por la acción de la selección purificadora. El segundo estudio se centra en la evolución codificadora de los genes *Hox*, una clase de factores de transcripción esenciales en el desarrollo temprano que están involucrados en la especificación de las regiones a lo largo del eje anteroposterior del cuerpo. Hemos medido las tasas de divergencia nucleotídica y de fijación de inserciones y deleciones en tres genes *Hox*, y las hemos comparado con las de tres genes derivados de *Hox* y un conjunto de genes no *Hox* para probar la hipótesis que los genes *Hox* evolucionan lentamente. Los resultados muestran que tanto el número de sustituciones no sinónimas como el grado de constreñimiento funcional no son significativamente distintos entre los genes *Hox* y los no *Hox*, y que los genes *Hox* y los derivados de *Hox* contienen significativamente más inserciones y deleciones que los genes no *Hox* en sus secuencias codificadoras. Así, los genes *Hox* evolucionan más rápidamente que otros genes esenciales expresados en el desarrollo temprano, con patrones de expresión complejos o con intrones largos ricos en elementos *cis*-reguladores.

En síntesis, los trabajos presentados en esta tesis cierran un ciclo completo de proyecto bioinformático, incluyendo todos los pasos necesarios desde la extracción de datos hasta la generación de nuevo conocimiento científico. Es más, el resultado de cada paso es la semilla para múltiples posibles estudios en el siguiente paso, y por lo tanto esta tesis tiene muchas aplicaciones para la comunidad científica.

Part 1

Introduction

1 Introduction

"… we are always slow in admitting great changes of which we do not see the steps… The mind cannot possibly grasp the full meaning of the term of even a million years; it cannot add up and perceive the full effects of many slight variations, accumulated during an almost infinite number of generations."

— C. DARWIN, *The Origin of Species* (1859)

1.1. THE VARIATIONAL PARADIGM OF BIOLOGICAL EVOLUTION

Darwin's work *The Origin of Species* (1859) is a 'long argument' in favor of *Biological Evolution* as the underlying process of life, and of *Natural Selection* as the fundamental mechanism responsible for the adaptation of species. These revolutionary ideas became only intelligible after a new way of considering natural variation emerged: *Population Thinking* (MAYR 1963; MAYR 1976). This term captures the Darwinian view that swept through systematics and evolutionary biology in the first half of the twentieth century (O'HARA 1998). In contrast to pre-Darwinian essentialism, species are not invariable molds where individual variations are merely noisy deviations from an archetypical phenotype. Individual variation is instead the very real stuff of the evolutionary process, from which adaptations are created and species are transformed. Phenotypic differences among individuals within populations become, through their magnification in time and space, biological evolution: new populations, new species and, by extension, all the biological diversity in the Earth, results from this elementary process (LEWONTIN 1974).

Population genetics provides the theoretical framework for explaining biological evolution from the variational paradigm. Evolution is here envisaged as a process of statistical transformation of Mendelian populations. These are the units of evolution, which consist of groups of interbreeding individuals that share a common genetic pool. Individuals are defined by their internal inheritable traits —their genotypes—, while the distribution of alleles and genotypes describe the populations. From this reference frame, population geneticists have developed an extensive theoretical body which describes the dynamics of the distribution of alleles and genotypes in Mendelian populations, beginning from the zero-force state (the Hardy-Weinberg principle) and considering the impact of different evolutionary forces on the genetic distributions (WRIGHT 1931). In summary, from the variational paradigm the distribution of alleles and genotypes define a population and evolution is the accumulative and irreversible change of these distributions on time. This is, in essence, the basic structure of evolutionary biology.

1.2. GENETIC DIVERSITY

1.2.1. THE GENETIC MATERIAL AS THE CARRIER OF THE EVOLUTIONARY PROCESS

There are two necessary conditions for *evolution* to occur: (i) traits of organisms vary within populations (*phenotypic variation*), and (ii) this variation needs to be at least partially genetically determined (*inheritance*). The DNA is the molecule that carries the genetic information (AVERY *et al.* 1944), and within other properties, it is intrinsically mutable, originating genetic variation. New variants can be accurately replicated and transmitted from generation to generation. If the genetic variants contribute differentially to the survival or reproductive success of individuals within the population (*fitness differences*), then *natural selection* occurs. Therefore, natural selection is not a necessary condition for evolution to occur, and it is only a subset of the evolutionary process (Figure 1).

Mutation is the ultimate source of genetic variation. Alterations of the genetic material range from single nucleotide substitutions to duplications of the whole genome, each occurring at their characteristic rates. Table 1 lists the main types of mutation at the DNA level. Most studies of genetic variation have focused on single-nucleotide differences among individuals. Although one single nucleotide is affected, their abundance in the genome makes them the most frequent source of inter-individual genetic variation. Until recently, they were believed to account for >90% of the genomic

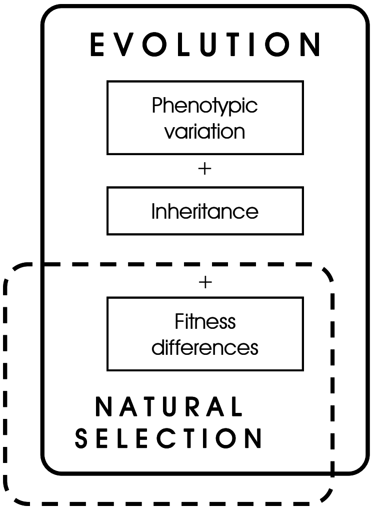


Figure 1
Evolution as a two-condition process

Note that natural selection is neither a *necessary* nor a *sufficient* condition for evolution to occur. It only accounts for a subset of the evolutionary process in which genetic variants differ in fitness.

variability in humans (COLLINS *et al.* 1998). However, recent studies have uncovered an unexpectedly large amount of structural variants in the DNA that span up to several mega bases (FEUK *et al.* 2006; EICHLER *et al.* 2007). Even though they are far less numerous than single-base substitutions, their longer length add up to a significant fraction of the genome, and copy number variations (CNVs) are believed to represent at least 12% of the human genome (REDON *et al.* 2006). These findings question previous estimations that every two human genomes are ~99.9% identical (KRUGLYAK and NICKERSON 2001). The frequency of CNVs in other species than humans still needs to be exhaustively addressed (FREEMAN *et al.* 2006).

1.2.2. THE DYNAMICS OF GENETIC VARIATION

In the early years of genetics, the fate of genetic variation in the populations was unclear. In 1908, G. H. Hardy and W. R. Weinberg independently formulated a mathematical model —the Hardy-Weinberg principle— to explain that allele frequencies in populations would remain the same generation after generation if it were not for a number of forces that may lead to the loss of existing alleles or the acquisition of new alleles. The forces that have an impact in the allele frequencies of populations are principally mutation, migration, natural selection, recombination and random drift.

Table 1 Types of mutation at the DNA level

Type of mutation	Description	Mutation rate*
1. Single-base substitutions	Differences in the sequence of nucleotides. Can be transitions or transversions. Coding-related mutations can be missense, nonsense, silent or splice-site mutations.	10^{-8} / bp / generation
a) Transition	Substitution of one purine (A or G) by another, or one pyrimidine (C or T) by another.	
b) Transversion	Substitution of a purine by a pyrimidine, or vice-versa.	
a) Missense	The new nucleotide alters the codon so as to produce an altered amino acid in the protein. Also called *nonsynonymous*.	
b) Nonsense	The new nucleotide changes a codon that specified an amino acid, to one that stops prematurely the transcription, and thus generates a truncated protein.	
c) Silent	Replacement of one nucleotide by another that does not alter the amino acid. Also called *synonymous*.	
d) Splice-site	Mutations that alter the splice-site signals so that the intron cannot be removed from the RNA molecule, what results in an altered protein product.	
2. Insertions and deletions (indels)	Extra base pairs that may be added (*insertions*) or removed (*deletions*) from the DNA. Mean sizes in *Drosophila* are 42 bp for deletions and 12 bp for insertions. Many large indels result from the activity of transposable elements (TEs). Note that indels not multiple of three within a coding sequence generate frame shifts when the RNA is translated.	0.115 (del.) and 0.028 (ins.) relative to single-base substitutions
3. Variable number of tandem repeats (VNTR)	A locus that contains a variable number of short (2-8 nt for *microsatellites*, 7-100 nt for *minisatellites*) tandemly repeated DNA sequences that vary in length and are highly polymorphic. Microsatellites are also called short sequence repeats (SSRs) or short tandem repeats (STRs).	9.3×10^{-6} / locus / generation for dinucleotide repeats
4. Copy number variation (CNV)	A structural genomic variant that results in confined copy number changes of DNA segments ≥1 kb (i.e. large duplications). They are usually generated by unequal crossing over between similar sequences.	-
5. Inversions	Change in the orientation of a piece of the chromosome. May include many genes.	-
6. Translocations	Transfer of a piece of a chromosome to a nonhomologous chromosome. Can often be reciprocal.	-

* Mutation rates estimated in *Drosophila* according to: 1. Li (1997); 2. Lynch (2007); 3. Kruglyak *et al.* (1998); the rate shown in the table is for dinucleotide repeats, but it is 6.4 and 8.4 times lower for tri- and tetranucleotide repeats respectively. Mutation rate estimates for structural variations are not so well characterized, but see Ranz *et al.* (2007) for inversion fixation rates between different *Drosophila* species.

Mutations are random or undirected events in the sense that they occur independently of whether they help or harm the individual in the environment in which it lives. Most mutations are lost in the first generation. Occasionally, some mutations may increase their frequency through generations, either because they are associated with higher fitness in the population or just by random genetic drift. These allelic variants contribute to the within-population common variability, referred to as *polymorphism* (i.e. any site in the DNA sequence that is present in the population in more than one state of appreciable frequency). We denote as *substitution* or *fixation* the process by which one of the alleles segregating as polymorphisms increases even more in frequency and replaces all the other alleles in the population.

At one stage, two different populations of the same species may become isolated and drive to speciation. Because species undergo independent evolution since they split from a common ancestor, the fixation of different alleles in different species contributes to the differentiation of species. This process is referred to as *divergence* and is ultimately responsible for the branching process of life, which is represented by the Tree of Life. Polymorphism and divergence convey different and complementary information of the genetic history of populations: while polymorphism provides detailed information on recent events, divergence is a window to an older history. Thus, the combined analysis of both within-species polymorphism and between-species divergence represents one of the most powerful approaches to understand the impact of different forces on the patterns of evolutionary change.

1.2.3. ONE CENTURY OF POPULATION GENETICS

The main aim of population genetics is the description and interpretation of genetic variation within and among populations (DOBZHANSKY 1937). The mathematical foundations of population genetics were set up by R. A. Fisher, J. B. S. Haldane and S. Wright between 1910 and 1930. They worked out the quantitative consequences of chance and selection in populations with Mendelian inheritance, and turned population genetics into the explanatory core of the evolutionary theory. In the late 1930s and 40s, the integration of theoretical population genetics with other evolutionary research fields such as experimental population biology, paleontology, systematics, zoology and botany gave rise to the Modern Synthesis of evolutionary biology (DOBZHANSKY 1937; MAYR 1942; SIMPSON 1944; STEBBINS 1950). The major difference between the modern synthetic theory —the neo-Darwinism— and that of natural selection as set forth by

Darwin is the addition of the Mendelian laws of heredity in a population genetics framework. In the neo-Darwinism, evolution is seen as a two-step process: (i) the random generation of variation, and (ii) the directional selection of the allelic variants produced in the first step. A combination of both the chance of mutations and selection drives evolution in natural populations.

A primary feature of the Synthesis period was the uttermost role of natural selection in the detriment of drift and other non-adaptive variation to explain evolution. In a first attempt to measure variation, two different models emerged. First, the so-called 'classical model' supported the role of natural selection in purging the population of most genetic variation, and thus predicted that most gene loci are homozygous for the wild-type allele (MULLER and KAPLAN 1966). Contrarily, the 'balance model' predicted that natural selection maintained high levels of genetic diversity in populations, and thus a large proportion of gene loci were polymorphic and individuals were heterozygous at many gene loci (DOBZHANSKY 1970; FORD 1971). Note that only the balance model is efficient in responding quickly to fluctuations in environmental conditions over time by selecting already existing individual variation and changing the population allelic frequencies. The controversy gained in impetus even after the estimation of the genetic diversity was first made possible.

Until now, three major stages define the molecular research of genetics diversity: the allozyme era (LEWONTIN 1974), the era of nucleotide sequences (KREITMAN 1983), and the current genomics era (LI *et al.* 2001), the main aim of which is still the description and interpretation of genetic variation within and between populations (LEWONTIN 2002). Population genetics entered the molecular age with the publication of seminal papers describing electrophoretically detectable variation —or allozymes (i.e. proteins differing in electrophoretic mobility as a result of allelic differences in the protein sequence)— in *Drosophila* (JOHNSON *et al.* 1966; LEWONTIN and HUBBY 1966) and humans (HARRIS 1966). Genetic diversity was measured in two ways: the average proportion of loci that are heterozygous in an individual (*heterozygosity* or *gene diversity*), and the average proportion of loci that are polymorphic in the population (*gene polymorphism*). The results of such electrophoretic surveys revealed a large amount of genetic variation in most populations (NEVO *et al.* 1984) (Table 2), much more than had been predicted, and seemed to unequivocally support the balance model rather than the classical model. Also, levels of genetic diversity were found to vary nonrandomly among populations, species, higher taxa and several ecological, demographic and life history parameters (NEVO *et al.* 1984). For example, most invertebrates appear to be highly polymorphic whereas the reptiles, birds and mammals are only about half as variable on average, and fish and amphibians are intermediate in their variability (see Table 2).

At the time when the genetic diversity of populations was beginning to be assessed by electrophoretic methods, a new theory was developed to explain the patterns of molecular genetic variation within and among species. In contrast to the selectionist argument of the balance hypothesis, the M. Kimura's Neutral Theory of molecular evolution suggests that most polymorphisms observed at the molecular level are either strongly deleterious or selectively neutral, and that their frequency dynamics in a

Table 2 Heterozygosity (H) and polymorphism (P) studied by protein electrophoresis in >1111 species

Taxa	*H*			*P*			*r (H,P)*	
	N	*Mean*	*S.d.*	*N*	*Mean*	*S.d.*		
Vertebrata	**551**	**0.054**	**0.059**	**596**	**0.226**	**0.146**	**0.792**	*********
Mammalia	184	0.041	0.035	181	0.191	0.137	0.821	***
Aves	46	0.051	0.029	56	0.302	0.143	0.497	***
Reptilia	75	0.083	0.119	84	0.256	0.148	0.814	***
exc. parthenogenetic	70	0.055	0.047	84	0.256	0.148		
Amphibia	61	0.067	0.058	73	0.254	0.151	0.735	***
Pisces	183	0.051	0.035	200	0.209	0.137	0.845	***
Invertebrata	**361**	**0.100**	**0.091**	**371**	**0.375**	**0.219**	**0.769**	*********
Echinodermata	15	0.126	0.083	17	0.505	0.181	0.836	***
Drosophila	34	0.123	0.053	39	0.480	0.143	0.552	***
Insecta exc. Dros.	122	0.089	0.060	130	0.351	0.187	0.753	***
Crustacea	122	0.082	0.082	119	0.313	0.224	0.879	***
Chelizerata	6	0.080	0.033	6	0.269	0.098	0.876	*
Mollusca	46	0.148	0.170	44	0.468	0.287	0.764	***
Brachiopoda	3	0.137	0.087	3	0.526	0.247	0.984	ns
Vermes	6	0.072	0.079	6	0.289	0.222	0.949	**
Coelenterata	5	0.140	0.042	5	0.481	0.191	0.840	ns
Plants	**56**	**0.075**	**0.069**	**75**	**0.295**	**0.251**	**0.842**	*********
Monocotyledoneae	7	0.116	0.091	12	0.378	0.275	0.985	**
Dicotyledoneae	40	0.052	0.049	56	0.235	0.204	0.751	***
Gymnospermeae	7	0.146	0.065	5	0.734	0.186	-0.948	ns

N is the number of species. Significance: * = p < 0.05; ** = p < 0.01; *** = p < 0.001; ns = p > 0.05
[Data compiled by NEVO et al. (1984) from different sources.]

population are determined by the rate of mutation and random genetic drift rather than natural selection (Kimura 1968) (Box 1). By extension, the hypothesis of selective neutrality would also apply to most nucleotide or amino acid substitutions that occur

Box 1 The neutral theory of molecular evolution

In the late 1960s, Motoo Kimura suggested that patterns of protein polymorphism seen in nature were consistent with the view that most polymorphisms and fixed differences between species are either strongly deleterious or selectively neutral. This proposition was called the Neutral Theory of molecular evolution (Kimura 1968). Some of the theory's principal implications (Kimura 1980; Kimura 1983) are:

1. Deleterious mutations are rapidly removed from the population, and adaptive mutations are rapidly fixed; therefore, most variation within species is the result of neutral mutations (Figure 2).
2. A steady-state rate at which *neutral* mutations are fixed in a population *(k)* equals the *neutral* mutation rate: $k = f_{neutral}\ \mu$, where $f_{neutral}$ is the proportion of all mutations that are neutral and μ is the mutation rate. Therefore, the average time between consecutive neutral substitutions is independent of population size *($1/\mu$)*.
3. The level of polymorphism in a population *(θ)* is a function of the neutral mutation rate and the effective population size *(N_e)*: $\theta = 4N_e\ \mu$.
4. Polymorphisms are transient (on their way to loss or fixation) rather than balanced by selection. Larger populations are expected to have a higher heterozygosity, as reflected in the greater number of alleles segregating at a time.

There have been some refinements to the neutral theory, specially the nearly-neutral and slightly deleterious mutation hypotheses of Tomoko Ohta (Ohta 1995), which have stimulated a resurgence of interest in natural selection. However, the consensus amongst population geneticists is that much of the variation at the DNA level is the result of effectively neutral mutations.

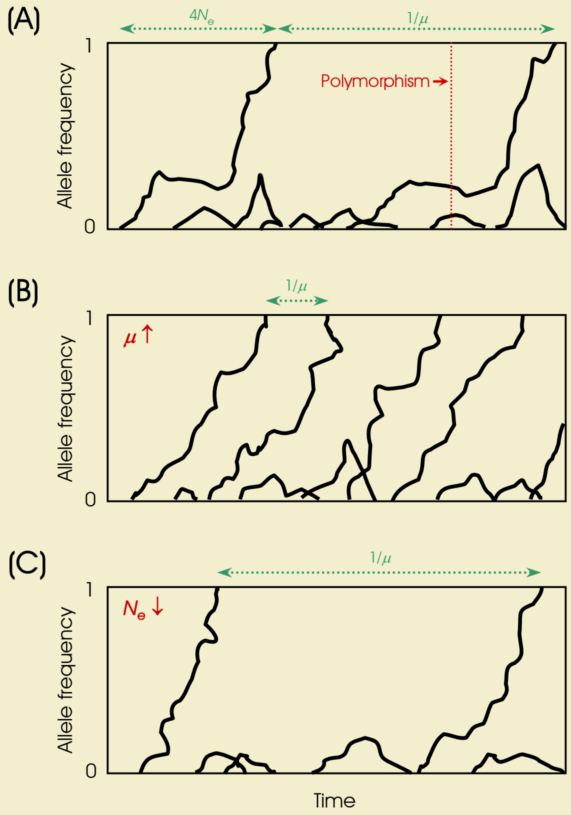


Figure 2
Diagram showing the trajectory of neutral alleles in a population

(A) New alleles arise by mutation with an initial allele frequency of $1/2N$. Their probability of fixation equals their initial frequency in the population ($1/2N$), and the time required for this to occur is $4N_e$ generations. (B) A higher mutation rate gives the same time to fixation, but less time between fixations. (C) In a smaller population, alleles that go to fixation become fixed more rapidly, but the time between fixations is still $1/\mu$. *[Figure redrawn from Hartl and Clark (1997).]*

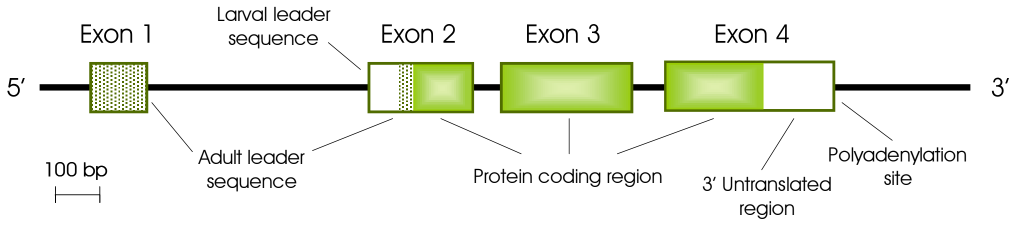
during the course of evolution. However, Kimura emphasized the compatibility of his neutral theory —mainly determined by mutation and drift— at the molecular level, with natural selection shaping patterns of morphological variation. Today, the Kimura's neutral theory is the theoretical foundation of all molecular population genetics.

A corollary of the neutral theory is the existence of a random molecular clock, which had already been previously inferred from protein sequence data by ZUCKERKANDL and PAULING (1962). According to neutralism, the rate at which neutral alleles are fixed in a population equals the neutral mutation rate. Thus, when two populations or species split, the number of genetic differences among them is proportional to the time of speciation. On that account, the number of differences among a set of sequences from different species can be used as a molecular clock to allow sorting the relative times of divergence among these species. The molecular clock is therefore a powerful approach to date ramification events in evolutionary trees.

Even though the major contribution to today's estimates of polymorphism are based upon electrophoretic studies (NEVO *et al.* 1984), the generality of such estimates is uncertain (BARBADILLA *et al.* 1996). One inevitable limitation of electrophoresis is the inability to detect variation in a nucleotide sequence that does not alter the amino acid sequence. Such variations can only be detected by analyses at the DNA level. The first study of nucleotide sequence variation was conducted by KREITMAN (1983) in the gene *Adh* of *D. melanogaster* (Figure 3), whose product had been previously studied by protein electrophoresis detecting two different allelic variants —fast *(Adh-f)* and slow *(Adh-s)*— in nearly all natural populations. This study was the key to uncover many types of nucleotide sequence variation that do not affect the amino acid sequence and that were previously invisible to protein electrophoresis. Furthermore, the availability of nucleotide sequence data allowed the development of more powerful statistical approaches to measure variation than did allozyme data before.

1.2.4. ESTIMATING GENETIC DIVERSITY FROM NUCLEOTIDE SEQUENCES

The *data desideratum* for population genetics studies is a set of homologous and independent sequences (or haplotypes) sampled in a DNA region of interest (Figure 4). From haplotypic sequences, one can estimate both the one-dimensional and multi-dimensional components of nucleotide diversity (Figure 4, Table 3). One-dimensional measures of variation estimate nucleotide diversity in a region taking each nucleotide site as an independent unit. For example, the distribution of π values along sliding windows, allows the detection of differently constrained regions (VILELLA *et al.* 2005). However, tests that only use information on the frequency distribution of segregating sites are clearly ignoring a significant source of information: associations between alleles, or the haplotype structure of the sample. It has been shown that nearby nucleotide sites are not independent of each other; instead, alleles are clustered in blocks of up to 2 kb in the



	5′ Flanking sequence	Adult leader (exon 1)	Intron 1 (adult intron, larval noncoding)	Larval leader	Translated region of exon 2	Intron 2	Exon 3	Intron 3	Translated region of exon 4	3′ UTR	3′ Flanking sequence
Reference sequence	CCG		CAATATGGG∇C∇G	C	T	AC	CCCC	GGAAT	CTCCACTAG	AA_{11}C	AGC∇CT_{09}TΔ
Wa-S	...		.....AT......	.	.	..	TT.A	CA.TA	AC.......	.A_{11}.	.....T_{09}.Δ
F1-1S	..C		....	.	.	..	TT.A	CA.TA	AC.......	.A_{11}.	.....T_{09}.Δ
Af-S	...		....	.	.	..	....	.....	........A	.A_{11}.	..T∇.T_{10}A.
Fr-S	...		....	.	.	GT	....	.....	........A	.A_{10}.	TA...T_{09}..
F1-2S	...		AG...A.TC....	A	G	GT	....	.....	....	CA_{14}.	.....T_{09}..
Ja-S	..C		....	.	G	..	....	.....	...T.T.CA	CA_{15}.	....TT_{09}..
F1-F	..C		....	.	G	..	....	.....	..GTCTCC.	CA_{15}.	.....T_{09}..
Fr-F	TGC		AG...A.TC∇G∇.	.	G	..	....	.....	..GTCTCC.	CA_{15}G	.....T_{09}..
Wa-F	TGC		AG...A.TC∇G∇.	.	G	..	....	.....	..GTCTCC.	CA_{15}G	.....T_{09}..
Af-F	TGC		AG...A.TC∇G∇.	.	G	..	....	.....	..GTCTCC.	CA_{16}G	.....T_{09}..
Ja-F	TGC		AGGGGA...∇..T	.	G	..	..A.	..G..	..GTCTCC.	CA_{15}.	.....T_{08}..
S	3	0	11	1	1	2	4	5	9	2	5
m	63	87	620	70	99	65	405	70	264	178	767
% sites polymorphic	4.7	0	1.8	1.4	1.0	3.1	1.0	7.1	3.5	1.1	0.6

Figure 3

Nucleotide sequence variation in the *Adh* locus of *D. melanogaster*

The sequencing of eleven cloned *D. melanogaster Adh* genes from five natural populations revealed a large number of silent polymorphisms (KREITMAN 1983). Only one of the 43 polymorphisms detected in the *Adh* region resulted in an amino acid change (red), the one responsible for the two electrophoretic variants —fast *(Adh-f)* and slow *(Adh-s)*— found in nearly all natural populations. Insertion/deletion polymorphisms (∇/Δ in blue) and homopolynucleotide runs (green) were only found outside the coding region. The reference sequence is the most common *Adh-s* nucleotide at each of the polymorphic sites. The dashed horizontal line separates *Adh-s* and *Adh-f* alleles. *S*, number of polymorphic sites; *m*, average number of nucleotides compared. *[Figure redrawn from* KREITMAN *(1983).]*

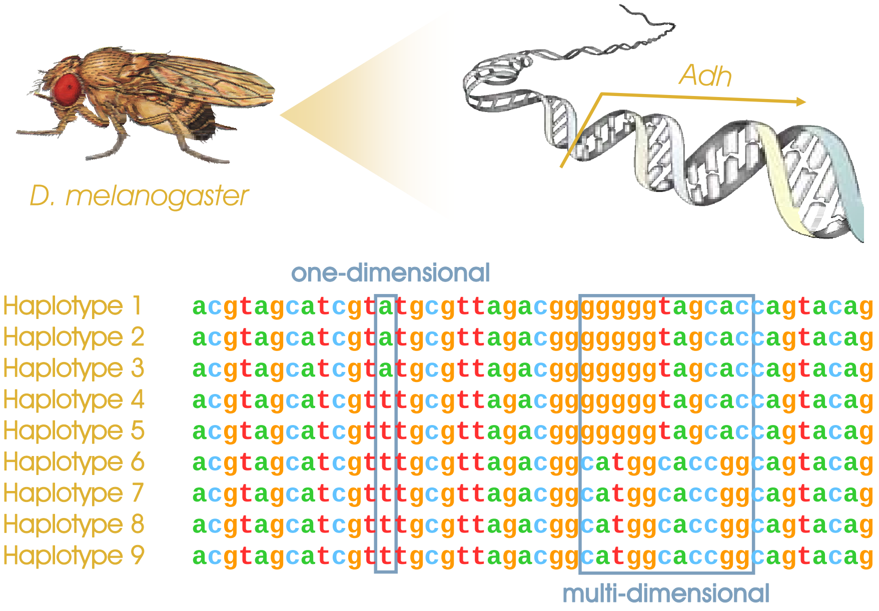

Adh
D. melanogaster
one-dimensional
Haplotype 1 acgtagcatcgtatgcgttagacgggggggtagcaccagtacag
Haplotype 2 acgtagcatcgtatgcgttagacgggggggtagcaccagtacag
Haplotype 3 acgtagcatcgtatgcgttagacgggggggtagcaccagtacag
Haplotype 4 acgtagcatcgtttgcgttagacgggggggtagcaccagtacag
Haplotype 5 acgtagcatcgtttgcgttagacgggggggtagcaccagtacag
Haplotype 6 acgtagcatcgtttgcgttagacggcatggcaccggcagtacag
Haplotype 7 acgtagcatcgtttgcgttagacggcatggcaccggcagtacag
Haplotype 8 acgtagcatcgtttgcgttagacggcatggcaccggcagtacag
Haplotype 9 acgtagcatcgtttgcgttagacggcatggcaccggcagtacag
multi-dimensional

Figure 4
Estimation of one-dimensional and multi-dimensional measures of nucleotide variation from a set of haplotypes

See text for details.

Table 3 **Basic measures of nucleotide diversity**

Nucleotide polymorphism (uni-dimensional measures):		
S, s	Number of segregating sites (per DNA sequence or per site, respectively).	NEI (1987)
H, η	Minimum number of mutations (per DNA sequence or per site, respectively)	TAJIMA (1996)
k	Average number of nucleotide differences (per DNA sequence) between any two sequences	TAJIMA (1983)
π	Nucleotide diversity: average number of nucleotide differences per site between any two sequences. Can be estimated at synonymous (π_s) and nonsynonymous (π_n) sites separately	NEI (1987); JUKES and CANTOR (1969); NEI and GOJOBORI (1986)
θ, θ_W	Nucleotide polymorphism: proportion of nucleotide sites that are expected to be polymorphic in any suitable sample	WATTERSON (1975); TAJIMA (1993; 1996)
Linkage disequilibrium (multi-dimensional measures of association among variable sites):		
D	The first and most common measure of linkage disequilibrium, dependent of allele frequencies	LEWONTIN and KOJIMA (1960)
D’	Another measure of association, independent of allele frequencies	LEWONTIN (1964)
R, R^2	Statistical correlation between two sites	HILL and ROBERTSON (1968)
ZnS	Average of R^2 over all pairwise comparisons	KELLY (1997)

Drosophila genome (MIYASHITA and LANGLEY 1988) and over tens of mega bases in the human genome (FRAZER *et al.* 2007) (Figure 4). This haplotype structure is influenced by recombination, as well as selective and demographic forces, and it can be described by the use of multi-dimensional measures of genetic variation, such as linkage disequilibrium estimators (Table 3). These multi-dimensional diversity measures provide key information on the history and evolution of a DNA region, including the effective recombination rate underlying the region (HUDSON 1987; NORDBORG and TAVARE 2002; MCVEAN *et al.* 2004). Both one-dimensional and multi-dimensional diversity components are necessary for a complete description of sequence variation, and thus haplotypic data provide the highest level of genetic resolution to make inferences about evolutionary history and about the evolutionary process.

1.2.5. THE POPULATION SIZE PARADOX: FROM GENETIC DRIFT TO GENETIC DRAFT

Despite the growing number of estimates of genetic diversity, we still lack a clear understanding about the contribution of different evolutionary forces to produce the patterns of molecular sequence variation we see today. The reason is that the forces that alter the genetic structure of populations tend to be very weak, operating on time scales of thousands to millions of years. According to the neutral theory, most genetic variation is modulated by two major forces: *mutation* and *random genetic drift* (Figure 5A). Mutation adds genetic variation at two times the rate of mutation μ (for diploid organisms), while drift removes each generation a fraction of the genetic diversity which depends on population size ($1/2N_e$). In small populations, drift removes variation faster than mutation can add it, but as populations get larger and larger, drift is less and less effective at removing variation. As a consequence, the extent of genetic diversity and population size are intimately related, and thus large populations are expected to be genetically more diverse than small ones.

This basic prediction of population genetics was first challenged by allozyme polymorphism studies of genetic diversity in the 1960s. However, the close range of diversity estimates obtained across distant species did not reflect the expected wide range in population sizes (LEWONTIN 1974). SMITH and HAIGH (1974) proposed *genetic*

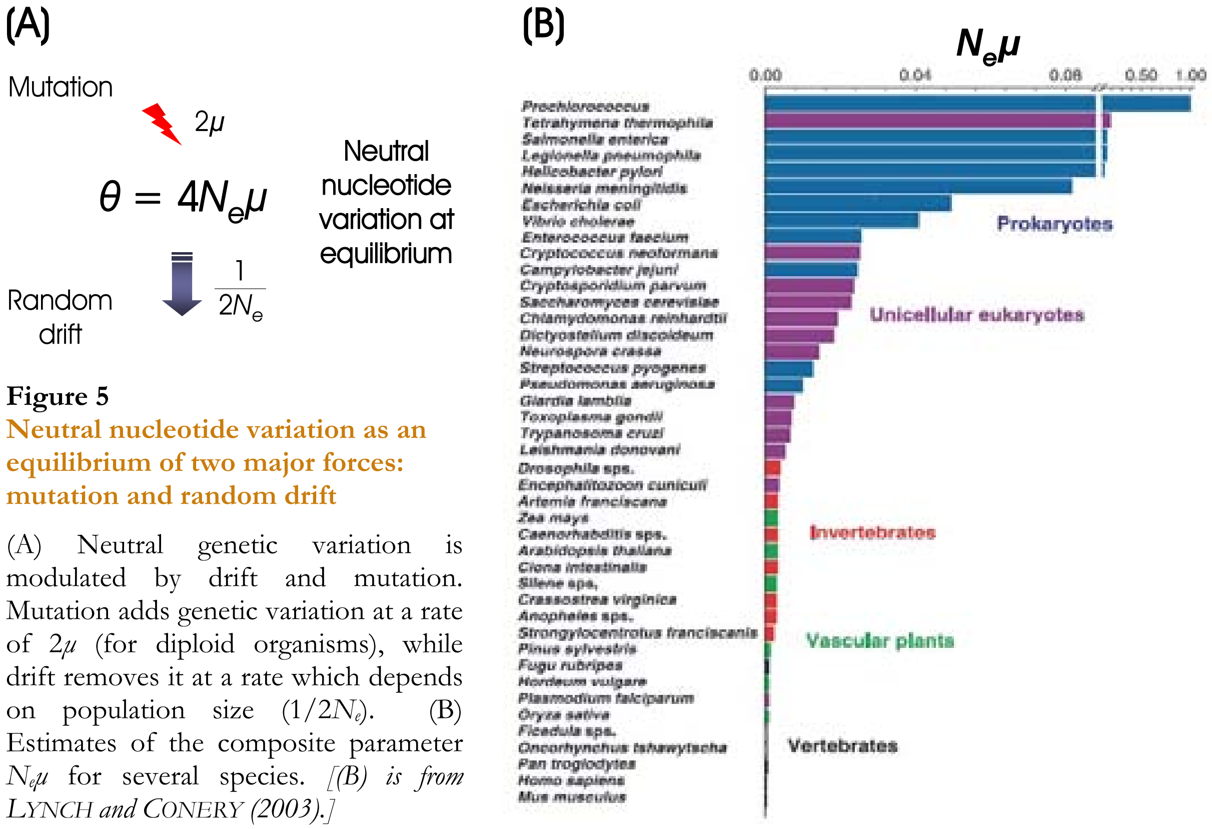


Figure 5

Neutral nucleotide variation as an equilibrium of two major forces: mutation and random drift

(A) Neutral genetic variation is modulated by drift and mutation. Mutation adds genetic variation at a rate of 2μ (for diploid organisms), while drift removes it at a rate which depends on population size ($1/2N_e$). (B) Estimates of the composite parameter $N_e\mu$ for several species. *[(B) is from* L*YNCH and* C*ONERY (2003).]*

hitchhiking as an explanation to the apparent population size paradox. In this process, neutral alleles sufficiently tightly linked go along to fixation together with a favorable mutation, resulting in a *selective sweep*. As a result, linked genetic variation is reduced, and this could explain the genetic homogeneity of large populations. However, Smith and Haigh's hitchhiking explanation relies on populations having high linkage disequilibrium, and it was not widely accepted.

In the late 1980s, when allozyme polymorphism studies were replaced by DNA-based markers —especially mitochondrial DNA (mtDNA) marker studies—, some data became available that showed that genetic variation was reduced in regions of low recombination in *Drosophila*, such as near the tip or near the base of each chromosome arm, or within chromosome rearrangements (AGUADE *et al.* 1989; STEPHAN and LANGLEY 1989; BERRY *et al.* 1991; BEGUN and AQUADRO 1992; MARTIN-CAMPOS *et al.* 1992; STEPHAN and MITCHELL 1992; LANGLEY *et al.* 1993). One possible reason for this correlation is that recombination itself is mutagenic, or that somehow the two processes are related mechanistically. If this were the case, then regions of low recombination should also have a low mutation rate, and hence lower interspecific divergence. However, levels of divergence were shown to be independent of local recombination rates. Therefore, the correlation between recombination rates and levels of polymorphism had to be due to more rapid elimination of the within-species variation in regions of low recombination. For this reason, John Gillespie revised the hitchhiking hypothesis and developed a stochastic model of the process he calls *genetic draft*. (GILLESPIE 2000a; GILLESPIE 2000b; GILLESPIE 2001). Draft produces drift-like dynamics in that it removes genetic variation from the population. However, while drift's ability to remove variation decreases with population size, the rate of hitchhiking increases with N because: (i) adaptive mutations occur more frequently at large populations since there are more alleles to mutate, and (ii) selection is more effective, so even adaptive mutations that are very weakly selected may become fixed in large populations. Then, as population size increases, genetic diversity tends to increase. But at the same time, the number of adaptive substitutions (and hence genetic hitchhiking events) increases, thus reducing the level of genetic diversity. Once N is sufficiently large, genetic draft dominates and genetic variation becomes insensitive to population size. Thus, through this innovative model, Gillespie was able to uncouple population size and levels of genetic diversity (GILLESPIE 2004; LYNCH 2007).

BAZIN *et al.* (2006) have recently resumed the connection between population size and genetic diversity against several ecological and phylogenetic factors, using a dataset of 417 species for nuclear DNA, 1,683 species for mtDNA and 912 species for allozymes. Their results were compatible with the idea that population size and levels of genetic variation are correlated for nuclear DNA, but not so for the mtDNA data. The genetic variation of mtDNA showed very similar levels across distant species, and thus appeared independent of population size. Bazin *et al.* explained the homogeneity of genetic variation for the mtDNA data by genetic draft. Two main characteristics make mtDNA prone to hitchhiking events of the sort Gillespie describes: (i) low levels of recombination in mtDNA relative to nuclear DNA, and (ii) evidence of selective substitutions occurring at mtDNA (58% of amino acid substitutions are selectively advantageous in invertebrate mtDNA, and 12% in vertebrate mtDNA) (BAZIN *et al.* 2006). Thus, mtDNA diversity is essentially unpredictable by population size and may only reflect the time since the last hitchhiking event, rather than population history and demography. In contrast, recombination reduces the effects of genetic hitchhiking in nuclear DNA, and nuclear diversity correlates with population size. However, the differences are remarkably small (Figure 5B): even though the total range in population sizes over all species certainly exceeds 20 orders of magnitude (LYNCH 2006), synonymous diversity varies by less that a factor of 10, and allozyme diversity by less that a factor of 4 (BAZIN *et al.* 2006). The lack of a strong correlation between diversity and population size in nuclear DNA may also reflect the effects of genetic hitchhiking, or the increased mutation rate in organisms with large genome sizes (and thus small populations) (LYNCH 2006).

1.2.6. TESTING THE NEUTRAL THEORY WITH POLYMORPHISM AND DIVERGENCE DATA

Looking for evidence of selection is a widely-used strategy for finding functional variants in the genome (BAMSHAD and WOODING 2003). Natural selection leaves signatures in the genome that can be used to identify the regions that have been selected (Figure 6). These signatures include: (i) a reduction in the genetic diversity, (ii) a skew towards rare derived alleles, and (iii) an increase in linkage disequilibrium (LD) (SCHLOTTERER 2003). It has already been shown that hitchhiking events reduce genetic diversity in the region by dragging linked neutral variation along with the selected site. During the process of fixation, new mutations accumulate in the region having initially low frequency in the population, and because common alleles present before the sweep have been removed, the result is an excess of rare derived alleles in the region, as shown in site frequency spectrum representations. Finally, the block-like nature of LD across the genome is another strategy to detect the signature of recent positive selection (NORDBORG and TAVARE 2002; SABETI et al. 2002). Consider a completely linked haplotype block in a neutral DNA region (long-range LD). Over time, both new mutation events and local recombination reduce the size of this haplotype block such that, on average, older and relatively common mutations will be found on smaller haplotype blocks (short-range LD). However, an allele influenced by recent positive selection might increase in frequency faster than local recombination can reduce the range of LD between the allele and linked markers. Thus, a signature of positive selection is indicated by an allele at high population frequency and with unusually long-range LD. This strategy has been formally implemented as the long-range haplotype (LRH) test (SABETI et al. 2002). Unfortunately, all these signatures quickly dissipate with time (KIM and STEPHAN 2002; NIELSEN et al. 2005) and this approach can only identify very recent and strong adaptive events. However, the wealth of nucleotide polymorphism data that has become available during the past few years has provided an exciting opportunity to carry out genome scans for selection (BAMSHAD and WOODING 2003; EYRE-WALKER 2006) and many cases of

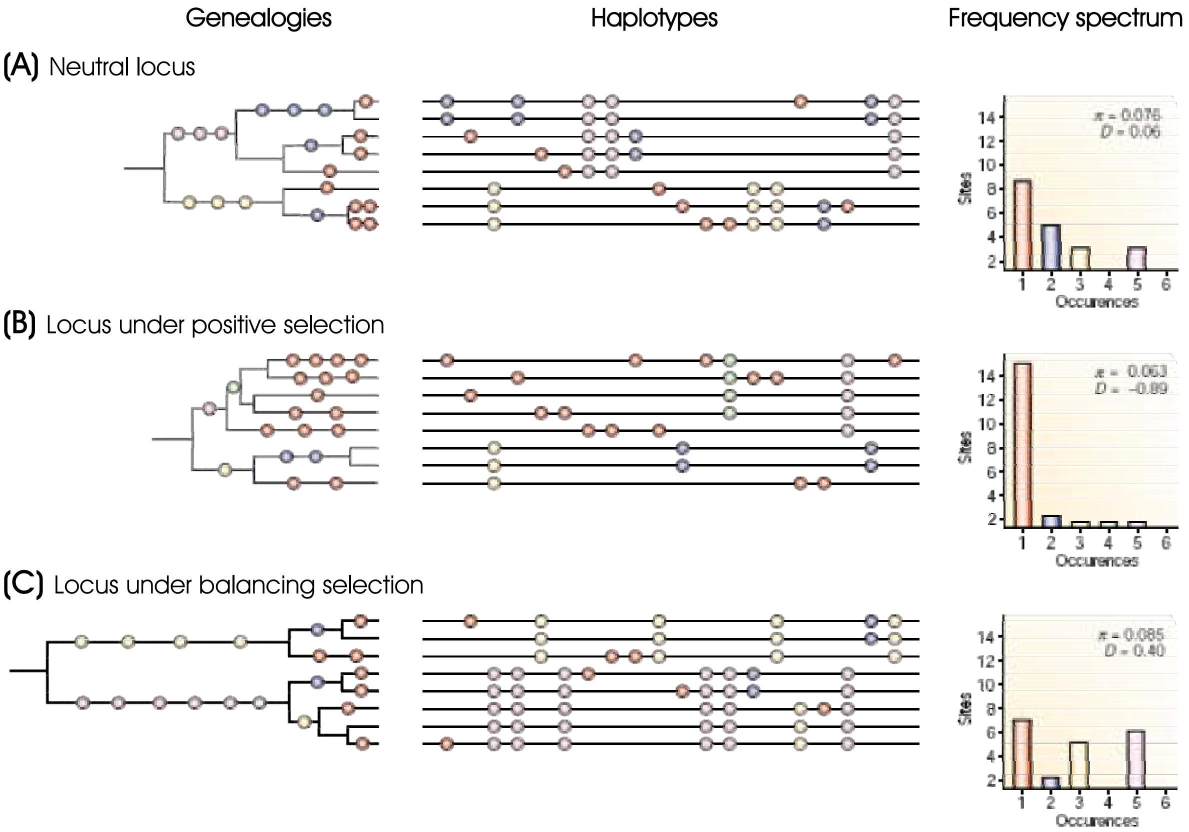


Figure 6

Effects of selection on the distribution of genetic variation

Genealogies typical of (A) a neutral locus, (B) a locus under positive selection, and (C) a locus under balancing selection. Mutations are represented by circles and colors are according to their final frequency in the sampled haplotypes. Note that positive selection results in a lower level of sequence diversity (π), an excess of low-frequency variants (red) and, consequently, a negative value of Tajima's D. Balancing selection results in a higher level of sequence diversity (π), an excess of intermediate-frequency variants (yellow, purple) and thus a positive value of Tajima's D. *[Figure from BAMSHAD and WOODING (2003).]*

selective sweeps have been found in *Drosophila*, humans and other species (SCHLOTTERER 2002; KAUER *et al.* 2003; AKEY *et al.* 2004; STORZ *et al.* 2004; WRIGHT *et al.* 2005; IHLE *et al.* 2006; VOIGHT *et al.* 2006; WIEHE *et al.* 2007).

There is another selective process that reduces the level of genetic variation in the region: *background selection* (i.e. the recurrent elimination of chromosomes carrying strongly deleterious mutations) (CHARLESWORTH *et al.* 1993; BRAVERMAN *et al.* 1995; CHARLESWORTH *et al.* 1995). However, this process can be distinguished from hitchhiking because it does not skew the distribution of rare polymorphisms neither generates blocks of LD. The effect in this case is to reduce the number of chromosomes that contribute to the next generation, and thus it is identical to that of a reduction in population size except that the reduction applies, not to the genome as a whole, but to a tightly linked region (CHARLESWORTH *et al.* 1993). Variations in the local rate of recombination along the genome also make the detection of selection difficult, since the signatures of selection highly depend on the local rate of recombination (HUDSON and KAPLAN 1995). In fact, it seems that about half of the variance in nucleotide diversity in the human genome might be explained just by differences in the local rate of recombination (NACHMAN 2001). On these grounds, the confounding effects of both demography and recombination in the patterns of genetic variation challenge the identification of regions in the genome showing truly signatures of adaptive evolution (TESHIMA et al. 2006).

Tests of selection

Several tests based on the level of variability and the distribution of alleles have been developed to summarize the previous signatures of selection and empirically spot regions in the genome with footprints of recent adaptive events (Table 4). In the d_n/d_s (or K_a/K_s) test (YANG and BIELAWSKI 2000), the rate of nonsynonymous substitution — d_n or K_a — is compared to the rate of silent substitution — d_s or K_s —. If we assume that all silent substitutions are neutral, then we can infer that the gene has undergone adaptive evolution only if d_n is significantly greater than d_s, because advantageous mutations have a higher probability of spreading through a population than do neutral mutations. On the contrary, if d_n is significantly lower than d_s, replacement substitutions are mainly removed

Table 4 **Commonly used tests of selection**

Test	Compares	References
Based on allelic distribution and/or level of variability:		
Tajima's *D*	The number of nucleotide polymorphisms with the mean pairwise difference between sequences	TAJIMA (1989)
Fu and Li's *D*, *D**	The number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants	FU and LI (1993)
Fu and Li's *F*, *F**	The number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences	FU and LI (1993)
Fay and Wu's *H*	The number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies	FAY and WU (2000)
Based on comparisons of divergence and/or variability between different classes of mutation:		
d_n/d_s, K_a/K_s	The ratios of nonsynonymous and synonymous nucleotide substitutions in protein coding regions	LI *et al.* (1985b); NEI and GOJOBORI (1986)
HKA	The degree of polymorphism within and between species at two or more loci	HUDSON *et al.* (1987)
MK	The ratios of synonymous and nonsynonymous nucleotide substitutions within and between species	MCDONALD and KREITMAN (1991)

HKA, Hudson-Kreitman-Aguade; MK, McDonald-Kreitman.
[Table from BAMSHAD and WOODING (2003).]

by negative selection because they are deleterious. Thus, the ratio $\omega = d_n/d_s$ is used as a common measure of functional constraint: d_n/d_s equals 1 under neutrality, is <1 under functional constraint, and is >1 under positive selection. Note that the method assumes that: (i) all synonymous substitutions are neutral, and (ii) all substitutions have the same biological effect, which might not be the case. This test is conservative because most nonsynonymous mutations are expected to be deleterious and d_n tends to be much lower than d_s. Thus, the proportion of adaptive substitutions needs to be high for adaptive evolution to be detectable using this method.

The McDonald-Kreitman (MK) test (MCDONALD and KREITMAN 1991) is a more powerful test that combines both between-species divergence and within-species polymorphism data, and also categorizes mutations into two separate classes (Table 4, Box 2). This test was developed as an extension of the Hudson-Kreitman-Aguade test (HUDSON *et al.* 1987) (Table 4). It compares the numbers of polymorphisms (*P*) to the numbers of fixed differences (*D*) at two classes of sites in protein-coding sequences:

synonymous (P_s, D_s) and nonsynonymous (P_n, D_n). If all mutations are either strongly deleterious or neutral, then D_n/D_s is expected to roughly equal P_n/P_s. By contrast, if positive selection is operating in the region, adaptive mutations rapidly reach fixation and

Box 2 The McDonald-Kreitman test

McDONALD and KREITMAN (1991) proposed a simple test of neutrality (the McDonald-Kreitman test, or MK test), which has become the basis of several methods to estimate the proportion of substitutions that have been fixed by positive selection rather than by genetic drift (FAY *et al.* 2001; SMITH and EYRE-WALKER 2002; SAWYER *et al.* 2003; BIERNE and EYRE-WALKER 2004; WELCH 2006). The test compares the amount of variation within species to the divergence between species at two types of site: synonymous and nonsynonymous sites. The test assumes that all synonymous mutations are neutral, and that nonsynonymous mutations are either strongly deleterious, neutral, or strongly advantageous.

It is expected that the effects on fitness of a mutation are the same whether within a species or at any time along the ancestral history of two species back to the common ancestor. If this is true, and if all mutations are neutral ($f_0 = 1$ and $a = 0$ in Table 5), then the ratio of synonymous to nonsynonymous polymorphisms (P_n/P_s) is expected to equal the ratio of synonymous to nonsynonymous substitutions (D_n/D_s) (see Table 5). This is the basis of the MK test. We can summarize the four values as a ratio of ratios termed the Neutrality Index (NI) (RAND and KANN 1996) as follows:

$$NI = \frac{P_n/P_s}{D_n/D_s}$$

Under neutrality, D_n/D_s equals P_n/P_s, and thus NI = 1. If NI < 1, there is an excess of fixation of amino acid replacements due to positive selection (D_n is higher than expected). If NI > 1, negative selection is preventing the fixation of harmful mutations (D_n is lower than expected). The test is therefore useful in assessing the relative importance of neutral drift and selection.

Assuming that adaptive mutations contribute little to polymorphism but substantially to divergence, the proportion of adaptive substitutions (α) can be estimated as follows:

$$\alpha = 1 - \frac{D_s P_n}{D_n P_s}$$

In the example of Table 5, NI = 0.116 and α = 0.884 (χ^2=8.20, df=1, p<0.01). Therefore, positive selection is inflating the expected number of nonsynonymous substitutions and, overall, adaptive substitutions account for 88.4% of all nonsynonymous substitutions.

Table 5 McDonald-Kreitman table

	Divergence	***Polymorphism***
Non-synonymous	$D_n = 2T\mu f_0 L_n + a$ 7	$P_n = 4N_e\mu f_0 L_n k$ 2
Synonymous	$D_s = 2T\mu L_s$ 17	$P_s = 4N_e\mu L_s k$ 42
	$D_n/D_s = (L_n/L_s) \cdot f_0 + a/D_s$ 0.412	$P_n/P_s = (L_n/L_s) \cdot f_0$ 0.048

[Data from 12 Adh *sequences in* D. melanogaster *from* MCDONALD *and* KREITMAN *(1991).]*

Abbreviations: ***P_s****, expected number of synonymous polymorphisms;* ***P_n****, expected number of nonsynonymous polymorphisms;* ***D_s****, expected number of synonymous substitutions;* ***D_n****, expected number of nonsynonymous substitutions;* ***L_s****, number of synonymous sites;* ***L_n****, number of nonsynonymous sites;* ***N_e****, effective population size;* ***μ****, nucleotide mutation rate;* ***T****, average time to coalescence;* ***f_0****, proportion of mutations that are neutral;* ***a****, number of adaptive substitutions;* ***k****, probability of observing a neutral variant (depends upon several factors, including the number of alleles sampled, the sampling strategy and the population history).*

thus contribute relatively more to divergence than to polymorphism when compared with neutral mutations, and then $D_n/D_s > P_n/P_s$. Furthermore, data from a MK test can be easily used to estimate the proportion of nonsynonymous substitutions that have been fixed by positive selection —α— (Box 2) (CHARLESWORTH 1994). However, this estimate can be easily biased by the segregation of slightly deleterious nonsynonymous mutations (EYRE-WALKER 2002). If the population size has been relatively stable, α is underestimated, because slightly deleterious mutations tend to contribute relatively more to polymorphism than they do to divergence when compared with neutral mutations. Because these slightly deleterious mutations tend to segregate at lower frequencies than do neutral mutations, they can be controlled for by removing low-frequency polymorphisms from the analysis (CHARLESWORTH 1994; FAY *et al.* 2001). However, slightly deleterious mutations can lead to an overestimate of α if population size has expanded, because those slightly deleterious mutations that could become fixed in the past by genetic drift due to the small population size only contribute to divergence (EYRE-WALKER 2002).

Quantifying the amount and strength of adaptive evolution

For over 30 years, population geneticists have debated the relative contributions of genetic drift and adaptation to the evolution at the molecular level. A recent increase in DNA sequence data and the development of new methods of analysis should soon shed light on the issue. So far, many attempts have been made to determine the extent of adaptive evolution in several species, and although the data is limited, it seems to be correlated with population size (EYRE-WALKER 2006). Then, hominids and land plants appear to have undergone very little adaptive evolution (probably <10%) compared with *Drosophila* (~40-50%), bacteria (>56%) and viruses (50-85%) (Table 6). This is expected since in large populations: (i) more advantageous mutations appear, and (ii) selection is more efficient on them. This indicates a better ability of large populations to adapt to the environment than small populations. In *Drosophila* for example, considering that 45% of the amino acid substitutions between *D. melanogaster* and *D. simulans* have been fixed by positive selection (SMITH and EYRE-WALKER 2002; BIERNE and EYRE-WALKER 2004), we can estimate that ~22,000 adaptive amino acid substitutions occur in these species per million years. Thus, even though these two *Drosophila* species are almost identical morphologically, their genomes differ from one another by an astonishing number of ~110,000 adaptive amino acid differences (Eyre-Walker 2006), or ~1.3 million adaptive nucleotide differences in the whole genome (including coding and noncoding regions) (Andolfatto 2005). This huge amount of adaptive differences between apparently similar species may indicate a higher influence of physiology, ecology, and adaptation to a varying environment than could be previously expected (Eyre-Walker 2006).

Table 6 Estimates of adaptive evolution

Species 1	Species 2	Gene region	Test	# loci	α[§]	References
Human	Mouse	Coding	MK	330	*0*	Zhang and Li (2005)
	Old-world monkey	Coding	MK	149	*0*	Zhang and Li (2005)
		Coding	MK	182/106[¥]	35	Fay *et al.* (2001)
	Chimpanzee	Coding	d_n/d_s	8,079	0.4	Nielsen *et al.* (2005)
		Coding	MK	13,500	0-9	Mikkelsen *et al.* (2005)
		Coding	MK	289	*20*	Zhang and Li (2005)
		5' flank	MK	305	*0.11*	Keightley *et al.* (2005b)
		3' flank	MK	305	*0.14*	Keightley *et al.* (2005b)
		Coding	MK	4,916	6	Bustamante *et al.* (2005)
	Chimp & mouse	Coding	d_n/d_s	7,645	0.08	Clark *et al.* (2003)
Arabidopsis thaliana	*A. lyrata*	Coding	MK	12	0	Bustamante *et al.* (2002)
		Coding	d_n/d_s	304	5	Barrier *et al.* (2003)
Drosophila simulans	*D. yakuba*	Coding	MK	35	45	Smith and Eyre-Walker (2002)
		Coding	MK	115	41	Welch (2006)
	D. melanogaster	Coding	MK	75	43	Bierne and Eyre-Walker (2004)
		Coding	MK	56	94	Sawyer *et al.* (2003)
Drosophila melanogaster	*D. simulans*	Coding	MK	44	45	Bierne and Eyre-Walker (2004)
		5' UTR	MK	18	61	Andolfatto (2005)
		3' UTR	MK	13	53	Andolfatto (2005)
		Intron	MK	72	19	Andolfatto (2005)
		Interg.	MK	50	15	Andolfatto (2005)
Escherichia coli	*Salmonella enterica*	Coding	MK	410	>56	Charlesworth and Eyre-Walker (2006)
HIV		Coding	MK	1	50	Williamson (2003)
		Coding	d_n/d_s	1	75[ς]	Nielsen and Yang (2003)
Influenza		Coding	d_n/d_s	1	85[ς]	Nielsen and Yang (2003)

[¥] Numbers of genes differ for divergence (182) and polymorphism (106).
[§] % of loci adaptively evolving. Estimates in italics are not significantly different from zero.
[ς] Proportion of codons showing evidence of adaptive evolution.
[Table from Eyre-Walker (2006).]

In principle, it is possible to estimate not only the number of adaptive substitutions that have occurred during the evolution of a species, but also the average strength of selection that has acted upon them (SAWYER and HARTL 1992; WIEHE and STEPHAN 1993; STEPHAN 1995; BUSTAMANTE *et al.* 2002; SAWYER *et al.* 2003; BUSTAMANTE *et al.* 2005), or the distribution of fitness effects of new mutations (PIGANEAU and EYRE-WALKER 2003; KRYUKOV *et al.* 2005). However, different methods with different assumptions have yielded disparate estimates of the strength of selection for the same species, with values of N_es in *Drosophila* ranging from 1-10 (SAWYER *et al.* 2003) to 350-3500 (ANDOLFATTO 2005). Thus, further work is clearly needed to resolve this issue.

1.3. BIOINFORMATICS OF GENETIC DIVERSITY

The deciphering of an explosive number of nucleotide sequences in a large number of genes and species, and the availability of complete genome sequences of model organisms have changed the approaches of genetic diversity studies. Population genetics has evolved from an insufficient empirical science into an interdisciplinary activity gathering together theoretical models and interpretation statistics, advanced molecular techniques of massive sequence production, and large-scale bioinformatic tools of data mining and management. As a result, population genetics is today an information-driven science in which hypotheses can be tested directly on the data sets stored in online databases and bioinformatics has emerged as a new cutting-edge approach to do science.

1.3.1. MANAGEMENT AND INTEGRATION OF MASSIVE HETEROGENEOUS DATA: THE NEED FOR RESOURCES

Bioinformatics is the computerized analysis of biological data, in which biology, computer science and information technology merge into a single discipline. The ultimate goal of bioinformatics is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. The evolution of bioinformatics has been marked by three main stages (VALENCIA 2002;

Kanehisa and Bork 2003). At the beginning of the genomic revolution, the major concern of bioinformatics was the creation and maintenance of databases to store raw biological data (or primary databases), such as nucleotide and amino acid sequences, as well as the creation of effective interfaces and new computational methods to facilitate the access and analysis of data at a large-scale. Later, bioinformatics became a powerful technology focused on creating biological databases of knowledge (or secondary databases) from previous unprocessed data (Galperin 2007). The major difficulty arises from the fact that there are almost as many formats to store and represent the data as the number of existing databases (Stein 2002). In this sense, bioinformatics has become essential to manage and integrate the torrent of raw data and transform it into biological knowledge (Searls 2000; Jackson *et al.* 2003). The ultimate goal of bioinformatics, however, is to combine all this information and create a comprehensive picture of complex biological systems (Di Ventura *et al.* 2006). The actual process of analyzing and interpreting data is often referred to as *computational biology*. Still, theory, modeling and data processing will continue to become more and more important as scientists working on model systems tend not to be limited by data (Stein 2003).

The first absolute requirement arising from the deluge of data in the genomic era is the establishment of computerized databases to store, organize and index massive and complex datasets, together with specialized tools to view and analyze the data. A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query and retrieve components of the data stored within the system. A simple database might be a simple flat file containing many records, each of which includes the same data fields. This strategy is still extensively used because of the standardization of formats and the existence of the Perl (Practical Extraction and Report Language) programming language (Stein 2001), which is very powerful in scanning, searching and manipulating textual data. However, relational databases (RDB) (Codd 1970) offer the best performance to complex and highly structured data, as is the case of biological data. In RDB, information is distributed into tables of rows and columns, and tables are related by one or several common fields. This system is especially useful for performing queries using the SQL (Structured Query Language) standard language.

Two more requirements are necessary for researchers to benefit from the data stored in a database: (i) easy access to the information, and (ii) a method for extracting only that information needed to answer a specific biological question. The web has played a very important role in genetics research by allowing a universal and free exchange of biological data (Guttmacher 2001). As a result of large projects such as the sequencing project of the human genome (Lander *et al.* 2001; Venter *et al.* 2001), powerful portals have been created to store and distribute a wide variety of data, which also include sophisticated web tools for its analysis (Wheeler *et al.* 2007). Indeed, the so announced milestone of having the complete sequence of the human genome would not have been possible without the arrival of the Internet and the development of information and communication technologies (ICTs).

1.3.2. Databases of nucleotide variation

Nowadays, four dominant resources allow free access to nucleotide variation data (Table 7). The largest and primary public-domain archive for simple genetic variation data is the *Entrez* dbSNP section of NCBI (Wheeler *et al.* 2007). It contains single nucleotide polymorphisms (SNPs), small-scale multi-base deletions or insertions (also called deletion-insertion polymorphisms or DIPs), and STRs (microsatellites) associated to genome sequencing projects of 43 different species, including human (>11.8 million SNPs, of which >5.6 million validated), mouse (>10.8 million SNPs, of which >6.4 million validated), dog (>3.3 million SNPs, of which 217,525 validated), chicken (>3.2 million SNPs, of which >3.2 million validated), rice (>3.8 million SNPs, of which 22,057 validated) and chimpanzee (>1.5 million SNPs, of which 112,654 validated). These SNPs can be browsed according to different criteria, such as heterozygosity or functional class. However, most maps have been developed under a medical or applied focus, and thus their application to evolutionary studies is limited. The non-random sampling of SNPs and/or individuals, the analysis of very few individuals (only those needed to position the SNP in the genome), or the inability to obtain haplotypic phases, together with the fact that only sequenced genomes have such a resource, make *Entrez* dbSNP an inappropriate source of data on which to carry out multi-species population evolutionary studies.

The HAPMAP project (CONSORTIUM 2003; CONSORTIUM 2004; THORISSON *et al.* 2005; FRAZER *et al.* 2007) is an international effort to catalog common genetic variants and their haplotype structure in different human populations, with the goal to identify genes affecting health, disease and responses to drugs and environmental factors. It is undoubtedly the most comprehensive description of nucleotide variation at any species (CONSORTIUM 2005; HINDS *et al.* 2005), but its medical focus limits its application to population genetics studies. For example, only common SNPs (rare variants at >5% frequency) were selected for HAPMAP Phase I, and the sampling methodology changed during the course of the project, which really hinders any evolutionary interpretation of the patterns found. The lack of a complete data collection and the biases underlying SNP sampling make virtually impossible to specify an evolutionary model of human genetic variation from the HAPMAP data (MCVEAN *et al.* 2005).

The *Entrez* POPSET database (WHEELER *et al.* 2007) is a collection of haplotypic sequences that have been collected from studies carried out within a populational perspective, with sequences coming either from different members of the same species or

Table 7 Data sources of nucleotide variation

Resource	Description	Amount of data$^{\zeta}$	Reference
Entrez dbSNP (NCBI)	Mapped SNPs associated to genome sequencing projects of eukaryotic species	>51.3 million SNPs (of which >22.2 million validated) from 43 species	WHEELER *et al.* (2007)
HAPMAP	Haplotype map of the human genome	>3.7 million genotyped SNPs in 270 independent samples from 4 human populations (CEU, CHB, JPT and YRI)	CONSORTIUM (2003; 2004); THORISSON *et al.* (2005)
Entrez POPSET (NCBI)	Haplotypic sequences from population studies of polymorphism or divergence	>52,000 eukaryotic entries	WHEELER *et al.* (2007)
GENBANK – *Entrez* NUCLEOTIDE (NCBI)	Public database of non-redundant nucleotide sequences from any species	>71.6 million eukaryotic sequences (CoreNucleotide: >7.2; EST: >46.1; GSS: >18.2)	BENSON *et al.* (2007); WHEELER *et al.* (2007)

$^{\zeta}$ HAPMAP Public Release #22 (March 2007); *Entrez* dbSNP Build 127 (March 2007); *Entrez* POPSET and GENBANK – *Entrez* NUCLEOTIDE as consulted on Sep 21st 2007 (excluding Whole Genome Shotgun (WGS) sequences and constructed (CON-division) sequences).

from organisms from different species. Even though it contains >52,000 eukaryotic entries, PopSet is a mere repository of DNA sequences, some of which have been aligned by the authors, but it does not give descriptive or comparative information of genetic diversity in any polymorphic set.

Finally, another potential resource for the study of nucleotide variation is the >76.1 million non-redundant sequences (>71.6 million from eukaryotes) of any region and/or species that are stored in major public DNA databases, such as *Entrez* Nucleotide (GenBank) (Benson *et al.* 2007; Wheeler *et al.* 2007) (see Figure 7). This dataset contains all the sequences from the *Entrez* PopSet database together with an extensive number of other heterogeneous sequences with respect to their origin and motivation for their sequencing. In principle, all these sequences could be used to estimate genetic diversity in a large number of genes and species and carry out a large-scale description of nucleotide variation patterns in any taxa (Pandey and Lewitter 1999). In such an approach, the reliability of the estimates depends on developing proper data mining and analysis tools that include accurate filtering criteria of the source data, as well as efficient checking procedures and quality parameters associated to any estimate.

1.4. Evolution of noncoding DNA

1.4.1. The amount of noncoding DNA and organismal complexity

A remarkably high proportion of the DNA in complex multicellular organisms seems not to perform an obvious function such as coding for proteins or RNAs (Britten and Davidson 1969; Taft and Mattick 2003). For example, >75% of the euchromatic portion of the *D. melanogaster* genome is contained in noncoding intronic and intergenic regions (Misra *et al.* 2002), and this percentage rises to 98.5% in humans (Venter *et al.* 2001). Interestingly, the fraction of noncoding DNA (ncDNA) —and not the number of protein-coding genes, as had been previously suggested (Bird 1995)— has shown to be positively correlated with biological complexity: ncDNA accounts for 5-24% of the genome in prokaryotes, 26-52% in unicellular eukaryotes and >62% in been complex multicellular organisms (Figure 8) (Taft and Mattick 2003). Therefore, it is

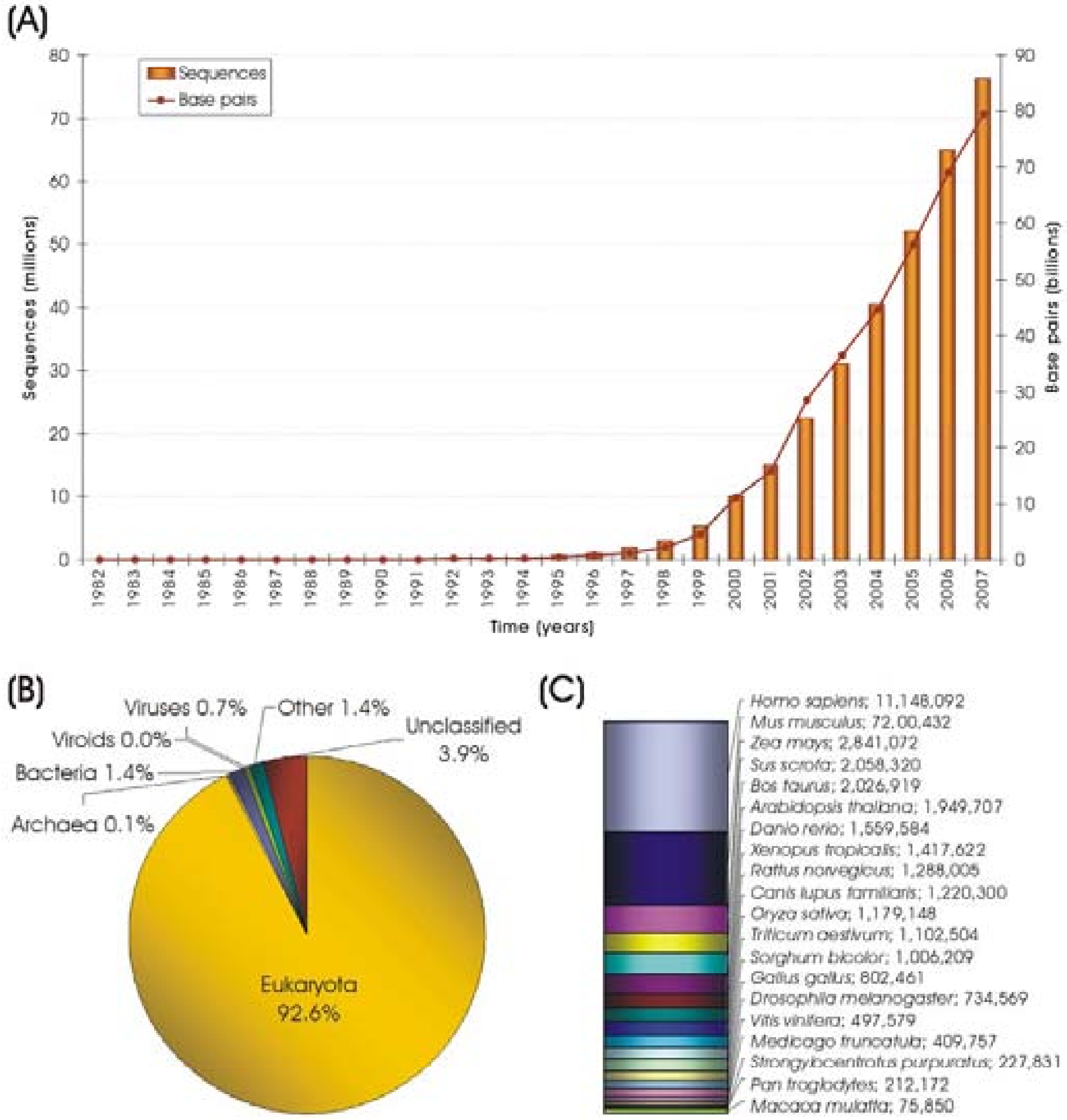


Figure 7

The GENBANK database

(A) Contents of the GENBANK database from 1982 to 2007 in terms of number of sequence records (bars) and number of bases (line) (data is from the last release of each year, and Release 161.0 for 2007). (B) Distribution of the number of entries among different taxonomic groups (as on Sep 21st 2007). (C) Number of entries of the twenty most sequenced organisms, excluding chloroplast, mitochondrial and metagenomic sequences (Release 161.0). All graphs exclude Whole Genome Shotgun (WGS) sequences and constructed (CON-division) sequences.

possible that introns, intergenic sequences, repetitive elements and other genomic features previously regarded as functionally inert (VENTER *et al.* 2001; APARICIO *et al.* 2002; DENNIS 2002) may be far more important to the evolution and functional repertoire of complex organisms than has been previously appreciated.

1.4.2. STRUCTURE OF THE EUKARYOTIC GENE

The enrichment in ncDNA in eukaryotes mainly comes from the 'embellished' structure of the eukaryotic gene, including the presence of introns embedded within coding sequences, transcribed but untranslated leader and trailer sequences (5' and 3'UTRs), modular regulatory elements controlling gene expression, and longer intergenic regions harboring additional control mechanisms (LYNCH 2006; LYNCH 2007) (Figure

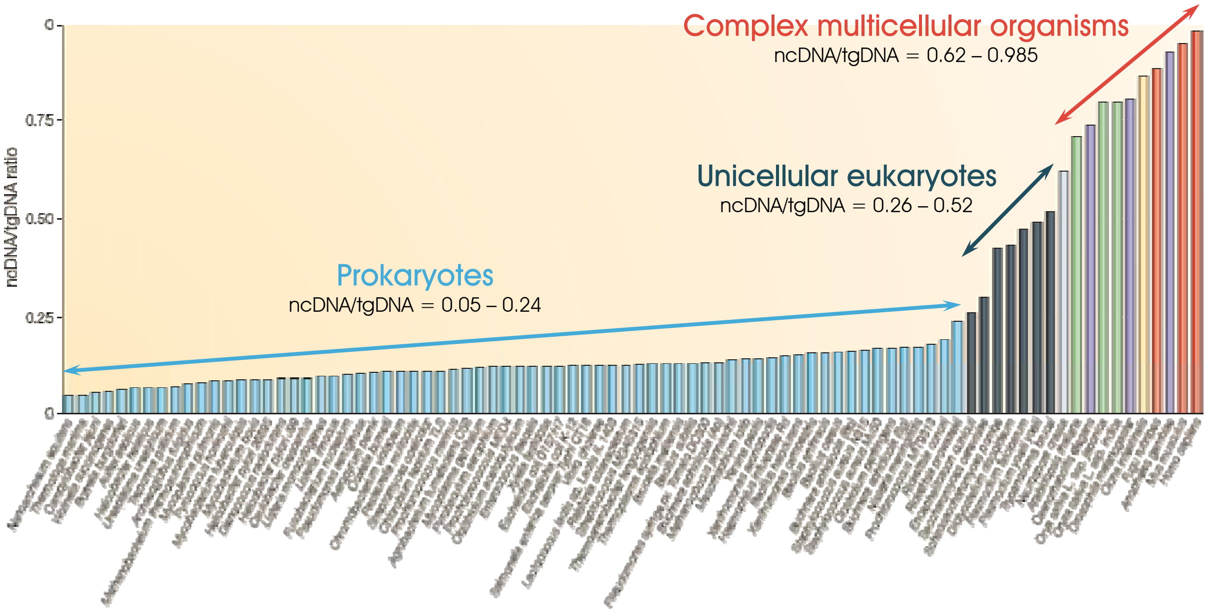


Figure 8

The percentage of noncoding DNA correlates with biological complexity

The increase in the ratio of noncoding DNA to total genomic DNA (ncDNA/tgDNA) is shown to correlate with increasing biological complexity (corrected for ploidy). Blue, prokaryotes (bacteria and archaea). Black, unicellular eukaryotes. Grey, the multicellular fungus *Neurospora crassa*. Green, plants. Purple, non-chordate invertebrates (nematodes and insects). Yellow, the urochordate *Ciona intestinalis*. Red, vertebrates. *[Figure modified from MATTICK (2004).]*

9A). Lynch argues that, because each of the previous features increases the genic mutation rate to defective alleles and is potentially harmful, their origin in eukaryotes must reflect the reduced efficacy of selection in this lineage due to their dramatically reduced population sizes —especially in multicellular species— compared to prokaryotes. The consequent increase in the intensity of random genetic drift appears to be sufficient to overcome the weak mutational disadvantages associated with the eukaryotic gene structure, and thus most eukaryotic gene novelties could simply result from semi-neutral processes rather than natural selection. However, in a second phase, the eukaryotic condition would promote a reliable resource from which natural selection could build novel forms of organismal complexity (LYNCH and CONERY 2003; LYNCH 2006).

Differences in the rates of polymorphism and divergence among different gene regions are usually attributed to differences in the intensity of purifying selection (i.e. selection against deleterious mutations) affecting these sites, such that regions facing stronger functional constraints are more sensitive to selection and evolve slower (FAY *et al.* 2001). Thus, the rate of evolution of a region or type of region gives an idea of their functional significance. Nonsynonymous substitutions at coding regions evolve at the slowest rate of all kinds of substitution (Figure 9B), implying that they face the strongest selective constraints (Table 8). This is expected since changes at these sites imply amino acid replacements that may impair protein function. On the contrary, synonymous substitutions evolve much faster (Figure 9B) because they do not alter the amino acid sequence, although they too face some form of constraint (Table 8) mainly due to codon preference (AKASHI 1995). As a result, while the rate of synonymous substitutions is more or less uniform across genes (resembling the neutral mutation rate), the rate of nonsynonymous substitutions varies widely among different genes according to the severity of the functional constraint to which they are exposed to (LI *et al.* 1985a).

ANDOLFATTO (2005) reports strong evidence of selection acting in *Drosophila* ncDNA, resembling general patterns of protein evolution in the same species (FAY *et al.* 2002; SMITH and EYRE-WALKER 2002) (Figure 9, Table 8). On the one hand, his results suggest that ~40-70% of the intergenic DNA, UTRs and introns is evolutionarily constrained relative to synonymous sites (Table 8). Noticeably, the level of selective constraint both in introns and intergenic regions is positively correlated with sequence

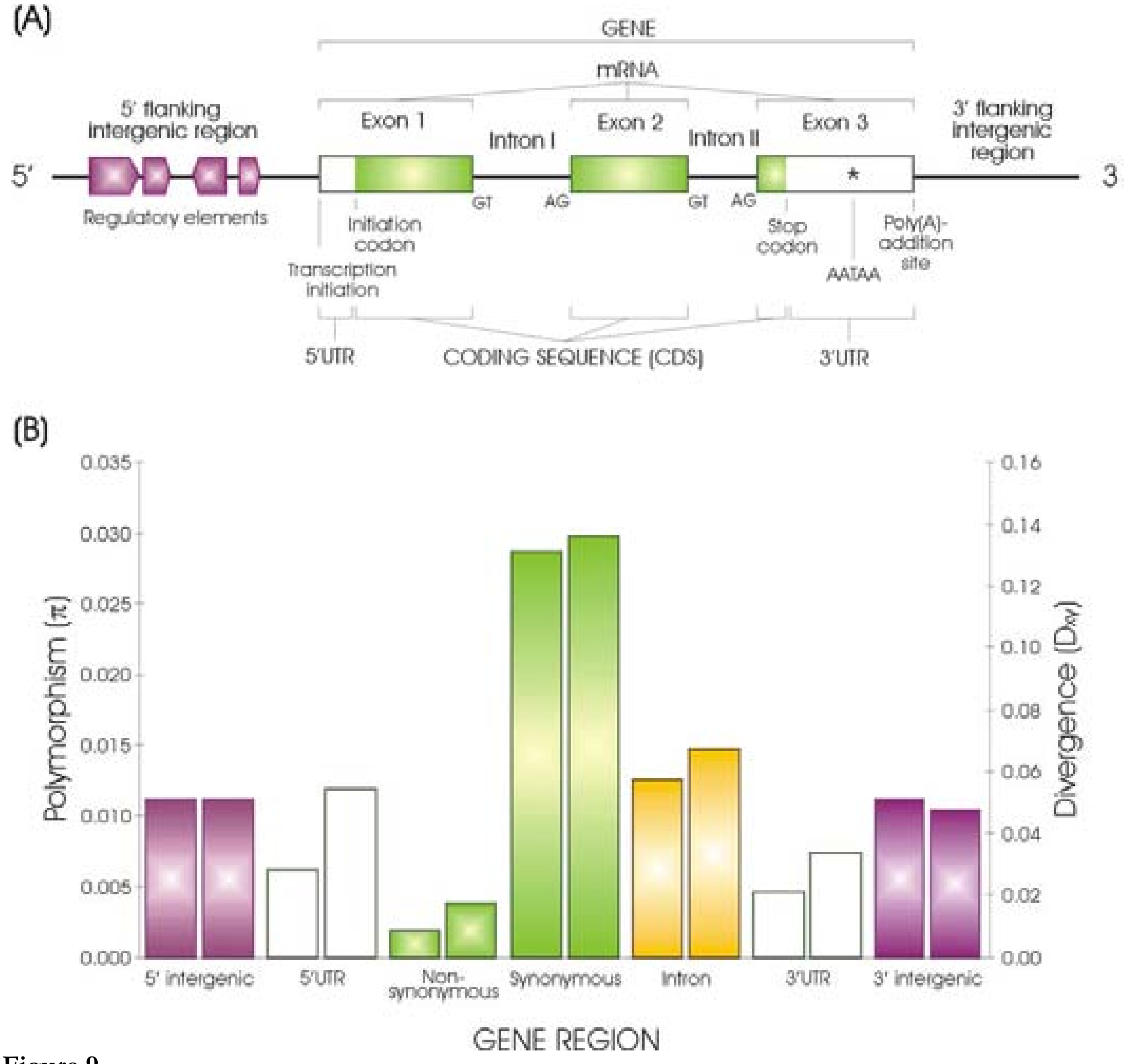


Figure 9
Polymorphism and divergence along the gene

(A) Different functional regions of the gene. Green rectangles denote exons; a white area in a green rectangle denotes a transcribed but untranslated region, while a shaded area denotes a translated region. (B) Estimates of polymorphism (π, left) and divergence (D_{xy}, right) in coding and noncoding DNA of *D. melanogaster*. π is the weighted average within-species pairwise diversity per site, according to ANDOLFATTO (2005); for intergenic regions, π is estimated for 5' and 3' regions altogether ($\pi_{\text{intergenic (5'+3')}}$ = 0.0111). D_{xy} for UTRs, coding sites and introns is the weighted average pairwise divergence per site between *D. melanogaster* and *D. simulans*, corrected for multiple hits (Jukes-Cantor), according to ANDOLFATTO (2005). D_{xy} for intergenic regions is the observed number of substitutions per site between *D. melanogaster* and *D. simulans*, according to HALLIGAN and KEIGHTLEY (2006).

Table 8 Estimates of constraint per base pair for different classes of site in *Drosophila*

Site class	*Constraint per site*	*Relative to*	*Reference*
Synonymous coding sites	0.126	FEI sites	*HALLIGAN and KEIGHTLEY (2006)*
Non-synonymous coding sites	0.862	FEI sites	*HALLIGAN and KEIGHTLEY (2006)*
UTRs	0.604	4-fold deg.	*ANDOLFATTO (2005)*
5'UTRs	0.529	4-fold deg.	*ANDOLFATTO (2005)*
3'UTRs	0.707	4-fold deg.	*ANDOLFATTO (2005)*
Introns	0.395	4-fold deg.	*ANDOLFATTO (2005)*
Introns (≤80 bp)	0.196	FEI sites	*HALLIGAN and KEIGHTLEY (2006)*
Introns (>80 bp)	0.531	FEI sites	*HALLIGAN and KEIGHTLEY (2006)*
Intergenic regions	0.493	4-fold deg.	*ANDOLFATTO (2005)*
Proximal intergenic (≤2 kb)	0.406	4-fold deg.	*ANDOLFATTO (2005)*
Distant intergenic (>4 kb)	0.546	4-fold deg.	*ANDOLFATTO (2005)*
5' intergenic	0.558	FEI sites	*HALLIGAN and KEIGHTLEY (2006)*
3' intergenic	0.585	FEI sites	*HALLIGAN and KEIGHTLEY (2006)*

Synonymous and non-synonymous coding sites are defined as fourfold degenerate and non-degenerate coding sites, respectively. Proximal and distant intergenic regions are according to their distance from the nearest gene. FEI = fastest evolving intronic sites. 4-fold deg = fourfold degenerate coding sites.

length (e.g. long introns (>80 bp) have more than twice selective constraints than short introns (≤80 bp); see Table 8) (HALLIGAN and KEIGHTLEY 2006). On the other hand, he estimates that ~20% of all intronic and intergenic substitutions and ~60% of UTR substitutions have been driven to fixation by positive selection (ANDOLFATTO 2005). Overall, ~47-63% of introns and intergenic regions, and >80% of UTRs might be subject to either positive or negative selection. Andolfatto's results emphasize the functional significance of *Drosophila* ncDNA, supporting their role in transcription initiation, termination and the regulation of expression.

1.4.3. CONSERVED NONCODING SEQUENCES AND THE FRACTION OF FUNCTIONALLY IMPORTANT DNA IN THE GENOME

The genomes of eukaryotic species have been shown to contain blocks of conservation when compared to other related species, part of which lie in regions of ncDNA (Figure 10) (SIEPEL *et al.* 2005). These conserved noncoding sequences (CNSs) are thought to represent the signature of functionally constrained elements maintained by purifying selection in a background of neutrally-evolving, possibly non-functional DNA (CLARK 2001). In fact, HALLIGAN and KEIGHTLEY (2006) have shown that most deleterious mutations occur in ncDNA, and that CNSs tend to be clustered in blocks of constrained nucleotides presumably involved in regulating gene expression. Indeed, CNSs have been extensively used to guide the prediction of *cis*-regulatory regions (BERGMAN *et al.* 2002; COSTAS *et al.* 2004; NEGRE *et al.* 2005) and functional noncoding RNAs (ncRNAs) (ENRIGHT *et al.* 2003; LAI *et al.* 2003). However, CLARK (2001) came up with the 'mutational cold-spot' hypothesis as an alternate hypothesis to explain the existence of CNSs without the action of purifying selection. This model proposes that extremely low mutation rates varying over short spatial scales (e.g. on the order of tens of base pairs) are

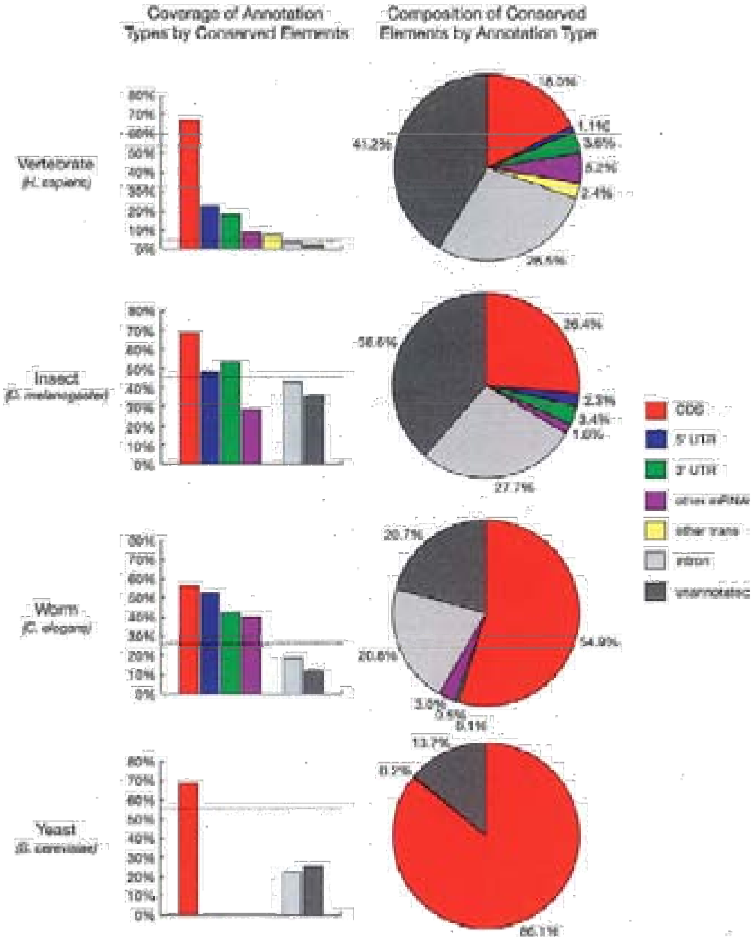


Figure 10
Conserved sequences in different types of DNA

Fractions of bases of various annotation types covered by predicted conserved elements (left) and fractions of bases in conserved elements belonging to various annotation types (right). Annotation types: coding regions (CDS), 5' and 3' UTRs of known genes, other regions aligned to mRNAs or spliced ESTs from GENBANK (other mRNA), other transcribed regions according to data from Phase 2 of Affymetrix/NCI Human Transcriptiome project (other trans), introns of known genes, and other unannotated regions (i.e. intergenic). Dashed lines in column graphs indicate expected coverage if conserved elements were distributed uniformly. *[Figure from* SIEPEL *et al. (2005).]*

responsible for the high conservation of these regions. Arguments against it include the nonrandom distribution of CNSs in flies (BERGMAN *et al.* 2002; HALLIGAN and KEIGHTLEY 2006) and worms (WEBB *et al.* 2002), which would require a nonrandom distribution of mutation rates as well. However, and despite that no molecular mechanism has been described to produce such localized mutation cold-spots, local variations in the mutation rates along ncDNA remains a formal possibility that must be investigated more thoroughly to demonstrate that CNSs are indeed maintained by the action of purifying selection and are not the result of mutational cold-spots.

According to the predictions of SIEPEL *et al.* (2005), only 3-8% of the human genome is conserved with other vertebrate species (Figure 10). Furthermore, current estimates of the fraction of functionally important segments in mammalian ncDNA range from ~10-15% (SHABALINA *et al.* 2001) to just ~3% (WATERSTON *et al.* 2002; SIEPEL *et al.* 2005). This supports the idea that mammalian genomes contain large amounts of 'junk' DNA (VENTER *et al.* 2001). Conversely, 37-53% of the *Drosophila* genome is conserved with other insect species (SIEPEL *et al.* 2005), and at least ~20-30% of the ncDNA is included in CNSs (BERGMAN and KREITMAN 2001; BERGMAN *et al.* 2002; SIEPEL *et al.* 2005) (Figure 10). This observation led KONDRASHOV (2005) to define two classes of eukaryotic genomes. First, compact and mostly functional *Drosophila*-like genomes are characterized by keeping 'junk' DNA at a reduced fraction by efficient selection: very few transposable element (TE) insertions (BARTOLOME *et al.* 2002; BERGMAN *et al.* 2002; QUESNEVILLE *et al.* 2005), short intronic sequences and few pseudogenes (HARRISON *et al.* 2003). Second, bloated and mostly neutrally-evolving mammal-like genomes contain long segments of 'junk' DNA due to very inefficient purifying selection in these species —presumably because of their low effective population sizes—.

BEJERANO *et al.* (2004) have recently discovered long segments (>200 bp) of DNA that are absolutely conserved (100% identity with no insertions or deletions (indels)) between orthologous regions of the human, rat and mouse genomes, and with 95-99% identity with the chicken and dog genomes. Most of these ultraconserved segments are noncoding, mainly located in introns or nearby genes involved in the regulation of transcription and development. Their extreme conservation since the divergence of mammals and birds >300 MYA suggests that they may perform functions that are indispensable for viability or reproduction. However, other recent studies show contradictory results. While KATZMAN *et al.* (2007) show that these regions are under strong negative selection (i.e. much stronger than protein coding genes), CHEN *et al.* (2007a) find numerous polymorphisms within these regions in the human population that may hold only subtle phenotypic consequences. More intriguingly, AHITUV *et al.* (2007) have recently reported knockout mice showing almost no ill effects at all. It has been speculated that ultraconserved segments could be mutational cold-spots, or regions where every site is under weak but still detectable negative selection. The true reason for their extreme conservation still remains a mystery.

1.4.4. THE ROLE OF NCDNA IN MORPHOLOGICAL DIVERSITY

Although it is feasible that some of the eukaryotic ncDNA could be truly non-functional, the fact that most of our genetic and morphological complexity originates in these noncoding regions is widely recognized (CARROLL *et al.* 2001; CARROLL 2005). It is indeed within these regions where complex regulatory signals orquestrate when, where and how much genes are translated, inducing from subtle to major pleiotropic changes in gene expression and, therefore, phenotype (MARKSTEIN *et al.* 2002; DE MEAUX *et al.* 2005). While orthologs of many developmental genes can even be identified at species that split during the bilaterian radiation >500 MYA (DE ROSA *et al.* 1999), the conservation of noncoding elements and RNAs is restricted to more related lineages (COOPER and SIDOW 2003; SIEPEL *et al.* 2005). Regulatory DNA may tolerate mutational change better than coding DNA does, and this would allow genetic interactions to evolve without changing the number of genes or even the protein sequences. But, which are the molecular components of this gene regulation in complex organisms? Do intronic and intergenic regions participate equally in this gene regulation? And how is this regulation related to changes in morphology? The recent availability of multiple complete genomes and powerful genome comparison tools open unlimited opportunities to unveil the real meaning of ncDNA.

1.5. *HOX* GENES AND THEIR ROLE IN THE EVOLUTION OF MORPHOLOGY

Development is the process through which an egg becomes an adult organism. During this process, an organism's genotype is expressed as a phenotype, and the latter is exposed to the action of natural selection. Studies of development are important to evolutionary biology for several reasons. First, changes in the genes controlling development can have major effects on the adult's morphology, and thus it is thought that changes in developmental genes have driven large-scale evolutionary transformations. These genes should thus explain how some hoofed mammals invaded the ocean, or how small, armored invertebrates evolved wings. Because of their major effects on morphology, developmental processes may also constrain evolutionary change, possibly preventing certain characters from evolving in certain lineages. Thus, for example, development may explain why there are no six-fingered tetrapods. And finally, an organism's development may also contain clues about its evolutionary history, which can be used to disentangle relationships among different lineages. Then, comparisons among different lineages should provide answers to general questions such as: Does morphological evolution occur gradually or in big steps? Are there trends in evolution? Why are some clades very diverse and some unusually sparse? How does evolution produce morphological novelties? Which are the genes and gene networks involved in morphological evolution? Are they shared among species? These and other questions are within the scope of the relatively new discipline *Evolutionary Developmental biology (Evo-Devo)* (GILBERT 2003).

1.5.1. ORIGIN OF THE BODY PLAN IN ANIMALS

Life during the first 3 billion years on the Earth consisted of single-celled organisms only. Multicellular animals arose from one of these single-celled organisms related to choanoflagellates, a group that originated ~1 billion years ago. The most primitive living animal phyla are the sponges, forms of which have been found in Neoproterozoic fossils dating back 565 MYA. The Earth is now populated by 1-20 million animal species, probably <1% of all animal species that have ever existed, but strikingly all of their diversity was originated >540 MYA from a common bilaterally symmetrical ancestor. These are indeed the most important milestones in early animal history: (i) the evolution of bilaterally symmetric animals, and (ii) the explosive radiation of these forms in the Cambrian period >500 MYA. The 'Cambrian explosion' signified a burst of biological creativity unprecedented in the Earth's history. Many of these animals are now extinct, but those that remained established all of the basic body plans we see today. As a consequence of this ancient origin of today's phyla, all living animals belong to a limited number of basic designs, referred to as *Bauplan* or 'body plans' (ERWIN *et al.* 1997).

1.5.2. HOMEOTIC GENES AND THE DISCOVERY OF THE HOMEOBOX

Early interest in the development of body pattern was largely motivated by curiosity on the origin of the diversity of living species. As early as 1859, Darwin noticed a common feature of many creatures: the repetition of elements along the length of the body, today known as *segmentation* (DARWIN 1859). Some years later, in 1894, W. Bateson described one of the most extraordinary phenotypes ever described in animals that affected indeed the patterning of the body plan and body parts. He catalogued several cases in nature in which one normal body part was replaced with another, such as a leg in place of an antenna in arthropods, or a thoracic vertebra in place of a cervical vertebra in vertebrates, and termed this phenomenon *homeosis* (BATESON 1894). In 1923, C. B. Bridges and T. H. Morgan showed that homeosis was heritable in flies, and that whatever was responsible for such inheritance was coded in the fly's third chromosome (BRIDGES and MORGAN 1923). But it was not until half a century later that the genetic basis of homeosis could be unveiled (GARCIA-BELLIDO 1975; LEWIS 1978). E. B. Lewis studied the relationship genotype-phenotype of homeotic mutations at the fly's Bithorax Complex of genes (BX-C), and reported that this cluster consisted of various genetic elements and that mutations mapped in an order that corresponded to the anteroposterior (A/P) body axis of the embryo (spatial collinearity). He already predicted that the identity of an individual body segment was produced by a combination of different BX-C genes, and that these were activated in response to an A/P gradient. Shortly later, T. C. Kaufman's lab described a second homeotic complex affecting anterior regions of the fly's body, the Antennapedia Complex (ANT-C), and made similar predictions to those by Lewis (KAUFMAN *et al.* 1980; LEWIS *et al.* 1980a; LEWIS *et al.* 1980b).

Two teams, one led by M. P. Scott and the other including W. McGinnis, M. S. Levine and W. J. Gehring, showed in 1984 that genes involved in homeotic mutations — called *homeotic* genes— share a highly conserved sequence of 180 nucleotides (LAUGHON and SCOTT 1984; MCGINNIS *et al.* 1984; SCOTT and WEINER 1984). This sequence — called *homeotic box* or *homeobox*— codes for a 60 amino acid protein domain —the *homeodomain*— that binds particular sequences in the DNA through a 'helix-turn-helix' structure (Figure 11). This highly conserved sequence, which is not exclusive of homeotic genes, was soon used in homology searches to pull out more homeobox-containing genes, which could be easily identified in such disparate organisms as hydra (SCHUMMER *et al.* 1992; GAUCHAT *et al.* 2000), nematodes (WANG *et al.* 1993), leech (NARDELLI-HAEFLIGER and SHANKLAND 1992), amphioxus (HOLLAND *et al.* 1992), zebrafish

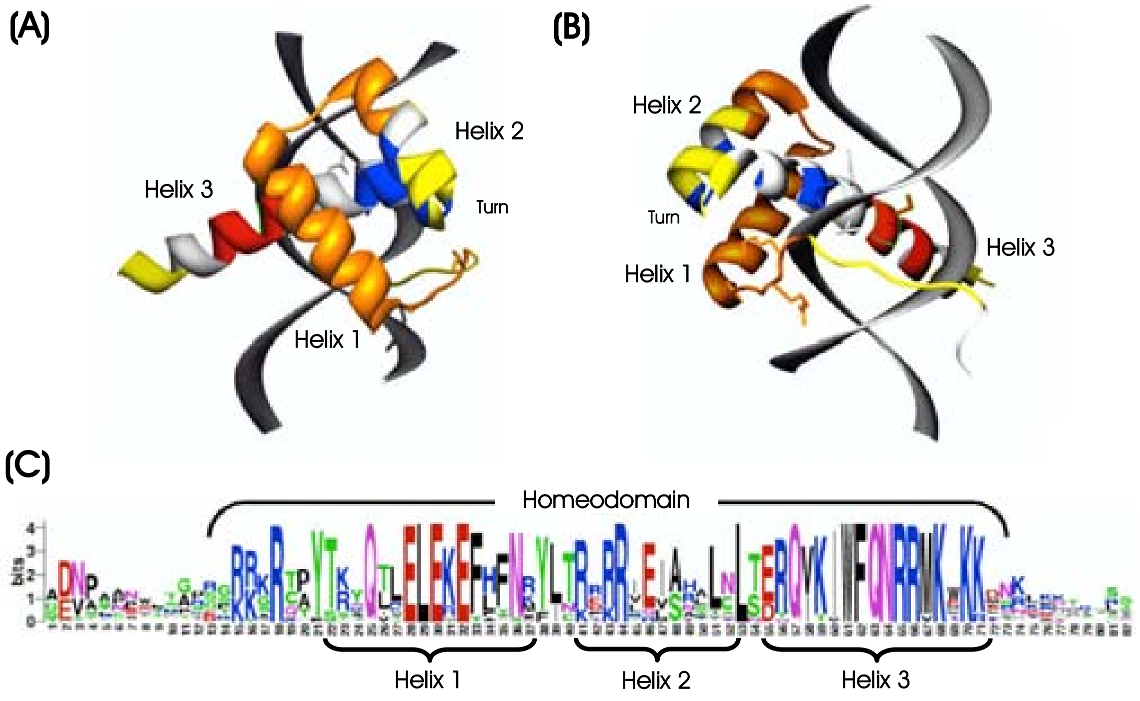


Figure 11
Structure and conservation of the homeodomain

(A,B) The structure of the homeodomain bound to DNA is shown as ribbon models. (C) Sequence logo of the homeodomain and surrounding amino acids for a set of ortholog and paralog *Hox* genes in several species of vertebrates. The overall height of the stacked amino acids indicates sequence conservation at each position, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position. *[Figure modified from* LYNCH *et al. (2006).]*

(NJOLSTAD and FJOSE 1988; NJOLSTAD *et al.* 1988) or humans (ACAMPORA *et al.* 1989). Indeed, this motif has been found in >200 non-homeotic genes in vertebrates and ~100 in invertebrates, all coding for DNA-binding proteins often involved in different aspects of animal development (NAM and NEI 2005).

1.5.3. THE *HOX* GENE COMPLEX

One of the most important biological discoveries of the past two decades is that most animals, no matter how divergent in form, share specific families of genes that regulate major aspects of body pattern, such as the determination of anterior *versus* posterior, or dorsal *versus* ventral (MCGINNIS 1994; ERWIN *et al.* 1997). The discovery of this common genetic 'toolkit' for animal development unveiled conserved molecular, cellular and developmental processes that were previously hidden by disparate anatomies. It also focused the study of the genetic basis of animal diversity on how the number, regulation and function of genes within the toolkit have changed over the course of animal history (DE ROSA *et al.* 1999; CARROLL *et al.* 2001).

Hox genes are an essential class of homeobox-containing genes involved in the specification of regional identities along the A/P body axis of the developing embryo (LEWIS 1978; KAUFMAN *et al.* 1980; MCGINNIS and KRUMLAUF 1992). They play as transcription factors (TFs) that modulate levels of expression of other genes located downstream in the regulatory cascade of development. Additionally, *Hox* genes have the following particularities: (i) they are usually clustered together in complexes (LEWIS 1978; KAUFMAN *et al.* 1980) (but see NEGRE and RUIZ (2007)), (ii) they are arranged in the chromosome in an order that corresponds to the A/P body axis of the embryo (*spatial collinearity*) (MCGINNIS and KRUMLAUF 1992) (Figure 12), (iii) they are expressed also in a temporal order that match their physical order on the chromosome (*temporal collinearity*) (DUBOULE 1994; KMITA and DUBOULE 2003), and (iv) they are universal in animals, suggesting that they are evolutionarily related (MCGINNIS and KRUMLAUF 1992; SLACK *et al.* 1993; GARCIA-FERNANDEZ 2005) (Box 3).

1.5.4. NEW FUNCTIONS FOR SOME INSECT *HOX* GENES

The fact that all animal species share the basic *Hox* gene content suggests that

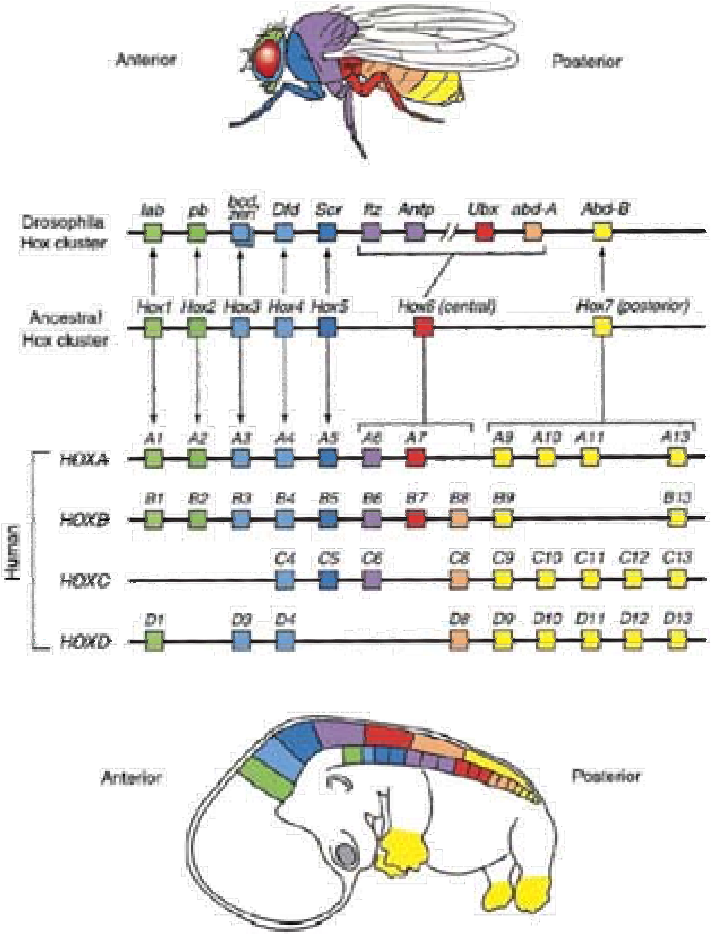


Figure 12
Conservation of the genomic structure and expression patterns of *Hox* genes

Hox gene complexes and expression patterns of *Drosophila* (top) and mammals (bottom). The hypothetical gene complement of the ancestral *Hox* cluster is shown in the middle. *[Figure from* VERAKSA *et al. (2000).]*

Box 3 Origin of the *Hox* gene complex

The reconstruction of the evolutionary history of the *Hox* gene family is a key issue to understand the evolution of body plans in bilateria and the relationships between genetic complexity and morphology. The *Hox* gene complex probably arose by tandem duplications and posterior divergence from an ancestral *Hox* gene (DE ROSA *et al.* 1999; FERRIER and MINGUILLON 2003; GARCIA-FERNANDEZ 2005). These duplicated genes have usually remained together in the genome; now, all metazoans show different configurations of the ancestral *Hox* gene cluster (Figure 13). Cnidarians, the most ancient animal phyla, have only one anterior and one posterior *Hox* genes. Before the bilaterian radiation, the *Hox* groups *Hox1-Hox5* were established and fixed. Then, early expansion of the *Hox* complex at the base of the bilaterian lineage generated many of the central *Hox* genes: *Hox6-Hox13* in deuterostomes, *Ubx* and *Abd-B* in ecdysozoans, and *Lox5*, *Lox2*, *Lox4*, *Post1* and *Post2* in lophotrochozoans. During early vertebrate evolution, the entire complex was duplicated; e.g. tetrapods have 4 complexes summing a total of 39 *Hox* genes. Teleost fish have undergone an additional round of tetraploidization, creating the seven *Hox* complexes found in zebrafish and at least five in *Fugu*.

Box 3 (continued)

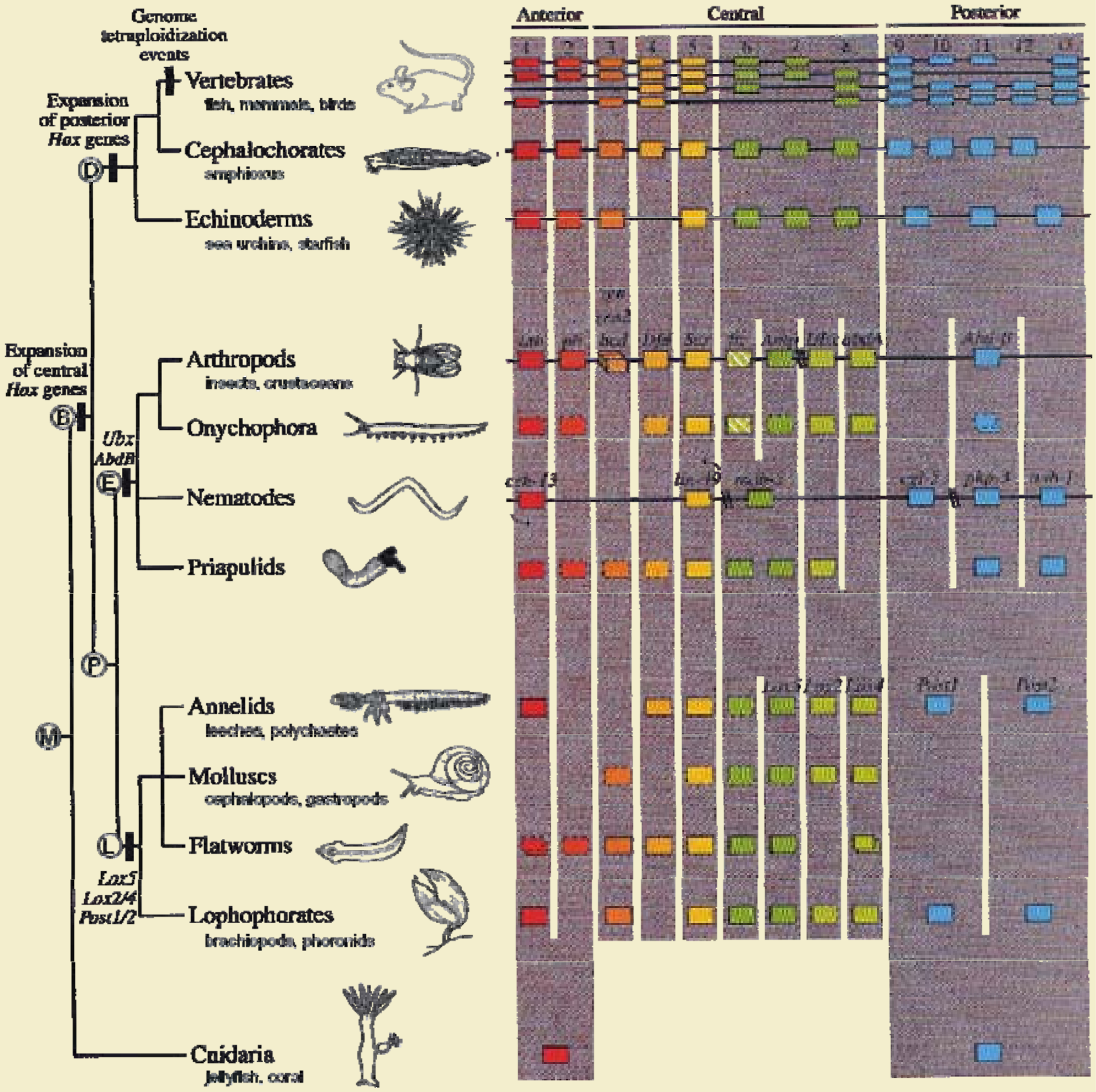


Figure 13

Evolution of metazoan *Hox* genes

The relative timing of *Hox* duplication events is mapped onto a phylogenetic tree of Metazoan phyla (left), as deduced from the distribution of *Hox* genes in the different species (right). Common ancestors: M, metazoan; B, bilaterian; D, deuterostome; E, stem ecdysozoan; L, stem lophotrochozoan; P, protostome. Striped colors indicate fast-evolving *Hox*-derived genes. *[Figure modified from* CARROLL *et al. (2001).]*

evolutionary changes in gene regulation —and not gene content— might have been the key in shaping large-scale changes in animal body plans and body parts. In particular, differences in the spatial and temporal regulation of *Hox* genes have been correlated to changes in axial morphology in many comparative analyses of *Hox* gene expression in arthropods, annelids and vertebrates (CASTELLI-GAIR *et al.* 1994; MANN 1994; CASTELLI-GAIR and AKAM 1995). Such differences in *Hox* expression domains during evolution are most probably be explained by changes in the *cis*-regulatory regions of *Hox* genes and/or changes in the expression of their *trans*-acting regulators (BELTING *et al.* 1998; DOEBLEY and LUKENS 1998; WEATHERBEE and CARROLL 1999). The logic behind this statement is related to the pleiotropy of mutations. In general, it is expected that mutations with far-reaching effects will have more deleterious consequences on organismal fitness and will be a less common source of variation than mutations with less widespread effects. While mutations in a single *cis*-regulatory element affect gene expression only in the domain governed by that element, changes in the coding region of a TF may directly affect all of the genes it regulates, and thus, have broad effects in the developing organism (CARROLL 2005). On these grounds, any modification in a *Hox* gene pathway might produce changes in animal morphology, and the extent of these morphological changes may correlate with the pleiotropy of the modification (GELLON and MCGINNIS 1998). Notably, the genome tetraploidization events at the base of vertebrates and further genome duplications in fish might have been responsible for the huge morphological diversity in these lineages. Yet, the basic components and the biochemical functions of the encoded proteins are surprisingly conserved across hundreds of millions of years.

However, some members of the insect *Hox* complex have shown a relaxation in their constraint and have evolved new functions. In winged insects, including *Drosophila*, *Hox3* (STAUBER *et al.* 1999; STAUBER *et al.* 2002; BONNETON 2003) and *fushi tarazu* (*ftz*) (TELFORD 2000) have lost their *Hox*-like role in regulating regional identity along the A/P body axis and acquired new functions in animal development (HUGHES *et al.* 2004). *Hox3* gained a novel extraembryonic function, and underwent two consecutive duplications that gave rise to *bicoid* (*bcd*), *zerknüllt* (*zen*) and *zerknüllt-related* (*zen2*) (hereafter called *Hox*-derived genes) (Figure 14). The first duplication took place in the cyclorrhaphan fly lineage and gave rise to *zen* and *bcd* (STAUBER *et al.* 1999; STAUBER *et al.* 2002). Afterwards, but before the *Drosophila* radiation, *zen* went through a second duplication that gave birth to *zen2* (NEGRE *et al.* 2005). The gene *zen* is expressed in extraembryonic tissue during early development in several different insects, indicating that the shift in *zen* function occurred fairly early in insect evolution (PANFILIO and AKAM 2007). The gene *bcd* codes for an important morphogen that establishes A/P polarity during oogenesis (BERLETH *et al.* 1988). *zen2* has the same expression pattern of *zen*, although its function is unknown. These duplicated copies of *Hox3* may have experienced a period of accelerated evolution following duplication during which they may have adopted part of the functions of their parental gene (*subfunctionalization*) and/or acquired new functions (*neofunctionalization*) (LYNCH and CONERY 2000; LYNCH and FORCE 2000; LONG *et al.* 2003; ZHANG 2003). This rapid evolution has already been demonstrated for the homeodomains of all of these genes (and especially that of *zen2*) (DE ROSA *et al.* 1999), which might have facilitated the rapid functional evolution of these genes in the development of insects.

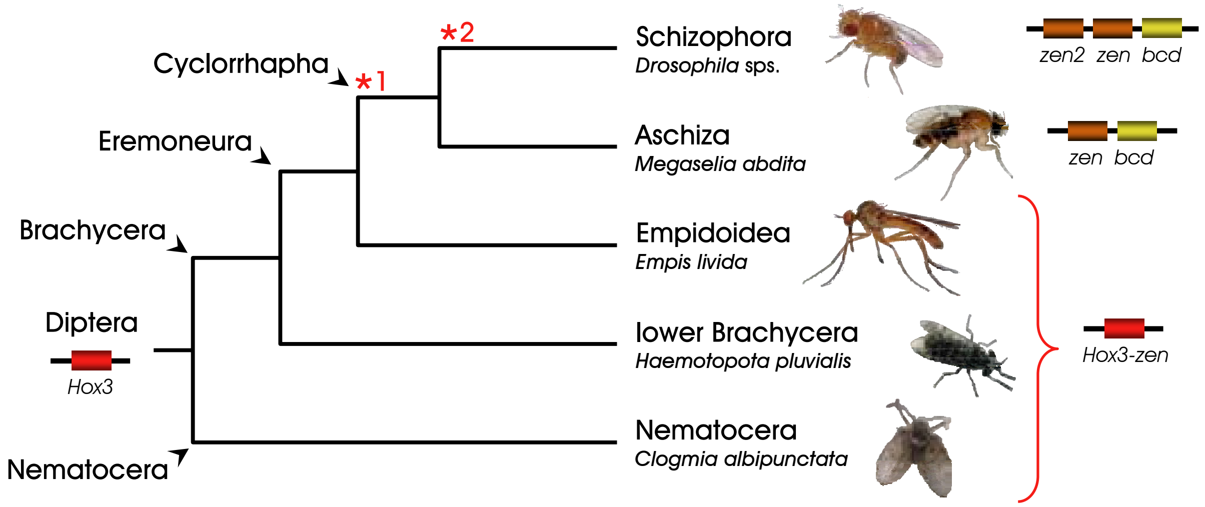


Figure 14

Origin of *bcd*, *zen* and *zen2* from *Hox3*

The composition of *Hox3*-related genes is shown for the major groups of diptera and for their inferred ancestor. Duplications are shown as asterisks. See text for details. (Note that the image for the Aschiza group corresponds to *Megaselia scalaris*).

1.6. *DROSOPHILA* AS A MODEL ORGANISM

Arthropods are certainly the most successful animal taxa, with insects alone accounting for ~75% of all known animal species. *Drosophila* encompasses ~2,000 recognized species that form a separate group within the insect lineage, although probably more species still have to be discovered (POWELL 1997). Phylogenetic analyses indicate the existence of two main lineages within the genus *Drosophila* which diverged 40-62 MYA (RUSSO *et al.* 1995; TAMURA *et al.* 2004) (Figure 15). One of the lineages led to the subgenus Sophophora, with ~330 recognized species among which *D. melanogaster* resides. The second lineage gave rise to the subgenera Drosophila and Idiomyia (Hawaiian Drosophila), which include ~1100 and ~380 identified species respectively.

The fruit fly *Drosophila melanogaster* is probably the most successful experimental model ever used in the lab (ROBERTS 2006). The *Drosophila* genome (~176 Mb) is on average 5% the size that of mammals, but it still shares with them most gene families and

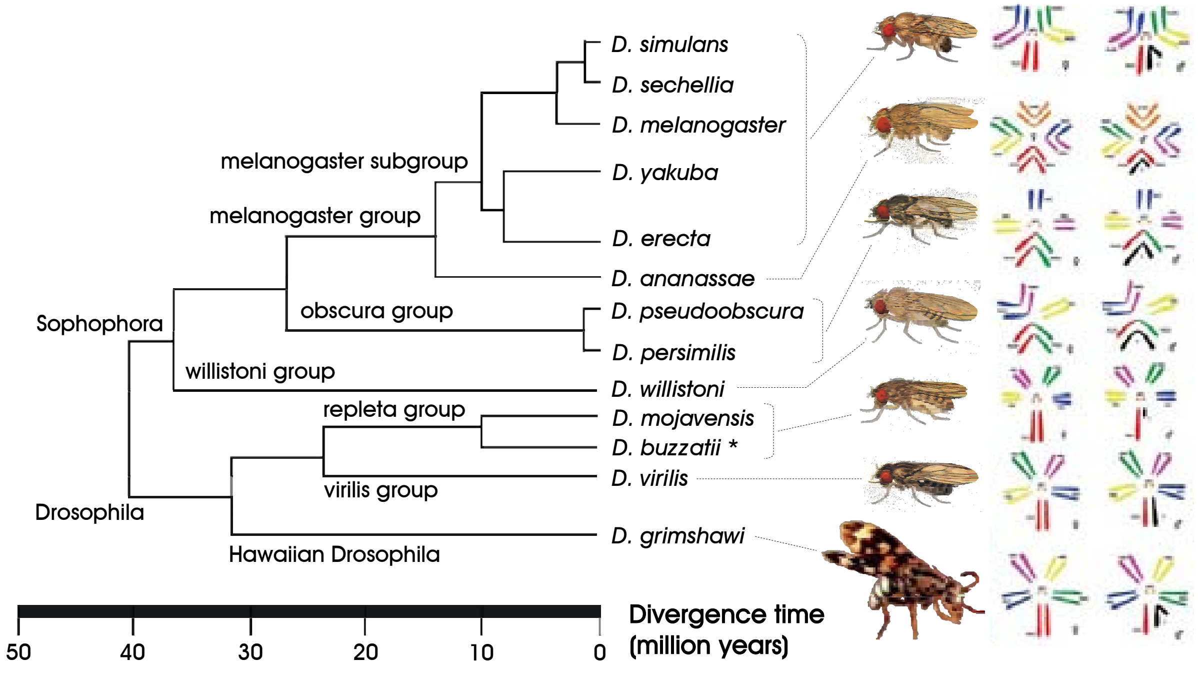


Figure 15
Phylogeny of the *Drosophila* genus

Phylogenetic tree of 12 *Drosophila* species according to AAA, plus *D. buzzatii* (* note that the complete genomic sequence is not yet available for this species). *[Fly drawings are from* FLYBASE*.]*

pathways, as well as some tissues and organ systems (De Velasco *et al.* 2004; Kida *et al.* 2004). It was introduced as a research tool in the early twentieth century, when it played a crucial role during the first steps of genetics (Morgan *et al.* 1915; Muller 1927). Since then, it has been at the leading edge of research into a wide range of fields, providing valuable information about the mechanisms of inheritance (Morgan *et al.* 1915), the forces affecting genetic variation in natural populations (Ayala *et al.* 1974; Singh and Rhomberg 1987), the construction of the animal body plan (Lewis 1978; Nusslein-Volhard and Wieschaus 1980), and the function of the nervous system (Ivanov *et al.* 2004). Notably, *Drosophila* has undoubtedly been the experimental model par excellence to test the basic assumptions of the population genetics theory, being crucial in three basic lines of research: (i) the study of chromosomal evolution (Dobzhansky and Sturtevant 1938), (ii) electrophoretic variability (Lewontin 1974), and (iii) nucleotide variation (Kreitman 1983). Few organisms have been so much exploited in research as has been *Drosophila*.

More recently, *D. melanogaster* has greatly contributed to major advances in the genomics field (Miklos and Rubin 1996; Rubin and Lewis 2000; Celniker and Rubin 2003). Its historical importance, including a long course in genetic research and powerful analysis tools, together with its modest genome size, contributed to the election of the little fruit fly to explore the application of complete genome sequencing by whole-genome shotgun (WGS) in eukaryotic genomes (Rubin 1996; Adams *et al.* 2000). *D. melanogaster* was also the third eukaryote and the second metazoan that was ever sequenced, after the yeast *Saccharomyces cerevisiae* (Goffeau *et al.* 1996) and the nematode *Caenorhabditis elegans* (Consortium 1998). Since then, the genomic era has markedly accelerated the development of research resources for *D. melanogaster* (Matthews *et al.* 2005), including a multitude of specialized databases and sequence analysis tools (Matthews *et al.* 2005; Fox *et al.* 2006; Galperin 2007). Many of these resources have allowed the improvement of the *D. melanogaster* initial genome sequence, both in quality and richness (Ashburner and Bergman 2005), including the closure of pre-existing gaps (Celniker *et al.* 2002; Hoskins *et al.* 2002) and the extensive definition of functional annotations (Kopczynski *et al.* 1998; Berger *et al.* 2001; Berman *et al.* 2002; Kaminker *et al.* 2002; Misra *et al.* 2002; Ohler *et al.* 2002; Carvalho *et al.* 2003; Bergman *et al.* 2005; Tupy *et al.* 2005). The current availability of many complete sequenced genomes of related species is leading to powerful studies of comparative genomics (Rubin *et al.* 2000; Lewis *et al.* 2003; Consortium 2007; Negre and Ruiz 2007; Zhu and Buell 2007). Within the *Drosophila* genus, a coordinated leading project to assemble, align and annotate another 11 *Drosophila* genomes (Figure 15, see AAA) is providing the fly lineage with a boundless resource (Ashburner and Bergman 2005). Furthermore, other non-*Drosophila* insect species have been also sequenced or their genomes are in progress, including those of *Anopheles gambiae* (malaria mosquito) (Holt *et al.* 2002; Kaufman *et al.* 2002; Mongin *et al.* 2004; Sharakhova *et al.* 2007), *Apis mellifera* (honey bee) (Consortium 2006), *Bombyx mori* (silkworm) (Mita *et al.* 2004; Xia *et al.* 2004), and *Tribolium castaneum* (red flour beetle). As a result, we are entering a fascinating era in which deep knowledge can be obtained from comparative genomics in *Drosophila*.

1.7. Objectives

This thesis is a comprehensive bioinformatics and population genetics project centered on nucleotide polymorphism and divergence. It is accomplished in three sequential steps (Table 9): (i) the development of tools for data mining, processing, filtering and quality checking of raw data, (ii) the generation of databases of knowledge from refined data obtained in the first step, and (iii) the testing of hypotheses which require the multi-species and/or multi-locus data that has been obtained. As a result, and in spite of the apparent heterogeneity of the works presented, they all round a complete bioinformatics project off, including all the necessary steps from mining the data to generating new scientific knowledge.

1.7.1. Development of tools for data mining, processing, filtering and quality checking of raw data

The first objective of this thesis is to develop an elaborated bioinformatic system to mine all the haplotypic sequences that are stored in the major DNA sequence repositories and transform them into solid population genetics data that can be used in

Table 9 The three steps of this bioinformatics project focused on genetic diversity

Step	*Outcome*	*Associated publications*
1. Tools for data mining, processing, filtering and quality checking of raw data	Pipeline Diversity Analysis A tool to explore and estimate nucleotide diversity at DNA databases	*CASILLAS and BARBADILLA (2004)* *CASILLAS and BARBADILLA (2006)*
2. Databases of knowledge from refined data		*CASILLAS et al. (2005)* *CASILLAS et al. (2007b)*
		EGEA et al. (2007) (see Appendix I)*
3. Multi-species and/or multi-locus analyses of genetic diversity using specific sets of data	Purifying selection maintains highly conserved noncoding sequences in *Drosophila*	*CASILLAS et al. (2007a)*
	Fast sequence evolution of *Hox* and *Hox*-derived genes in the genus *Drosophila*	*CASILLAS et al. (2006)*
	Protein polymorphism is negatively correlated with conservation of intronic sequences and complexity of expression patterns in *D. melanogaster*	*PETIT et al. (2007)* (see Appendix II)*

* These works are part of other theses in our group, thus reflecting the many applications that the system created in the first step has to generate new knowledge other than that presented here

large-scale studies of genetic diversity. To this end, we have created a pipeline to automate the extraction of haplotypic sequences from GENBANK for any gene or species, align all the homologous regions and describe their levels of polymorphism. The creation of such a bioinformatic system requires the development of new algorithms to solve difficulties associated with sequence grouping and alignment, methods to validate the source data and the obtained estimates, efficient modules to manage and represent polymorphic data, and optimized processes for the extraction and analysis of large amounts of data. Thus, this is a fully entitled scientific objective, with a problem definition, search for creative solutions and important contributions to knowledge, as is any other research with more empirical objectives or focused on the analysis and interpretation of data.

All the extracted and analyzed data resulting from the first objective of this thesis is stored in structured relational databases in order to facilitate flexible data retrieving and subsequent data analysis. Thus, the previous system is able to create comprehensive databases of knowledge containing estimates of polymorphism for any biological species that have haplotypic sequences stored in GENBANK.

1.7.2. Generation of databases of knowledge

The second objective of this thesis is to use the system to create a comprehensive on-line resource that provides searchable collections of polymorphic sequences with their associated diversity measures in the genus *Drosophila*, by developing a robust platform to manage the data and distribute it to the scientific community through the Web. The database is updated daily, feeding on new sequences as they are introduced in GenBank and recalculating the corresponding estimates of nucleotide diversity. This resource is an ambitious pledge to test the efficiency of the system created in the first step.

1.7.3. Testing of hypotheses

The modules of data mining and analysis developed in the first step are a useful resource for other genomic analyses. One of these analyses, which is another contribution to this thesis, concerns the study of patterns of sequence evolution to infer constraint and adaptation in *Drosophila* CNSs. For this study we have used population genetics re-sequencing data from *D. melanogaster* together with comparative genomic data from other *Drosophila* species. The main issues considered are: (i) to investigate the evolutionary forces governing the evolution of CNSs in the *Drosophila* genome and determine whether they are functionally constrained (and thus potential *cis*-regulatory regions or noncoding RNAs), or rather a result of mutation cold spots; (ii) if CNSs are selectively constrained, determine the proportion of CNS sites that are under selection and quantify the amount of selective pressure acting on them; and (iii) determine whether or not positive selection is a main force driving the evolution of ncDNA as a means of organism adaptation. The already availability of our analytic system allowed the gathering and management of the data, even though further modules were needed to enable the system to cope with noncoding sequences without genic annotations and to solve specific tasks for the project.

Finally, another multi-locus study that also required the use of our bioinformatic tools for evolutionary research is presented. Before starting the work presented here, we had already obtained the sequences of two regions of the *D. buzzatii* genome (Figure 15) containing, among others, three *Hox* genes (*labial* (*lab*), *abdominal-A* (*abd-A*) and *proboscipedia* (*pb*)) and the three fast-evolving genes derived from *Hox3* (*bcd*, *zen* and *zen2*), adding up to a total of 257 kb (NEGRE *et al.* 2005) (see Appendix III). The objective of the previous study was twofold: (i) to determine the breakpoints of two HOM-C splits present in *D. buzzatii*, and (ii) to investigate the functional consequences of these splits on the three *Hox* genes. The comparison of the *D. buzzatii* sequence to the homologous regions in *D. melanogaster* and *D. pseudoobscura* revealed that both breakpoints had conserved all the *Hox* and *Hox*-derived gene structures intact, keeping every transcriptional unit with its *cis*-regulatory regions as permanent blocks. As a result, *Hox* genes did not show any functional alteration in *D. buzzatii* compared to its relatives. These results argued in favor of the most common sneaking feeling that, besides the strong conservation of *Hox* genes' function and their regulatory regions, their coding sequences are also highly conserved in evolution, probably due to their important function in early development (RIEDL 1978; POWELL *et al.* 1993; DAVIS *et al.* 2005). However, developmental biologists have long noticed that a large portion of the sequence of *Hox* proteins diverges so fast that it is difficult to align homologues from different arthropod classes, and several microsatellites have also been described in the coding sequences of these genes. The main goals of the work presented here are: (i) to measure the rates of coding evolution in the three *Hox* genes and resolve whether they are evolving fast despite their essential function in development, or rather their functional conservation is also reflected at the sequence level; and (ii) to measure the effect of the homeobox and the repetitive regions on the overall estimates of nucleotide divergence and indel fixation of these genes.

It is worth noting that the outcome of every step presented in this thesis is the seed of multiple possible studies in the next step, such that it generates the necessary tools and data from which other studies can be set out. For example, tools generated in the first step can be used to generate multiple databases of knowledge (e.g. databases of genetic diversity in different taxa, such as *Drosophila* or mammals), and these databases can in turn be used to answer many interesting questions in population genetics, either in a specific taxon or combining data from many taxa. Thus, this thesis has many applications for the scientific community and some works have already been elaborated by other members of our group during the development of this thesis (see Table 9).

Part 2

Results

2
Results

2.1. Pipeline Diversity Analysis (PDA): a bioinformatic system to explore and estimate polymorphism in large DNA databases

This first part of the results includes two publications corresponding to the two released versions of PDA –Pipeline Diversity Analysis–. PDA is a web-based tool, mainly written in PERL, that automatically extracts heterogeneous haplotypic sequences from GENBANK for any gene or species, aligns all the homologous regions and describes their levels of polymorphism. The first release of PDA was published in the 2004 Web Server issue of *Nucleic Acids Research*. The paper describes the collection of modules that make up the tool and how data flows through a pipeline structure, the output of a typical query, and the main limitations of this first version of the software. Two years later, we published an improved version of PDA in the special issue of 2006. In this second release, PDA incorporated new methods for data mining and grouping of sequences, new criteria for data quality assessment, additional implementations of aligning software, and a completely renewed interface with more functionality. The second paper describes these improvements made on PDA v.2. The most recent version of PDA is available on the web at http://pda.uab.cat/.

- **Article 1:** CASILLAS, S. and A. BARBADILLA (2004) PDA: a pipeline to explore and estimate polymorphism in large DNA databases. *Nucleic Acids Research* **32:** W166-169.

- **Article 2:** CASILLAS, S. and A. BARBADILLA (2006) PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. *Nucleic Acids Research* **34:** W632-634.

W166–W169 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh428

PDA: a pipeline to explore and estimate polymorphism in large DNA databases

Sònia Casillas and Antonio Barbadilla*

Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

Received February 15, 2004; Revised and Accepted April 13, 2004

ABSTRACT

Polymorphism studies are one of the main research areas of this genomic era. To date, however, no available web server or software package has been designed to automate the process of exploring and estimating nucleotide polymorphism in large DNA databases. Here, we introduce a novel software, PDA, *Pipeline Diversity Analysis*, that automatically can (i) search for polymorphic sequences in large databases, and (ii) estimate their genetic diversity. PDA is a collection of modules, mainly written in Perl, which works sequentially as follows: unaligned sequence retrieved from a DNA database are automatically classified by organism and gene, and aligned using the ClustalW algorithm. Sequence sets are regrouped depending on their similarity scores. Main diversity parameters, including polymorphism, synonymous and non-synonymous substitutions, linkage disequilibrium and codon bias are estimated both for the full length of the sequences and for specific functional regions. Program output includes a database with all sequences and estimations, and HTML pages with summary statistics, the performed alignments and a histogram maker tool. PDA is an essential tool to explore polymorphism in large DNA databases for sequences from different genes, populations or species. It has already been successfully applied to create a secondary database. PDA is available on the web at http://pda.uab.es/.

INTRODUCTION

Molecular data is growing dramatically and the need to develop efficient large-scale software to deal with this huge amount of information has become a high priority in this genomic era (1). Polymorphic studies are one of the main focuses of genomic research because of their promise to unveil the genetic basis of phenotypic diversity, with all their potential implications in basic biology, health and society. So far, several software programs have been developed that successfully analyze local data in terms of nucleotide variability [DnaSP (2), Arlequin http://lgb.unige.ch/arlequin/, SITES http://lifesci.rutgers.edu/~heylab/ProgramsandData/Programs/WH/WH_Documentation.htm], but they usually require that input sequences are previously aligned, which assumes that sequences are known to be polymorphic. None of these programs include a first step that permits to explore for potential polymorphic sequences from a large source of heterogeneous DNA, and then to extract and sort them out by gene, species and extent of similarity. Finally, for each group of two or more sequences already aligned, the main diversity parameters can be estimated.

With this prospect in mind we have developed PDA, *Pipeline Diversity Analysis*, a web-based tool which retrieves information from large DNA databases and provides a consistent (3), user-friendly interface to explore and estimate nucleotide polymorphisms. PDA can deal with large sets of unaligned sequences, which can be retrieved automatically from DNA databases given a list of organisms, genes or accession numbers. Even though it is web based, the source code can also be downloaded and installed locally.

A typical user of this site is a researcher who wants to know how many polymorphic sequences are available in Genbank (4) for one or several species of interest and how much variation there is in such sequences. Then, the researcher addresses to the PDA main page, writes the species names and chooses Genbank as the data to search for. Additionally, the user defines some parameter values such as the minimum ClustalW pairwise similarity score from which the sequences or the different gene regions to be analyzed will be grouped. The researcher will receive as output a database containing all the sequences and measures of DNA diversity, as well as HTML pages with summary statistics, the performed alignments and a histogram maker tool for graphical display of the results.

PDA has already been successfully used to explore the amount of polymorphism in the *Drosophila* genus and to create the DNA secondary database DPDB, *Drosophila Polymorphism Database* (http://dpdb.uab.es). This is the first

*To whom correspondence should be addressed. Tel: +34 935 812 730; Fax: +34 935 812 387; Email: Antonio.Barbadilla@uab.es

database that allows the search of DNA sequences by genes, species, chromosome, etc., according to different parameter values of nucleotide diversity. PDA is available on the web at http://pda.uab.es.

PROGRAM OVERVIEW

PDA is a *pipeline* made of multiple programs written in Perl (http://www.perl.com). This language was chosen for the development of PDA because of its initial orientation to the search, extraction and formatting of sequence strings, its support for object-oriented programming, the existence of a public repository of reusable Perl modules [the Bioperl project, http://www.bioperl.org (5)], and the ease of Perl commands to control and execute external programs in other languages (6).

Pipeline design

PDA runs sequentially several modules in a pipeline process as illustrated in Figure 1. Initially, sequences and their annotations are extracted from the input source defined by the user in the PDA home page. Input sources include DNA databases such as Genbank (4) (http://www.ncbi.nlm.nih.gov/Genbank/index.html), EMBL-Bank (http://www.ebi.ac.uk/embl/index.html) or the DPDB database (http://dpdb.uab.es). Low quality sequences coming from large-scale sequencing projects (i.e. *working draft*) are excluded from the analysis. Searches to these databases are done according to a list of accession numbers, organisms and/or genes. Alternatively, sequences can be introduced manually in Fasta or Genbank formats. All the retrieved sequences are introduced into a database (Figure 1: 1a) and passed to the next module (Figure 1: 1b). The second module organizes the sequences by organism and gene and filters these groups according to a minimum number of sequences per group set by the user (Figure 1: 2). Then, every group is aligned using the ClustalW algorithm (7) (Figure 1: 3). Default values have been fitted for the optimal alignments obtained in DPDB, but they can be alternatively defined by the user. The percentage of similarity between each pair of sequences (*ClustalW score*) is taken into account to group again the sequences in subgroups having a higher score than the minimum defined (Figure 1: 4). The value of this score can also be defined by the user and is set to 90% by default. Later on, the alignments are input into the Diversity Analysis module (Figure 1: 5–6), where the main nucleotide diversity, linkage disequilibrium and codon bias measures can be estimated. Finally, the results of the analyses are presented in four formats: a complete output database (in MySQL or MS-Access format) which can be downloaded as a compressed *.gz* file, a web-based output with summary statistics and the estimators, all the performed alignments, and a histogram maker tool for graphic display (Figure 1: 7).

Different gene regions can be analyzed separately. In this case, some additional steps are taken before presenting the results (Figure 1: 8–10). First, a module reads the annotations of the gene corresponding to the sequences on each alignment resulting from previous analyses. The fragments of the sequences from every gene region to analyze (e.g. exon, intron, etc., defined by the user) are extracted from the initial sequence according to the annotations and reversed-complemented if needed. Finally, the resulting sequences fragments are aligned and analyzed as before (Figure 1: 3–7).

Limitations

The heterogeneous nature of the source sequences is intrinsically problematic because the grouping module can lump

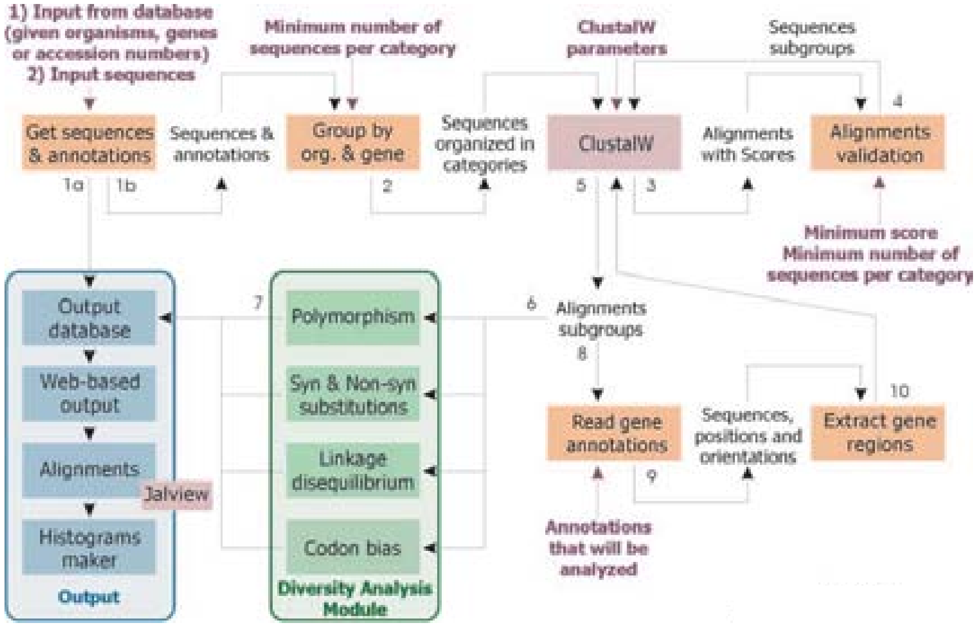


Figure 1. PDA program design and data flow. Independent Perl modules are represented by color boxes, and data flow by arrows and numbers. Lettering in purple corresponds to user-defined parameters. Meanings of color boxes: orange, sequences manipulations; green, nucleotide diversity analysis; blue, output; purple, external programs implemented in PDA. See text for details.

together sequences that are fragmented, or paralogous, or coming from different populations or arrangements, or simply incorrectly annotated, among other reasons. This can distort, to different degrees, the estimated diversity values and therefore, a first analysis must be seen as preliminary. To minimize this problem it would be useful to define an appropriate similarity score between each pair of sequences (*ClustalW score*) or to repeat the analysis with different values. High values of this score would make more restrictive the grouping of sequences. Nevertheless, after a first analysis it is always advisable to inspect visually the alignments, mainly those that yield extreme values, that have a high proportion of gaps or ambiguous bases, or whose sequence lengths vary widely. Two parameters, the percentage of excluded sites due to gaps or ambiguous bases within the aligned sequences and the relative and absolute differences between the longest and shortest sequences are estimated. A warning message appears in the output when the percentage of excluded sites is >30%. In addition, sequences with lengths <100 nt are excluded from the analysis. Both values are set by default and can be modified by the user. Since every sequence from an alignment is linked to its annotation, it is easy to trace the origin of the sequence and to assess its suitability to be included in the analysis. After this inspection, dubious, incorrect or unequal sequences can be manually eliminated via the *Jalview* editor (8), implemented in the *Alignments* section of the output and a reanalysis performed.

PDA has been optimized in terms of speed analysis. However, the process is intensive by nature and the analysis is run in a batch queue. We are putting our effort into parallelizing different instances of PDA using a large cluster of computers through the Condor batch queues specialized management system (http://www.cs.wisc.edu/condor/). However, we encourage users aiming to conduct large and frequent analyses to download, install and use PDA locally in their computers.

DIVERSITY PARAMETERS ESTIMATED

PDA provides a wide range of polymorphic estimations (with their respective variances and SD measures) and statistical tests for polymorphism, codon bias and linkage disequilibrium analyses. Table 1 lists all estimated parameters that have been implemented. All the algorithms have been checked with specific examples or by comparing the results with other available software such as DnaSP (2). Future improvements of the program will include the implementation of typical measures of divergence between different species and the reconstruction of phylogenetic trees. In this way, PDA should be seen as a general tool for large-scale DNA diversity analysis, both for within and among species gene variation.

OUTPUT

The results of PDA are stored in the PDA server and can be accessed through an HTML page using a unique ID that is assigned to every job. The output includes: (i) a MySQL or MS-Access 2002 database with all the retrieved sequences and the results of the analyses, which can be downloaded as a compressed *.gz* file or searched directly through the PDA server in the case of MySQL; (ii) a set of HTML pages with most of the contents of the database and summary statistics both for the whole gene length and for gene regions; (iii) the performed alignments in Fasta and Clustal formats, and the alignments visualization java applet Jalview (8); and (iv) a histogram maker tool for graphic display of personalized histograms and frequency representations of all the estimations. A sample output can be seen at http://pda.uab.es/pda/pda_example.asp.

PDA has already been used on all the sequences of the *Drosophila* genus. The results have been introduced in a relational database which is integrated in the web bioinformatics platform DPDB (http://dpdb.uab.es). Using the DPDB interface, these estimations and the original sequences analyzed can be searched and retrieved according to different parameter values of nucleotide diversity, and many tools can be used online with the users input, including the PDA itself.

Table 1. List of estimators implemented in PDA for DNA polymorphism, codon bias and linkage disequilibrium analysis

Nucleotide polymorphism	
Number of segregating sites (S, s)	Nei (9)
Minimum number of mutations (H, η)	Tajima (10)
Nucleotide diversity (π) (with and without Jukes and Cantor correction)	Nei (9); Jukes and Cantor (11)
Theta (θ) per DNA sequence from S	Tajima (12)
Theta (θ) per site from S	Nei (9)
Theta (θ) per site from Eta (η)	Tajima (10)
Theta (θ) per site from π, from S and from η under the Finite Sites Model	Tajima (10)
Average number of nucleotide differences (k)	Tajima (13)
Tajima statistic test (D)	Tajima (14)
Total number of synonymous and non-synonymous sites	Nei and Gojobori (15)
Number of non-synonymous substitutions per non-synonymous site (Ka) and number of synonymous substitutions per synonymous site (Ks)	Nei and Gojobori (15)
Codon bias	
Relative Synonymous Codon Usage (RSCU)	Sharp (16)
Effective Number of Codons (ENC)	Wright (17)
Codon Adaptation Index (CAI)	Sharp and Li (18)
Scaled Chi Square	Shields (19)
G + C content in second, third and total positions	Wright (17)
Linkage disequilibrium	
Nucleotide distance (Dist) between a pair of polymorphic sites	
D	Lewontin and Kojima (20)
D′	Lewontin (21)
R and R^2	Hill and Robertson (22)
ZnS statistic	Kelly (23)
Chi-square test	
Fisher's exact test	

AVAILABILITY

PDA can be accessed on the web at site http://pda.uab.es together with examples and documentation. In addition, the source code to PDA is distributed as a package of programs to be downloaded and run locally (http://pda.uab.es/pda/pda_download.asp) under the GNU General Public License (GPL).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR online.

ACKNOWLEDGEMENTS

The authors would like to thank Jordi Pijoan and Helena Norén for help in developing the software, and Rosemary Thwaite, Marta Puig, Natalia Petit and Alfredo Ruiz for their critical reading of the manuscript. This work was funded by the Ministerio de Ciencia y Tecnología (Grant PB98-0900-C02-02). S.C. was supported in part by the bioinformatics company e-Biointel and the Ministerio de Ciencia y Tecnología (Grant BES-2003-0416).

REFERENCES

1. Collins,F.S., Green,E.D., Guttmacher,A.E. and Guyer,M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
2. Rozas,J., Sanchez-DelBarrio,J.C., Messeguer,X. and Rozas,R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.
3. Stein,L. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119–120.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
5. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
6. Stein,L.D. (2001) Using Perl to facilitate biological analysis. In Ouellette,B.F.F. (ed.), Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Wiley-Liss, Inc., pp. 413–449.
7. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
8. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
9. Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
10. Tajima,F. (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, **143**, 1457–1465.
11. Jukes,T.H. and Cantor,C.R. (1969) Mammalian protein metabolism. In Munro,H.N. (ed.), *Evolution of Protein Molecules*. Academic Press, New York, pp. 21–132.
12. Tajima,F. (1993) Mechanisms of molecular evolution. In Takahata,N. and Clark,A.G. (eds), *Mesurement of DNA Polymorphism*. Sinauer Associates Inc., Suderland, Massachusetts.
13. Tajima,F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.
14. Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
15. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
16. Sharp,P.M., Tuohy,T.M. and Mosurski,K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.*, **14**, 5125–5143.
17. Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23–29.
18. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
19. Shields,D.C., Sharp,P.M., Higgins,D.G. and Wright,F. (1988) 'Silent' sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.*, **5**, 704–716.
20. Lewontin,R.C. and Kojima,K. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution*, **14**, 458–472.
21. Lewontin,R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, **49**, 49–67.
22. Hill,W.G. and Robertson,A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, **38**, 226–231.
23. Kelly,J.K. (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.

W632–W634 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl080

PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA

Sònia Casillas and Antonio Barbadilla*

Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

Received February 14, 2006; Revised and Accepted March 3, 2006

ABSTRACT

Pipeline Diversity Analysis (PDA) is an open-source, web-based tool that allows the exploration of polymorphism in large datasets of heterogeneous DNA sequences, and can be used to create secondary polymorphism databases for different taxonomic groups, such as the *Drosophila* Polymorphism Database (DPDB). A new version of the pipeline presented here, PDA v.2, incorporates substantial improvements, including new methods for data mining and grouping sequences, new criteria for data quality assessment and a better user interface. PDA is a powerful tool to obtain and synthesize existing empirical evidence on genetic diversity in any species or species group. PDA v.2 is available on the web at http://pda.uab.es/.

INTRODUCTION

The first version of Pipeline Diversity Analysis (PDA), was announced in the Web Server Issue of this journal (1) as a web-based tool that allowed the exploration for polymorphism in large datasets of heterogeneous DNA sequences. The pipeline automatically extracts a set of sequences from a DNA database given a list of organisms, genes or accession numbers, and sorts them by gene, species and extent of similarity. Then it aligns the homologous sequences and calculates the standard population genetic diversity parameters on the generated alignments. PDA is not aimed to provide exhaustive measures of DNA diversity (2), but rather to be an exploratory tool to transform the huge amounts of sequences available in public databases into information that can be analyzed from a population genetic perspective. PDA gives an overview of the empirical evidence on genetic diversity in any species or group of species.

PDA has already been used successfully to explore the amount of polymorphism in the *Drosophila* genus and to create the secondary database DPDB, *Drosophila* Polymorphism Database (http://dpdb.uab.es) (3). This is the first database that allows the search of DNA sequences and polymorphic alignments by diversity values, in addition to filter the results by organism, gene region or data quality criteria. At present, PDA is being used to create a database for mammalian sequences (MamPol, http://pda.uab.es/mampol/) of nuclear and mitochondrial genes that will include new features with respect to DPDB. A modified version of PDA is also being developed to extend the DPDB database to include sequences from non-coding regions.

In this paper we introduce a new version of the pipeline, PDA v.2, which incorporates novel features and substantial improvements with respect to the original version, including new methods for data mining and grouping, new criteria for data quality assessment and a much better interface usability.

NEW METHODS FOR DATA GROUPING AND ANALYSIS

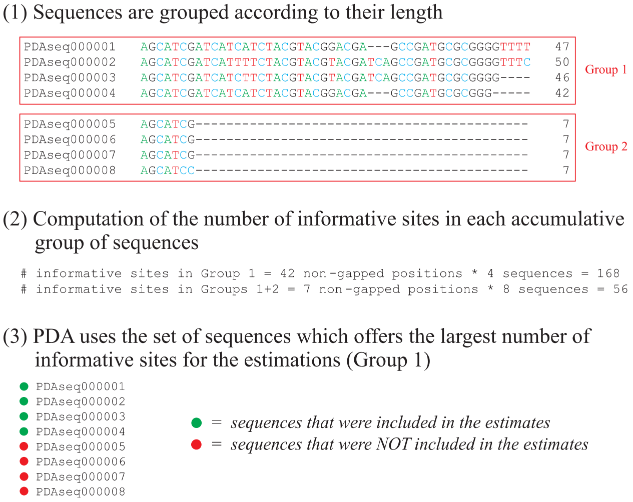
The input raw data for PDA are polymorphic sets formed by groups of orthologous sequences (alleles or haplotypes) for a given species and DNA region. Sequences belonging to a polymorphic set can come either: (i) from previous polymorphism studies, or (ii) from independent studies of the same gene and species, possibly not primarily focused on polymorphism. This second subset of sequences increases significantly the amount of polymorphic sets, although it raises the question whether the estimations are reliable. Due to the heterogeneous origin of the source sequences, PDA can mix together fragmented sequences coming from different regions of the same gene that do not align together, paralogous sequences or sequences coming from different populations or arrangements that have very distinct haplotypes. These cases were already resolved in PDA v.1 using a minimum similarity score for each

*To whom correspondence should be addressed. Tel: +34 935 812 730; Fax: +34 935 812 387; Email: Antonio.Barbadilla@uab.es

pair of sequences in the alignment that is customizable by the user. The default score is 95%, so sequences differing in more than the 5% of the sequence (excluding gaps) are split into separate alignments. PDA v.2 includes new features to handle the heterogeneity of the source sequences and to improve the quality of the alignments.

Algorithm for maximization of the number of informative sites

Although sequences from a given alignment are usually very similar in terms of sequence identity, they can vary widely in length. Because estimates of genetic diversity usually exclude gapped sites, a significant amount of information can be lost if large and short sequences are aligned together, since only the sites included in the shortest sequences will be used in the analyses. To maximize the amount of information that can be used in such estimates, we have implemented an algorithm that works as follows (Figure 1). First, sequences from an alignment are grouped according to their length, so that sequences in a group cannot differ in more than 20% of their length. After that, the amount of informative sites in each accumulative group of sequences is calculated, starting with the group of the longest sequences (group 1) and adding in each step the next group of sequences ordered by their length (groups 1 + 2, groups 1 + 2 + 3, etc.). By informative sites we mean the number of non-gapped positions multiplied by the number of sequences in the set (note that this differs from the definition of 'informative site' typically used in phylogenetics). Finally, PDA v.2 uses the set of sequences which offers the largest number of informative sites, in some cases discarding the shortest sequences. This algorithm can be used optionally in PDA v.2.



(1) Sequences are grouped according to their length

```
PDAseq000001   AGCATCGATCATCATCTACGTACGGACGA---GCCGATGCGCGGGGTTTT   47
PDAseq000002   AGCATCGATCATTTTCTACGTACGTACGATCAGCCGATGCGCGGGGTTTC   50
PDAseq000003   AGCATCGATCATCTTCTACGTACGTACGATCAGCCGATGCGCGGGG----   46
PDAseq000004   AGCATCGATCATCATCTACGTACGGACGA---GCCGATGCGCGGG-----   42

PDAseq000005   AGCATCG-------------------------------------------    7
PDAseq000006   AGCATCG-------------------------------------------    7
PDAseq000007   AGCATCG-------------------------------------------    7
PDAseq000008   AGCATCC-------------------------------------------    7
```

(2) Computation of the number of informative sites in each accumulative group of sequences

```
# informative sites in Group 1 = 42 non-gapped positions * 4 sequences = 168
# informative sites in Groups 1+2 = 7 non-gapped positions * 8 sequences = 56
```

(3) PDA uses the set of sequences which offers the largest number of informative sites for the estimations (Group 1)

```
● PDAseq000001
● PDAseq000002
● PDAseq000003
● PDAseq000004
● PDAseq000005
● PDAseq000006
● PDAseq000007
● PDAseq000008
```

● = *sequences that were included in the estimates*
● = *sequences that were NOT included in the estimates*

Figure 1. Example showing the new algorithm for maximizing the number of informative sites. (1) Input sequences are grouped according to their length, so that sequences in a group cannot differ in more than the 20% of their length. In this example, the eight input sequences are split into two different groups (group 1 and group 2). (2) Assuming that an 'informative site' is the number of non-gapped positions multiplied by the number of sequences in the set (note that this differs from the definition of 'informative site' typically used in phylogenetics), PDA v.2 calculates the amount of informative sites in each accumulative group of sequences, starting with the group of the longest sequences (group 1 = 168 informative sites) and adding in each step the next group of sequences ordered by their length (groups 1 + 2 = 56 informative sites). (3) Finally, PDA v.2 shows the alignment with all the sequences, but uses the set of sequences which offer the largest number of informative sites for the estimations, in some cases discarding the shortest sequences. In this case, PDA v.2 would use only the four longest sequences for the estimations (group 1). To distinguish which sequences were used in the analyses from those which were discarded, PDA v.2 uses a color code: green for sequences that were included in the estimates, and red for sequences that were not included.

Filtering raw sequences for well annotated genes

PDA v.1 analyzed raw sequences directly from GenBank regardless of the annotation quality or the number of genes included in the sequence. So, large genomic fragments including more than one gene could be aligned together with sequences of single genes. To avoid these noisy data, only well annotated sequences for the different functional regions of the genes (genes, CDSs, exons, introns, UTRs, promoters, etc.), as defined in the Features section of the GenBank format files, are now analyzed in PDA v.2. Note that sequences lacking these annotations, even coming from polymorphic studies, will not be included in the analyses. Thus, in PDA v.2 raw data is more appropriately pre-processed by functional category, and the main unit for storing information in the database is not the raw sequence coming from GenBank but the corresponding polymorphic sets for each organism and gene region [see Figure 1 in (3)].

Additional alignment programs

We have incorporated two new programs within PDA that can be used to align the polymorphic sequences in addition to ClustalW (4,5): Muscle (6) and T-Coffee (7). These programs have been shown to achieve better accuracy than the commonly used ClustalW for sequences with a high proportion of gaps, such as non-coding sequences (see the Help section of the Web site). We suggest using these alternative programs when analyzing non-coding regions (introns, promoters, UTRs, etc.).

DATA QUALITY ASSESSMENT

In PDA v.2 we provide several measures concerning the quality of each dataset so that the user can assess the confidence on the data source and the estimations. A quick guide is also supplied explaining how to use these quality measures and how to easily reanalyze the data.

Quality assessment of the alignments

To assess the quality of an alignment we use three criteria: (i) the number of sequences included in the alignment; (ii) the percentage of gaps or ambiguous bases within the alignment; and (iii) the percent difference between the shortest and the longest sequences. Three qualitative categories are defined for each criterion: high, medium and low quality, which are shown in the main output table to quickly visualize the confidence on the results (further details are given in the Help section of the Web site).

Quality assessment of the data sources

According to the data source, we use four criteria to determine if the sequences from a polymorphic set come from a population study: (i) one or more sequences from the alignment are stored in the PopSet database; (ii) all the sequences have consecutive GenBank accession numbers; (iii) all the sequences share at least one reference; and (iv) one or more references are from journals that typically publish polymorphism studies (*Genetics*, *Molecular Biology and Evolution*, *Journal of Molecular Evolution*, *Molecular Phylogenetics and Evolution* or *Molecular Ecology*). This information is shown in the main output table by means of a confirmatory tick where the dataset satisfies the corresponding criterion.

Origin of the sequences

PDA v.2 reports the origin of each sequence (country, strain and population variant) when this information is available in the GenBank annotations. This allows the user to trace the origin of the source sequences and to assess the suitability of each sequence to be included in the dataset.

INTERFACE AND NEW UTILITIES

Important improvements in the text and graphic interface and other new features make PDA a much more useful tool.

Completely renewed interface

PDA v.2 offers a more intuitive and visually improved interface for both data input and output. For example, the page for job submission is designed in layers, which substantially facilitates the understanding of the available options. The output is more clearly displayed, and is based on the design of the DPDB database (3).

Management of previous analyses

On submitting a job, PDA v.2 can optionally store user information to allow them enter the ‘Previous IDs’ section and manage their previous analyses, either to revisit or to delete them. This new feature extends the previous ‘Request by ID’ option of PDA v.1, which is still available.

Improved database structure

The database has been extended to store the new data gathered by PDA, e.g. the storage of polymorphism datasets by functional categories (see above). Moreover, existing tables have been redefined, improving the performance of the search responses.

Tools for extraction and representation of polymorphic sites

A new module for extraction of SNPs from the aligned sequences has been incorporated. It lists the position of each SNP in the alignment and the frequency of the different alleles. Moreover, the data can be directly submitted to the SNPs-Graphic tool of the DPDB database to perform sliding windows and graphs for detailed analyses of polymorphism.

Improved sections of the web site

We have extended the Help section of the Web to provide a more complete and detailed description of PDA and to explain the new features of PDA v.2. We have also included links to the polymorphic databases created with this software.

AVAILABILITY

PDA v.2 can be accessed on the web at http://pda.uab.es/, together with examples and documentation. Jobs are run in a batch queue. Although at present the number of sequences that can be analyzed on the Web is limited to 500, we are working to have ready a parallel version of PDA to extend the number of sequences that can be analyzed. In addition, the source code of PDA is distributed under the GNU General Public License (GPL) as a package of Perl programs to be downloaded and run locally without limitations (http://pda.uab.es/pda2/pda_download.asp).

ACKNOWLEDGEMENTS

The authors would like to thank Casey Bergman for valuable discussions and guidance in implementing the alignment programs, and for his critical reading of the manuscript. The authors also thank Francesco Catania for helpful comments on this manuscript. S.C. was supported by the Ministerio de Ciencia y Tecnología (Grant BES-2003-0416) and a Marie Curie fellowship from the European Commission (HPMT-GH-01-00285-13). The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Casillas,S. and Barbadilla,A. (2004) PDA: a pipeline to explore and estimate polymorphism in large DNA databases. *Nucleic Acids Res.*, **32**, W166–W169.
2. Rozas,J., Sanchez-DelBarrio,J.C., Messeguer,X. and Rozas,R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.
3. Casillas,S., Petit,N. and Barbadilla,A. (2005) DPDB: a database for the storage, representation and analysis of polymorphism in the *Drosophila* genus. *Bioinformatics*, **21**, ii26–ii30.
4. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
5. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
6. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
7. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

2.2. Drosophila Polymorphism Database (DPDB): a secondary database of nucleotide diversity in the genus *Drosophila*

This section includes two publications of DPDB –Drosophila Polymorphism Database–. DPDB is the first comprehensive secondary database that provides searchable collections of polymorphic sequences in the genus *Drosophila* with their associated diversity measures. It was created by using the bioinformatic system PDA and offered to the scientific community through an interactive Web platform. In the first paper, we explain the data model for the storage, representation and analysis of haplotypic data under which both PDA and DPDB were developed. We also describe technical aspects of the software and the updating process, specific terminologies of the database, and additional utilities of the Web site, including a list of implemented tools for the analysis of genetic diversity. The second publication provides a general description of DPDB, including recent improvements, a step-by-step guide to all its searching and analytic capabilities, and a simple multi-species analysis performed by making simple queries to the DPDB interface that illustrates the power of this database. Both publications give an overview of the amount and quality of data in DPDB at the time of writing. DPDB is available at http://dpdb.uab.cat/ and a daily-updated MySQL database can be downloaded from the Web site.

- **Article 3:** Casillas, S., N. Petit and A. Barbadilla (2005) DPDB: a database for the storage, representation and analysis of polymorphism in the *Drosophila* genus. *Bioinformatics* **21:** ii26-ii30.

- **Article 4:** Casillas, S., R. Egea, N. Petit, C. M. Bergman and A. Barbadilla (2007) Drosophila Polymorphism Database (DPDB): a portal for nucleotide polymorphism in *Drosophila*. *Fly* **1(4):** 205-211.

BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii26–ii30
doi:10.1093/bioinformatics/bti1103

Databases

DPDB: a database for the storage, representation and analysis of polymorphism in the *Drosophila* genus

Sònia Casillas, Natalia Petit and Antonio Barbadilla*

Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona,
08193 Bellaterra (Barcelona), Spain

ABSTRACT

Motivation: Polymorphism studies are one of the main research areas of this genomic era. To date, however, no comprehensive secondary databases have been designed to provide searchable collections of polymorphic sequences with their associated diversity measures.

Results: We define a data model for the storage, representation and analysis of genotypic and haplotypic data. Under this model we have created DPDB, '*Drosophila* Polymorphism Database', a web site that provides a daily updated repository of all well-annotated polymorphic sequences in the *Drosophila* genus. It allows the search for any polymorphic set according to different parameter values of nucleotide diversity, linkage disequilibrium and codon bias. For data collection, analysis and updating we use PDA, a pipeline that automates the process of sequence retrieval, grouping, alignment and estimation of nucleotide diversity from Genbank sequences in different functional regions. The web site also includes analysis tools for sequence comparison and the estimation of genetic diversity, a page with real-time statistics of the database contents, a help section and a collection of selected links.

Availability: DPDB is freely available at http://dpdb.uab.es and can be downloaded via FTP.

Contact: antonio.barbadilla@uab.es

1 INTRODUCTION

Drosophila is the most intensively studied genus for DNA polymorphism, since current population genetics models on nucleotide variation have been tested using the extensive sequence data gathered for this genus (Aquadro *et al*., 2001; Powell, 1997). Each polymorphic study releases groups of homologous sequences (or haplotypes) for a given DNA region and species. The haplotypic information of a polymorphic set allows the estimation of both the one-dimensional and multi-dimensional components of nucleotide diversity in the studied regions. One-dimensional measures, such as the distribution of PI values [Nei's diversity index, (Nei, 1987)] along sliding windows, allow the detection of differently constrained regions (Vilella *et al*., 2005). Multi-dimensional diversity measures search for association among variable sites, as summarized by linkage disequilibrium estimators, and provide key information on the history and evolution of a DNA region, including the effective recombination rate underlying the region (Hudson, 1987; McVean *et al*., 2004; Nordborg and Tavare, 2002). Both diversity components are necessary for a complete description of nucleotide variation at the DNA level. To date, however, no comprehensive secondary database provides searchable collections of polymorphic sequences with their associated diversity measures. '*Drosophila* Polymorphism Database' (DPDB) is a database aimed to fill this vacuum, and allows the search of polymorphic sequences in the *Drosophila* genus according to different measures of nucleotide diversity.

2 DPDB APPROACH

The creation of a secondary database on DNA variation requires the development of a set of modules of data mining and analysis which operate together to automatically extract the available sequences from public databases, align them and compute the diversity estimates. A priori, the automation of this process seems destined to fail, since variation estimates usually require a careful manual inspection. Especially critical is the alignment of sequences (which is sensitive to the input parameters and the intrinsic characteristics of the sequences) and the sample stratification of aligned sequences (because any non-controlled heterogeneity will invalidate the estimates). On facing this 'manual versus automatic' dilemma, a first option would consist of giving up the automation and limiting the analyses to our own data. However, automation is nowadays an aspiration that cannot be waived. Therefore, while conscious of the limitations, we have tackled the bioinformatics automation of genetic diversity.

Our approach to build DPDB is outlined in Figure 1. We define a data model for the storage, representation and analysis of haplotypic variability based on the 'polymorphic set' as the basic storing unit: a group of homologous sequences for a given gene and species. Polymorphic sets are created by grouping by gene and species all the *Drosophila* sequences available in Genbank that are well annotated. From the sequence annotations, homologous subgroups are created for each polymorphic set corresponding to different functional regions (genes, CDSs, exons, introns, UTRs and promoters). Every subgroup is then aligned and selected according to different quality criteria. The selected alignments form the 'analysis units' of DPDB, on which the commonly used diversity parameters are computed. The results of the estimations are annotated together with the corresponding polymorphic set.

Besides such filtering during the processing, information on the data source and the quality of the alignments is given with the query output to allow the user assessment of the confidence on the estimated values. Furthermore, any subset can be directly reanalyzed using PDA (Casillas and Barbadilla, 2004) by adding or deleting sequences, or changing default parameters.

*To whom correspondence should be addressed.

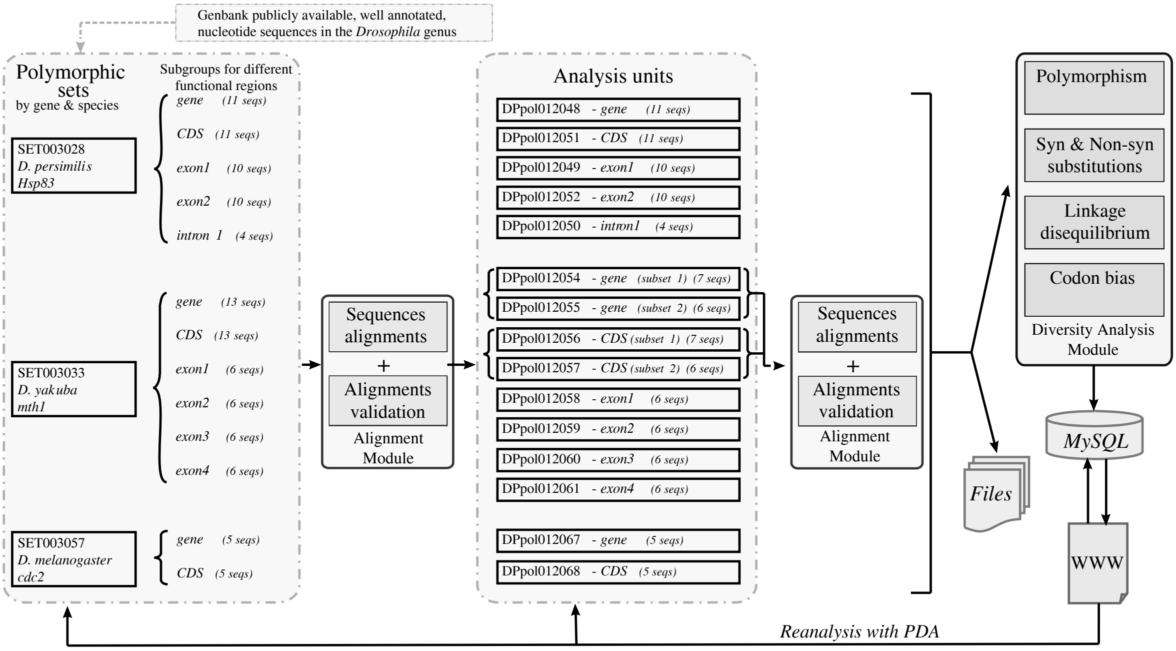

Genbank publicly available, well annotated,
nucleotide sequences in the *Drosophila* genus
Polymorphic sets
by gene & species
Subgroups for different functional regions
SET003028
D. persimilis
Hsp83
gene (11 seqs)
CDS (11 seqs)
exon1 (10 seqs)
exon2 (10 seqs)
intron 1 (4 seqs)
SET003033
D. yakuba
mth1
gene (13 seqs)
CDS (13 seqs)
exon1 (6 seqs)
exon2 (6 seqs)
exon3 (6 seqs)
exon4 (6 seqs)
SET003057
D. melanogaster
cdc2
gene (5 seqs)
CDS (5 seqs)
Sequences alignments
+
Alignments validation
Alignment Module
Analysis units
DPpol012048 - *gene* (11 seqs)
DPpol012051 - *CDS* (11 seqs)
DPpol012049 - *exon1* (10 seqs)
DPpol012052 - *exon2* (10 seqs)
DPpol012050 - *intron1* (4 seqs)
DPpol012054 - *gene* (subset 1) (7 seqs)
DPpol012055 - *gene* (subset 2) (6 seqs)
DPpol012056 - *CDS* (subset 1) (7 seqs)
DPpol012057 - *CDS* (subset 2) (6 seqs)
DPpol012058 - *exon1* (6 seqs)
DPpol012059 - *exon2* (6 seqs)
DPpol012060 - *exon3* (6 seqs)
DPpol012061 - *exon4* (6 seqs)
DPpol012067 - *gene* (5 seqs)
DPpol012068 - *CDS* (5 seqs)
Sequences alignments
+
Alignments validation
Alignment Module
Polymorphism
Syn & Non-syn substitutions
Linkage disequilibrium
Codon bias
Diversity Analysis Module
Files
MySQL
WWW
Reanalysis with PDA

Fig. 1. The DPDB approach defines two basic data units: the 'polymorphic set' (a group of homologous sequences for a given gene and species) and the 'analysis unit' (on which diversity estimates are carried out). An analysis unit is a subset of sequences from a polymorphic set obtained by functional annotation and alignment filtering. Sequences are extracted from Genbank, and any subset can be directly reanalyzed using PDA. See text for details.

3 IMPLEMENTATION OF THE DATA MODEL IN DPDB

3.1 Overview of DPDB

DPDB is a secondary database which provides the collection of well-annotated polymorphic sequences in the *Drosophila* genus. DPDB allows, for the first time, the search for any polymorphic set according to different parameter values of nucleotide diversity, such as the PI value, the degree of linkage disequilibrium or the codon bias.

DPDB is searchable through a web site and also includes: (1) an Analysis section with tools for sequence comparison and the estimation of genetic diversity; (2) a daily updated Statistics page with the database contents; (3) a comprehensive Help section for the whole site; and (4) a page with selected links for the study of *Drosophila* polymorphism.

DPDB aims to be the reference site for DNA polymorphism in *Drosophila* (Galperin, 2005; Matthews *et al*., 2005), spanning studies that try to describe and explain the underlying causes of polymorphic patterns found in these species, such as recombination rate (Begun and Aquadro, 1992; Betancourt and Presgraves, 2002), gene density in different genomic regions (Payseur and Nachman, 2002), chromosomal inversions (Navarro *et al*., 2000), sequence complexity (Nelson *et al*., 2004) or demographic history (Glinka *et al*., 2003). DPDB has already been successfully used to study the association between coding polymorphism levels and gene structure in *Drosophila melanogaster* (Petit *et al*., unpublished data).

We want to guarantee long-term support for this database by including updating of the interface and new data processing and representation. We also aim to extend the database to other species groups.

3.2 Primary data source and processing

For data collection, diversity measures and updating we use PDA (Casillas and Barbadilla, 2004), a pipeline made of a set of Perl modules that automates the process of sequence retrieval, grouping, alignment and estimation of diversity parameters from sequences in large DNA databases. Using PDA we get all the publicly available *Drosophila* nucleotide sequences (excluding ESTs, STSs, GSSs, working draft and patents) with their annotations and references from Genbank (Benson *et al*., 2005), additional information of genes and aberrations from Flybase (Drysdale *et al*., 2005) and the cross-references to Popset [from NCBI (Wheeler *et al*., 2005)].

Polymorphic sets of two or more sequences are created by grouping sequences by gene and species. For each polymorphic set, subgroups of homologous sequences are created for the different functional regions (genes, CDSs, exons, introns, UTRs and promoters), as defined in the Features section of the Genbank format files. Note that those sequences lacking these annotations, even though coming from polymorphic studies, are not included in the analyses, so only well-characterized sequences are used. Every subgroup is then aligned with ClustalW (Chenna *et al*., 2003). After a manual inspection of hundreds of ClustalW alignments, we defined an optimal ClustalW parameter setting for *Drosophila* polymorphic data. Likewise, we fixed 95% as the minimum percentage of similarity between each pair of sequences within an analysis unit (excluding gaps), so that different analysis units can be obtained for a given functional region. Diversity measures are estimated on these analysis units, including polymorphism at synonymous and non-synonymous sites, linkage disequilibrium and codon bias [see Table 1 in Casillas and Barbadilla (2004) for a detailed description of all the estimations].

Sequences belonging to a polymorphic set can be either: (1) from previous polymorphism studies, or (2) from independent studies of the same gene and species, possibly not primarily focused on polymorphism. This second subset of sequences increases significantly the amount of polymorphic sets (figures are given in the Statistics section of the Web site); although it raises the question whether the estimations are reliable. We assess the confidence on each polymorphic set by taking into account the data source and the quality of the alignment. According to the data source, we use the following four criteria to determine if the study had a polymorphism goal: (1) one or more sequences from the alignment are stored in the Popset database; (2) all the sequences have consecutive Genbank accession numbers; (3) all the sequences share at least one reference; and (4) one or more references are from journals that typically publish polymorphism studies. To assess the quality of an alignment we use three other criteria: (1) the number of sequences included in the alignment; (2) the percentage of gaps or ambiguous bases within the alignment; and (3) the percentage of difference between the shortest and the longest sequences. Three qualitative categories are defined for each criterion: high, low and medium quality (for details on these criteria see the Help section in the Web site). Finally, alignments giving extreme polymorphism values are routinely checked. This allows us to continue improving the default parameters and to check data consistency.

3.3 Database structure, querying and output

The storage of diversity estimates in databases makes them permanently available and allows the reanalysis of all or part of the sequences. With this perspective in mind, we have created a relational MySQL database (see its structure in the Help section of the Web site) to store the results of the analyses. This database is centered on the two main storing units: the polymorphic set and the analysis unit (Fig. 1), and all the subsequent diversity data are annotated into different joined tables. The database also includes the *Drosophila* primary information retrieved from different external sources (Genbank, Flybase and Popset).

The database contents are updated daily, and records are assigned unique and permanent DPDB identification numbers to facilitate cross-database referencing: an increasing six-digit number is preceded by the string *SET* for polymorphic sets, *DPpol* for analysis units, *DPseq* for individual sequences, or *DPref* for references. Earlier analysis units are stored in separate tables when they are updated, and the later ones are assigned new identification numbers, so that the user can trace the history of a polymorphic set.

DPDB is accessible via web at http://dpdb.uab.es using a query interface based on SQL (Structured Query Language) searches (Fig. 2). The interface facilitates data interrogation by diversity estimates and the results can be filtered according to different confidence criteria established in DPDB (Fig. 2A). The first output page lists all the polymorphic sets by organism, gene, analyzed region and analysis unit showing additional information about the quality of the alignment, the confidence on the data source and the date of the last update (Fig. 2B). A complete report for each analysis unit can then be obtained through the corresponding link (Fig. 2C), as well as access to the primary database (individual sequences, genes, aberrations, references and polymorphic studies in the Popset database). Note that the alignment can be obtained in different formats, as well as the DND tree file, so that the user can revise it and decide if the estimates are reliable.

Furthermore, any analysis unit can be interactively reanalyzed using PDA. On using this option, the set of sequences is taken as input in the PDA submission page. Any subset of sequences can then be included or excluded from the analysis or the default parameters modified.

A Graphical search can also be performed for the different diversity values. A histogram is displayed on which any category can be queried to the database.

3.4 Analysis tools

The DPDB web site includes a set of analysis tools organized in different modules for sequence comparison and the estimation of genetic diversity. On the first module, three programs are available: (1) the Blast package (McGinnis and Madden, 2004) is implemented to search for homologous sequences in the primary DPDB database or in the *D.melanogaster* genome; (2) the ClustalW software (Chenna *et al*., 2003) is available with default parameters optimized for alignments of *Drosophila* polymorphic sequences (as manually checked); and (3) Jalview (Clamp *et al*., 2004) is implemented on the web to visualize and edit sequences alignments. The second module includes two other tools: (1) SNPs-Graphic allows performing analyses by the sliding window method, obtaining both the estimations in different regions of the alignment and graphic representations; and finally, (2) the PDA pipeline (Casillas and Barbadilla, 2004).

3.5 Statistics

The Statistics section summarizes the contents of both the primary and secondary databases. It is updated on a daily base, and includes tabular and graphic information.

The distributions of polymorphic sets according to different parameters, such as the species, genes and classes of genes [GO categories (Ashburner *et al*., 2000)] are shown. Then, the analysis units are classified according to the gene region, the quality of the alignments and the confidence on the data source. Average diversity estimates by gene region are also shown. The number of analysis units per taxon can be viewed in the 'Phylogeny of the *Drosophila* genus' graph (categories are based on the NCBI's taxonomy browser). Finally, some important statistics on the primary database are displayed, such as the total number of sequences, genes, aberrations and references, in different classifications.

At the time of writing this article, DPDB contained 1082 polymorphic sets, corresponding to 119 different species of the *Drosophila* genus and 587 different genes. A total of 2879 analysis units on these polymorphic sets were analyzed, most of them corresponding to the gene (1177), CDS (769), exon (473) or intron (435) regions.

The statistics on the quality of the alignments show that a high percentage of analysis units have <6 sequences, but that most of them have few gaps within the alignment, and that sequences are generally of similar length. Finally, according to the data source confidence, ~50% of the analysis units come from sequences where polymorphism was the primary focus of the study.

3.6 Software details

DPDB is stored locally in a MySQL relational database running on a Windows 2003 Server, using the software IIS (Internet Information Server). It can be freely downloaded via ftp at ftp://dpdb.uab.es.

The web interface is mainly implemented in ASP and offers constant interfaces, standard file formats and *ad hoc* queries. Programs for data manipulation and search are all implemented in Perl modules, and search and analysis results are given in HTML formats.

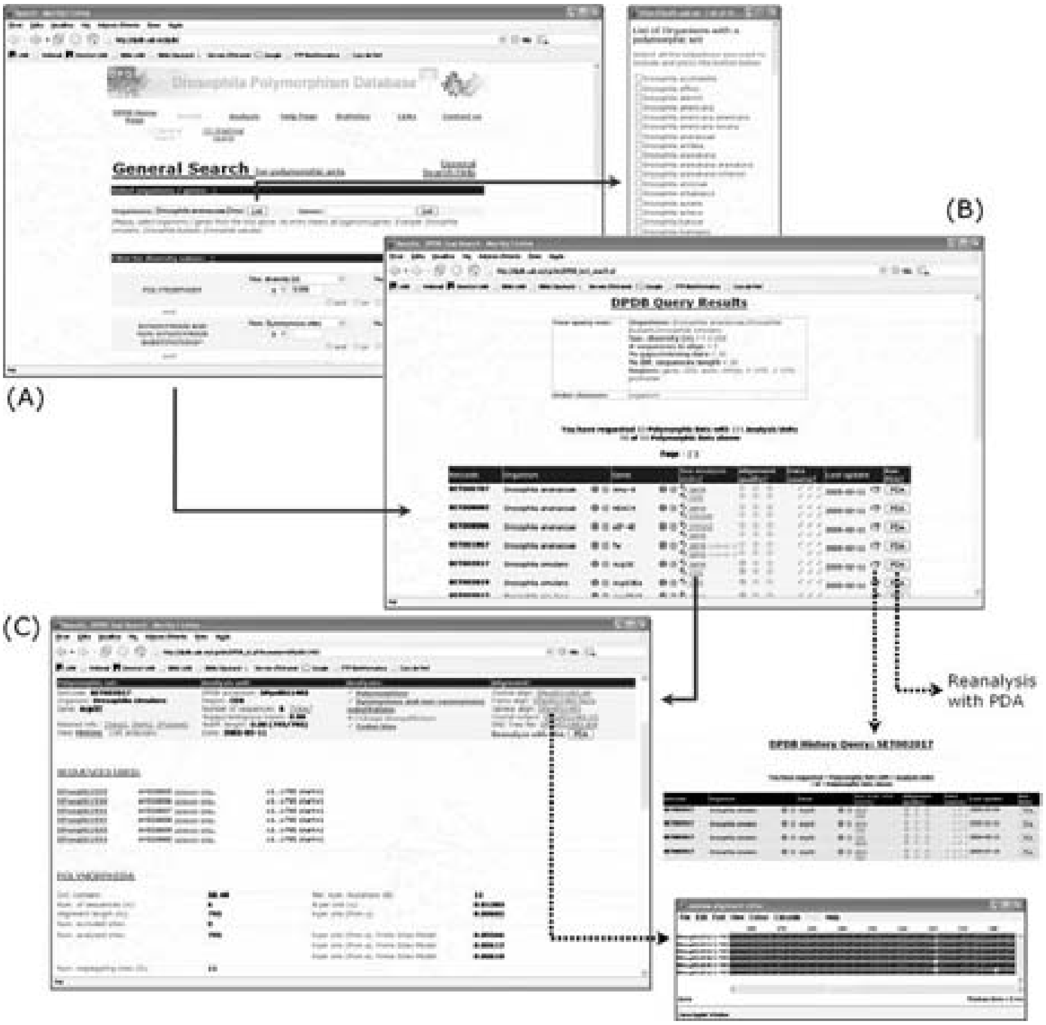

(A)
(B)
(C)
Drosophila Polymorphism Database
General Search
DPDB Query Results
Reanalysis with PDA
DPDB History Query: SET002017

Fig. 2. DPDB interface: (**A**) the General Search page (with the species selector window), (**B**) the first output page of a query, and (**C**) a full report for an analysis unit. In this example we quired all the analysis units from the *Drosophila ananassae, Drosophila buzzatii* and *Drosophila simulans* species, having a nucleotide diversity value <0.006 and excluding lower quality alignments. At the time of making the figure, 55 polymorphic sets were found, each of them with different analysis units for the different gene regions. Part of the full report corresponding to the *CDS* region of gene *Acp32* in *D.simulans* in shown in (C), including general information about the analysis, the alignment in different formats, the corresponding sequences and some of the estimations. Sequences from the analysis units can be directly reanalyzed with PDA. Note that the history of a polymorphic set (the series of previous analysis that have been updated) can be queried from (B).

ACKNOWLEDGEMENTS

Authors would like to thank Raquel Egea and Jordi Pijoan for help in the manual revision of the alignments and testing the database, and Alfredo Ruiz for helpful discussions and comments. This work was funded by the Ministerio de Ciencia y Tecnología (Grants PB98-0900-C02-02 and BMC2002-01708). S.C. was supported by the Ministerio de Ciencia y Tecnología (Grant BES-2003-0416) and partially by the bioinformatics company Ebiointel, and N.P. by a grant from the Departament de Genètica i Microbiologia of the Universitat Autònoma de Barcelona.

Conflict of Interest: none declared.

REFERENCES

Aquadro,C.F. *et al.* (2001) Genome-wide variation in the human and fruitfly: a comparison. *Curr. Opin. Genet. Dev.*, **11**, 627–634.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Begun,D.J. and Aquadro,C.F. (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D.melanogaster*. *Nature*, **356**, 519–520.

Benson,D.A. *et al*. (2005) GenBank. *Nucleic Acids Res.*, **33** (Database issue), D34–D38.

Betancourt,A.J. and Presgraves,D.C. (2002) Linkage limits the power of natural selection in *Drosophila*. *Proc. Natl Acad. Sci. USA*, **99**, 13616–13620.

Casillas,S. and Barbadilla,A. (2004) PDA: a pipeline to explore and estimate polymorphism in large DNA databases. *Nucleic Acids Res.*, **32**, W166–W169.

Chenna,R. *et al*. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.

Clamp,M. *et al*. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.

Drysdale,R.A. *et al*. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33** (Database issue), D390–D395.

Galperin,M.Y. (2005) The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Res.*, **33** (Database issue), D5–D24.

Glinka,S. *et al*. (2003) Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics*, **165**, 1269–1278.

Hudson,R.R. (1987) Estimating the recombination parameter of a finite population model without selection. *Genet. Res.*, **50**, 245–250.

Matthews,K.A. *et al*. (2005) Research resources for *Drosophila*: the expanding universe. *Nat. Rev. Genet.*, **6**, 179–193.

McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.

McVean,G.A. *et al*. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Navarro,A. *et al*. (2000) Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in *Drosophila*. *Genetics*, **155**, 685–698.

Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nelson,C.E. *et al*. (2004) The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.*, **5**, R25.

Nordborg,M. and Tavare,S. (2002) Linkage disequilibrium: what history has to tell us. *Trends Genet.*, **18**, 83–90.

Payseur,B.A. and Nachman,M.W. (2002) Gene density and human nucleotide polymorphism. *Mol. Biol. Evol.*, **19**, 336–340.

Powell,J.R. (1997) *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, New York.

Vilella,A.J. *et al*. (2005) VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*, **21**, 2791–2793.

Wheeler,D.L. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33** (Database issue), D39–D45.

[Fly 1:4, 205-211, July/August 2007]; ©2007 Landes Bioscience

Research Paper

Drosophila Polymorphism Database (DPDB)

A Portal for Nucleotide Polymorphism in Drosophila

Sònia Casillas[1,*]

Raquel Egea[1]

Natalia Petit[1]

Casey M. Bergman[2]

Antonio Barbadilla[1]

[1]Departament de Genètica i de Microbiologia; Universitat Autònoma de Barcelona; Bellaterra (Barcelona), Spain

[2]Faculty of Life Sciences; University of Manchester; Manchester, UK

*Correspondence to: Sònia Casillas; Universitat Autònoma de Barcelona; Departament de Genètica i de Microbiologia; Bellaterra (Barcelona) 08193 Spain; Tel.: 34.935812730; Fax: 34.935812387; Email: sonia.casillas@uab.cat

Original manuscript submitted: 08/10/07
Revised manuscript submitted: 09/14/07
Manuscript accepted: 09/17/07

Previously published online as a *Fly* E-Publication:
http://www.landesbioscience.com/journals/fly/article/5043

KEY WORDS

nucleotide variation, DNA polymorphism, population genetics, Drosophila, database, large scale analyses, bioinformatics of genetic diversity

ABBREVIATIONS & ACRONYMS

DPDB	Drosophila Polymorphism Database
CDS	coding sequence
CNS	conserved noncoding sequence
CON	constructed
EST	expressed sequence tag
GO	gene ontology
GSS	genome sequence scan
HTC	high throughput cDNA sequencing
HTG	high throughput genome sequencing
PAT	patents
PDA	Pipeline Diversity Analysis
SNP	single nucleotide polymorphism
STS	sequence tagged site
SYN	synthetic
TPA	third party annotation
UTR	untranslated region
WGS	whole genome shotgun

ABSTRACT

As a growing number of haplotypic sequences from resequencing studies are now accumulating for Drosophila in the main primary sequence databases, collectively they can now be used to describe the general pattern of nucleotide variation across species and genes of this genus. The Drosophila Polymorphism Database (DPDB) is a secondary database that provides a collection of all well-annotated polymorphic sequences in Drosophila together with their associated diversity measures and options for reanalysis of the data that greatly facilitate both multi-locus and multi-species diversity studies in one of the most important groups of model organisms. Here we describe the state-of-the-art of the DPDB database and provide a step-by-step guide to all its searching and analytic capabilities. Finally, we illustrate its usefulness through selected examples. DPDB is freely available at http://dpdb.uab.cat.

INTRODUCTION

Biological evolution is essentially a process by which genetic variation among individuals within populations is converted into variation between groups in space and time.[1] Genetic variation is the real material of the evolutionary process, and the main aim of population genetics is thus the description and explanation of the forces controlling genetic variation within and between populations.[2] The allozyme era,[1] the era of nucleotide sequences[3] and the current genomics era[4] represent the three major stages of the evolutionary research of genetic diversity. The deciphering of an explosive number of new nucleotide sequences in different genes and species has changed radically the scope of population genetics, transforming it from an empirically insufficient science into a powerfully explanatory interdisciplinary endeavour, where high-throughput instruments generating new sequence data are integrated with bioinformatic tools for data mining and management, and interpreted using advanced theoretical and statistical models.

Drosophila has been the experimental model par excellence to inspire and to test the new developments in molecular population genetics theory.[5,6] Nucleotide studies in this genus involve the resequencing of homologous sequences (haplotypes) for a given DNA region and species. Most of these studies are limited to a few species and genes, although a few studies report tens or hundreds of loci.[7-10] As a growing number of haplotypic sequences from individual studies are now accumulating for this genus in the main molecular biology databases,[11] they can opportunistically be used to describe the pattern of nucleotide variation in many species and genes of this genus.[12] A database describing nucleotide diversity estimates in Drosophila is a necessary resource that greatly facilitates both multi-locus and multi-species diversity studies. The database to be described here is such a bioinformatic resource.

The Drosophila Polymorphism Database (DPDB)[13] is a secondary database designed to provide a collection of all the existing polymorphic sequences in the Drosophila genus together with their associated diversity measures. Estimates of diversity on single nucleotide polymorphisms (SNPs) are provided for each set of haplotypic homologous sequences, including polymorphism at synonymous and non-synonymous sites, linkage disequilibrium and codon bias. Data gathering from GenBank,[11] calculation of diversity measures and daily updates are automatically performed using PDA.[14,15] The DPDB website (http://dpdb.uab.cat) includes several interfaces for browsing the contents of the database and making customizable comparative searches of different species or taxonomic groups. It also contains a set of tools for the reanalysis of data and a statistics section that summarizes the contents of the database. As a result, DPDB aims to be a reference site for DNA polymorphism in Drosophila,[16,17] encompassing studies that try to describe and explain the underlying causes of polymorphic patterns found in these species, such as recombination rate,[18,19] sequence structure and complexity[20,21] or demographic history.[8]

Here we describe the state-of-the-art of the DPDB database and provide a step-by-step guide to all its searching and analytic capabilities. Finally, we illustrate its usefulness by testing a selected population genetics hypothesis, which is solved by performing simple queries using the DPDB interface.

THE CHALLENGE: AUTOMATING THE ESTIMATION OF GENETIC DIVERSITY

The large-scale estimation of genetic diversity from sources of heterogeneous sequences requires the development of elaborate modules of data mining and analysis, which operate together to automatically extract the available sequences from public databases, align them and compute diversity measures. A priori, the automation of this process seems difficult, since variation estimates usually require a careful manual inspection. The main limitation of this process is undoubtedly the heterogeneous nature of the sequences, because such an automatic process can lump together sequences that are fragmented, paralogous, from different populations or chromosome arrangements, or simply incorrectly annotated sequences. Also critical is the multiple alignments of sequences, which is sensitive to the choice of algorithm, the input parameters and the intrinsic characteristics of the sequences. However, millions of haplotypic sequences, including those of complete chromosomes, that are today stored in public databases are an outstanding resource for the estimation of genetic diversity that cannot be neglected. Therefore, while conscious of the limitations, we have tackled the bioinformatic automation of genetic diversity and developed both appropriate methods for data grouping and analysis, and rigorous controls for data quality assessment, to generate the first database of diversity measures in the Drosophila genus. Quality reports considering the source of the sequences and the alignments are provided to check the reliability of the estimates, as well as options for the reanalysis of any set of data.

THE DPDB APPROACH

Data model. A key step in the process of large-scale management of sequence data is to define appropriate bioinformatic data objects that facilitate the storage, representation and analysis of genetic diversity from raw data. DPDB introduces two novel data objects based on two basic storing units: the 'polymorphic set' and the 'analysis unit'. The polymorphic set is a group of homologous sequences for a given gene and species obtained from the public databases. Polymorphic sets are identified by unique set codes in DPDB (e.g., SET000033 corresponds to the set of polymorphic sequences for the gene *Adh* in *D. melanogaster* see Fig. 1). Homologous subgroups are then created for each polymorphic set corresponding to the different annotated functional regions (i.e., CDS, each different exon and intron, 5'UTR, 3'UTR and promoter) with sequences within a subgroup having ≥ 95% sequence identity (otherwise, sequences are split into different subsets). These subgroups are the analysis units on which the commonly used diversity parameters are estimated (e.g., DPpol000025 identifies the current analysis for the CDS region of SET000033, see Fig. 1). Since analysis units within a polymorphic set may be added, removed or changed during daily updates, up-to-date identifiers for the analysis units are not stable (e.g., DPpol001600 is a deprecated analysis unit for the CDS region of the gene *tim* in *D. americana*). Thus, the DPDB contents should be normally linked through set code identifiers (e.g., when linking DPDB from an external database). However, old analysis units can be recovered from the DPDB interface and they may be cited in studies that use specific datasets from DPDB. All the data is stored in a relational MySQL database which was designed according to the DPDB data model (see the Help section in the DPDB website). For a complete description of the DPDB approach and implementation readers are referred to the original publication.[13]

Data gathering and processing. Data collection, alignment and calculation of diversity measures are performed by PDA,[14,15] a pipeline made up of a set of Perl modules that automates the mining and analysis of sequences stored in GenBank.[11] Using PDA we get all the publicly available Drosophila nucleotide sequences from the Entrez Nucleotide database (GenBank) that are well annotated (we exclude sequences from divisions CON, EST, GSS, HTC, HTG, PAT, STS, SYN, TPA and WGS, as well as sequences without gene annotations). We also obtain their cross-references to the NCBI PopSet[22] database and additional information including Gene Ontology (GO)[23] terms from FlyBase.[24] In this last version of the DPDB database, the annotated sequences of the complete chromosomes of *D. melanogaster*[25] are also used for the estimation of genetic diversity. As a result, the number of analysis units in this species has increased by ~50%, since many genes with a single sequence in GenBank in addition to the genome sequence that were previously discarded can now be analyzed together with its corresponding allele in the genomic sequence.

One serious problem in large-scale studies of genetic diversity is the automatic detection of homologous DNA regions. According to the original DPDB data model,[13] homologous sequences were determined based on gene name. However, sequences stored in GenBank use sometimes different names for the same gene, and thus homologous sequences could eventually be grouped into different polymorphic sets in DPDB. To cope with this problem, all gene synonyms recorded for each accepted gene symbol in Drosophila have been downloaded from FlyBase and gene names from GenBank are replaced by their accepted gene symbol before being introduced into the DPDB database. Following this procedure, the fraction of redundant polymorphic sets in the current release of DPDB is expected to be low (~98% of the *D. melanogaster* genes that are currently analyzed in DPDB match an accepted gene symbol in FlyBase).

Once the homologous sequences are determined, sequences are aligned. DPDB originally aligned homologous sequences with ClustalW.[26] However, Muscle[27] and T-Coffee[28] have been shown to achieve a better accuracy, especially in alignments with a high proportion of gaps.[15,29] Thus, in the current release of DPDB all polymorphic sets have been realigned with Muscle and the corresponding diversity measures recalculated. DPDB deals with the problem of non-homology in alignments by grouping sequences by similarity (a 95% minimum identity must exist between each pair of sequences within an alignment). On the other hand, given that sites with gaps are not used for the estimation of single nucleotide polymorphism, inclusion of short sequences tends to reduce the

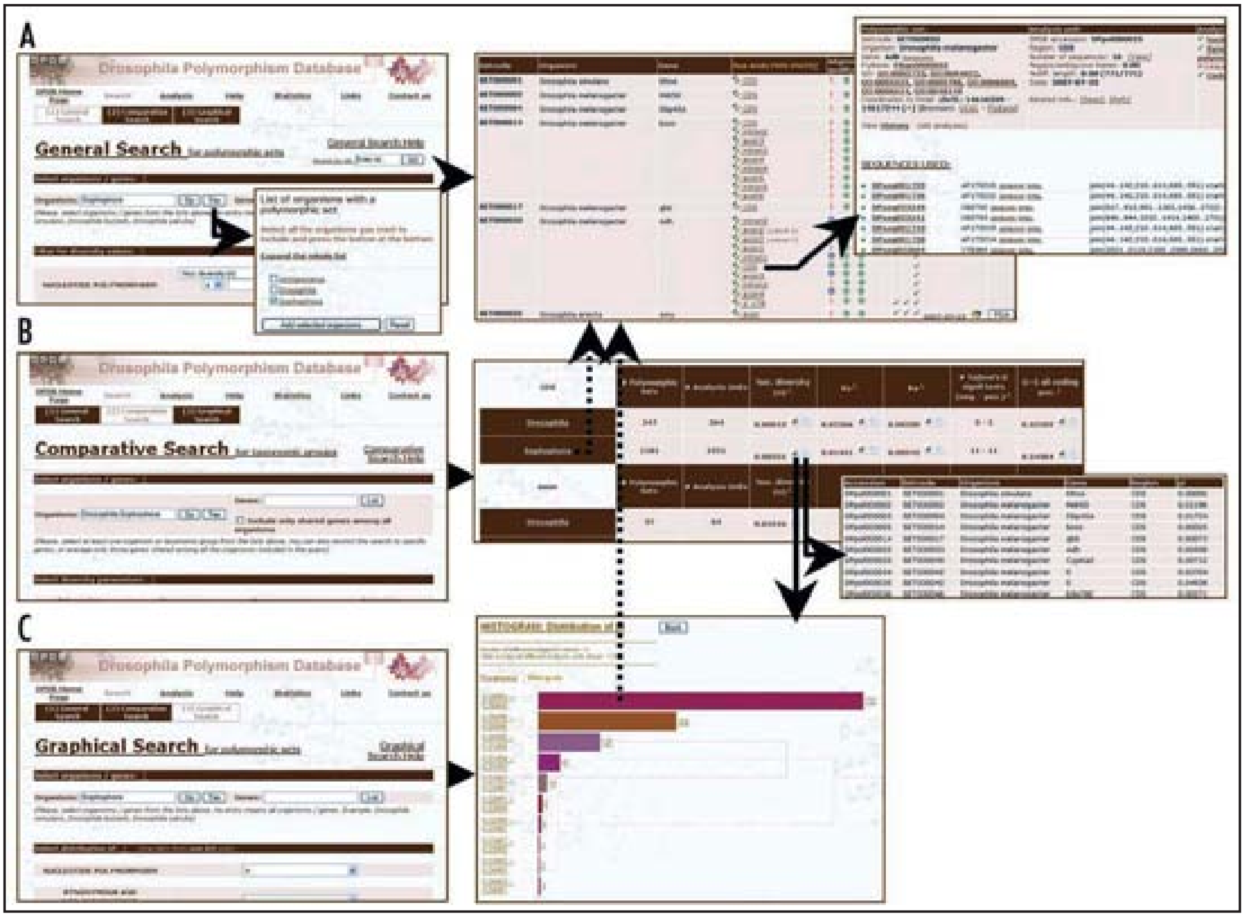

A
Drosophila Polymorphism Database
General Search
List of organisms with a polymorphic set
B
Drosophila Polymorphism Database
Comparative Search
Drosophila
Sophophora
C
Drosophila Polymorphism Database
Graphical Search
NUCLEOTIDE POLYMORPHISM

Figure 1. Example queries using the DPDB interface. (A) General Search (with the taxa selector pop-up window). In the example, all polymorphic sets from the subgenus Sophophora are queried and a part of the complete report for an analysis unit is shown. (B) Comparative Search. In the example, nucleotide diversity is compared for the two Drosophila subgenus (Drosophila and Sophophora). From the results, graphical distributions and lists of data can be obtained, as well as browsing all averaged data within each taxon. Note that queries from the comparative search are always performed by gene region. (C) Graphical Search. In the example, nucleotide diversity for all CDSs from the subgenus Sophophora are displayed graphically. Dashed arrows: these links would display only a subset of the data shown in the image (i.e., only CDSs from the comparative search, and only CDSs with $\pi < 0.00498$ from the graphical search).

amount of informative content in an alignment. As a result, DPDB has recently implemented an algorithm for the maximization of the number of informative sites of an alignment[15] in which sequences are grouped by length in order to score the largest number of informative sites for the calculation of diversity measures.

Finally, each alignment is mapped to the genome sequence of *D. melanogaster*.[25] First, a consensus sequence is obtained from the multiple sequence alignment. The consensus is then aligned to the *D. melanogaster* genome using BLAT[30] and the corresponding coordinates are obtained and provided with the alignments. These coordinates are used to link each analysis unit to the genome browsers in FlyBase[24] and UCSC.[31] This allows users to integrate analyses of polymorphism within species with other comparative or functional genomic resources that are aligned to the reference genome sequence. Additional links to FlyBase based on gene name, and related GO terms are also provided.

Confidence assessment of each polymorphic set. DPDB provides several measures to assess the confidence of each polymorphic set, according to both the data source and the quality of the alignment. For the data source, we provide four criteria to help determining if the sequences initially were reported as part of a polymorphism study: (1) one or more sequences from the alignment are stored in the NCBI PopSet database, (2) all the sequences have consecutive GenBank accession numbers, (3) all the sequences share at least one reference, and (4) one or more references are from journals that typically publish polymorphism studies (i.e., Genetics, Journal of Molecular Evolution, Molecular Biology and Evolution, Molecular Phylogenetics and Evolution and Molecular Ecology). To assess the quality of an alignment we use three other criteria: (1) the number of sequences included in the alignment, (2) the percentage of gaps or ambiguous bases within the alignment, and (3) the percentage of difference in length between the shortest and the longest sequences. Users of DPDB are advised to: (1) revise all the previous parameters, (2) check the alignments and phylogenetic trees provided for each alignment, (3) revise the origin of the sequences, (4) pay special attention to estimates of polymorphism giving extreme values, and (5) reanalyze the data when needed. See a detailed guide to assess the reliability of the estimations in the Help section of the DPDB website.

USING DPDB

DPDB allows browsing the contents of the database and reanalyzing any subset of data by using different searching and analytic tools available from the website (http://dpdb.uab.cat). For example, the General Search is used to browse the DPDB database by organism, gene and gene region (Fig. 1A) and allows filtering of results by diversity estimations or quality of the data. The Comparative Search is used to compare average nucleotide diversity estimates in different species or taxonomic groups (Fig. 1B), and the Graphical Search generates graphical distributions for any diversity measure on any taxon or gene of interest (Fig. 1C). Notably, users can easily move from one tool to another within the same dataset (Fig. 1). Table 1 summarizes all different searching and analytic options to help navigating around the DPDB website. Each option is illustrated by a step-by-step example. To perform complex queries to the DPDB database or do any meta-analysis of the available data, users are encouraged to download the complete MySQL database (Table 1).

CONTENTS OF THE DPDB DATABASE AND QUALITY OF THE DATA

At the time of writing, the DPDB database contained > 40,000 sequences from GenBank, corresponding to 392 species and 15,177 different genes (Fig. 2). When these sequences were filtered and analyzed, DPDB could gather informative data for 1,898 polymorphic sets (from 145 species and 1,184 different genes), and estimations were calculated on 3,741 analysis units, mostly corresponding to the functional regions CDS, exon and intron. The best-represented species was *D. melanogaster* (53.2% of all analysis units), and the gene with the highest number of alignments was *Adh* (5.3% of all analysis units), which Drosophilists should be proud to note is the first gene in any species whose population genetics was studied using resequencing methods.[3] In terms of quality of the alignments, many estimates were performed on alignments with < 6 sequences (45.2%), but most of the alignments contain < 10% of gaps or ambiguous bases (95.5%) and small differences in sequence length (84.8%). In terms of quality of the data source, only 26% of the analysis units contain sequences from the NCBI PopSet database, which means that DPDB contains an additional 3-fold more genomic regions that would otherwise be overlooked if only sequences from polymorphism studies deposited in NCBI PopSet were searched. The PDA retrieval system used in the construction of the DPDB database has thus provided a notable enrichment of the available diversity data. Daily-updated statistics of the DPDB database can be monitored at the Statistics section of the DPDB website.

DPDB IN ACTION

As we have detailed above, DPDB provides estimates of nucleotide diversity for a large number of genes and species of Drosophila, which in conjunction with the web interface greatly facilitate both multi-species and multi-locus genetic diversity analyses by providing options to make totally customizable queries. Some examples of simple queries to the database involving one or more taxa have already been proposed in Table 1. More interestingly, subsets of data from DPDB have already been used in large-scale analyses of nucleotide diversity. For example, Petit et al.[21] have recently studied the association between coding polymorphism levels (estimates obtained from DPDB), intron content and expression patterns in *D. melanogaster*. The study reports that genes with low nonsynonymous polymorphism contain long introns with a high content of conserved noncoding sequences (CNSs), and that genes with CNSs in their introns have more complex regulation. Also, Casillas et al.[29] show the action of purifying selection maintaining highly conserved noncoding sequences by combining genomic data from recently completed insect genome projects with population genetic data in *D. melanogaster*. For this study, a selected set of noncoding data from genome scans was gathered and analyzed both for point mutations and insertions/deletions (indels) using PDA. As another example of use of DPDB, here we consider a simple study investigating a potential association between levels of synonymous polymorphism in different groups of Drosophila and the length of the genetic map of the genome (the total recombination rate).

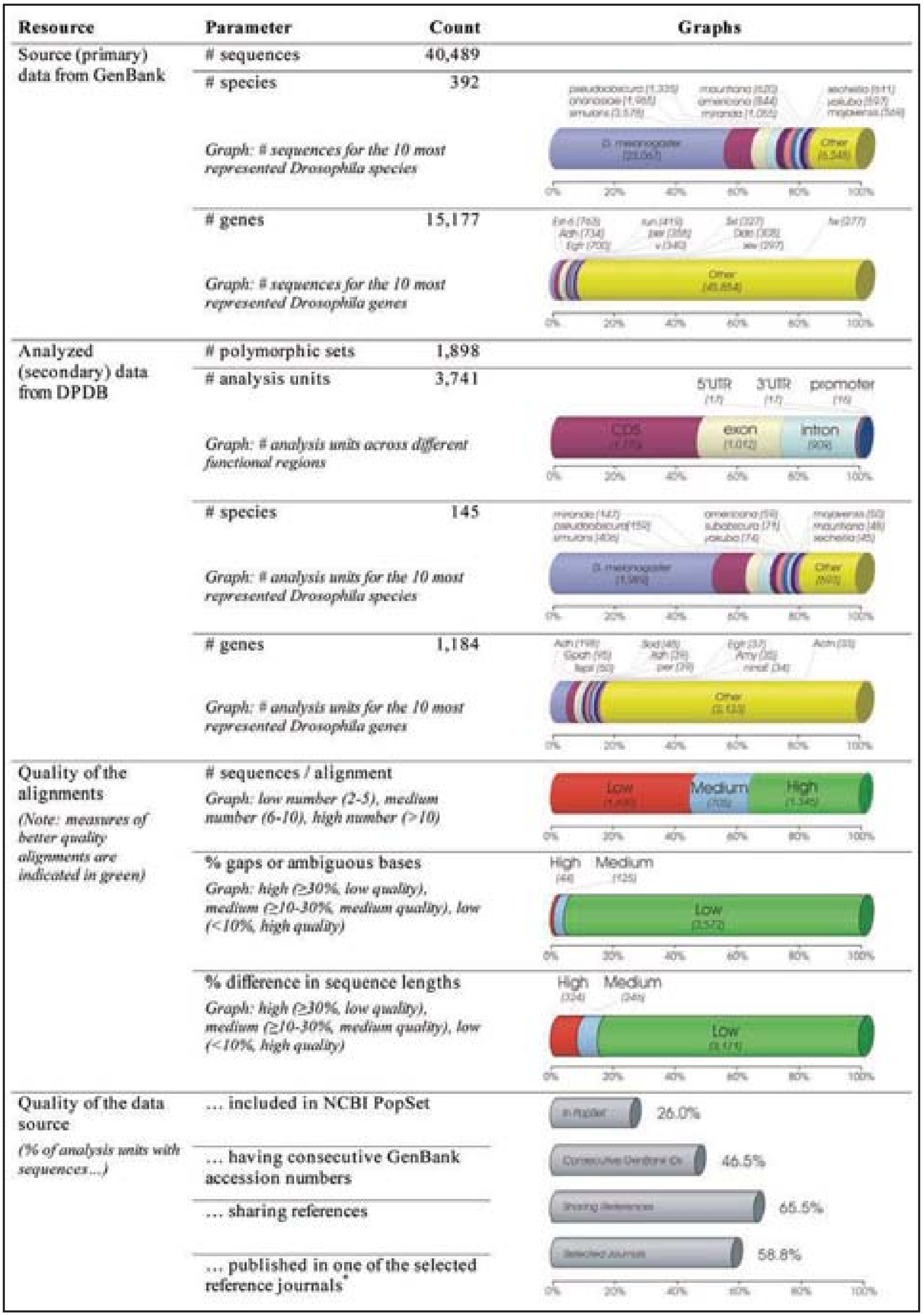
In regions of low recombination, neutral polymorphism is swept away by the action of both positive and background selection on linked selected mutations.[18,35] As a result, the level of nucleotide polymorphism is expected to be positively correlated with the level of recombination rate in Drosophila.[18] Average recombination rate is significantly different among groups of species in Drosophila,[36] and thus levels of synonymous polymorphism are also expected to vary among groups according to the rate of recombination. To test this hypothesis, we have performed comparative searches using the DPDB website.

We have obtained genomic recombination estimates for six species' groups of *Drosophila* (*funebris*, *melanogaster*, *obscura*, *repleta*, *saltans* and *tripunctata*) from Cáceres et al.[36] and correlated them to the estimates of synonymous polymorphism in coding sequences (CDS) for the same groups reported in DPDB (see Supplementary Material). The resulting correlation was positive but non-significant ($r_{\text{Pearson}} = 0.00178$, $p = 0.94790$, $N = 1352$). However, the same correlation became highly significant when *D. melanogaster* was excluded from the analysis ($r_{\text{Pearson}} = 0.13693$, $p = 0.00075$, $N = 603$). One possible explanation is that *D. melanogaster*, which accounts for ~70% of the data in its group, has an unusually small effective population size as a consequence of a bottleneck suffered after its dispersion out of Africa.[37] Small effective population size is known to cause low levels of synonymous polymorphism in the genome, and this may explain why the effect of recombination rate on synonymous polymorphism is undetectable at the group level when *D. melanogaster* is included in the analysis.

This example illustrates the power of DPDB to reveal new knowledge about the evolutionary process in Drosophila without the need for labor-intensive sequence retrieval and data processing on the part of the user. The wide range of potential queries that can be performed using the searching capabilities of the DPDB website remarkably facilitate comprehensive multi-species and multi-locus analyses of nucleotide diversity. Future improvements to DPDB will include the integration of divergence data (i.e., outgroup sequences to each polymorphic set), additional tests of neutrality such as the McDonald-Kreitman test[38] and derived allele frequency distributions, as well as the estimation of indel polymorphism. We will also create a specific section within DPDB that will include all the

Table 1 **Step-by-step guide to DPDB**

Option	Function	Usage
General Search	Browse the contents of the DPDB database by polymorphic sets *E.g., retrieve all analysis units from* D. melanogaster *having high-quality alignments (> 5 sequences, < 30% gaps, < 30% difference in sequences length and ≥ 500 analyzed sites) with low values of nucleotide diversity (p ≤ 0.001)*	1. Go to Search → General Search 2. Select organisms/genes: click the 'Sp' button; in the pop-up window, check *Drosophila melanogaster* and click the button 'Add selected organisms' at the bottom of the pop-up window 3. Filter for diversity values: on 'Nucleotide polymorphism', choose 'Nuc. Diversity (π)'≤0.001 4. Filter for degree of confidence on the polymorphic set: choose 'Num. of sequences' > 5, 'Perc. gaps/ambiguous bases'< 30%, 'Perc. difference in seqs. length'< 30% and 'Exclude alignments with less than 500 analyzed sites' 5. Click the button 'Run Search' 6. See detailed reports for each analysis unit (fourth column)
Comparative Search	Summarize and compare diversity measures across species or taxonomic groups[32] *E.g., compare nucleotide diversity between the cosmopolitan species* D. simulans *and the endemic species* D. sechellia *including only those genes that have been analyzed in both species*	1. Go to Search → Comparative Search 2. Select organisms/genes: click the 'Sp' button; in the pop-up window, check *Drosophila simulans* and *Drosophila sechellia*, and click the button 'Add selected organisms' at the bottom of the pop-up window. Then check the box 'Include only shared genes among all organisms' 3. Select diversity parameters: leave defaults 4. Filter for degree of confidence on the polymorphic set: remove default values to include all alignments 5. Click the button 'Run Search'
Graphical Search	Generate graphical distributions for any diversity measure on any taxon or gene of interest *E.g., generate the distribution of π values for any coding sequence in* D. melanogaster, *including only high-quality alignments*	1. Go to Search → Graphical Search 2. Select organisms/genes: click the 'Sp' button, check *Drosophila melanogaster*, and click the button 'Add selected organisms' at the bottom of the page 3. Select distribution of: leave default ('π' distribution) 4. Filter for degree of confidence on the polymorphic set: see General Search above 5. Advanced options: in 'Regions to be displayed', unselect all regions but 'CDS' 6. Click the button 'Run Search'
Search by DPDB or GenBank accession number	Search the DPDB database by any DPDB or GenBank accession number *E.g., display all analysis units which use the GenBank sequence AF175215*	1. Go to Search → General Search 2. Search by Id: enter the accession number AF175215 in the box 3. Click the button 'Go' *Note: this option accepts any DPDB accession (e.g., SET000033, DPpol000025, DPseq001739) or GenBank accession (e.g., AF175215, AF175215.1, 6002968)*
Search by sequence similarity	Search the DPDB database by sequence similarity *E.g., find homologous sequences to a query sequence in DPDB which can be used later to estimate nucleotide diversity*	1. Go to Analysis → Sequence comparison → Blast 2. Choose your custom Blast parameters or leave defaults 3. Paste your sequence in the appropriate box 4. Click the button 'Run Blast' *Note: the Blast package implemented in DPDB also allows sequence similarity searches to any of the 12* Drosophila *sequenced genomes or* Anopheles gambiae
Other analysis tools (all tools available from the Analysis section in the website)	Standard sequence comparison tools (Blast,[33] Clustal,[26] Jalview[34]), and specific software for the estimation of nucleotide diversity (SNPs-Graphic,[13] PDA[15]) *E.g., following the previous example, perform a nucleotide diversity study*	1. Once you have a set of homologous sequences (see above), align them with Clustal and revise/edit the alignment with Jalview 2. Then perform a sliding-window analysis with SNPs-Graphic 3. Alternatively, perform the analysis automatically with PDA *Note: SNPs-Graphic and PDA can be used independently with custom data (as in the previous example) or be executed from searches to the DPDB database to reanalyze specific datasets*
Local installation of DPDB	Perform complex queries to the DPDB database or do any meta-analysis of the data	1. Go to DPDB Home Page → Database download 2. Download and install the database following the instructions



Resource	Parameter	Count	Graphs
Source (primary) data from GenBank	# sequences	40,489	
	# species	392	
	Graph: # sequences for the 10 most represented Drosophila species		
	# genes	15,177	
	Graph: # sequences for the 10 most represented Drosophila genes		
Analyzed (secondary) data from DPDB	# polymorphic sets	1,898	
	# analysis units	3,741	
	Graph: # analysis units across different functional regions		
	# species	145	
	Graph: # analysis units for the 10 most represented Drosophila species		
	# genes	1,184	
	Graph: # analysis units for the 10 most represented Drosophila genes		
Quality of the alignments *(Note: measures of better quality alignments are indicated in green)*	# sequences / alignment *Graph: low number (2-5), medium number (6-10), high number (>10)*		
	% gaps or ambiguous bases *Graph: high (≥30%, low quality), medium (≥10-30%, medium quality), low (<10%, high quality)*		
	% difference in sequence lengths *Graph: high (≥30%, low quality), medium (≥10-30%, medium quality), low (<10%, high quality)*		
Quality of the data source *(% of analysis units with sequences...)*	... included in NCBI PopSet		
	... having consecutive GenBank accession numbers		
	... sharing references		
	... published in one of the selected reference journals*		

Figure 2. DPDB contents and quality of the data. Statistics are according to August 1, 2007. *Reference journals include: *Genetics, Journal of Molecular Evolution, Molecular Biology and Evolution, Molecular Phylogenetics* and *Evolution and Molecular Ecology.*

SNPs discovered using the PDA system. Finally, we will develop new methods to deal with unannotated noncoding sequences from genome scans[8-10] and data coming from SNP mapping studies (http://flysnp.imp.ac.at/), whole genome shotgun and tiling array resequencing (http://www.dpgp.org/). It is thus our goal that DPDB becomes a comprehensive reference site for intraspecific genetic variation in Drosophila, describing different types of genetic variation (e.g., SNPs and indels), distinct functional regions (e.g., coding and noncoding) and accepting diverse sources of data (e.g., resequencing data, SNP typing and whole genome sequencing).

Acknowledgements

The authors would like to thank Rolf Apweiler, Fernando Ayllón, Oriol Calvete, Mar Marzo, Alejandra Delprat, Marta Puig, Alfredo Ruiz and two anonymous reviewers for valuable comments on this manuscript and/or the DPDB website. This work was funded by the Ministerio de Ciencia y Tecnología (Grants BMC2002-01708 and BFU2006-08640/BMC). Sònia Casillas was supported by the Ministerio de Ciencia y Tecnología (Grant BES-2003-0416) and by a Marie Curie fellowship from the European Commission (HPMT-GH-01-00285-13). Raquel Egea was supported by the Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya i el Fons Social Europeu (Grant 2005FI-00328). Natalia Petit was supported by a grant from the Departament de Genètica i de Microbiologia of the Universitat Autònoma de Barcelona.

Note

Supplementary Material can be found at http://www.landesbioscience.com/supplement/CasillasFLY1-4-sup.pdf

References

1. Lewontin RC. The Genetic Basis of Evolutionary Change. New York: Columbia University Press, 1974.
2. Lewontin RC. Directions in evolutionary biology. Annu Rev Genet 2002; 36:1-18.
3. Kreitman M. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 1983; 304:412-7.
4. Li WH, Gu Z, Wang H, Nekrutenko A. Evolutionary analyses of the human genome. Nature 2001; 409:847-9.
5. Aquadro CF, Bauer DuMont V, Reed FA. Genome-wide variation in the human and fruitfly: A comparison. Curr Opin Genet Dev 2001; 11:627-34.
6. Powell JR. Progress and Prospects in Evolutionary Biology: The Drosophila Model. NY: Oxford University Press, 1997.
7. Andolfatto P. Adaptive evolution of non-coding DNA in Drosophila. Nature 2005; 437:1149-52.
8. Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: A multi-locus approach. Genetics 2003; 165:1269-78.
9. Ometto L, Glinka S, De Lorenzo D, Stephan W. Inferring the eff-ects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol 2005; 22:2119-30.
10. Orengo DJ, Aguade M. Detecting the footprint of positive selection in a european population of *Drosophila melanogaster*: Multilocus pattern of variation and distance to coding regions. Genetics 2004; 167:1759-66.
11. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2007; 35:D21-5.
12. Pandey A, Lewitter F. Nucleotide sequence databases: A gold mine for biologists. Trends Biochem Sci 1999; 24:276-80.
13. Casillas S, Petit N, Barbadilla A. DPDB: A database for the storage, representation and analysis of polymorphism in the Drosophila genus. Bioinformatics 2005; 21:ii26-ii30.
14. Casillas S, Barbadilla A. PDA: A pipeline to explore and estimate polymorphism in large DNA databases. Nucleic Acids Res 2004; 32:W166-9.
15. Casillas S, Barbadilla A. PDA v.2: Improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. Nucleic Acids Res 2006; 34: W632-4.
16. Matthews KA, Kaufman TC, Gelbart WM. Research resources for Drosophila: The expanding universe. Nat Rev Genet 2005; 6:179-93.
17. Galperin MY. The molecular biology database collection: 2007 update. Nucleic Acids Res 2007; 35:D3-4.
18. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 1992; 356:519-20.
19. Betancourt AJ, Presgraves DC. Linkage limits the power of natural selection in Drosophila. Proc Natl Acad Sci USA 2002; 99:13616-20.
20. Nelson CE, Hersh BM, Carroll SB. The regulatory content of intergenic DNA shapes genome architecture. Genome Biol 2004; 5:R25.
21. Petit N, Casillas S, Ruiz A, Barbadilla A. Protein polymorphism is negatively correlated with conservation of intronic sequences and complexity of expression patterns in *Drosophila melanogaster*. J Mol Evol 2007; 64:511-8.
22. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the national center for biotechnology information. Nucleic Acids Res 2007; 35:D5-12.
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: Tool for the unification of biology: The gene ontology consortium. Nat Genet 2000; 25:25-9.
24. Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM. FlyBase: Genomes by the dozen. Nucleic Acids Res 2007; 35:D486-91.
25. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, Woodage T, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC. The genome sequence of *Drosophila melanogaster*. Science 2000; 287:2185-95.
26. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003; 31:3497-500.
27. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32:1792-7.
28. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000; 302:205-17.
29. Casillas S, Barbadilla A, Bergman CM. Purifying selection maintains highly conserved noncoding sequences in Drosophila. Mol Biol Evol 2007.
30. Kent WJ. BLAT-the BLAST-like alignment tool. Genome Res 2002; 12:656-64.
31. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ. The UCSC genome browser database: Update 2006. Nucleic Acids Res 2006; 34:D590-8.
32. Egea R, Casillas S, Fernandez E, Senar MA, Barbadilla A. MamPol: A database of nucleotide polymorphism in the Mammalia class. Nucleic Acids Res 2007; 35:D624-9.
33. McGinnis S, Madden TL. BLAST: At the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 2004; 32:W20-5.
34. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. Bioinformatics 2004; 20:426-7.
35. Charlesworth B. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet Res 1996; 68:131-49.
36. Caceres M, Barbadilla A, Ruiz A. Recombination rate predicts inversion size in Diptera. Genetics 1999; 153:251-9.
37. Andolfatto P. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and Drosophila simulans. Mol Biol Evol 2001; 18:279-90.
38. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature 1991; 351:652-4.

Supplementary Material

1. Comparative Search to DPDB: Query options

1. Select organisms/genes:

- Taxonomic groups: funebris group, obscura group, repleta group, tripunctata group, melanogaster group, saltans group

2. Select diversity parameters: defaults

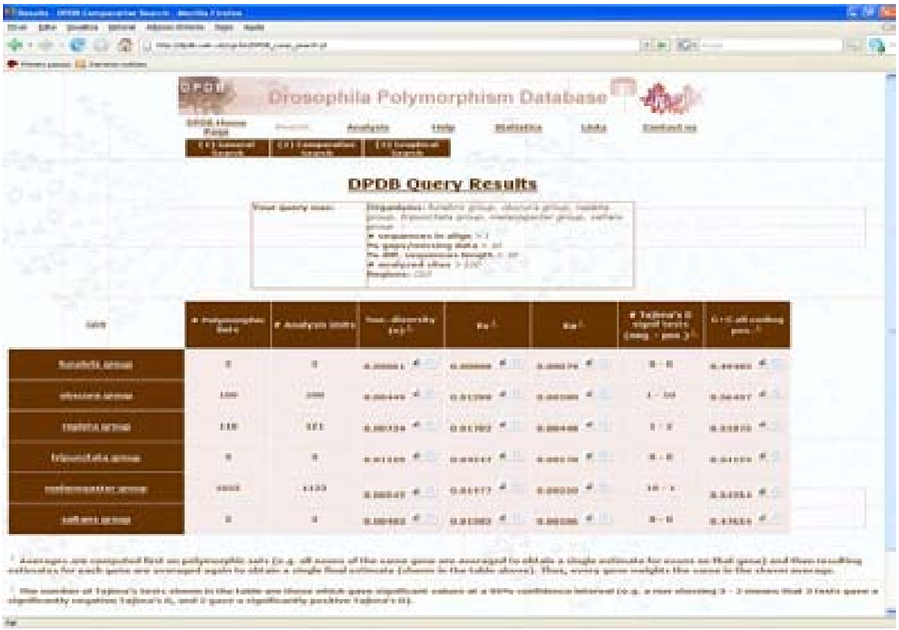
3. Filter for degree of confidence on the polymorphic set:

- Num. of sequences > 1
- Perc. gaps / ambiguous bases < 30%
- Perc. difference in seqs. length < 30%
- Exclude alignments with less than 100 analyzed sites

4. Advanced options:

- Regions to be displayed: CDS

2. Results: Screen-shot



DPDB Query Results

Your query was: Organisms: funebris group, obscura group, repleta group, tripunctata group, melanogaster group, saltans group
# sequences in align. > 1
% gaps/missing data < 30
% diff. sequences length < 30
# analyzed sites > 100
Regions: CDS

Gene	# Polymorphic Sets	# Analysis Units	Nuc. diversity (π)[1]	Ks[1]	Ka[1]	# Tajima's D signif. tests (neg. - pos.)[2]	G+C all coding pos.[1]
funebris group	0	0	[illegible]	[illegible]	[illegible]	0 - 0	[illegible]
obscura group	100	100	[illegible]	[illegible]	[illegible]	1 - 10	[illegible]
repleta group	118	121	[illegible]	[illegible]	[illegible]	1 - 2	[illegible]
tripunctata group	0	0	[illegible]	[illegible]	[illegible]	0 - 0	[illegible]
melanogaster group	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	16 - 1	[illegible]
saltans group	0	0	[illegible]	[illegible]	[illegible]	0 - 0	[illegible]

[1] Averages are computed first on polymorphic sets (e.g. all scores of the same gene are averaged to obtain a single estimate for scores on that gene) and then resulting estimates for each gene are averaged again to obtain a single final estimate (shown in the table above). Thus, every gene weights the same in the shown average.

[2] The number of Tajima's tests shown in the table are those which gave significant values at a 95% confidence interval (e.g. a row showing 3 - 2 means that 3 tests gave a significantly negative Tajima's D, and 2 gave a significantly positive Tajima's D).

3. Results: Data

[This table is available in the accompanying CD]

4. Results: Correlations

Pearson's product-moment correlations:

	$r_{Pearson}$	t	df	p-value
All data	0.00178	0.0654	1350	0.94790
Excl. *D. melanogaster*	0.13693	3.3887	601	0.00075

Amount of data analyzed in each group:

	All data	Excl. *D. melanogaster*
funebris	2	2
obscura	100	100
repleta	121	121
tripunctata	3	3
melanogaster	1123	374
(of which D. melanogaster)	*749*	*0*
saltans	3	3
TOTAL	**1352**	**603**

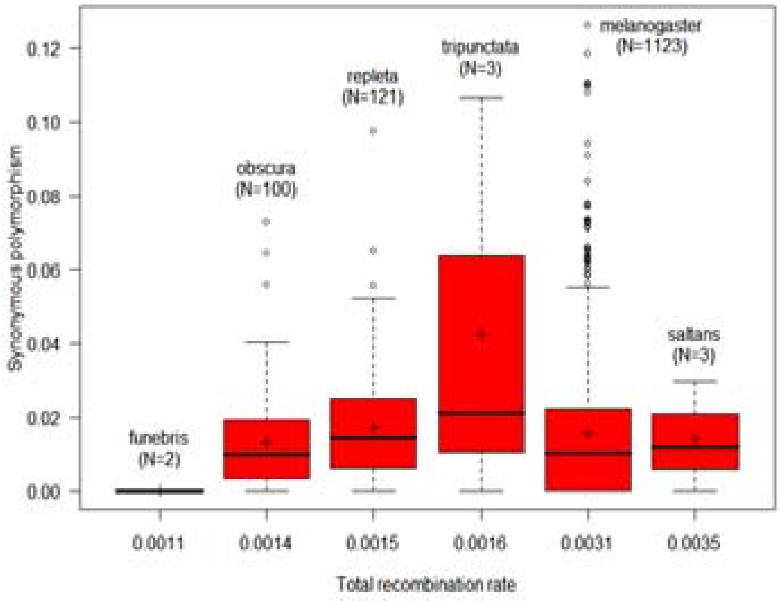


Figure 1.

Relationship between synonymous polymorphism and total recombination rate among different groups of species in *Drosophila*. Horizontal bar: median. Star: mean. Box: interquartile range. Whiskers: range of the data up to 1.5 times the interquartile range. Open dots: extreme values.

2.3. Using patterns of sequence evolution to infer functional constraint in *Drosophila* noncoding DNA

Here, patterns of sequence evolution are used to infer functional constraint and adaptation in conserved noncoding regions of *Drosophila*. We have compiled published re-sequencing polymorphic data in *D. melanogaster* with comparative genomic data from recently completed insect genome projects using the bioinformatic system PDA and tested basic predictions of the mutational cold-spot model of CNS evolution. We present strong evidence that both intronic and intergenic CNSs harbor deleterious alleles in natural populations that are prevented from going to fixation by the action of purifying selection. Furthermore, we estimate the strength of this selection and show that a large proportion of CNS sites are evolving under a moderate negative selection. In addition, we find that non-CNS regions also show some evidence of purifying selection stronger than four-fold synonymous coding sites. Controlling for the effects of negative selection, we find no evidence of positive selection acting on *Drosophila* CNSs, although we do find evidence for the action of recurrent positive selection in the spacer regions between CNSs. Intriguingly, results for SNPs and indels show different trends and provide further evidence that the non-CNS regions are under purifying selection, which may be stronger in the case of indels. Our results that CNSs are selectively constrained support similar findings in other mammalian genomes and argue against the general likelihood that CNSs are generated by mutational cold-spots in any metazoan genome. Also, they provide concrete evidence to support a widely-held assumption that underpins comparative genomic methods to detect noncoding *cis*-regulatory sequences and RNA genes from sequence conservation across related species.

- **Article 5:** Casillas, S., A. Barbadilla and C. M. Bergman (2007) Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Molecular Biology and Evolution* **24:** 2222-2234.

Purifying Selection Maintains Highly Conserved Noncoding Sequences in *Drosophila*

Sònia Casillas,† Antonio Barbadilla,† and Casey M. Bergman**

*Faculty of Life Sciences, University of Manchester, Michael Smith Building, Manchester M13 9PT, UK; and
†Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

The majority of metazoan genomes consist of nonprotein-coding regions, although the functional significance of most noncoding DNA sequences remains unknown. Highly conserved noncoding sequences (CNSs) have proven to be reliable indicators of functionally constrained sequences such as *cis*-regulatory elements and noncoding RNA genes. However, CNSs may arise from nonselective evolutionary processes such as genomic regions with extremely low mutation rates known as mutation "cold spots." Here we combine comparative genomic data from recently completed insect genome projects with population genetic data in *Drosophila melanogaster* to test predictions of the mutational cold spot model of CNS evolution in the genus *Drosophila*. We find that point mutations in intronic and intergenic CNSs exhibit a significant reduction in levels of divergence relative to levels of polymorphism, as well as a significant excess of rare derived alleles, compared with either the nonconserved spacer regions between CNSs or with 4-fold silent sites in coding regions. Controlling for the effects of purifying selection, we find no evidence of positive selection acting on *Drosophila* CNSs, although we do find evidence for the action of recurrent positive selection in the spacer regions between CNSs. We estimate that ~85% of sites in *Drosophila* CNSs are under constraint with selection coefficients ($N_e s$) on the order of 10–100, and thus, the estimated strength and number of sites under purifying selection is greater for *Drosophila* CNSs relative to those in the human genome. These patterns of nonneutral molecular evolution are incompatible with the mutational cold spot hypothesis to explain the existence of CNSs in *Drosophila* and, coupled with similar findings in mammals, argue against the general likelihood that CNSs are generated by mutational cold spots in any metazoan genome.

Introduction

A largely unexplained structural feature common to metazoan genomes is the presence of vast amount of nonprotein-coding DNA (Britten and Davidson 1969; Taft and Mattick 2003). For example, over 75% of the euchromatic portion of the *Drosophila melanogaster* genome sequence is found in noncoding intronic and intergenic regions (Misra et al. 2002). Unlike large mammalian genomes, the overwhelming majority of noncoding DNA in the *D. melanogaster* genome sequence is unique, with less than 6% of the genome confidently identified as repetitive transposable element sequences (Quesneville et al. 2005; Bergman et al. 2006). Of these 80+ Mb of unique noncoding sequences in *D. melanogaster*, a minimum of ~20–30% has been shown to be highly conserved among other insect species (Bergman and Kreitman 2001; Bergman et al. 2002; Siepel et al. 2005). These highly conserved noncoding sequences (CNSs) are typically interpreted as representing the signature of functionally constrained elements maintained by purifying selection and have been successfully used to guide the prediction of *cis*-regulatory sequences (Bergman et al. 2002; Costas et al. 2004) and functional noncoding RNAs (Enright et al. 2003; Lai et al. 2003).

Despite the widespread assumption that CNSs are maintained by the action of purifying selection, the first results on the length distribution of CNSs reported in Bergman and Kreitman (2001) were shown to be compatible with a nonselective mode of evolution that invokes only mutation and genetic drift (Clark 2001), albeit by assuming extremely low mutation rates that vary over spatial scales on the order of tens of base pairs. This lead Clark (2001) to propose the "mutational cold spot" hypothesis to explain the existence of CNSs as an alternative to the functional constraint hypothesis (see also Shabalina and Kondrashov 1999). Subsequently, arguments against the mutational cold spot hypothesis were levied on the grounds of the nonrandom spatial distribution of CNSs in *Drosophila* (Bergman et al. 2002) and *Caenorhabditis* (Webb et al. 2002), a pattern which would further require a nonrandom organization of mutational cold spots to explain the pattern of CNS evolution. Although no molecular mechanism has been shown to produce such localized mutation cold spots, local variation in mutation rates remains a formal possibility that must be investigated more thoroughly to demonstrate the general property that CNSs are indeed selectively constrained.

Recently, the mutational cold spot hypothesis has been tested directly using population genetic data from single-nucleotide polymorphism (SNP) projects in the human genome (Keightley, Kryukov et al. 2005; Drake et al. 2006). Both the mutational cold spot hypothesis and the functional constraint hypothesis predict that levels of within-species variation in CNSs should be reduced relative to flanking non-CNS sequences. This is because either low mutation rates or the elimination of deleterious alleles by purifying selection can reduce the number of segregating polymorphisms observed in CNSs. In contrast to levels of variation, the 2 hypotheses make distinct predictions about the distribution of allele frequencies in CNS and non-CNS regions. Under the mutation cold spot hypothesis, CNS and non-CNS regions should not differ from one another in their allele frequency spectra because their dynamics are both governed only by mutation and genetic drift. In contrast, the functional constraint hypothesis predicts that CNSs should harbor more rare derived alleles relative to non-CNS regions if CNSs are maintained by purifying

Key words: noncoding DNA, comparative genomics, mutation, constraint, adaptation, *Drosophila*.

E-mail: casey.bergman@manchester.ac.uk.

Mol. Biol. Evol. 24(10):2222–2234. 2007
doi:10.1093/molbev/msm150
Advance Access publication July 23, 2007

selection that confines weakly deleterious alleles within low population frequencies. Tests of these alternative models based on differences in derived allele frequency (DAF) distributions have concluded that CNSs in the human genome exhibit an excess of rare derived alleles, consistent with functional constraint acting to maintain CNSs in mammals (Keightley, Kryukov et al. 2005; Drake et al. 2006). Similar results have also been reported for SNPs in predicted microRNA binding sites in the 3′ untranslated regions (UTRs) of human genes (Chen and Rajewsky 2006). Moreover, analysis of the distribution of fitness effects on mutations in mammalian CNSs has revealed the action of weak purifying selection (Kryukov et al. 2005), a mode of evolution which would allow deleterious SNPs to be observed in nature but prevent most of them from reaching high population frequency or going to fixation.

Here we investigated the evolutionary forces governing the evolution of CNSs in the *Drosophila* genome, using recently available comparative genomic data from insect genome projects (Holt et al. 2002; Richards et al. 2005; The Honeybee Genome Sequencing Consortium 2006; http://rana.lbl.gov/drosophila) combined with population genetic data from published nucleotide polymorphism studies in noncoding regions of the X chromosome in *D. melanogaster* (Glinka et al. 2003; Orengo and Aguade 2004; Ometto, Glinka et al. 2005). We extend previous within-species analysis of evolutionary dynamics in CNSs by testing a second prediction of the mutational cold spot hypothesis, based on a modified version of the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991), similar to that used to test differences between transcription factor binding sites and spacer regions in *cis*-regulatory sequences (Jenkins et al. 1995; Ludwig and Kreitman 1995). Specifically, we test whether the ratio of polymorphism within *D. melanogaster* relative to divergence with its sister species *Drosophila simulans* is the same in CNS and non-CNS regions, as is expected under a model of strictly neutral evolution required by the mutational cold spot hypothesis. By using resequencing data rather than data from SNP studies as used in the studies in mammals cited above, we have access to more comprehensive, unbiased population genetic data from contiguous genomic sequences. This permits us to test the mutational cold spot hypothesis more rigorously using both the MK test as well as the DAF test and further allows us to investigate the impact of both single nucleotide as well as insertion/deletion (indel) mutations on the evolution of CNSs.

Materials and Methods

Compilation of Published Sequence Data

We have used population genetic data from genome scans of noncoding regions homogeneously distributed across the X chromosome of *D. melanogaster* (Glinka et al. 2003; Orengo and Aguade 2004; Ometto, Glinka et al. 2005). These data include ~12 alleles per locus from each of 3 distinct data sets: AFR is an African sample from Lake Kariba, Zimbabwe; EUR1 is a European sample from Leiden, The Netherlands; and EUR2 is another European sample from Sant Sadurní d'Anoia, Catalonia, Spain. We used a modified version of Pipeline Diversity Analysis (PDA) (Casillas and Barbadilla 2006) to automatically retrieve noncoding sequences from Genbank, classify and align them by population and locus (see below), and estimate diversity measures. Data for coding regions to calculate levels of silent site polymorphism and divergence was obtained from Andolfatto (2005).

We selected one allele from each data set and used BLAT (Kent 2002) to map each alignment to the *D. melanogaster* genome sequence, which also provided an additional allele to each locus that was used only for reference purposes but which was never included in the polymorphism analyses, thus conserving the unique origin of the sequences in each population data set. Based on preliminary results indicating unusual properties of indels in the *D. simulans* composite whole-genome shotgun assembly, we used the orthologous *D. simulans* allele provided by the authors in the original papers. For loci where this sequence was not available, we obtained the orthologous *D. simulans* regions using whole-genome alignments from the VISTA browser (Couronne et al. 2003) using the genomic coordinates of *D. melanogaster*. *D. simulans* alleles were used to polarize polymorphic SNPs and indel polymorphisms (IPs) and to define single-nucleotide fixed differences (SNFs) and indel fixed differences (IFs).

Multiple Sequence Alignment

We evaluated the performance of several programs to generate multiple alignments of alleles by correlating estimates of polymorphism and divergence obtained automatically here with estimates derived from manually curated alignments previously reported in the primary publications. Based on this analysis, we chose to use MUSCLE (Edgar 2004) for multiple sequence alignment, which yielded correlations for polymorphism with $r > 0.96$ for the 3 populations and correlations for divergence with $r > 0.82$ (Supplementary File 1, Supplementary Material online).

Masking Exons, Repeats, and Low-Quality Regions

Each alignment was visually inspected but left unmodified. Unreliable alignments were discarded (7 loci from AFR and EUR1 and 10 loci from EUR2), and regions at the ends of the alignments were trimmed because of differences in the extent of sequence available for each allele. Coding exons, UTR exons, interspersed, and low-complexity repetitive sequences were also masked using the annotations from the University of California Santa Cruz (UCSC) Genome Browser (Hinrichs et al. 2006).

Defining Evolutionary CNSs

CNSs were obtained from the phastConsElements9-way track for the *D. melanogaster* genome at the UCSC browser, which identifies highly conserved elements using a phylogenetic hidden Markov model (Siepel et al. 2005) applied to multiple alignment of 7 species of *Drosophila*, mosquito, and honeybee. The version of the genome assemblies used for the phastConsElements9way track are *D. melanogaster* (dm2), *D. simulans* (droSim1), *Drosophila yakuba* (droYak1), *Drosophila ananassae* (droAna1), *Drosophila pseudoobscura* (dp2), *Drosophila virilis* (droVir1), *Drosophila mojavensis* (droMoj1), *Anopheles gambiae* (anoGam1) and *Apis melifera* (apiMel1). Similar results were obtained when CNSs were defined as conserved blocks of >20 bp and $>90\%$ identity in pairwise alignments of *D. melanogaster* with *D. pseudoobscura* or *D. virilis*, respectively, using the VISTA browser (Couronne et al. 2003).

Identification of Point and Indel Mutations

SNPs and IPs in *D. melanogaster* were polarized using the orthologous region in *D. simulans* assuming standard parsimony criteria. Therefore, only those well-characterized SNPs and IPs that could be polarized were kept for the analyses. SNPs were discarded when: 1) more than 2 different alleles were segregating at a polymorphic site, 2) polarization was not possible because the corresponding site in the outgroup species was missing (gapped site) or unknown (N), or 3) the *D. simulans* allele did not mach either *D. melanogaster* allele. Correspondingly, IPs were discarded when: 1) they were located at the boundaries of the alignments, 2) they could not be derived because 2 or more indels were overlapped (either at the polymorphism level and/or with the outgroup species), or 3) they were spanning different categories of sites (e.g., a single indel was laying part in a CNS and part in a non-CNS region).

We considered SNFs as those nucleotide sites that were identical in all the *D. melanogaster* sequences, but different in the outgroup species. Similarly, IFs were determined as those indels that were identical in all the *D. melanogaster* sequences, but different in the outgroup species (insertion in *D. melanogaster*/deletion in *D. simulans*, or vice versa). As above, SNFs were discarded when the corresponding site in the outgroup species was a gap or an N, and IFs were discarded when: 1) they were located at the boundaries of the alignments, 2) indels in the 2 species were overlapped, or 3) they were spanning different categories of sites.

Multilocus MK Test for Noncoding DNA

For comparisons of polymorphism and divergence between non-CNSs and CNSs in order to detect selection, we applied a modification of the MK test (McDonald and Kreitman 1991) for noncoding DNA. In this test, non-CNS sites were used in place of synonymous coding sites, and CNS sites were tested for the action of natural selection in place of nonsynonymous coding sites. Tests were performed both for point mutations (SNPs and SNFs) and indels (IPs and IFs) in all 3 populations. In each case, sites were pooled across all loci, and the significances were tested according to a χ^2 test with 1 degrees of freedom (df). For a single fragment, the assumption that CNS and non-CNS sites share the same genealogy with little or no recombination is as valid as similar modifications to the MK test used for *cis*-regulatory sequences (Jenkins et al. 1995; Ludwig and Kreitman 1995).

DAF Test

DAF analyses were performed both for point mutations and indels. Frequency distributions were created for sets of SNPs and IPs based on the data set and whether they were within or outside of CNSs. Significance was assessed using the Kolmogorov–Smirnov tests (Sokal and Rohlf 1995) to test for differences across the entire allele frequency distribution. Similar results were obtained using a χ^2 test comparing the DAF distribution for SNPs within and outside CNSs, using 10% as a frequency cutoff to separate rare from common SNPs as in Drake et al. (2006).

Estimating the Effects of Natural Selection

We estimated the selection coefficients operating on CNSs using 2 independent methods (Piganeau and Eyre-Walker 2003; Kryukov et al. 2005) and the proportion of constrained sites and sites undergoing positive selection in CNSs and non-CNSs using the methods of Halligan and Keightley (2006) and Smith and Eyre-Walker (2002), respectively. For the method of Kryukov et al. (2005), possible distributions of selection coefficients (s) were modeled by a 5- and 10-column histogram containing bins representing a fraction of sites under a given selection coefficient and where all bins together represent all sites within CNSs. For the 5-bin histograms, columns corresponded to s equal to -10^{-7}, -10^{-6}, -10^{-5}, -10^{-4}, and -10^{-3}, and for the 10-bin histograms, columns corresponded to s equal to $-10^{-7.5}$, -10^{-7}, $-10^{-6.5}$, -10^{-6}, $-10^{-5.5}$, -10^{-5}, $-10^{-4.5}$, -10^{-4}, $-10^{-3.5}$, and -10^{-3}. We applied the same theoretical measures as Kryukov et al. (2005), with the exception that we used 10% as a frequency cutoff to separate rare from common SNPs for the theoretical value of the fraction of rare alleles. Furthermore, we did not use any downweighting coefficient for this value when calculating the measure of dissimilarity of the theoretical values to the observed ones because we used high-quality resequencing data.

Results

Data Sets and Definition of CNSs

We compiled 3 population genetic data sets of noncoding regions on the X chromosome in *D. melanogaster* (Glinka et al. 2003; Orengo and Aguade 2004; Ometto, Glinka et al. 2005) using a modified version of the PDA pipeline (Casillas and Barbadilla 2006). AFR is an African sample from Lake Kariba, Zimbabwe; EUR1 is a European sample from Leiden, The Netherlands; and EUR2 is another European sample from Sant Sadurní d'Anoia, Catalonia, Spain. Each data set consists of ~12 independently sampled alleles from ~100–250 noncoding regions each of length ~500 bp and includes both intronic and intergenic regions. Our automated pipeline employing the MUSCLE multiple alignment tool (Edgar 2004) can obtain almost the same results on a locus-by-locus basis as previously reported estimates of nucleotide diversity and divergence based on manually curated alignments, indicating that our alignments are of high quality (Supplementary File 1, Supplementary Material online). Alignments of the

Table 1
Summary of Data Sets and Conserved Noncoding Sequences

	AFR			EUR1			EUR2		
	IT	IG	ALL	IT	IG	ALL	IT	IG	ALL
No. loci total	160	89	249	166	90	256	26	75	101
No. loci non-CNS	160	89	249	166	90	256	26	75	101
No. loci CNS	143	84	227	152	85	237	26	75	101
Average no. chromosome total	11.79	11.66	11.74	11.92	11.76	11.86	12.58	12.72	12.68
Average no. chromosome non-CNS	11.78	11.64	11.73	11.91	11.77	11.86	12.58	12.72	12.68
Average no. chromosome CNS	11.80	11.68	11.76	11.92	11.75	11.86	12.58	12.72	12.68
No. aligned sites[a] total	75,652	48,438	124,090	82,463	51,663	134,126	21,481	62,019	83,500
No. aligned sites[a] non-CNS	60,039	32,208	92,247	65,257	34,337	99,594	14,779	39,115	53,894
No. aligned sites[a] CNS	15,613	16,230	31,843	17,206	17,326	34,532	6702	22,904	29,606
No. ungapped sites[b] total	67,139	42,695	109,834	74,514	46,082	120,596	19,557	56,736	76,293
No. ungapped sites[b] non-CNS	52,068	27,108	79,176	57,858	29,449	87,307	13,065	34,508	47,573
No. ungapped sites[b] CNS	15,071	15,587	30,658	16,656	16,633	33,289	6492	22,228	28,720
Average GC content[c] total	38.33	42.15	39.80	38.24	42.25	39.76	40.69	42.66	42.16
Average GC content[c] non-CNS	37.57	41.55	38.92	37.48	41.60	38.86	39.48	41.95	41.28
Average GC content[c] CNS	40.98	43.18	42.09	40.87	43.40	42.13	43.09	43.75	43.60
Average no. of CNS/locus	2.73	3.40	2.96	2.89	3.61	3.14	4.65	5.09	4.98
Average CNS length (bp)	34.2	53.6	42.0	34.4	53.3	41.9	55.4	60.0	58.9
Average % CNS[d]	20.6	33.5	25.7	20.9	33.5	25.7	31.2	36.9	35.5
Average % CNS, ungapped[e]	22.4	36.5	27.9	22.4	36.1	27.6	33.2	39.2	37.6

NOTE.—Values for all 3 populations are given for introns (IT), intergenic (IG), and all noncoding (ALL) regions.

[a] Total number of aligned sites (gapped + ungapped).

[b] Total number of ungapped sites (ungapped in polymorphism and divergence).

[c] Averages weighted by the number of chromosomes sampled and the number of analyzed sites (ungapped in polymorphism and divergence).

[d] Calculated using the total number of aligned sites (gapped + ungapped).

[e] Calculated using the total number of ungapped sites (ungapped in polymorphism and divergence).

noncoding regions analyzed in this study can be found at http://www.bioinf.manchester.ac.uk/bergman.

After filtering exonic, low-complexity, and poor quality regions at the ends of alignments, we analyzed a total of 249 loci for the AFR sample (160 intronic and 89 intergenic, spanning a total of >124 Kb), 256 loci for the EUR1 sample (166 intronic and 90 intergenic, spanning a total of >134 Kb), and 101 loci for the EUR2 sample (26 intronic and 75 intergenic, spanning a total of 83.5 Kb) (Table 1). Using an aligned reference sequence from the *D. melanogaster* Release 4 genome assembly (http://www.fruitfly.org/annot/release4.html), we partitioned noncoding DNA into CNS and non-CNS regions using the UCSC dm2 phastConsElements9way track (Hinrichs et al. 2006), which identifies the most conserved regions among 9 insect species (Siepel et al. 2005). On average, ~20–35% of each noncoding alignment is found in conserved ~3–5 CNSs per locus, each of ~30–60 bp in length. We note that highly conserved blocks detected by phastCons permit indels to be included within them, and thus, CNSs defined by this method are longer than those estimated previously from ungapped blocks (Bergman and Kreitman 2001).

Intergenic regions have a higher proportion of phastCons CNSs relative to intronic regions, resulting from both an increased number and length of CNSs relative to intronic regions across all populations (Table 1). The AFR and EUR1 samples are enriched for intronic loci, whereas the EUR2 sample is enriched for intergenic loci, and therefore, EUR2 contains a higher proportion of CNS sequences than AFR or EUR1. Consistent with previous analyses of these data that treat bulk noncoding sequences as a single class of sites (Glinka et al. 2003; Ometto, Glinka et al. 2005), we find that both CNS and non-CNS sites in the AFR data set are more polymorphic than in EUR1 and EUR2, whereas the EUR1 and EUR2 data sets are more diverged from *D. simulans* than loci in the AFR (Table 2). For both intergenic and intronic regions, we found that CNS regions are slightly more GC rich than non-CNS regions (Table 1). We also confirmed that intergenic regions are more GC rich overall than intronic regions overall (Ometto et al. 2006), potentially because intergenic regions contain a higher proportion of GC-rich CNSs relative to introns. Elevated GC content in slowly evolving CNSs may explain the negative correlation between GC content and rate of evolution previously observed in *Drosophila* intronic regions (Haddrill et al. 2005).

Sequence Variation in CNS and Non-CNS Regions

We computed standard measures of nucleotide variation within *D. melanogaster* as well as divergence between *D. melanogaster* and *D. simulans* as shown in table 2. Table 2 also summarizes the numbers and density of SNPs, SNFs, IPs, and IFs in each population by genomic compartment (intronic vs. intergenic) and category of sites (CNS vs. non-CNS). Detailed information for each locus is provided in Supplementary File 2 (Supplementary Material online). Preliminary analysis revealed that the density of SNPs is ~2-fold higher in alignment columns that are deleted in the outgroup species in all 3 data sets (χ^2 test, $P < 2 \times 10^{-8}$; Supplementary File 4 (Supplementary Material online), Control Test A, see also [Ometto et al. 2006]). Thus, for the purposes of this study, we excluded all SNPs in IFs to ensure that we are studying the orthologous set of

Table 2
Summary of Polymorphism and Divergence in CNS and Non-CNS Regions

	AFR			EUR1			EUR2		
	IT	IG	ALL	IT	IG	ALL	IT	IG	ALL
π[a] Total	0.012	0.011	0.011	0.005	0.005	0.005	0.005	0.004	0.004
π[a] non-CNS	0.014	0.015	0.014	0.006	0.007	0.006	0.007	0.005	0.006
π[a] CNS	0.004	0.003	0.004	0.002	0.002	0.002	0.002	0.002	0.002
K[a] Total	0.053	0.045	0.050	0.061	0.054	0.058	0.057	0.049	0.051
K[a] non-CNS	0.064	0.065	0.064	0.074	0.078	0.075	0.077	0.072	0.074
K[a] CNS	0.015	0.010	0.012	0.017	0.011	0.014	0.018	0.012	0.013
Tajima's *D*[a] Total	−0.525	−0.694	−0.590	−0.061	0.019	−0.031	0.123	−0.117	−0.056
Tajima's *D*[a] non-CNS	−0.527	−0.572	−0.542	−0.073	0.157	0.004	0.108	0.003	0.032
Tajima's *D*[a] CNS	−0.517	−0.905	−0.714	−0.019	−0.225	−0.121	0.154	−0.304	−0.200
No. SNP total	2337	1434	3771	980	606	1586	271	595	866
No. SNP non-CNS	2121	1213	3334	871	516	1387	230	470	700
No. SNP CNS	216	221	437	109	90	199	41	125	166
SNP density[b] total	0.035	0.034	0.034	0.013	0.013	0.013	0.014	0.01	0.011
SNP density[b] non-CNS	0.041	0.045	0.042	0.015	0.018	0.016	0.018	0.014	0.015
SNP density[b] CNS	0.014	0.014	0.014	0.007	0.005	0.006	0.006	0.006	0.006
No. SNF total	3380	1848	5228	4312	2361	6673	1061	2629	3690
No. SNF non-CNS	3166	1688	4854	4042	2175	6217	948	2366	3314
No. SNF CNS	214	160	374	270	186	456	113	263	376
SNF density[b] total	0.050	0.043	0.048	0.058	0.051	0.055	0.054	0.046	0.048
SNF density[b] non-CNS	0.061	0.062	0.061	0.07	0.074	0.071	0.073	0.069	0.07
SNF density[b] CNS	0.014	0.010	0.012	0.016	0.011	0.014	0.017	0.012	0.013
No. IP total	287	159	446	157	78	235	36	94	130
No. IP non-CNS	253	127	380	136	64	200	29	78	107
No. IP CNS	34	32	66	21	14	35	7	16	23
IP density[c] total	0.004	0.003	0.004	0.002	0.002	0.002	0.002	0.002	0.002
IP density[c] non-CNS	0.004	0.004	0.004	0.002	0.002	0.002	0.002	0.002	0.002
IP density[c] CNS	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001
No. IF total	690	318	1008	849	412	1261	210	501	711
No. IF non-CNS	626	275	901	782	358	1140	174	437	611
No. IF CNS	64	43	107	67	54	121	36	64	100
IF density[c] total	0.009	0.007	0.008	0.01	0.008	0.009	0.01	0.008	0.009
IF density[c] non-CNS	0.010	0.009	0.010	0.012	0.01	0.011	0.012	0.011	0.011
IF density[c] CNS	0.004	0.003	0.003	0.004	0.003	0.004	0.005	0.003	0.003

NOTE.—Values for all 3 populations are given for introns (IT), intergenic (IG), and all noncoding (ALL) regions.

[a] Averages weighted by the number of chromosomes sampled and the number of analyzed sites.

[b] Calculated using the total number of ungapped sites (ungapped in polymorphism and divergence).

[c] Calculated using the total number of aligned sites (gapped + ungapped).

nucleotides when estimating polymorphism and divergence in point mutations. Accordingly, densities of point mutations (either SNPs or SNFs) were calculated as the number of point mutations per ungapped site, whereas densities of indels (either IPs or IFs) were calculated as the number of indels per aligned site (Table 2).

As expected under both the mutational cold spot and functional constraint hypotheses, we found reduced polymorphism (π) and divergence (*K*) (Nei 1987) in CNS compared with non-CNS regions in all 3 data sets in both intronic and intergenic regions (Table 2). Densities of both point and indel variation were reduced in CNS compared with non-CNS regions, and this reduction was higher for point mutations relative to indels and stronger in intergenic regions relative to introns (Table 2). These results alone cannot differentiate between the mutational cold spot and functional constraint hypotheses because both predict a reduction in polymorphism and divergence in CNSs. However, evidence for differential selective constraints operating on CNS and non-CNS regions can be found in the facts that CNSs exhibit a more negative Tajima's *D* (Tajima 1989) (Table 2) and a stronger reduction of divergence relative to polymorphism (~60% in SNPs vs. ~80% in SNFs; ~50% in IPs vs. ~70% in IFs) (Table 2). Although the magnitude of the reduction of variation in CNSs relative to non-CNS regions is dependent on our use of the phastCons blocks to define CNSs, alternative definitions of CNS provide similar results (see Materials and Methods for details). We note that the systematic reduction in variation specifically observed in CNSs cannot be explained by sequencing error in the population samples because the same sequencing strategies were applied to both CNS and non-CNS regions.

CNSs Are Selectively Constrained

To test whether the ratios of polymorphism (SNPs) to divergence (SNFs) differ significantly in CNS and non-CNS regions, we applied a modified MK test to point mutations in CNS and non-CNS regions. Separate tests were performed for all 3 data sets and for intronic and intergenic regions (Fig. 1). We present results here for all fragments

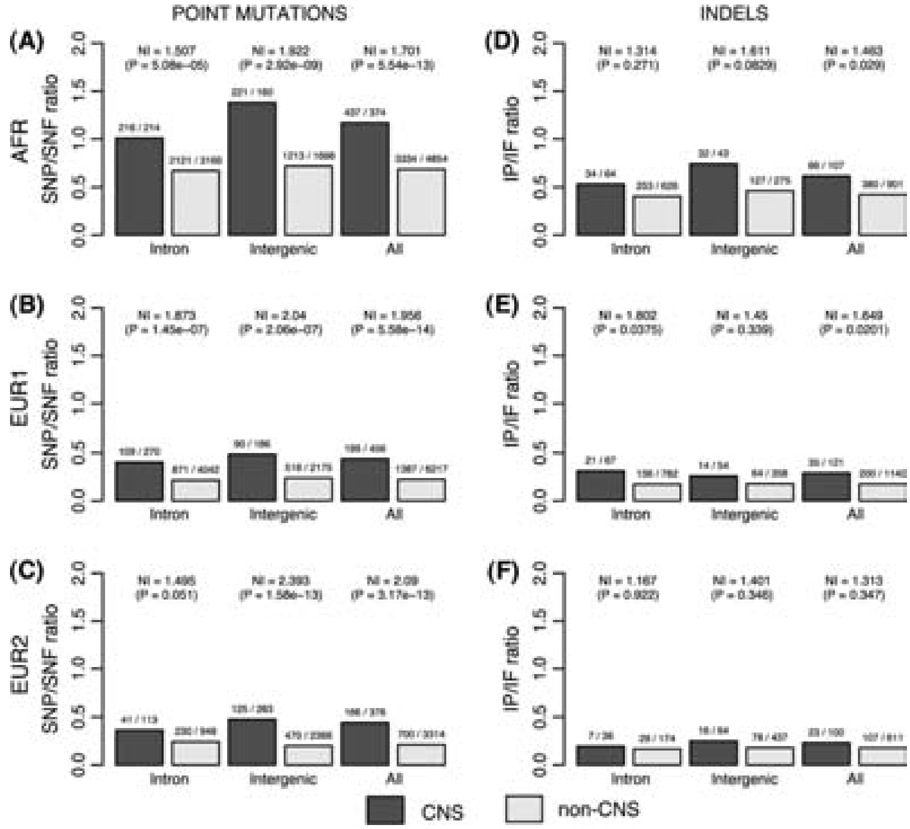


FIG. 1.—Ratios of polymorphism to divergence for CNS and non-CNS sites for both point mutations and indels in the AFR, EUR1, and EUR2 data sets. Dark gray and light gray bars represent data for CNSs and non-CNS regions, respectively. CNS/non-CNS ratios for polymorphism and divergence are summarized by the neutrality index as NI = $(\mathrm{SNP}_{\mathrm{CNS}}/\mathrm{SNP}_{\text{non-CNS}})/(\mathrm{SNF}_{\mathrm{CNS}}/\mathrm{SNF}_{\text{non-CNS}})$, and P values are computed using a χ^2 test of independence.

available on the X-chromosome, which did not differ from analyses using only noncoding sequences from regions of high recombination as defined in Andolfatto (2005) (results not shown). For point mutations (Fig. 1A–C), we observed very strong deviations from the null hypothesis (χ^2 tests; all $P < 10^{-5}$) with a significant excess of polymorphisms relative to fixed differences in CNSs in all cases but on one with the least amount of data (intronic regions in the EUR2 sample, which exhibit the same trend). One can summarize an entire MK table as a ratio of ratios termed the neutrality index (Rand and Kann 1996) which is defined as NI = $(\mathrm{SNP}_{\mathrm{CNS}}/\mathrm{SNP}_{\text{non-CNS}})/(\mathrm{SNF}_{\mathrm{CNS}}/\mathrm{SNF}_{\text{non-CNS}})$. Assuming the ratio of polymorphism to divergence in non-CNS regions is closer to that expected under neutrality than in CNS regions, the NI attempts to quantify whether the levels of divergence in CNSs are too low (NI $>$ 1) or too high (NI $<$ 1) relative to levels of polymorphism. In all data sets, we observe NI $>$ 1 overall and in both intronic and intergenic regions, although NI for intergenic regions is greater than NI for intronic regions, indicating that point mutations in intergenic CNSs are less likely to fix relative to those in intronic CNS. These results reject the mutational cold spot model to explain the mode of CNS evolution in intronic and intergenic regions and support the interpretation that a significant proportion of new point mutations in CNSs are deleterious and do not contribute to divergence between species. It is important to emphasize that under the null hypothesis, the definition of CNS and non-CNS regions does not affect the predictions of equal ratios of polymorphism and divergence in both categories of sites. However, because CNSs are defined a priori by low divergence, our test is biased against detecting excess divergence, and thus, this analysis does not eliminate the action of positive selection on CNSs.

Evidence for functional constraint acting on CNSs can also be found in differences between the DAF spectra for CNS and non-CNS regions (Keightley, Kryukov et al. 2005; Drake et al. 2006). We observed a highly significant excess of low-frequency derived alleles for SNPs within CNSs relative to non-CNS regions in all 3 populations (Fig. 2A–C). For example, in the AFR population, 63% of SNPs inside CNSs are singletons, compared with only 46% of SNPs outside CNSs (χ^2 test = 47.5, df = 1, $P < 5.45 \times 10^{-12}$). When data are partitioned into intronic and intergenic regions, all tests show a significant excess of low-frequency SNPs in CNSs relative to non-CNSs for all data sets, except for intronic regions in EUR2, which has the least amount of data (Supplementary File 3, Supplementary Material online, panels A–F). The excess of rare alleles in CNSs was more prominent for intergenic than intronic SNPs, consistent with the results of the MK test which indicate that fewer SNPs in intergenic CNSs go to fixation.

Similar results were obtained by testing for differences in the unpolarized minor allele frequency spectrum between CNS and non-CNS regions (results not shown), indicating that differences in the polarized DAF spectrum between these classes of sites are not strongly influenced by misinference of ancestral states (Hernandez et al. 2007). Overall, we find a strong signal that SNPs in CNS regions are specifically maintained at low frequency relative to SNPs in non-CNS regions, a finding which is inconsistent with the mutational cold spot hypothesis but is compatible with the presence of deleterious SNPs segregating at low frequency in functionally constrained CNSs.

One difficulty that arises from using non-CNS regions to detect purifying selection on CNSs is the fact that base composition differs in these 2 categories of sites (Table 1) (Drake et al. 2006). For example, recent changes in mutation biases could mimic the signature of selective differences between CNS and non-CNS regions and affect the tests of neutrality used here (Eyre-Walker 1997). In particular, a recent increase in the rate of G:C → A:T mutation has recently been suggested to explain nonequilibrium patterns of base composition evolution in *D. melanogaster* introns and intergenic regions (Kern and Begun 2005; Ometto et al. 2006) and may potentially cause an excess of SNPs in CNSs relative to non-CNS regions that would be restricted to low population frequency. Under this model of a recent increase in the rate of G:C → A:T mutation, differences in base composition between CNS and non-CNS regions cannot explain the reduction in polymorphism and divergence in CNSs because CNSs are more GC rich than non-CNS regions and G:C → A:T mutations occur at a higher rate than A:T → G:C mutations. Nevertheless, to control for any potential effects of biased mutation patterns that result from differences in base composition between CNS and non-CNS regions, we performed DAF tests for G:C → A:T and A:T → G:C mutations separately. We found an excess of rare alleles in CNSs when both G:C → A:T mutations and A:T → G:C mutations were considered separately in all samples except for A:T → G:C mutations in the EUR2 sample, which had the least amount of data but still shows the same trend (Supplementary File 4, Supplementary Material online, Control Test B). Thus, we can detect evidence for purifying selection on CNSs even when potential changes in base composition are controlled for, indicating that a recent increase in G:C → A:T mutation rate is unlikely to confound the conclusion that CNSs are selectively constrained. Moreover, if our interpretation that CNSs are constrained is correct, we suggest that the excess of low-frequency G:C → A:T mutations and other aspects of non-equilibrium base composition evolution in *Drosophila* may in fact be a consequence of the preservation of functional GC nucleotides in noncoding DNA by purifying selection, rather than evidence for a change in mutation rate or biased segregation as suggested previously (Kern and Begun 2005; Galtier et al. 2006; Ometto et al. 2006).

We also investigated 2 possible alternatives for the striking differences we observed between CNS and non-CNS regions that might result from alignment error. The first possibility is that indels may create low-quality regions of multiple alignments, causing SNPs and SNFs to accumulate in the vicinity of gaps and that the differences we observe in point mutations may be a byproduct of differences in indel rates between CNS and non-CNS regions. To control for this possibility, we repeated our analyses excluding CNS and non-CNS regions that contain indels (Supplementary File 4, Supplementary Material online, Control Test C). All MK tests remained highly significant when all regions with either IPs or IFs were excluded. Likewise, all DAF tests remained highly significant when regions with IPs were excluded. We also explored a related source indel-associated alignment error that might result from 2 indels of exactly the same length in essentially the same position of a single sequence, one an insertion and one a deletion. Two such indels may collapse in the alignment and thus result in a run of consecutive substitutions. Consecutive SNPs or SNFs may also occur through complex mutational events that replace more than a single nucleotide (Averof et al. 2000; Haag-Liautard et al. 2007). We repeated our analyses excluding consecutive SNPs or SNFs (all SNPs followed or preceded by another SNP in the alignment or all SNFs followed or preceded by another SNF; Supplementary File 4, Supplementary Material online, Control Test D). All tests remained highly significant, showing no evidence of this type of alignment error. These results demonstrate that indel-associated misaligment cannot explain the differences in point mutations between CNS and non-CNS regions.

Indels themselves, in contrast to point mutations, show less striking differences in the ratios of polymorphism to divergence in CNSs relative to non-CNS regions (Fig. 1D–F). In contrast to previous analysis of indels in an MK framework that contrast SNPs in silent sites with either insertions or deletions in introns (Presgraves 2006), here we directly contrast levels of polymorphism and divergence for all indels in CNSs to all indels in non-CNS regions. The NI values we observe for indels are still above 1, consistent with purifying selection on indels in CNS regions but are always lower than the corresponding values for SNPs. MK tests are only significant for the combined intronic plus intergenic regions in the 2 larger data sets (AFR and EUR1) and intronic regions in EUR1. Likewise, we observed no strong differences within species in the DAF spectra between indels in CNS and non-CNS regions, with nonsignificant DAF tests for all populations overall (Fig. 2D–F) and for intronic and intergenic regions separately (Supplementary File 3, Supplementary Material online, panels G–L). We note that potential differences in CNS and non-CNS regions are not diluted or obscured by a high rate of indels due to simple slippage because low-complexity repeat regions were filtered from the data.

Because there are 8-fold fewer IPs than SNPs and 5-fold fewer IFs than SNFs in our data set, the lack of strong differences in CNS and non-CNS regions for indels may simply result from low power to reject the null hypothesis. We tested if differences as strong as those observed for point mutations could also be observed in the smaller sample of indels by rescaling the point mutation data (which show significant results) to the sample sizes observed for indels and repeated the MK and DAF tests (Supplementary File 4, Supplementary Material online, Control Test E). Specifically, numbers of SNPs and SNFs were reduced to the numbers of observed IPs and IFs, while maintaining the observed ratio of SNPs to SNFs in contingency tables of

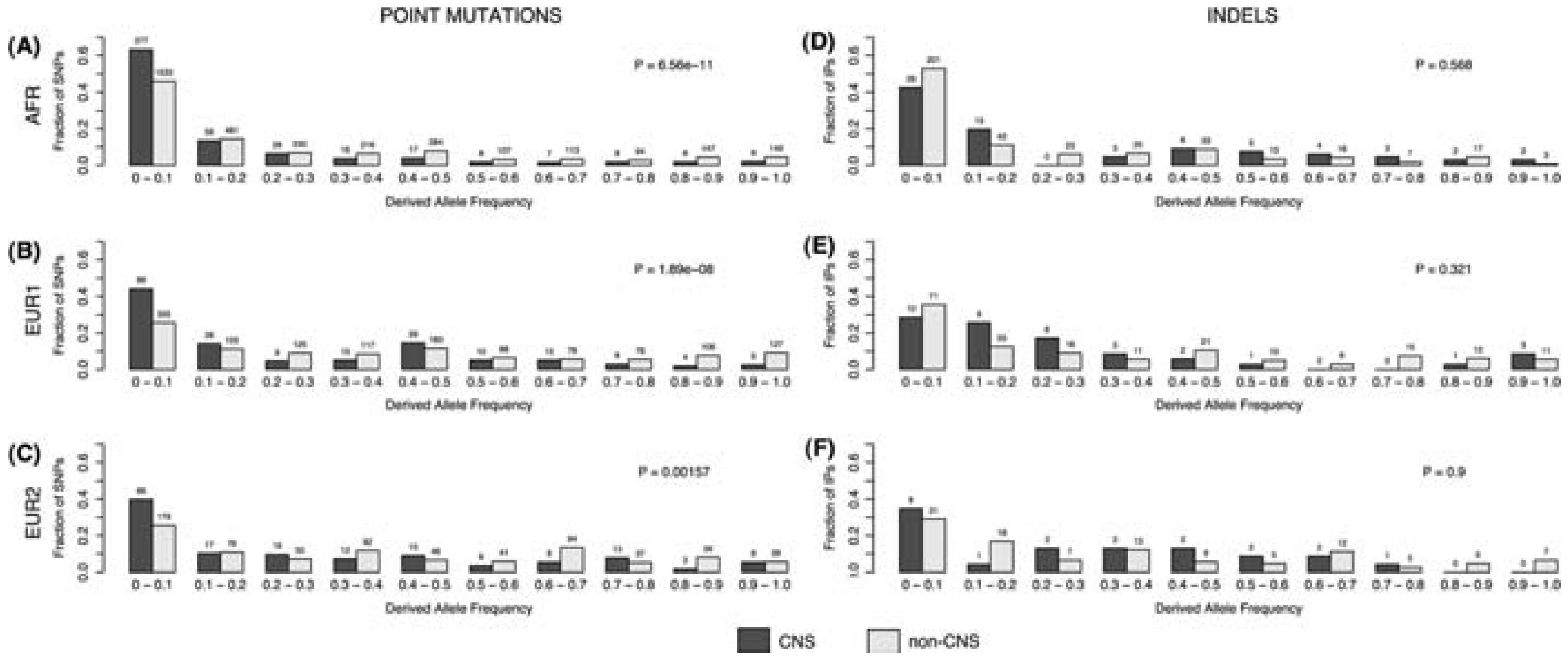


FIG. 2.—DAF distributions for polymorphic point mutations and indels in the AFR, EUR1, and EUR2 data sets. Dark gray and light gray bars represent data for all SNPs/IPs within CNSs and non-CNS regions, respectively. Reported *P* values are for 2-sample Komogorov–Smirnov tests.

the MK tests or the frequency bins of the DAF distributions. MK tests and DAF tests using rescaled point mutation data sets for combined intronic and intergenic were still highly significant in all 3 populations ($P < 0.01$). Assuming the same degree of purifying selection is acting on point mutations and indels, it is unlikely that low power is the main cause of the lack of significance for indels in MK and DAF tests between CNS and non-CNS regions. Given the fact that we find strong evidence that CNSs are constrained for point substitutions, we do not interpret these results as support for an indel cold spot hypothesis to explain the mode of CNS evolution. Rather, we interpret the lack of strong differences in indel evolution between CNS and non-CNS regions as evidence for spatial constraints acting on both CNS and non-CNS sequences (see Discussion).

Estimating the Effects of Deleterious Mutations on CNSs

Deleterious alleles that are not immediately purged by natural selection are expected to be maintained at low frequencies and not go to fixation in natural populations. The results that CNSs show a significant excess of polymorphism to divergence and an enrichment of low-frequency alleles indicate that CNSs do indeed harbor more deleterious SNPs relative to the non-CNS spacer sequences between them. Can we infer at what frequency deleterious SNPs in CNSs are segregating or the magnitude of fitness effects acting on SNPs in CNS regions? To address these questions, we restricted our analysis to the AFR data set, which is taken from a population of *D. melanogaster* that is assumed to be closest to ancestral conditions (Glinka et al. 2003; Thornton and Andolfatto 2006) and for which we have the largest sample of polymorphism to estimate properties of deleterious alleles in CNSs.

We found evidence that deleterious SNPs were restricted almost exclusively to the singleton class in the AFR data set by an analysis of the effects of removing singletons on the results of the MK test. This procedure has been used in tests for positive selection assuming that if deleterious mutations are restricted to sites with low population frequencies, derived alleles present only once in the sample should preferentially be enriched for deleterious mutations (Fay et al. 2002; Andolfatto 2005). In all cases, NIs remained greater than one after the removal of singleton SNPs but decreased relative to values for the total data set (Supplementary File 4, Supplementary Material online, Control Test F). For introns, intergenic regions and the combined data set, the resulting MK tests yielded nonsignificant differences in the ratio of nonsingleton polymorphism to divergence ratio, suggesting that common SNPs in CNSs in the AFR data set are effectively neutral. This result indicates that virtually all signals of deleterious alleles segregating in CNSs are restricted to rare SNPs. In addition, this result also indicates that there is no evidence for positive selection acting on CNSs, even when the confounding effects of purifying selection are taken into consideration (see Discussion). We note that similar tests on the EUR1 and EUR2 populations yielded NI > 1 which remained significant in the absence of singletons, suggesting that purifying selection may be acting even among common SNPs in these data sets (Supplementary File 4, Supplementary Material online, Control Test F). Further evidence that SNPs are restricted to low frequency can be found in the ratio of SNPs in CNSs relative to non-CNSs at different derived allele frequencies. We find that in the AFR population, the $\mathrm{SNP_{CNS}/SNP_{non\text{-}CNS}}$ ratio is significantly heterogeneous across all 10 DAF classes (χ^2 test; $P < 4.17 \times 10^{-9}$), consistent with categorical differences in the complete DAF spectra shown above. However, when low-frequency SNPs are removed, the remaining 9 DAF classes show no significant heterogeneity in their $\mathrm{SNP_{CNS}/SNP_{non\text{-}CNS}}$ ratio ($P > 0.07$). This result indicates that common SNPs in CNS and non-CNS regions have similar DAF spectra and that the majority of deleterious SNPs in CNSs are restricted to a DAF of less than 10% in the population samples.

To quantify the observed differences in selective pressure acting on CNSs relative to non-CNS regions, we estimated the distribution of selection coefficients in these regions using an exhaustive computational search method developed by Kryukov et al. (2005) (see Materials and Methods). Unlike other methods to estimate selection coefficients from population genetic data, no explicit distribution of selection coefficients is assumed by this method. Rather, distributions of selection coefficients (s) are modeled by histograms, where bins represent the fraction of sites under a given selection coefficient. All possible distributions are enumerated under a model of weakly deleterious evolution, and the fit of the data is evaluated for each possible distribution. Assuming an effective population size (N_e) for *D. melanogaster* of 10^6 (Kreitman 1983), our results indicate that best fit of the data is to a distribution where 80–85% of sites in CNSs are subject to weak purifying selection ($s \sim 10^{-5}$) and the remaining 15–20% of CNS sites are effectively neutral ($s \sim 10^{-7}$). Likewise, using the method of Piganeau and Eyre-Walker (2003), which assumes an underlying gamma distribution of selection coefficients, we obtain an average strength of selection on CNSs of $N_e s = 30.7$ (95% confidence interval: 13–117) with a shape parameter of $\beta = 0.31$ (95% confidence interval: 0.22–0.42). These results indicate that purifying selection on *Drosophila* CNSs exceeds the boundaries of nearly neutral evolution. However, the strength of purifying selection for bulk noncoding DNA in *Drosophila* may be on the boundaries of nearly neutral evolution because non-CNS regions are more abundant but less constrained than CNS regions. Thus, the evolution of *Drosophila* noncoding DNA in general may be sensitive to changes in N_e, both across time through changes in census population size (Keightley, Kryukov et al. 2005; Keightley, Lercher, and Eyre-Walker 2005) or across the genome such as in regions of reduced recombination (Haddrill et al. 2007). Our results also indicate that purifying selection is stronger (Chen et al. 2007) and affects more sites (Kryukov et al. 2005) in *Drosophila* CNSs than for mammalian CNSs.

Discussion

The major conclusion of this work is that highly CNSs in *Drosophila* are maintained by purifying selection and are not simply regions of the genome with extremely low mutation rate as predicted by the mutation cold spot hypothesis. In addition, we find that the strength of purifying selection acting to maintain CNSs is moderately strong, with most nucleotides in CNSs being preserved by selection coefficients 10- to 100-fold greater than the reciprocal of the effective population size. The conclusion that *Drosophila* CNSs are maintained by purifying selection supports previous analyses that have made this assumption based on reduced rates of molecular evolution (Bergman and Kreitman 2001; Siepel et al. 2005; Halligan and Keightley 2006). Specifically, our results support the UCSC phastCons highly conserved track (Siepel et al. 2005) as being able to identify selectively constrained regions of the *D. melanogaster* genome. These findings in *Drosophila* closely parallel those recently found for mammalian CNSs and predicted micro-RNA binding sites using population genetic data from human SNP studies (Keightley, Kryukov et al. 2005; Kryukov et al. 2005; Chen and Rajewsky 2006; Drake et al. 2006; Chen et al. 2007). Thus, purifying selection may be a general force acting to maintain highly CNSs in metazoan genomes. As no population genomic evidence (or molecular mechanism) has yet been put forth to support the mutation cold spot hypothesis, similar results in disparate organisms such as flies and mammals (together with the nonrandom spacing of CNSs in flies and worms [Bergman et al. 2002; Webb et al. 2002]) argue against the general likelihood that CNSs will be shown to be mutational cold spots in any organism. Further studies in disparate taxa including plants and other metazoans will be necessary to confirm the generality of this conclusion.

Selective constraint on CNSs is consistent with the large body of evidence from experimental studies that highly CNSs in *Drosophila* are often associated with regulatory function, such as *cis*-regulatory elements or noncoding RNAs (Bergman et al. 2002; Enright et al. 2003; Lai et al. 2003; Costas et al. 2004). Furthermore, several facts reported over the last decade collectively point to widespread selective constraint operating on *Drosophila* noncoding DNA. First, unconstrained noncoding DNA is quickly deleted from the *Drosophila* genome (Petrov et al. 1996; Petrov and Hartl 1998). This process is predicted to purge the fly genome of "junk" DNA, making nonfunctional sequence-like pseudogenes rare (Petrov et al. 1996; Harrison et al. 2003) and enriching noncoding DNA that remains in the genome for functional elements. Second, genes with complex transcriptional regulation have longer flanking intergenic regions (Nelson et al. 2004), suggesting that the mere presence of noncoding DNA in *Drosophila* may imply function. Third, for both intronic and intergenic DNA, the rate of molecular evolution between closely related *Drosophila* species decreases with increasing noncoding sequence length (Haddrill et al. 2005; Halligan and Keightley 2006), consistent with the interpretation that long noncoding regions may have increased functional constraints. Fifth, long introns have a higher proportion of CNS sequences and genes with CNSs in their introns have more complex regulation (Petit et al. 2007). Finally, adaptive substitutions may be commonplace in both intronic and intergenic regions (Andolfatto 2005), which can only occur if the density of functional nucleotide sites in noncoding DNA is high.

Given that constraints on noncoding sequences are widespread in *Drosophila*, and the possibility that adaptive substitution occurs in noncoding DNA, it is worth considering whether flanking non-CNSs are appropriate control sequences to detect selection on CNSs. Halligan and Keightley (2006) report a method to measure constraints on regions of genomic DNA as the reduction in the rate of substitution relative to that expected based on putatively unconstrained sequences, such as 4-fold degenerate silent sites. Despite widespread evidence for weak selection on silent sites in *Drosophila* (Shields et al. 1988; Akashi 1995), we applied this method to evaluate if stronger primary sequence constraints act on non-CNS regions relative to 4-fold silent sites, and, if so, what effect this may have on our conclusion that CNSs are selectively constrained and not mutational cold spots. As shown in table 3, we find that non-CNS regions exhibit an ~20% reduced rate of sequence evolution relative to 4-fold silent sites, indicating that primary sequence constraints act on non-CNS regions. Levels of constraint on CNSs are estimated to be ~85% by the Halligan and Keightley (2006) method (consistent with results above using the Kryukov et al. (2005) method), and thus, we infer that selective constraints operating on non-CNS regions affect ~4 times fewer sites than in CNS regions. Constraints on non-CNS regions are perhaps not surprising because even the most rigorous definition of CNSs (Siepel et al. 2005) is unlikely to capture all functionally constrained noncoding DNA, especially those which arise through lineage-specific gain-of-function events. Nevertheless, as constraints on non-CNS regions would only tend to obscure differences between CNS and non-CNS categories by making their patterns of evolution more similar, our conclusion that CNSs are selective constrained is conservative with respect to the null hypothesis that they are mutational cold spots. However, by using non-CNS regions as putatively unconstrained control sequences, the proportion of sites under constraint in CNSs and the magnitude of their selective effects are likely to be underestimated in our analysis.

Conversely, if adaptive substitutions preferentially occur in non-CNS regions, it may be possible that the rate of substitution in non-CNS regions is elevated relative to the unconstrained neutral substitution rate, which could cause us spuriously to reject the mutational cold spot hypothesis. To evaluate if any signature of adaptive substitution is detected in our data set, we conducted MK tests on CNS and non-CNS regions as selected classes of sites using 4-fold silent sites as putatively unconstrained controls. For these analyses, we reprocessed the sequence data reported in Andolfatto (2005) using PDA to extract SNPs and SNFs for 4-fold degenerate silent sites only. As is observed using sites from linked non-CNS regions above (Fig. 1), the NI for CNSs remains significantly greater than one when using partially linked 4-fold silent sites as controls (Table 3). As before, when we removed singletons to reduce the confounding effects of deleterious mutations present in low-frequency alleles, we found no departure from neutral expectations between CNS and 4-fold silent sites (Table 3). Additionally, SNPs in CNSs are more skewed to lower frequencies than the SNPs in 4-fold silent sites (Supplementary File 5, Supplementary Material online) as has been

Table 3
Summary of Constraint and Adaptation (α) on CNS and Non-CNS Regions Relative to 4-fold Degenerate Silent Sites in AFR Data Set

			All polymorphisms		Excluding singletons	
	Constraint	α	NI	*P*	NI	*P*
CNS vs. 4-fold silent	0.84525	−0.40214	1.402	0.00216	0.782	0.07567
Non-CNS vs. 4-fold silent	0.18360	0.17577	0.824	0.02623	0.684	0.00012
CNS + non-CNS vs. 4-fold silent	0.36616	0.13443	0.866	0.09808	0.691	0.00017

NOTE.—The reported α is based on all polymorphisms including singletons. CNS/non-CNS ratios for polymorphism and divergence are summarized by the neutrality index as NI = $(SNP_{CNS}/SNP_{non\text{-}CNS})/(SNF_{CNS}/SNF_{non\text{-}CNS})$, and *P* values are computed using a χ^2 test of independence. Tests are performed both including and excluding singletons.

shown recently for bulk noncoding DNA in *D. melanogaster* (Andolfatto 2005; Mustonen and Lassig 2007) and *Drosophila miranda* (Bachtrog and Andolfatto 2006). These results confirm our main claims that CNSs are selectively constrained with deleterious SNPs restricted to low frequencies and clearly demonstrate that the conclusion that CNSs are functionally constrained does not depend on our use of linked non-CNS regions as controls.

Intriguingly, we find evidence for positive selection in non-CNS regions when 4-fold silent sites are used as unconstrained controls, both when all sites are used or when singletons are removed (Table 3). The same trends are observed when CNS and non-CNS regions are pooled into bulk noncoding DNA, as in Andolfatto (2005). Thus, we confirm the results of Andolfatto (2005) for the putative signature of adaptive substitution on noncoding DNA when 4-fold silent sites are used as controls, even despite differences in methods of estimating polymorphism and divergence. Somewhat counterintuitively, perhaps, the signature of adaptive substitution in *Drosophila* noncoding DNA does not appear to occur in CNSs, which might be expected to contain the functional elements that are targets for positive selection. Conversely, the signal of excess substitution in bulk noncoding DNA relative to silent sites appears to be restricted to the non-CNS regions that are divergent among other insect species. Using the method of Smith and Eyre-Walker (2002), we estimate that ~18% of substitutions in non-CNS regions are driven to fixation by positive selection relative to neutral expectations (Table 3), which is compatible with previous estimates for bulk noncoding DNA (Andolfatto 2005). Selective constraints on CNSs coupled with the signature of adaptive substitution in non-CNS regions might be expected under the model of stabilizing selection proposed by Ludwig et al. (2000) for the *even-skipped* stripe 2 enhancer, whereby loss of ancestral transcription factor binding sites in CNS regions may lead to compensatory adaptive fixations of lineage-specific binding sites in non-CNS regions (e.g., bcd-3) that restore *cis*-regulatory function.

In summary, using 4-fold silent sites as another class of putatively unconstrained neutrally evolving sequences, we find evidence that both constraint and adaptation influence rates of substitution in non-CNS regions. Assuming additive influences of these 2 opposing forces (Andolfatto 2005; Halligan and Keightley 2006), we find that the estimated proportion of mutations purged by purifying selection in non-CNS regions (C ~18%) is approximately the same as the estimated proportion of substitutions that have been driven to fixation by positive selection (α ~17%). Assuming no adaptive evolution in CNS regions, these results also imply that the proportion of functionally relevant nucleotides in non-CNS regions is FRN = C + (1 − C)α ≈ 33%, or ~2.5-fold less than the proportion of functional sites in CNS regions (85%). Thus, given that the majority of *Drosophila* noncoding DNA is found in non-CNS regions (65–80%, Table 1), the number of functional sites in CNS and non-CNS regions is approximately equivalent. Although less densely packed than in CNS regions, the greater number of functional sites in non-CNS regions coupled with their relaxed selective constraints may explain why these regions appear to be the most likely targets for positive selection in noncoding DNA. Further work will be necessary to determine if (and how) the distribution of positive and negative selection coefficients acting on polymorphisms in non-CNS regions affects their utility in testing the mutational cold spot hypothesis. Likewise, the influence of alternative CNS definitions on quantitative estimates of constraint and adaptation in CNS and non-CNS regions needs to be investigated further. Nevertheless, the direct result for a significant constraint and skew toward rare alleles in CNSs using 4-fold silent sites as controls (see above) indicates that our main claim that CNSs are selectively constrained and not mutational cold spots is unaffected by the potentially confounding effects of either constraint or adaptive substitution in non-CNS regions.

Selective constraints may also operate on the length of noncoding DNA as well as on primary sequence. This possibility may explain why differences in the ratio of polymorphism to divergence or DAF spectrum between CNS and non-CNS regions are not as strong for indels as they are for point mutations, assuming that spatial constraints act on both CNS and non-CNS regions. Mechanistically, this might be expected to occur if CNSs represent the constraints imposed by transcription factor binding sites that could be disrupted by both point and indel mutations, whereas non-CNS regions that act to position neighboring binding sites would be affected only by indel mutations (Ondek et al. 1988). Spatial constraints have been argued previously in *Drosophila* noncoding DNA based on the non-random distribution of CNSs and the strong correlation in the length between neighboring CNSs across divergent species (Bergman et al. 2002). Ometto, Stephan, and De Lorenzo et al. (2005) have also argued for spatial constraints acting within *Drosophila* noncoding DNA based on the ratio of insertions to deletions and the size distribution of deletions segregating in natural populations. More recently, Lunter et al. (2006) have inferred that spatial constraints act on human noncoding DNA based on the distribution of indel positions in alignments with other mammalian species, and Sun et al. (2006) have argued for spatial constraints between neighboring vertebrate ultraconserved regions. A lack of strong differences in indel evolution between CNS and non-CNS regions may alternatively arise because of technical reasons such as the fact that phastCons permits indels in CNSs, blurring real differences in the pattern of indel evolution between these categories. Another possible explanation is that CNSs are mutation cold spots for indels, although this seems unlikely if CNSs are selectively constrained for point mutations as we argue here. If selective constraints are indeed operating on indels in both CNS and non-CNS regions, the lack of strong differences in the ratio of polymorphism to divergence or DAF spectrum suggests that the strength of selection against indels in noncoding DNA may be stronger than for point mutations because only relatively weak purifying selection allows an accumulation of low-frequency variants in nature (as is observed for SNPs). Future progress in detecting evidence for spatial constraints and quantifying the mode and strength of selection acting on both indels and point mutations in noncoding DNA will shed light on the functions encoded in this most abundant, yet least explored, territory of metazoan genomes.

Supplementary Material

Supplementary Files 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Douda Bensasson, Francesco Catania, Manolis Dermitzakis, and Dan Halligan for valuable discussions and Peter Andolfatto, Francesco Catania, Dan Halligan, Martin Kreitman, Gregory Kryukov, Alfredo Ruiz, Josh Shapiro, Kevin Thornton, and 3 anonymous reviewers for comments on the manuscript. We also thank Gregory Kryukov for guidance in implementing his method to estimate selection coefficients, Adam Eyre-Walker for providing programs and assistance to estimate the strength of selection, and Dan Halligan for providing programs and data sets to estimate constraint using 4-fold silent sites. We are grateful to the Washington University Genome Sequencing Center, the Broad Institute, Agencourt Biosciences, and the laboratories of Monserrat Aguade, Peter Andolfatto, David DeLorenzo, and Wolfgang Stephan for producing the sequence data used in this work. S.C. was supported by a Marie Curie fellowship from the European Commission (CHPMT-GH-01-00285-13), by the Ministerio de Ciencia y Tecnología (BES-2003-0416), and grant BFU2006-08640/BMC from the Ministerio de Educación y Ciencia, Spain.

Literature Cited

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at “silent” sites in *Drosophila* DNA. Genetics. 139:1067–1076.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature. 437:1149–1152.

Averof M, Rokas A, Wolfe KH, Sharp PM. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science. 287:1283–1286.

Bachtrog D, Andolfatto P. 2006. Selection, recombination and demographic history in *Drosophila miranda*. Genetics. 174:2045–2059.

Bergman CM, Kreitman M. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res. 11:1335–1345.

Bergman CM, Pfeiffer BD, Rincon-Limas DE, et al. (17 co-authors). 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. Genome Biol. 3:RESEARCH0086.

Bergman CM, Quesneville H, Anxolabehere D, Ashburner M. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. Genome Biol. 7:R112.

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. Science. 165:349–357.

Casillas S, Barbadilla A. 2006. PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. Nucleic Acids Res. 34:W632–W634.

Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in the human genome. Am J Hum Genet. 80:692–704.

Chen K, Rajewsky N. 2006. Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet. 38:1452–1456.

Clark AG. 2001. The search for meaning in noncoding DNA. Genome Res. 11:1319–1320.

Costas J, Pereira PS, Vieira CP, Pinho S, Vieira J, Casares F. 2004. Dynamics and function of intron sequences of the wingless gene during the evolution of the *Drosophila* genus. Evol Dev. 6:325–335.

Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin E, Pachter L, Dubchak I. 2003. Strategies and tools for whole-genome alignments. Genome Res. 13:73–80.

Drake JA, Bird C, Nemesh J, et al. (11 co-authors). 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet. 38:223–227.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. 2003. MicroRNA targets in *Drosophila*. Genome Biol. 5:R1.

Eyre-Walker A. 1997. Differentiating between selection and mutation bias. Genetics. 147:1983–1987.

Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature. 415:1024–1026.

Galtier N, Bazin E, Bierne N. 2006. GC-biased segregation of noncoding polymorphisms in *Drosophila*. Genetics. 172:221–228.

Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics. 165:1269–1278.

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Charlesworth B, Keightley PD. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. Nature. 445:82–85.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. Genome Biol. 6:R67.

Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. Genome Biol. 8:R18.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res. 16:875–884.

Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. 2003. Identification of pseudogenes in the *Drosophila melanogaster* genome. Nucleic Acids Res. 31:1033–1037.

Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol. 24:1792–1800.

Hinrichs AS, Karolchik D, Baertsch R, et al. (27 co-authors). 2006. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 34:D590–D598.

Holt RA, Subramanian GM, Halpern A, et al. (123 co-authors). 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science. 298:129–149.

Jenkins DL, Ortori CA, Brookfield JF. 1995. A test for adaptive change in DNA sequences controlling transcription. Proc R Soc Lond B Biol Sci. 261:203–207.

Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. 2005. Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res. 15:1373–1378.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol. 3:e42.

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. Genome Res. 12:656–664.

Kern AD, Begun DJ. 2005. Patterns of polymorphism and divergence from noncoding sequences of *Drosophila melanogaster* and *D. simulans*: evidence for nonequilibrium processes. Mol Biol Evol. 22:51–62.

Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature. 304:412–417.

Kryukov GV, Schmidt S, Sunyaev S. 2005. Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet. 14:2221–2229.

Lai EC, Tomancak P, Williams RW, Rubin GM. 2003. Computational identification of *Drosophila* microRNA genes. Genome Biol. 4:R42.

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature. 403:564–567.

Ludwig MZ, Kreitman M. 1995. Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*. Mol Biol Evol. 12:1002–1011.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol. 2:e5.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature. 351:652–654.

Misra S, Crosby MA, Mungall CJ, et al. (30 co-authors). 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol. 3:RESEARCH0083.

Mustonen V, Lassig M. 2007. Adaptations to fluctuating selection in *Drosophila*. Proc Natl Acad Sci USA. 104:2277–2282.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nelson CE, Hersh BM, Carroll SB. 2004. The regulatory content of intergenic DNA shapes genome architecture. Genome Biol. 5:R25.

Ometto L, De Lorenzo D, Stephan W. 2006. Contrasting patterns of sequence divergence and base composition between *Drosophila* introns and intergenic regions. Biol Lett. 2:604–607.

Ometto L, Glinka S, De Lorenzo D, Stephan W. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol. 22:2119–2130.

Ometto L, Stephan W, De Lorenzo D. 2005. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. Genetics. 169:1521–1527.

Ondek B, Gloss L, Herr W. 1988. The SV40 enhancer contains two distinct levels of organization. Nature. 333:40–45.

Orengo DJ, Aguade M. 2004. Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. Genetics. 167:1759–1766.

Petit N, Casillas S, Ruiz A, Barbadilla A. 2007. Protein polymorphism is negatively correlated with conservation of intronic sequences and complexity of expression patterns in *Drosophila melanogaster*. J Mol Evol. 64:511–518.

Petrov DA, Hartl DL. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol Biol Evol. 15:293–302.

Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature. 384:346–349.

Piganeau G, Eyre-Walker A. 2003. Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. Proc Natl Acad Sci USA. 100:10335–10340.

Presgraves DC. 2006. Intron length evolution in *Drosophila*. Mol Biol Evol. 23:2203–2213.

Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D. 2005. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 1:e22.

Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol Biol Evol. 13:735–748.

Richards S, Liu Y, Bettencourt BR, et al. (52 co-authors). 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and *cis*-element evolution. Genome Res. 15:1–18.

Shabalina SA, Kondrashov AS. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genet Res. 74:23–30.

Shields DC, Sharp PM, Higgins DG, Wright F. 1988. Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol. 5:704–716.

Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature. 415:1022–1024.

Sokal RR, Rohlf FJ. 1995. Biometry. New York: W.H. Freeman and Co.

Sun H, Skogerbo G, Chen R. 2006. Conserved distances between vertebrate highly conserved elements. Hum Mol Genet. 15:2911–2922.

Taft RJ, Mattick JS. 2003. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. Genome Biol. 5:P1.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.

The Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature. 444:512.

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics. 172:1607–1619.

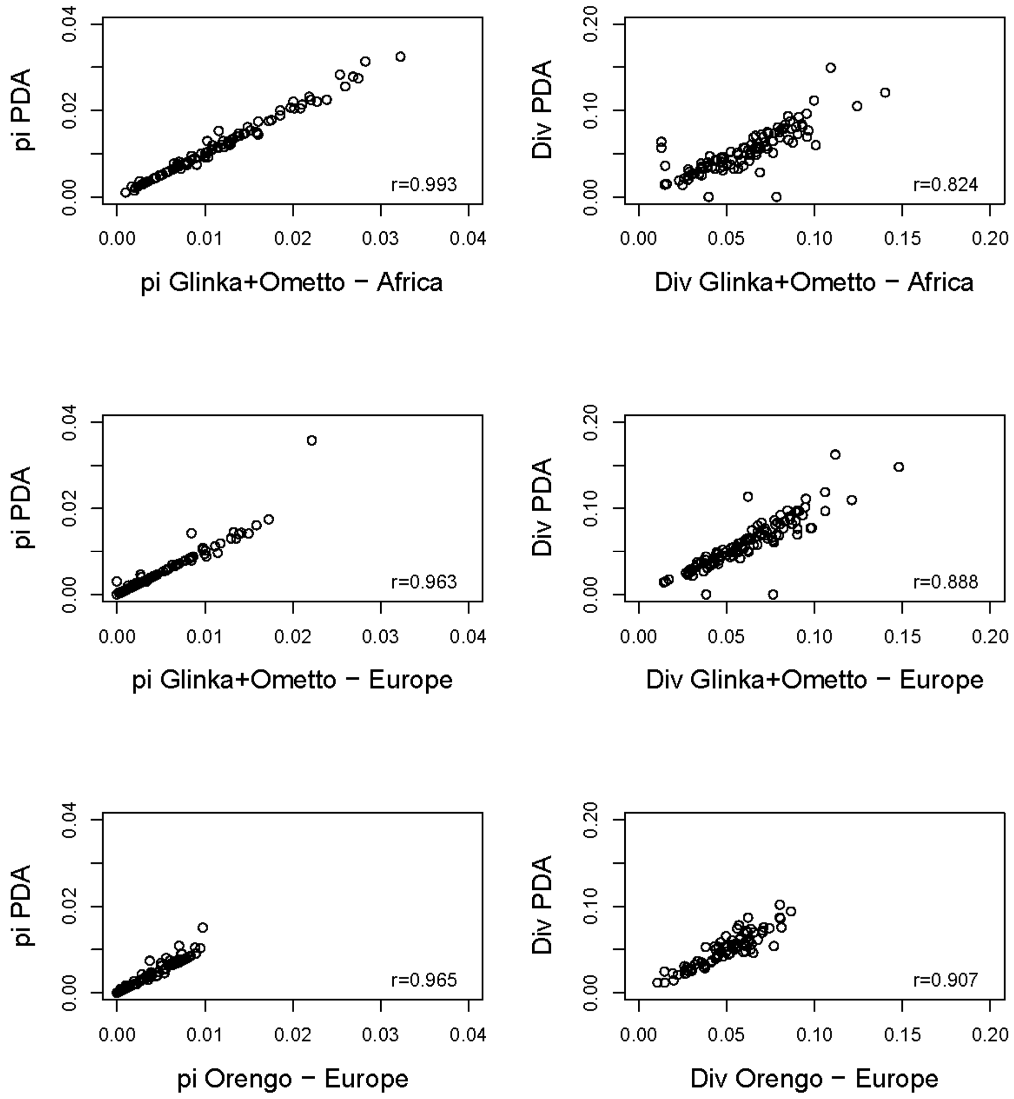
Webb CT, Shabalina SA, Ogurtsov AY, Kondrashov AS. 2002. Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Nucleic Acids Res. 30:1233–1239.

William Martin, Associate Editor

Accepted July 16, 2007

Supplementary Material

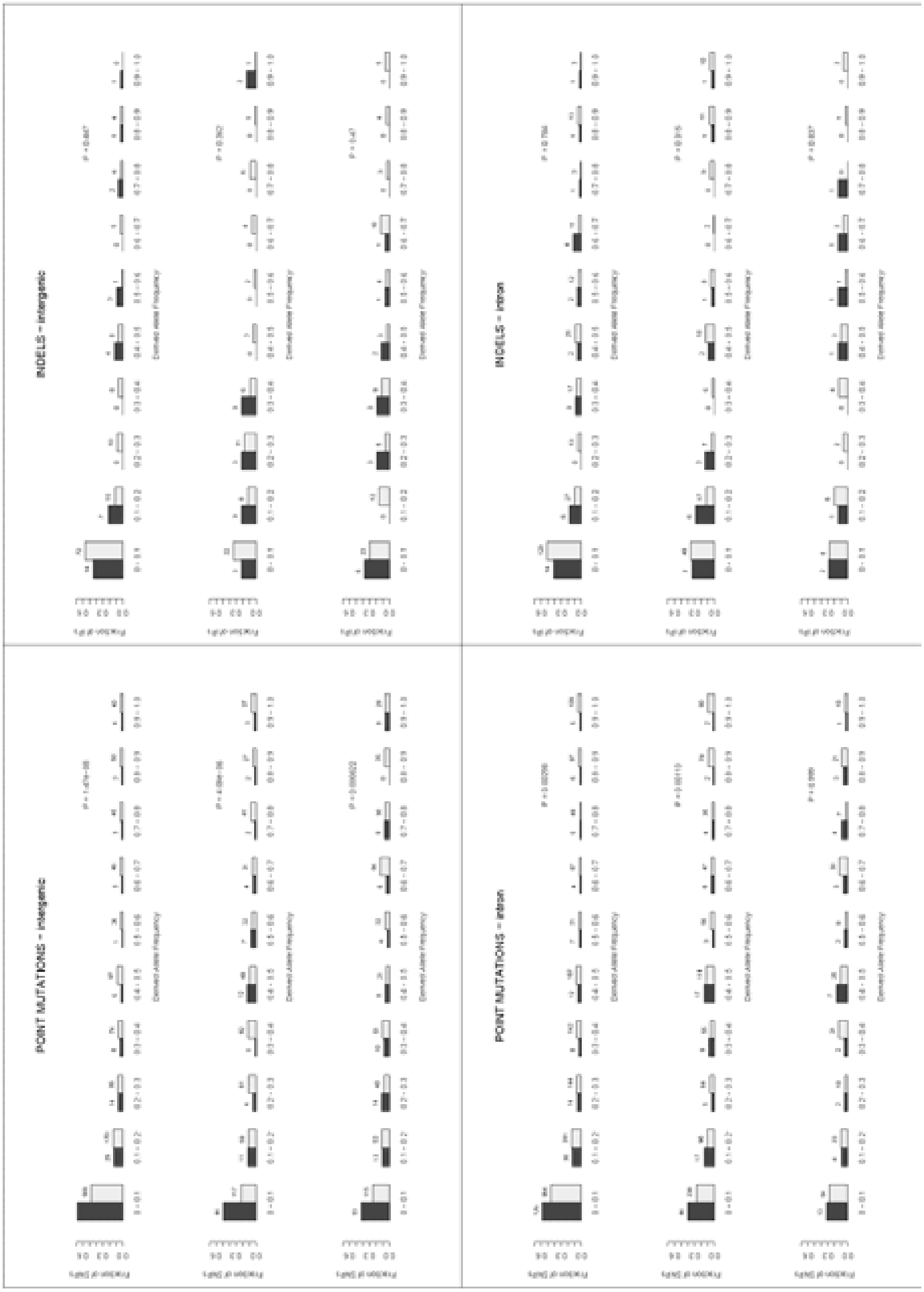
Supplementary File 1



Supplementary File 2

[This table is available in the accompanying CD]

SUPPLEMENTARY FILE 3



SUPPLEMENTARY FILE 4

Control Test A: SNPs in IFs

Problem: Noncoding regions which accumulate indel fixed differences (IFs) may also have a higher SNP density, which may affect McDonald-Kreitman tests if different alignment columns are used to calculate SNPs and SNFs.

Method: Test whether IF columns contain more SNPs than non-IF columns, or rather SNPs are distributed homogeneously among IFs and non-IFs.

CNS + Non-CNS:

AFR	Non-IF columns	IF columns	Total
bp	109834	5511	115345
SNP # Obs	3771	350	4121
Density	0.034	0.064	0.036
SNP # Exp	3924.1	196.9	-
$(O-E)^2/E$	5.97	119.05	-
Chi-squared =	125.03	**, df = 1, P =**	5.02E-29

EUR1	Non-IF columns	IF columns	Total
bp	120596	6482	127078
SNP # Obs	1586	140	1726
Density	0.013	0.022	0.014
SNP # Exp	1638.0	88.0	-
$(O-E)^2/E$	1.65	30.67	-
Chi-squared =	32.31	**, df = 1, P =**	1.31E-08

EUR2	Non-IF columns	IF columns	Total
bp	76293	2823	79116
SNP # Obs	866	67	933
Density	0.011	0.024	0.012
SNP # Exp	899.7	33.3	-
$(O-E)^2/E$	1.26	34.13	-
Chi-squared =	35.39	**, df = 1, P =**	2.69E-09

Non-CNS:

AFR	Non-IF columns	IF columns	Total
bp	79176	5012	84188
SNP # Obs	3334	336	3670
Density	0.042	0.067	0.044
SNP # Exp	3451.5	218.5	-
(O-E)^2/E	4.00	63.20	-

Chi-squared = 67.20 **, df = 1, P =** 2.45E-16

EUR1	Non-IF columns	IF columns	Total
bp	87307	5747	93054
SNP # Obs	1387	134	1521
Density	0.016	0.023	0.016
SNP # Exp	1427.1	93.9	-
(O-E)^2/E	1.12	17.09	-

Chi-squared = 18.21 **, df = 1, P =** 1.98E-05

EUR2	Non-IF columns	IF columns	Total
bp	47573	2580	50153
SNP # Obs	700	64	764
Density	0.015	0.025	0.015
SNP # Exp	724.7	39.3	-
(O-E)^2/E	0.84	15.52	-

Chi-squared = 16.36 **, df = 1, P =** 5.23E-05

CNS:

AFR	Non-IF columns	IF columns	Total
bp	30658	499	31157
SNP # Obs	437	14	451
Density	0.014	0.028	0.014
SNP # Exp	443.8	7.2	-
(O-E)^2/E	0.10	6.36	-

Chi-squared = 6.46 **, df = 1, P =** 0.01102

EUR1	Non-IF columns	IF columns	Total
bp	33289	735	34024
SNP # Obs	199	6	205
Density	0.006	0.008	0.006
SNP # Exp	200.6	4.4	-
(O-E)^2/E	0.01	0.56	-

Chi-squared = 0.57 **, df = 1, P =** 0.45027

EUR2	Non-IF columns	IF columns	Total
bp	28720	243	28963
SNP # Obs	166	3	169
Density	0.006	0.012	0.006
SNP # Exp	167.6	1.4	-
(O-E)²/E	0.01	1.77	-

Chi-squared = 1.78 **, df = 1, P =** 0.18212

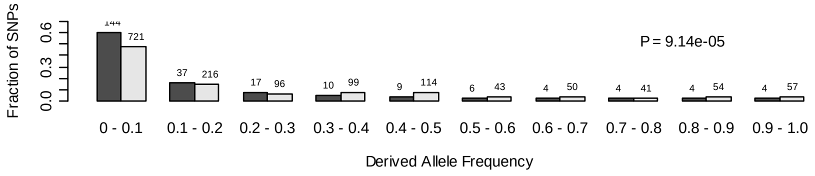
Control Test B: Base composition

Problem: CNSs are more GC-rich than non-CNSs, and GC sites are known to mutate more frequently than AT sites. These differences in the mutation rate between different classes of sites may cause differences in the distribution of derived allele frequencies, independent of selection pressure.

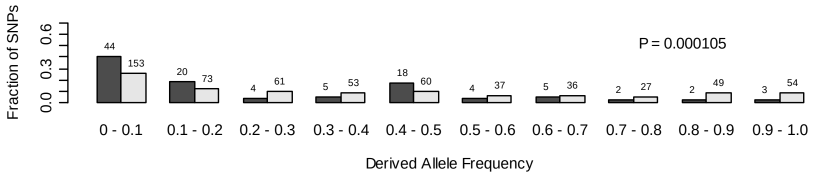
Method: Perform DAF distributions separately for G:C → A:T and for A:T → G:C mutations.

G:C → A:T DAF test:

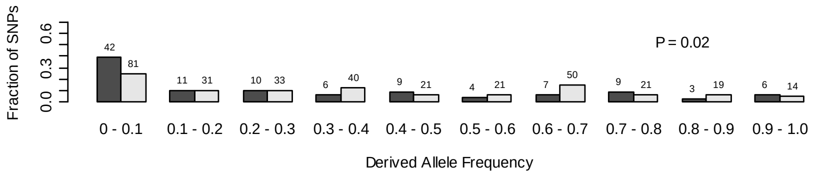
AFR



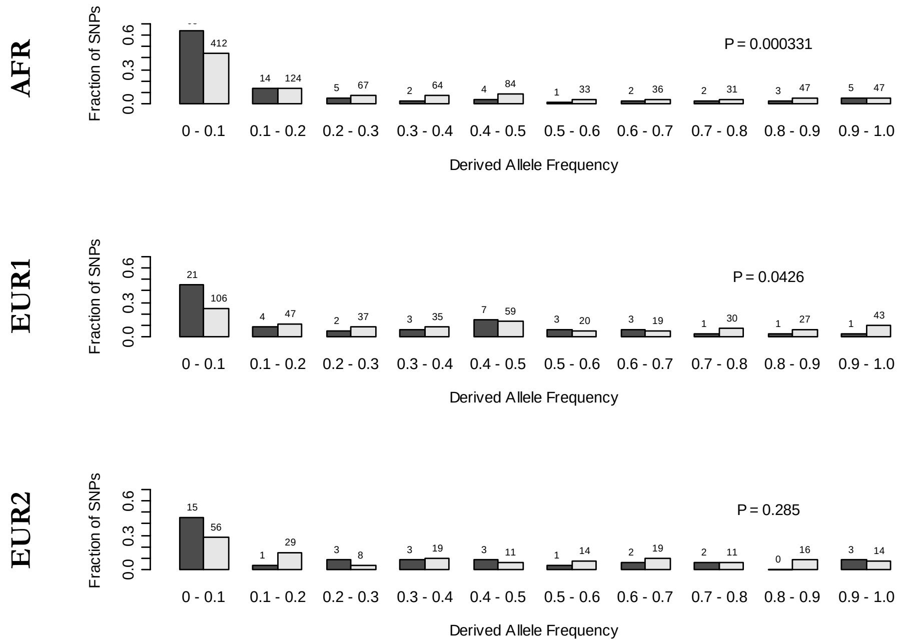
EUR1



EUR2



A:T → G:C DAF test:



Control Test C: Indel errors

Problem: SNPs/SNFs may tend to accumulate near gaps in multiple alignments and therefore differences in SNP density may be an artifact of differences in indel density.

Method: Exclude from the analyses CNS and non-CNS regions that contain either IPs or IFs for McDonald-Kreitman tests, and exclude regions with IPs for DAF tests.

MK test:

AFR

	P	D
C	262	194
NC	349	530
Σ	**611**	**724**
Reduc.	**24.93%**	**63.40%**

P = 9.585e-10

EUR1

	P	D
C	132	257
NC	151	637
Σ	**283**	**894**
Reduc.	**12.58%**	**59.65%**

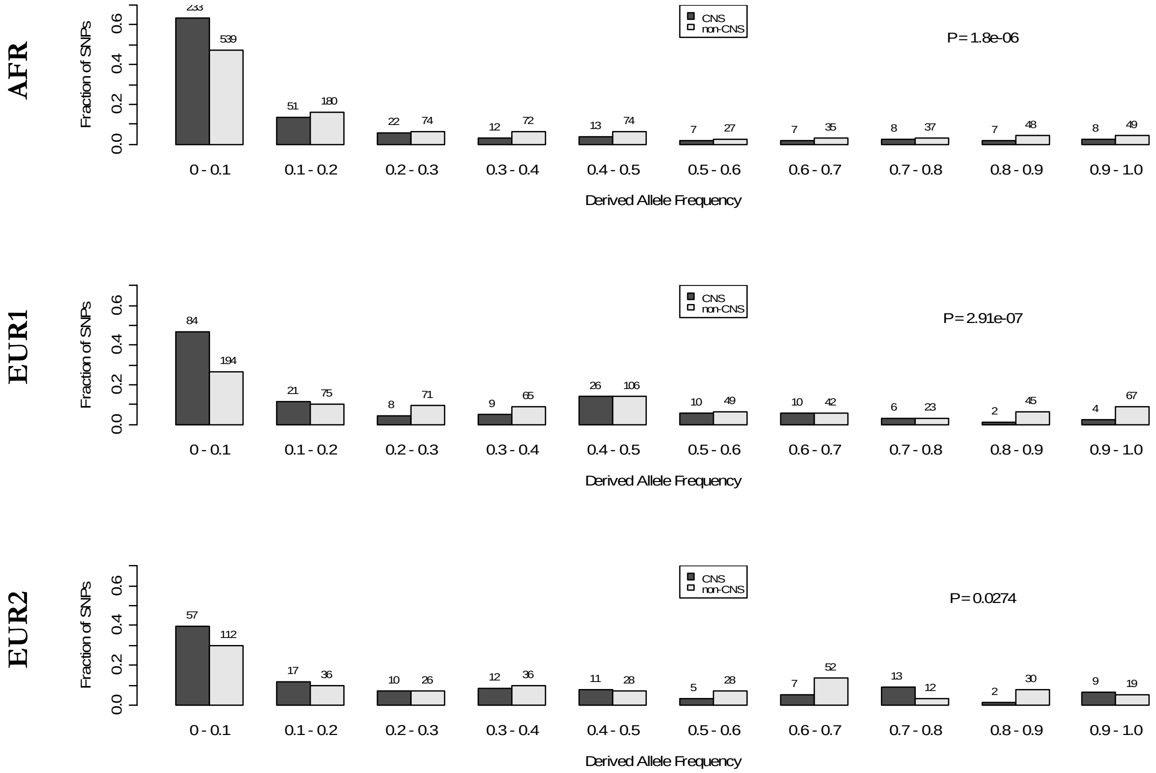
P = 3.685e-08

EUR2

	P	D
C	107	217
NC	97	403
Σ	**204**	**620**
Reduc.	**-10.31%**	**46.15%**

P = 1.401e-05

DAF test:



Control Test D: Consecutive SNP/SNFs

Problem: If two indels of exactly the same length occur in the essentially the same position, one an insertion and one a deletion, the alignment program can force the two indels to collapse in the alignment, and thus result in a run of consecutive substitutions. Thus differences in SNP density may be an artifact of differences in indel density.

Method: All SNPs followed or preceded by another SNP in the alignment, and all SNFs followed or preceded by another SNF, have been excluded from MK and DAF tests.

MK tests:

AFR

	P	D
C	390	332
NC	2756	3816
Σ	**3146**	**4148**
Reduc.	85.85%	91.30%

P = 6.325e-10

EUR1

	P	D
C	188	406
NC	1192	4752
Σ	**1380**	**5158**
Reduc.	84.23%	91.46%

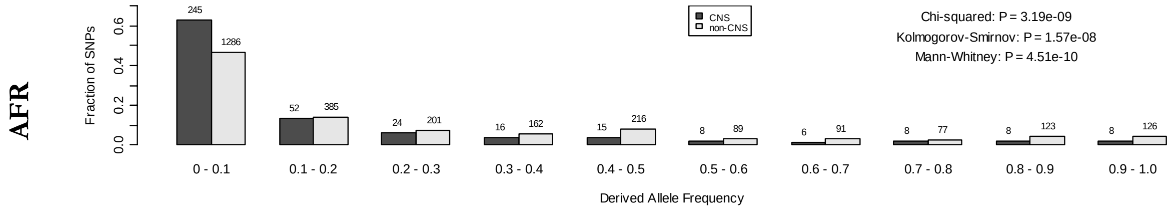
P = 5.719e-11

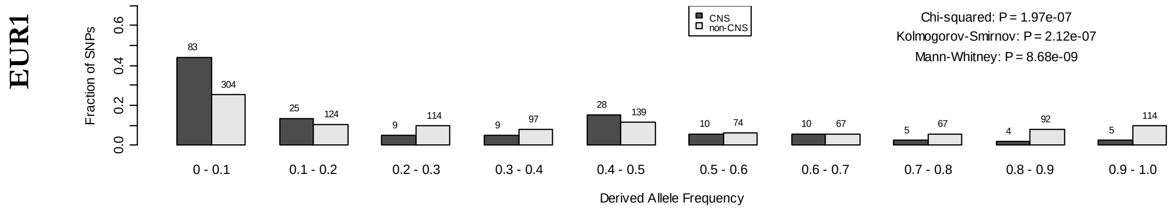
EUR2

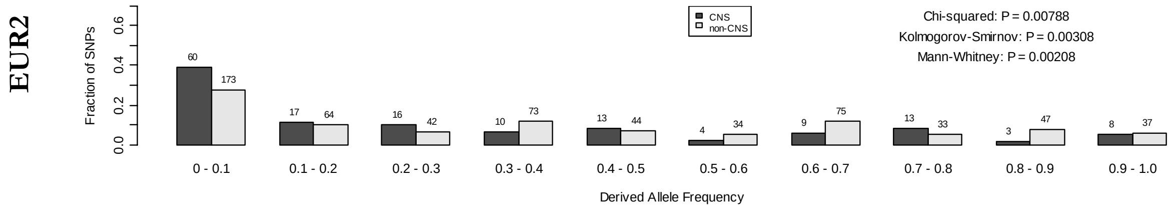
	P	D
C	153	339
NC	622	2508
Σ	**775**	**2847**
Reduc.	75.40%	86.48%

P = 2.339e-08

DAF tests:







Control Test E: Low power of indel data

Problem: The lack of significance of MK and DAF tests with indel data may be due to small sample size of IPs/IFs compared to SNPs/SNFs.

Method: Rescale SNP/SNF data so that the numbers of SNPs and SNFs equal the numbers of observed IPs and IFs, while maintaining the proportions of the observed contingency tables.

Observed SNP/SNF:

	Polym	**Diverg**
C	P_C	D_C
NC	P_{NC}	D_{NC}
	P	**D**

Rescaled SNP/SNF:

	Polym	**Diverg**
N	$P_C/P * P_{indels}$	$D_C/D * D_{indels}$
NC	$P_{NC}/P * P_{indels}$	$D_{NC}/D * D_{indels}$
	$\mathbf{P_{indels}}$	$\mathbf{D_{indels}}$

MK tests:

SNPs Observed

AFR

	P	**D**
C	437	374
NC	3334	4854
Σ	**3771**	**5228**
Reduc.	**86.89%**	**92.30%**

P = 5.54e-13

EUR1

	P	**D**
C	199	456
NC	1387	6217
Σ	**1586**	**6673**
Reduc.	**85.65%**	**92.67%**

P = 5.58e-14

EUR2

	P	**D**
C	166	376
NC	700	3314
Σ	**866**	**3690**
Reduc.	**76.29%**	**88.65%**

P = 3.17e-13

INDELS Observed

AFR

	P	**D**
C	66	107
NC	380	901
Σ	**446**	**1008**
Reduc.	**82.63%**	**88.12%**

P = 0.029

EUR1

	P	**D**
C	35	121
NC	200	1140
Σ	**235**	**1261**
Reduc.	**82.50%**	**89.39%**

P = 0.0201

EUR2

	P	**D**
C	23	100
NC	107	611
Σ	**130**	**711**
Reduc.	**78.50%**	**83.63%**

P = 0.347

SNPs RESCALED

AFR

	P	**D**
C	51.684	72.11
NC	394.32	935.89
Σ	**446**	**1008**
Reduc.	**86.89%**	**92.30%**

P = 0.007112

EUR1

	P	**D**
C	29.486	86.171
NC	205.51	1174.8
Σ	**235**	**1261**
Reduc.	**85.65%**	**92.67%**

P = 0.003992

EUR2

	P	**D**
C	24.919	72.449
NC	105.08	638.55
Σ	**130**	**711**
Reduc.	**76.29%**	**88.65%**

P = 0.005221

DAF distribution:

DAF:	0 - .1	.1 - .2	.2 - .3	.3 - .4	.4 - .5	.5 - .6	.6 - .7	.7 - .8	.8 - .9	.9 - 1	p-values
AFR											
SNPs Observed											
C	277	59	28	16	17	8	7	8	8	9	**Chi:** 1.07e-11
NC	1533	481	230	216	264	107	113	94	147	149	**KS:** 6.56e-11
Total	1810	540	258	232	281	115	120	102	155	158	
Indels Observed											
C	28	13	0	3	6	5	4	3	2	2	**Chi:** 0.151
NC	201	42	23	25	33	13	16	7	17	3	**KS:** 0.568
Total	229	55	23	28	39	18	20	10	19	5	
SNPs RESCALED											
C	35.05	6.01	2.50	1.93	2.36	1.25	1.17	0.78	0.98	0.28	**Chi:** 0.02354
NC	193.95	48.99	20.50	26.07	36.64	16.75	18.83	9.22	18.02	4.72	**KS:** 0.00206
Total	229	55	23	28	39	18	20	10	19	5	
EUR1											
SNPs Observed											
C	88	28	9	10	29	10	10	6	4	5	**Chi:** 6.96e-08
NC	355	155	125	117	160	88	78	76	106	127	**KS:** 1.89e-08
Total	443	183	134	127	189	98	88	82	110	132	
Indels Observed											
C	10	9	6	3	2	1	0	0	1	3	**Chi:** 0.547
NC	71	25	18	11	21	10	6	15	12	11	**KS:** 0.321
Total	81	34	24	14	23	11	6	15	13	14	
SNPs RESCALED											
C	16.09	5.20	1.61	1.10	3.53	1.12	0.68	1.10	0.47	0.53	**Chi:** 0.05510
NC	64.91	28.80	22.39	12.90	19.47	9.88	5.32	13.90	12.53	13.47	**KS:** 0.00061
Total	81	34	24	14	23	11	6	15	13	14	
EUR2											
SNPs Observed											
C	66	17	16	12	15	6	9	13	3	9	**Chi:** 0.00038
NC	179	76	50	82	46	41	94	37	56	39	**KS:** 0.00157
Total	245	93	66	94	61	47	103	50	59	48	

Indels Observed

C	8	1	3	3	3	2	2	1	0	0
NC	31	18	7	13	6	5	12	3	5	7
Total	39	19	10	16	9	7	14	4	5	7

Chi: 0.763
KS: 0.9

SNPs RESCALED

C	10.51	3.47	2.42	2.04	2.21	0.89	1.22	1.04	0.25	1.31
NC	28.49	15.53	7.58	13.96	6.79	6.11	12.78	2.96	4.75	5.69
Total	39	19	10	16	9	7	14	4	5	7

Chi: 0.24599
KS: 0.00206

Control Test F: Effect of rare alleles

Problem: Rare alleles (singletons) are likely to be enriched for slightly deleterious alleles and may mask the effect of positive selection.

Method: Discard singletons from the MK-tests.

AFR

ALL ALLELES

INTRON

	P	**D**
C	216	214
NC	2121	3166

NI: 1.506
P: 5.08E-05

INTERGENIC

	P	**D**
C	221	160
NC	1213	1688

NI: 1.923
P: 2.92E-09

ALL

	P	**D**
C	437	374
NC	3334	4854

NI: 1.701
P: 5.54E-13

EXCLUDING RARE ALLELES

INTRON

	P	**D**
C	90	214
NC	1157	3166

NI: 1.151
P: 0.3113

INTERGENIC

	P	**D**
C	70	160
NC	644	1688

NI: 1.147
P: 0.405

ALL

	P	**D**
C	160	374
NC	1801	4854

NI: 1.153
P: 0.1623

EUR1

ALL ALLELES

INTRON

	P	D
C	109	270
NC	871	4042

NI: 1.873
P: 1.45E-07

INTERGENIC

	P	D
C	90	186
NC	516	2175

NI: 2.041
P: 2.06E-07

ALL

	P	D
C	199	456
NC	1387	6217

NI: 1.957
P: 5.58E-14

EXCLUDING RARE ALLELES

INTRON

	P	D
C	65	270
NC	633	4042

NI: 1.536
P: 0.0036

INTERGENIC

	P	D
C	46	186
NC	399	2175

NI: 1.348
P: 0.1023

ALL

	P	D
C	111	456
NC	1032	6217

NI: 1.466
P: 6.64E-04

EUR2

ALL ALLELES

INTRON

	P	D
C	41	113
NC	230	948

NI: 1.495
P: 0.051

INTERGENIC

	P	D
C	125	263
NC	470	2366

NI: 2.392
P: 1.58E-13

ALL

	P	D
C	166	376
NC	700	3314

NI: 2.092
P: 3.17E-13

EXCLUDING RARE ALLELES

INTRON

	P	D
C	28	113
NC	166	948

NI: 1.414
P: 0.1584

INTERGENIC

	P	D
C	72	263
NC	355	2366

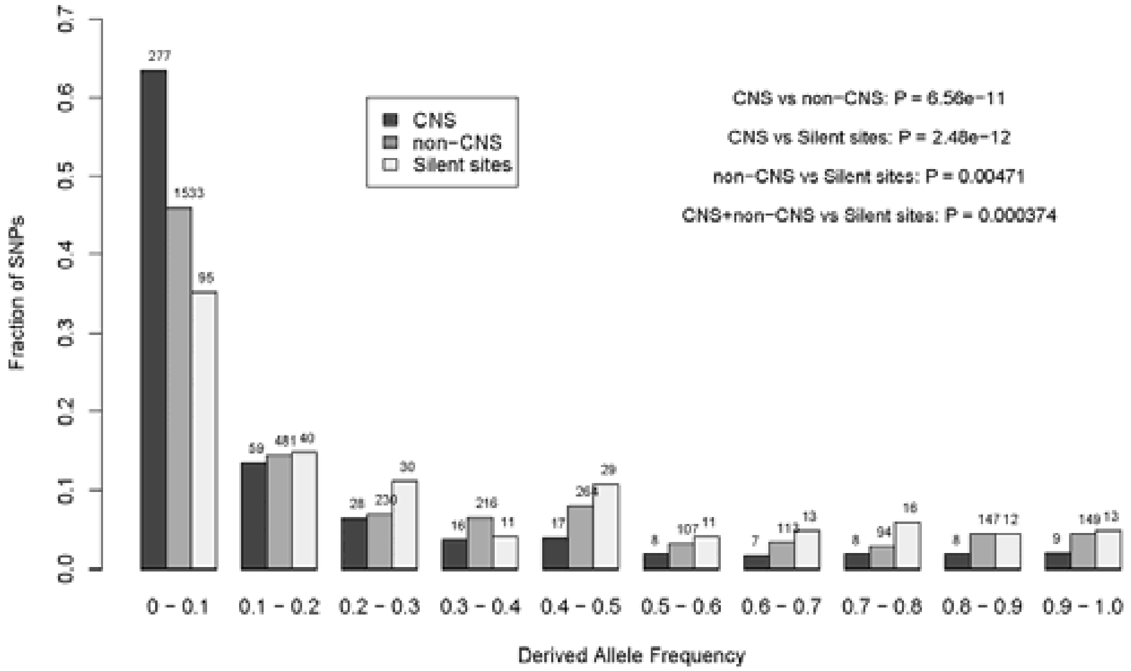
NI: 1.825
P: 3.73E-05

ALL

	P	D
C	100	376
NC	521	3314

NI: 1.692
P: 1.86E-05

SUPPLEMENTARY FILE 5



2.4. Coding evolution of *Hox* and *Hox*-derived genes in the genus *Drosophila*

In this last part of the thesis, the sequence evolution of *Hox* and *Hox*-derived genes is studied with the purpose of determining whether or not the high functional conservation of *Hox* genes is also met at the DNA level. In the appended publication, we measure the rates of nucleotide divergence and indel fixation of three *Hox* genes and compare them with those of three *Hox*-derived genes and 15 non-*Hox* genes in sets of orthologous sequences of three species of the genus *Drosophila*. Our results show that *Hox* genes in fruit flies are evolving rapidly despite their conserved role in development and their complex expression patterns. Their evolutionary rate is even higher than that of non-*Hox* genes when both amino acid differences and indels are taken into account: 43.39% of the amino acid sequence is altered in *Hox* genes, *versus* 30.97% in non-*Hox* genes and 64.73% in *Hox*-derived genes. Surprisingly, microsatellites scattered along the coding sequence of *Hox* genes explain partially, but not fully, their fast sequence evolution. Overall, these results show that *Hox* genes have a higher evolutionary dynamics than other developmental genes and emphasize the need to take into account indels in addition to nucleotide substitutions in order to accurately estimate evolutionary rates.

- **Article 6:** CASILLAS, S., B. NEGRE, A. BARBADILLA and A. RUIZ (2006) Fast sequence evolution of *Hox* and *Hox*-derived genes in the genus *Drosophila*. *BMC Evolutionary Biology* **6**: 106.

BMC Evolutionary Biology

BioMed Central

Research article

Open Access

Fast sequence evolution of *Hox* and *Hox*-derived genes in the genus *Drosophila*

Sònia Casillas[1], Bárbara Negre[1,2], Antonio Barbadilla[1] and Alfredo Ruiz*[1]

Address: [1]Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain and [2]Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK

Email: Sònia Casillas - Sonia.Casillas@uab.es; Bárbara Negre - bn219@cam.ac.uk; Antonio Barbadilla - Antonio.Barbadilla@uab.es; Alfredo Ruiz* - Alfredo.Ruiz@uab.es

* Corresponding author

Published: 12 December 2006

BMC Evolutionary Biology 2006, **6**:106 doi:10.1186/1471-2148-6-106

Received: 13 June 2006
Accepted: 12 December 2006

This article is available from: http://www.biomedcentral.com/1471-2148/6/106

Abstract

Background: It is expected that genes that are expressed early in development and have a complex expression pattern are under strong purifying selection and thus evolve slowly. *Hox* genes fulfill these criteria and thus, should have a low evolutionary rate. However, some observations point to a completely different scenario. *Hox* genes are usually highly conserved inside the homeobox, but very variable outside it.

Results: We have measured the rates of nucleotide divergence and indel fixation of three *Hox* genes, *labial* (*lab*), *proboscipedia* (*pb*) and *abdominal-A* (*abd-A*), and compared them with those of three genes derived by duplication from *Hox3*, *bicoid* (*bcd*), *zerknüllt* (*zen*) and *zerknüllt-related* (*zen2*), and 15 non-*Hox* genes in sets of orthologous sequences of three species of the genus *Drosophila*. These rates were compared to test the hypothesis that *Hox* genes evolve slowly. Our results show that the evolutionary rate of *Hox* genes is higher than that of non-*Hox* genes when both amino acid differences and indels are taken into account: 43.39% of the amino acid sequence is altered in *Hox* genes, versus 30.97% in non-*Hox* genes and 64.73% in *Hox*-derived genes. Microsatellites scattered along the coding sequence of *Hox* genes explain partially, but not fully, their fast sequence evolution.

Conclusion: These results show that *Hox* genes have a higher evolutionary dynamics than other developmental genes, and emphasize the need to take into account indels in addition to nucleotide substitutions in order to accurately estimate evolutionary rates.

Background

Hox genes are homeobox containing genes involved in the specification of regional identities along the anteroposterior body axis and, thus, play a fundamental role in animal development [1]. They encode transcription factors that regulate the expression of other genes downstream in the regulatory cascade of development and have been found in all metazoans, including flies, worms, tunicates, lampreys, fish and tetrapods. A particular feature of these genes is that they are usually clustered together in complexes and arranged in the chromosome in the same order as they are expressed along the anteroposterior body axis of the embryo [2,3]. Ten genes arranged in a single complex comprised the ancestral *Hox* gene complex of arthropods (HOM-C) [4-6]. However, at least three different HOM-C splits have occurred during the evolution of diptera [7-10], and several non-homeotic genes and other genes derived from ancestral *Hox* genes are interspersed among the *Drosophila Hox* genes.

The stability of *Hox* gene number and the conservation of *Hox* ortholog sequences prompted the notion that *Hox* proteins have not significantly diverged in function. However, it is now known that several arthropod *Hox* proteins have changed in sequence and/or function, including those encoded by *Hox3* [11-13], *fushi tarazu* (*ftz*) [14], *Ultrabithorax* (*Ubx*) [15] and *Antennapedia* (*Antp*) [16]. In winged insects, including *Drosophila*, *Hox3* and *ftz* lost their homeotic function, that is, their ability to transform the characteristics of one body part into those of another body part [17,18], and their expression domains are no longer arranged along the anteroposterior axis of the embryo. Therefore, only eight *Hox* genes remain in these species [6]. *Hox3* gained a novel extraembryonic function, and underwent two consecutive duplications that gave rise to *bicoid* (*bcd*), *zerknüllt* (*zen*) and *zerknüllt-related* (*zen2*). The first duplication took place in the cyclorrhaphan fly lineage and gave rise to *zen* and *bcd* [12,13]. Afterwards, but before the *Drosophila* radiation, *zen* went through a second duplication that gave birth to *zen2* [19]. Seemingly, *bcd* and *zen* have specialized and perform separate functions in the establishment of the embryo's body plan: the maternal gene *bcd* codes for an important morphogen that establishes anteroposterior polarity [20] and *zen* is a zygotic gene involved in dorsoventral differentiation [21]. *zen2* has the same expression pattern of *zen*, although its function is unknown. Despite its high sequence divergence across species, it has been maintained for more than 60 Myr [19].

Hox proteins contain a highly conserved domain of 60 amino acids (coded by the homeobox) that binds DNA through a '*helix-turn-helix*' structure. This motif is very similar in terms of sequence and structure to that of many DNA binding proteins. Functional comparisons of *Hox* orthologs have largely focused on their highly conserved homeodomain sequences and have demonstrated their functional interchangeability between species [22-26]. *Hox*-derived genes, although having lost their homeotic function, still retain the homeobox.

It has been shown that housekeeping genes, which are expressed in all cells and at all times, are under strong purifying selection and thus evolve slowly (e.g. histones, or genes involved in the cell cycle) [27,28]. *Hox* genes, on the contrary, are expressed early in development and have a complex regulated expression pattern. Mutations in such genes will on average have more deleterious fitness consequences than mutations occurring in genes expressed later on, because they may have cascading consequences for the later steps in development and thus may broadly alter the adult phenotype [29-31]. Therefore, we also expect *Hox* genes to be highly constrained and thus evolve slowly. In fact, Davis, Brandman, and Petrov [29] found a highly significant relationship between the developmental timing of gene expression and their nonsynonymous evolutionary rate: genes expressed early in development are likely to have a slower rate of evolution at the protein level than those expressed later. Surprisingly, the strongest negative relationship between expression and evolutionary rate occurred only after the main burst of expression of segment polarity and *Hox* genes in embryonic development, so these genes could be evolving differently from other developmental genes. However, only one segment polarity gene, *wingless* (*wg*), and two *Hox* genes, *Antp* and *abdominal-A* (*abd-A*), were analyzed.

Furthermore, Marais *et al*. [32] found a negative correlation between evolutionary rate at the protein level (as measured by the number of nonsynonymous substitutions per nonsynonymous site, d_N) and intron size in *Drosophila*, likely due to a higher abundance of *cis*-regulatory elements in introns (especially first introns) in genes under strong selective constraints. We know from a previous study that the *Hox* genes used in this study contain a long intron replete with regulatory elements [19]. Therefore, we would expect these genes to be strongly constrained.

However, other studies seem to point to a completely different scenario. Developmental biologists noticed a long time ago that a large portion of the sequence of *Hox* proteins diverges so fast that it is difficult to align homologues from different arthropod classes [33]. In fact, nucleotide sequences outside the homeobox in *labial (lab)* and *Ubx* have been reported to diverge significantly [8,15]. These sequence differences may be neutral with respect to protein function or, more intriguingly, they could be involved in the functional divergence of *Hox* proteins and the evolutionary diversification of animals [15]. Moreover, Karlin and Burge [34] have shown that many essential developmental genes, including *Hox* genes, contain long microsatellites within their coding sequence (e.g. trinucleotide repeats that do not disrupt the open reading frame). The vast majority of these genes function in development and/or transcription regulation, and are expressed in the nervous system. Due to the particular mutation mechanism acting on these repetitive sequences by replication slippage [35,36], microsatellites are subject to frequent insertions and deletions. Thus, these repetitive sequences could be responsible for a higher than expected evolutionary rate of *Hox* genes. However, and despite all the previous contributions, no quantification of the rates of nucleotide and indel evolution has been reported so far for a set of *Hox* genes.

On the other hand, the origin by duplication and the functional evolution of *Hox*-derived genes suggest that they might be evolving fast at the sequence level as well. Duplicated genes are known to undergo a period of accelerated evolution where: they may degenerate to a pseudogene (pseudogenization), each daughter gene may adopt part of the functions of their parental gene (subfunctionalization), or they may acquire new functions (neofunctionalization) [37-40]. The only divergence estimate reported in a *Hox*-derived gene was calculated between two close species (*D. melanogaster* and *D. simulans*) in *bcd* [41]. A recent study found an increased sequence polymorphism in *bcd* in comparison to *zen*, which was ascribed to a relaxation of selective constraint on this maternal gene resulting from sex-limited expression [42]. Therefore, *bcd* is expected to evolve faster than *zen* under this model. The evolutionary rates of *zen* and *zen2*, however, have not been reported so far.

We have measured the rates of nucleotide substitution and indel fixation of three *Hox* genes, *lab*, *proboscipedia* (*pb*) and *abd-A*, and compared them with those of *bcd*, *zen* and *zen2*, which were derived by duplication from *Hox3*, and a sample of 15 non-*Hox* genes, in the genus *Drosophila*. These rates were compared to test the hypothesis that *Hox* genes, similar to other genes with complex expression patterns and that are essential in the early development, evolve slowly. We have also evaluated the contribution of the homeobox and the repetitive regions within *Hox* and *Hox*-derived genes to the evolutionary rates.

The sequences compared comprise all the complete genes available in *D. buzzatii* (representative of the Drosophila subgenus), and their orthologs in *D. melanogaster* and *D. pseudoobscura* (both species in the Sophophora subgenus). *D. buzzatii* belongs to the *repleta* species group, a group comprising ~100 species that has been widely used as a model in studies of genome evolution, ecological adaptation and speciation. Negre *et al.* [19] have recently compared the genomic organization of the HOM-C complex in *D. buzzatii* to that of *D. melanogaster* and *D. pseudoobscura*, and studied the functional consequences of two HOM-C splits present in this species. When our study began, this was the largest set of orthologous *Hox* genes in species from both subgenera of the *Drosophila* genus, and this allowed the exploration of evolutionary rates throughout the *Drosophila* phylogeny. Due to the high divergence of *Hox* genes [8], the inclusion of more distant species outside the *Drosophila* genus (such as mosquito or honeybee) would probably not be appropriate for the estimation of genetic distances. Moreover, these species do not contain the *Hox*-derived genes studied here.

Results

Nucleotide evolution of Hox, Hox-*derived and non*-Hox *genes*

Nucleotide substitution parameters were calculated for the coding nucleotide alignments independently for each gene [see Additional file 1]. We then tested for differences between the three groups of genes (*Hox*, *Hox*-derived and non-*Hox*) (top section of Table 1) [see Additional file 2]. Our results showed that *Hox*-derived genes are evolving much faster and with less functional constraint than *Hox* and non-*Hox* genes. Differences among the three groups are significant for the number of nonsynonymous substitutions per nonsynonymous site, d_N (P = 0.022), and the level of functional constraint, ω (P = 0.000) (see Methods). The gene *zen2* is the main gene responsible for the high values of nucleotide substitutions (both synonymous and nonsynonymous) in its group [see Additional file 1]. On the contrary, *Hox* and non-*Hox* genes have a similar number of nucleotide substitutions, *t* (P > 0.1). However the level of functional constraint is even higher (lower ω) in non-*Hox* genes than in *Hox* genes (ω = 0.04156 versus ω = 0.06094, respectively), although differences are only marginally significant (P = 0.063). Therefore, *Hox* genes do not seem to be evolving more slowly than other non-homeotic genes, despite their essential function in early development.

Then, we plotted d_N and ω in sliding windows along the coding sequences of *Hox* and *Hox*-derived genes to see whether or not these parameters behave homogeneously along the sequence. Figure 1 shows that, in all genes except *zen2*, there is a substantial decrease of both d_N and ω near the homeobox. *zen2* contains a rapidly evolving homeobox with high ω values. Contrarily, we have observed that peaks of d_N tend to lie within repetitive regions (data not shown).

To control for a possible effect on the overall nucleotide evolution of both the homeobox and the repetitive regions (see Methods) of these *Hox* and *Hox*-derived genes, we tested again for differences among the three groups of genes excluding these regions. Removing the homeobox in *Hox* and *Hox*-derived coding sequences (second section of Table 1) elevated the number of nucleotide substitutions in these two groups, and decreased further their level of functional constraint. Again, differences among groups were significant for d_N (P = 0.005) and ω (P = 0.000), and the same tendency of the previous analysis with complete coding sequences was observed. In contrast, removing repetitive regions (third section of Table 1) decreased the number of nucleotide substitutions, especially in *Hox* genes, where all the genes in the group contain this type of region. Therefore, the elimination of repetitive regions slightly increases the difference between *Hox* and non-*Hox* genes in terms of nucleotide substitu-

Table 1: Mean nucleotide substitution parameters and ANOVAs for the three groups of genes.

		t	d_N	d_S	ω
Complete coding sequences	*Hox*	2.10917	0.15964	2.59066	0.06094
	Hox-derived	3.86336	0.39380	4.27598	0.09226
	Non-*Hox*	2.91160	0.15802	3.80668	0.04156
	ANOVA	n.s.	*	n.s.	***
Coding sequences excluding the homeobox	*Hox*	2.27653	0.18257	2.65921	0.06673
	Hox-derived	5.04914	0.54809	5.26666	0.11320
	Non-*Hox*	2.91160	0.15802	3.80668	0.04156
	ANOVA	n.s.	**	n.s.	***
Coding sequences excluding repetitive regions	*Hox*	1.81997	0.12399	2.35029	0.05310
	Hox-derived	3.71981	0.37759	4.14242	0.09042
	Non-*Hox*	2.85593	0.15444	3.76458	0.04035
	ANOVA	n.s.	*	n.s.	***
Coding sequences excluding the homeobox and repetitive regions	*Hox*	1.94286	0.14684	2.33783	0.06146
	Hox-derived	4.88928	0.53011	5.12014	0.11245
	Non-*Hox*	2.85593	0.15444	3.76458	0.04035
	ANOVA	n.s.	**	n.s.	***

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001)

tions, and reduces the difference in functional constraint. Once more, differences among groups were significant for d_N (P = 0.030) and ω (P = 0.001). Excluding both the homeobox and the repetitive regions (bottom section of Table 1) gave intermediate results. Therefore, we can conclude that: (1)*Hox* and non-*Hox* genes are evolving similarly in terms of nucleotide substitutions, (2) *Hox*-derived genes are evolving much faster and with less functional constraint than the other two groups of genes, and (3) neither the homeobox nor the repetitive regions alter the estimates significantly, and thus are not entirely responsible for the two previous conclusions.

An excess of nonsynonymous over synonymous substitutions is a robust indicator of positive selection at the molecular level. Therefore, we searched for values of nonsynonymous/synonymous rate ratio (d_N/d_S = ω) greater than 1 to investigate whether Darwinian selection has been acting on any of the coding sequences analyzed in this study. However, no evidence of positive selection in any coding sequence or region of it was found.

Amino acid and structural changes at the protein level

We used the protein alignments to calculate the proportion of amino acid differences and indels. In the first case (Table 2, Figure 2), differences among the three groups – *Hox*, *Hox*-derived and non-*Hox* – were not significant (P = 0.101). However, the proportion of amino acid differences was substantially higher for *Hox*-derived genes (40.43%) than for *Hox* and non-*Hox* genes (22.80% and 23.77%, respectively). This result is in full agreement with our previous estimates of d_N (Table 1), which showed high values of this parameter for *Hox*-derived genes, but very similar values for *Hox* and non-*Hox* genes.

Second, we analyzed the proportion of indels in the alignments (Table 3, Figure 2). In this case, differences among the three groups of genes were highly significant (P = 0.000). Surprisingly, differences were due to the low indel proportion in non-*Hox* genes (8.73%) compared to the high values for *Hox* and *Hox*-derived genes (25.77% and 37.53%, respectively). Furthermore, we tested for differences in indel length using a nested ANOVA. The results indicated that, although the variation in indel length between genes within groups is significant (P = 0.021), the difference between groups is even more significant (P = 0.001). Mean indel length for *Hox*, *Hox*-derived and non-*Hox* genes is 4.22, 5.99 and 3.55 amino acids, respectively. Non-*Hox* genes not only have on average shorter indels, but also their longest indel is only 23 amino acids, in comparison with 43 and 40 amino acids for *Hox* and *Hox*-derived genes, respectively. In all groups, the indel

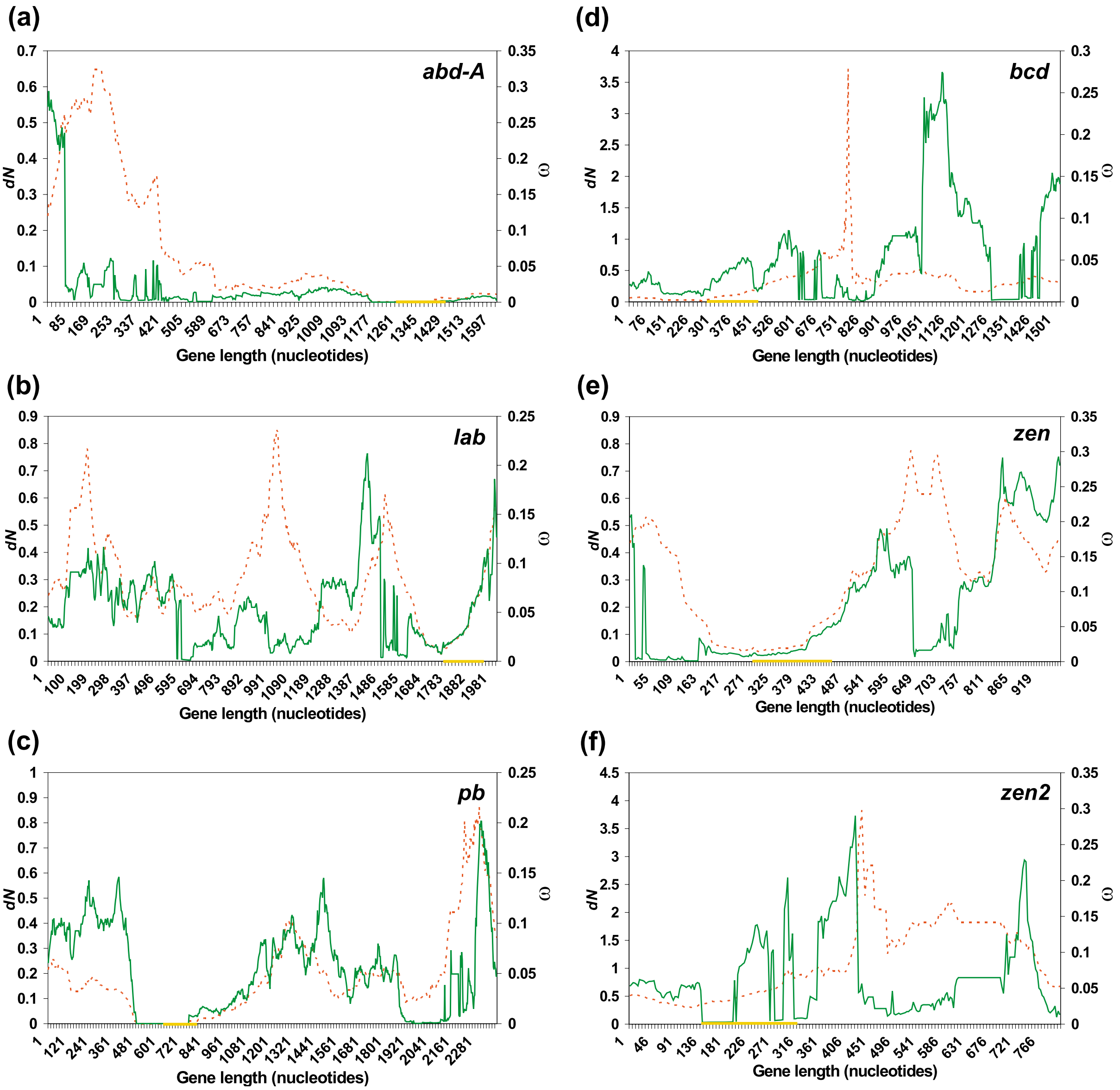


Figure 1
Distribution of d_N and ω in sliding windows along the coding sequence of genes. Distribution of d_N (broken line) and ω (solid line) in sliding windows of 240 nucleotides. (a) *abd-A*, (b) *lab*, (c) *pb*, (d) *bcd*, (e) *zen* and (f) *zen2*. In each case, the position of the homeobox is represented by a yellow box within the X axis.

length distribution follows a negative exponential curve: short indels are common and their abundance declines as length increases (data not shown).

Finally, we tested whether the proportions of amino acid differences and indels are correlated. The Pearson correlation indicated that these two variables are positively but not significantly correlated ($r_{Pearson}$ = 0.307, P = 0.175).

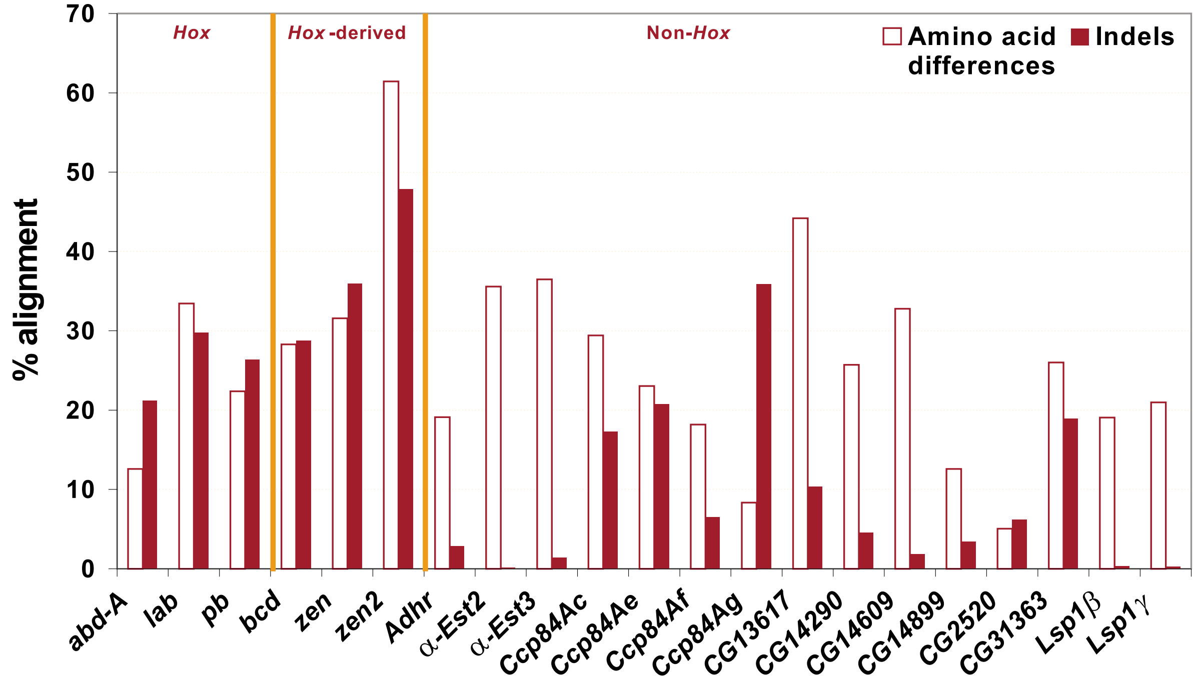


Figure 2
Proportion of amino acid differences and indels in the set of genes analyzed in this study.

Therefore, genes with a high proportion of indels do not necessarily have a high proportion of amino acid substitutions. This probably points to different causal mechanisms for amino acid substitutions and indels.

Effect of long repetitive tracks in the percentages of amino acid differences and indels of Hox *and* Hox*-derived proteins*

Most *Hox* and *Hox*-derived proteins contain large repetitive regions present throughout the protein except the region near the homeobox and other highly conserved regions (see for instance the amino acid sequence of ABD-A in Figure 3). Predominant repetitions are polyglutamine (poly-Q), poly-alanine (poly-A) and serine-rich regions (S-rich). These repetitive regions seem to include most of the indels and amino acid differences, and therefore they might be responsible for the surprisingly high evolutionary rate of *Hox* and *Hox*-derived proteins.

To test this hypothesis, we repeated the analyses of amino acid differences and indels inside and outside these repetitive regions (see Methods), and compared these two kinds of sequences (repetitive and unique). In the case of amino acid differences (Table 2), the percentage of aligned, non-conserved amino acids is higher in repetitive regions than in unique sequence in all the three groups. The T-test for paired samples (unique versus repetitive) on proteins having both types of regions showed significant differences between unique and repetitive sequences (P = 0.001), the mean of repetitive sequences being more than twice that for unique sequences (51.01% versus 23.19%, respectively). Despite this higher percentage of amino acid differences in repetitive than in unique sequence, the three groups of genes behave in a similar manner in both types of regions (note that the ranking is the same in both unique and repetitive regions).

Finally, we wanted to determine whether or not repetitive regions accumulate a larger number of indels than unique sequence (Table 3). The results show that in all the three groups, the percentage of indels in repetitive regions is much higher than that in unique sequence. These differences are significant (P = 0.006) according to a T-test for paired samples, giving an average value of 42.32% in repetitive regions versus 15.53% in unique sequence. Nevertheless, the ANOVA computed after removing repetitive regions remained highly significant (P = 0.003). Thus repetitive regions are not entirely responsible for the high percentage of indels in *Hox* and *Hox*-derived proteins. Therefore, *Hox* and *Hox*-derived genes have a tendency to

Table 2: Percentage of amino acid differences in the alignment (± SD) in the three groups of proteins.

	TOTAL	UNIQUE	REPETITIVE	T-test§
Hox	22.80 ± *10.44*	18.22 ± *10.50*	37.11 ± *12.33*	***
***Hox*-derived**	40.43 ± *18.26*	39.00 ± *19.64*	62.97 ± *24.08*	
Non-*Hox*	23.77 ± *10.81*	23.38 ± *10.93*	55.46 ± *31.35*	
ANOVA	n.s.	n.s.	n.s.	

n.s. (P > 0.05), * (P < 0.05), ** (P < 0.01), *** (P < 0.001).

§ T-test for paired samples (unique *vs*. repetitive) on proteins having both types of regions [ABD-A, LAB, PB, BCD, ZEN, Ccp84Ac, CG13617, CG14290 and LAP (product of *CG2520*)].

accumulate indels even outside of repetitive regions, which does not seem to be allowed in non-*Hox* genes.

Discussion

Evolutionary rates of Hox *genes*

This study shows that *Hox* genes seem to be evolving differently from other essential genes expressed in early development, with complex expression patterns or with long introns rich in *cis*-regulatory elements. Both the number of nonsynonymous substitutions and the degree of functional constraint are not significantly different between *Hox* and non-*Hox* genes, and this remains true even when the most peculiar regions (the homeobox and the repetitive regions) are excluded (Table 1). Therefore, *Hox* genes do not seem to be evolving more slowly than other non-homeotic genes, despite their essential function in the early development and even though their interchangeability between species has been proven to be functional in some cases [22-26].

Differences in the evolutionary rate among the three groups of genes (*Hox*, *Hox*-derived and non-*Hox*) could be mediated by some properties of genes that are correlated with the number of nucleotide substitutions (t). One possibility is that *Hox* and *Hox*-derived genes experience similar background rates of mutation that are different from those of non-*Hox* genes. We can use the number of synonymous substitutions per synonymous site (d_S) as a measure of the mutation rate of a gene. This variable is not significantly different among the three groups of genes (P = 0.530), and thus we can consider that mutation rate is constant across groups [see Additional file 2]. Another possibility is that genes within a group may have correlated levels of synonymous codon bias. Given that genes with higher codon bias tend to evolve more slowly [28,43], codon bias may contribute to spurious differences in the rates of protein evolution among groups. We have measured codon bias for each gene using the Effective Number of Codons, N_C [44]. There are no significant differences in the codon bias among groups, and the average N_C value for non-*Hox* genes is the lowest among the three groups (the highest codon bias) [see Additional files 1 and 2].

Some *Hox* and *Hox*-derived genes considered here have been included in previous studies [29,41]. Davis *et al.* [29] showed that the strongest negative relationship between expression profile and evolutionary rate occurs at a late stage in embryonic development, soon after the main burst of expression of segment polarity and *Hox* genes. However, they also show that the most constrained transcription factors and signal transducers, the functional class that contains many developmentally essential genes, are expressed precisely at the same time as the segment polarity and *Hox* genes. One of the two *Hox* genes included in their study has also been analyzed here (*abd-A*), and it is incidentally the gene with the lowest number of nonsynonymous substitutions and the one that is most constrained in our sample of *Hox* genes. On the other hand, *bcd*, although being one of the first genes acting in *Drosophila* development, was reported in the same study

Table 3: Percentage of indels in the alignment (± SD) in the three groups of proteins.

	TOTAL	UNIQUE	REPETITIVE	T-test§
Hox	25.77 ± *4.31*	16.21 ± *8.40*	44.82 ± *2.38*	**
***Hox*-derived**	37.53 ± *9.63*	34.88 ± *12.40*	75.64 ± *34.45*	
Non-*Hox*	8.73 ± *10.24*	8.46 ± *10.28*	23.79 ± *25.66*	
ANOVA	***	**	n.s.	

n.s. (P > 0.05), * (P < 0.05), ** (P < 0.01), *** (P < 0.001).

§ T-test for paired samples (unique *vs*. repetitive) on proteins having both types of regions [ABD-A, LAB, PB, BCD, ZEN, Ccp84Ac, CG13617, CG14290 and LAP (product of *CG2520*)].

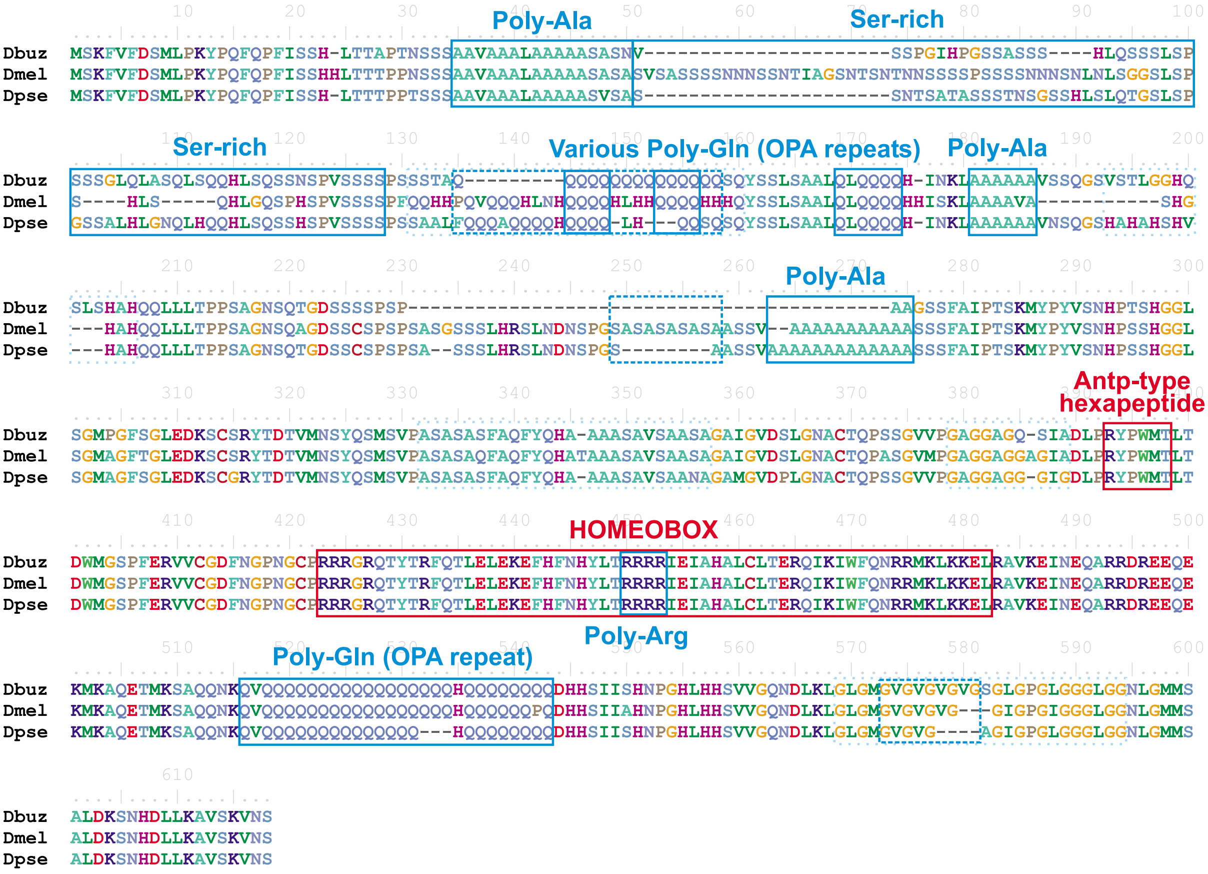


Figure 3

Alignment of a *Hox* protein (ABD-A) showing multiple long repeats spacing functional domains. Functional domains are represented by red boxes, and repeats by blue boxes as follows: repetitive regions annotated in UniProt are represented by solid boxes, simple repeats by dashed boxes and complex repeats by dotted light boxes (see Methods). Notation: Dbuz = *D. buzzatii*; Dmel = *D. melanogaster*; Dpse = *D. pseudoobscura*.

as an exceptional case of a gene acting in the earliest stages of development but evolving surprisingly fast [29].

Furthermore, *Hox* genes depart from a negative correlation found in previous studies between evolutionary rate at the protein level and intron size, number of conserved noncoding sequences within introns, or regulatory complexity [32]. In this respect, all *Hox* genes used in this study contain a total intron size >10 Kb [see Additional file 3], which corresponds to the longest intron size category used in [32]. Therefore, *Hox* genes are expected to evolve slowly as they contain long intronic sequences. Both *Hox*-derived and non-*Hox* genes contain shorter intron lengths than *Hox* genes [see Additional file 3], and thus would be expected to evolve faster.

Amino acid differences and indels

The percentages of amino acid differences and of indels in *Hox* proteins also depart from the initial expectations. While the percentage of amino acid differences is not significantly different among the three groups compared (Table 2), the percentages of indels in *Hox* and *Hox*-derived proteins are much higher than that in non-*Hox* proteins (Table 3). Therefore, *Hox* proteins are as divergent as non-*Hox* proteins in terms of amino acid changes, but they are much more divergent in terms of indels. A visual inspection of the alignments pointed out a possible explanation to these results (Figure 3). *Hox* and some *Hox*-derived proteins contain large repetitive regions, mostly homopeptides, present all along the protein except the region near the homeodomain and other highly conserved regions. It is within these repetitive regions where most indels and amino acid differences seem to accumulate, in some cases resulting in poor alignment, and therefore they could be responsible for the surprisingly high amino acid and indel evolution of *Hox* and *Hox*-derived proteins.

Although repetitive regions have been shown to be richer in amino acid differences and indels than unique sequence, they do not fully explain the high variation found in *Hox* and *Hox*-derived proteins. Even excluding repetitive regions, *Hox* and *Hox*-derived genes contain many more indels than non-*Hox* genes, although the percentage of amino acid substitutions is not significantly different between *Hox* and non-*Hox* genes. Therefore, taking amino acid differences and indels altogether we can state that the overall rate of evolution of *Hox* and *Hox*-derived genes is faster than that of non-*Hox* genes. The percentage of the alignment that has changed is 43.39% in *Hox* proteins, 64.73% in *Hox*-derived proteins and 30.97% in non-*Hox* proteins (the percentage of amino acid differences has been recalculated before being added to the percentage of indels to account for the total number of sites, both gapped and non-gapped, in order to make both percentages comparable). Finally, a lack of correlation between the proportion of indels and amino acid differences in the set of genes used in this study highlights the different evolutionary mechanisms that regulate both types of changes.

Homopeptides and other repetitions in Hox *and* Hox*-derived proteins*

Multiple long homopeptides are found in 7% of *Drosophila* proteins, most of which are essential developmental proteins expressed in the nervous system and involved in transcriptional regulation [34,45]. What is the role of these homopeptides? They could be tolerated, non-essential insertions that may play a role as transcriptional activity modulators. Some examples have been described in *Hox* and *Hox*-derived proteins [15] that illustrate the acquisition of new functions in the insect lineage while maintaining their homeotic role. In these examples, selection against coding changes might have been relaxed because of functional redundancy among *Hox* paralogs. These sequence differences could be involved in the functional divergence of *Hox* proteins and the evolutionary diversification of animals [15].

The large effects of *Hox* genes on morphology suggest that they regulate, directly or indirectly, a large number of genes. It would be expected that such pleiotropic proteins would be constrained in their sequence variation and, hence, their contribution to morphological variation. However, it has been shown that microsatellite sequences in developmental genes are a source of variation in natural populations, affecting visible traits by expanding or contracting at very high rates [46]. One intrinsic characteristic of microsatellites is their hypervariability, resulting from a balance between slippage events and point mutations [35,36]. Their mutation rate has been estimated to be 1.5×10^{-6} per locus per generation in the case of trinucleotide repeats in *D. melanogaster* [47], and is even greater in the case of dinucleotides. These values contrast with the general mutation rate of $\sim 10^{-8}$ per site per generation of base pair substitutions [48]. These repeats typically generate regions in the alignment with high variability in sequence and length, and that are difficult to align.

A potential role for homopeptides is to serve as spacer elements between functional domains, to provide flexibility to the three-dimensional conformation, and fine-tuning domain orientation of the protein in its interactions with DNA and other proteins. To that effect, changes in nucleotide distances between target binding sites might be accompanied by complementary changes in the sequences spacing the binding domains of transcription factors (mostly homopeptides). This would produce a coordinated evolution between transcription factors and their target binding sites. Excessive expansions of homopeptides, however, have often been associated with disease in humans [49-52]. Amazingly, essential developmental proteins like homeotic proteins that apparently need such homopeptides for their correct functioning have to suffer the consequences of their quick and apparently unpredictable evolution, and sacrifice in this way the conservation that would be expected in proteins of this type.

Among non-*Hox* genes, the cluster of cuticular genes (*Ccp84Ac*, *Ccp84Ae*, *Ccp84Af* and *Ccp84Ag*) behave similarly to *Hox* and *Hox*-derived genes and account for the vast majority of indels in their group (Figure 2). These short proteins share a conserved C-terminal section [53] and include a 35–36 amino acid motif known as the R&R consensus, present in many insect cuticle proteins, an extended form of which has been shown to bind chitin (chitin-bind 4; PF00379) [54]. Outside these conserved domains, cuticular proteins share hydrophobic regions dominated by tetrapeptide repeats (A-A-P-A/V), which are presumed to be functionally important [55,56] and are responsible for the high percentage of indels found in these proteins. These repeats are usually complex repeats that are not annotated in UniProt, nor detected as runs of identical amino acid repetitions (see Methods), and thus contribute to the percentage of indels in unique sequence in non-*Hox* genes (Table 3). When complex repeats were annotated and considered as repetitive sequence (see Methods), the percentage of indels in the unique portion of all classes of genes decreased substantially, but especially in non-*Hox* genes [see Additional file 4]. The elimination of complex repeats in cuticular genes was crucial in this reduction, and further increased the differences among groups.

Therefore, our results show that long repetitive sequences are not enough to explain all the differences found between *Hox* or *Hox*-derived genes and non-*Hox* genes. *Hox* and *Hox*-derived genes have a tendency to accumulate indels outside these repetitive regions that is not observed in non-*Hox* genes. We propose that spontaneous deletions between short repeated sequences could be the mechanism responsible for this difference [57]. Such deletions have been described in phages [58,59], *Escherichia coli* [60-65] and humans [66,67], and predominate between short sequence similarities of as few as 5–8 base pairs [68]. Two different models can explain the generation of spontaneous deletions: slipped mispairing during DNA synthesis, and recombination events mediated by enzymes that recognize these sequence similarities. In either case, the repetitive and compositionally biased nature of several regions within *Hox* and *Hox*-derived sequences might explain the major incidence of indels in these two groups. This would also explain the large differences in protein lengths among species that have been observed in some *Hox* proteins [8]. This higher probability of mutation would presumably be accompanied by a higher tolerance to indels of *Hox* and *Hox*-derived proteins outside their binding domains.

For a correct interpretation of our results, the set of non-*Hox* genes should be an unbiased sample of genes, both in terms of protein expression and structure. We have gathered this information from the literature, and verified that our non-*Hox* sample comprises a variable group of genes that are expressed through the fly life cycle (from young embryo to adult) and contains a wide variety of protein domains [see Additional file 5]. Therefore, we assume that, although small, it represents an unbiased sample of all non-*Hox* genes in the genome, and that results presented here are reliable.

The fate of **Hox***-derived genes after their origination by duplication*

The three *Hox*-derived genes used in this study (*bcd*, *zen* and *zen2*) originated from two consecutive duplications of the ancestral *Hox3* gene. Seemingly, *bcd* and *zen* have specialized and perform separate functions in the establishment of the embryo's body plan [11-13]. This is supported by our data, as these two genes have a moderate evolutionary rate but low level of functional constraint (high d_N/d_S rate ratio). However, the finding of Barker *et al*. [42] that genes with a maternal effect experience relaxed selective constraint resulting from sex-limited expression is not supported by our data. Our results show that *bcd* and *zen* are evolving at very similar rates in the *Drosophila* lineage, and *bcd* is even more constrained than *zen* [see Additional file 1].

The function of *zen2* is unclear. It has the same expression pattern as *zen* and, despite its high divergence across species, it has been maintained for more than 60 Myr [19]. Conservation of two paralogous genes maintaining the same function is unlikely, and could only be explained under some peculiar conditions (e.g. two strongly expressed genes whose products are in high demand [40]). It could be that this gene is experiencing a process of pseudogenization, supported by the fact that the evolutionary rate of *zen2* is more than twice that of *bcd* and *zen*, and that it has also the highest percentage of the alignment represented by indels. If so, we would expect to see a relaxation of the functional constraint. However, the relatively high level of functional constraint of *zen2* (ω = 0.09144) rather indicates a process of neofunctionalization, even though positive selection was not detected. The fact that this gene does not show an explicit pattern of variation of ω along its sequence (Figure 1) further supports the progressive loss of its original homeotic function and the acquisition of new functions.

Compared to the other two groups (*Hox* and non-*Hox* genes), *Hox*-derived genes are evolving significantly much faster and with less functional constraint. It is also the group with the highest proportion of amino acid differences and indels. These results reflect their relatively recent origin by duplication, which was followed by extensive changes in their role during the development of insects.

Conclusion

Many studies so far have largely focused on *Hox* gene homeobox sequences, and have demonstrated that they are highly conserved across species. However, *Hox* genes and in general all transcription factors share a particular structure where different highly conserved modules are interspersed with long repetitive regions, mostly microsatellites. Our results show that both *Hox* and *Hox*-derived genes have an overall high rate of evolution, especially in terms of indels. Moreover, although repetitive regions are richer in both amino acid differences and indels than the rest of the coding sequence, they do not seem to fully explain the differences in evolutionary rates found between *Hox* or *Hox*-derived genes and non-*Hox* genes. Therefore, by using complete gene sequences rather than their conserved modules, we observe that the *Hox* gene evolutionary rate is as high as that of non-*Hox* genes in terms of nucleotide evolution, and even higher in terms of indels. *Hox*-derived genes constitute the group with the highest evolutionary rate by all criteria. These results emphasize the need to take into account indels in addition to nucleotide substitutions in order to estimate evolutionary rates accurately. This study is the first quantification of the rates of nucleotide and indel evolution in these groups of genes, and shows that *Hox* and *Hox*-derived genes have a higher evolutionary dynamics than other developmental genes.

Methods

Genes analyzed and their classification

All the completely sequenced genes in *D. buzzatii* with a clear ortholog in *D. melanogaster* and *D. pseudoobscura* (23) were included in our analysis: *abd-A*, *lab*, *pb*, *bcd*, *zen*, *zen2*, *Dbuz\Ccp3* (ortholog of *Dmel\Ccp84Ac*), *Dbuz\Ccp6* (ortholog of *Dmel\Ccp84Ae*), *Dbuz\Ccp7* (ortholog of *Dmel\Ccp84Af*), *Dbuz\Ccp8* (ortholog of *Dmel\Ccp84Ag*), *CG1288*, *CG14290*, *CG14609*, *CG14899*, *CG17836*, *CG2520* and *CG31363* from Negre *et al*. [19]; *Adh-related* (*Adhr*) from Betran and Ashburner [69]; *α-Esterase-2* (*α-Est2*) and *α-Esterase-3* (*α-Est3*) from Robin *et al*. [70]; *CG13617* from Puig, Caceres, and Ruiz [71]; and *Larval serum protein 1 β* (*Lsp1β*) and *Lsp1γ* from Gonzalez, Casals and Ruiz [72]. Sequences of *D. melanogaster* orthologs were collected from Flybase [73,74], and those of *D. pseudoobscura* were annotated on the scaffolds from the whole genome shotgun sequencing project [75,76]. We identified the *D. pseudoobscura* orthologs by using the alignment of this species with the *D. melanogaster* genome generated by the Berkeley Genome Pipeline [77], and annotated the target sequences with the aid of ARTEMIS v. 7 [78] and BIOEDIT v. 7.0.4.1 [79]. A complete list of all genes, accession numbers (from Genbank or Flybase) and chromosomal locations is provided [see Additional file 3]. The longest transcript of each gene was used for the analyses. Genes were classified into three categories: 1) *Hox* genes (*abd-A*, *lab* and *pb*); 2) *Hox*-derived genes (*bcd*, *zen* and *zen2*); and 3) non-*Hox* genes (the remaining 17 genes). Results in each group were produced by calculating the average of all the genes within the group.

Sequence annotation and alignment

A set of Perl scripts, together with modules from PDA v. 1.4 [80] and BIOPERL v. 1.2.3 [81], were used to automatically check sequence annotations, extract the coding sequences (CDSs) of the selected transcripts and calculate basic gene structure and base composition parameters (gene and protein lengths; codon bias measured by the Effective Number of Codons (N_C); and G+C content in second, third and all codon positions) [see Additional file 1]. Differences among the three groups of genes were tested with one-way ANOVAs and pairwise contrast tests [82], assuming homogeneity of variances for those variables that gave non-significant P values for the Levene test [83] [see Additional file 2]. Orthologous coding sequences in *D. buzzatii*, *D. melanogaster* and *D. pseudoobscura* were aligned according to their translation to protein using RevTrans 1.3 Server [84] with some manual editing using BIOEDIT v. 7.0.4.1 [79]. Two non-*Hox* genes of the initial sample (*CG1288* and *CG17836*) showed a doubtful alignment, containing many gaps and few residue matches, and thus were excluded from the analyses to avoid unreliable estimates. A total of 15 non-*Hox* genes were therefore used in this study.

Estimation of evolutionary rates

The numbers of synonymous and nonsynonymous substitutions per site (d_S and d_N, respectively) were estimated on the nucleotide alignments of each gene using maximum likelihood methods with the program *codeml* of the PAML v. 3.14 package [85] [see Additional file 1]. We used an unrooted tree and the codon equilibrium frequencies (π_i) estimated from the nucleotide frequencies of the three codon sites (F3X4 option of *codeml*). Differences among the three groups of genes were tested using one-way ANOVAs and pairwise contrast tests as before. Furthermore, we visualized differences along the genes by plotting d_N and ω in sliding windows of 240 nucleotides and a step size of three nucleotides (one codon).

Measurement of amino acid differences and indels

We measured the proportion of amino acid differences and indels in the protein alignments (translated from the previous nucleotide alignments) using in-house Perl scripts. The methodology was based on measuring the number of non-conserved positions due to either amino acid differences (point changes) or indels (structural changes) in the protein multiple alignments (e.g. the minimum indel length is one amino acid, corresponding to three nucleotides in the nucleotide sequence). We can estimate in this way the percentage of the protein which has been changed in our set of species. We think that this is a simple (yet somewhat rough) measure to estimate the degree of constraint relaxation of proteins.

Specifically, the number of amino acid differences was computed as the number of non-gapped positions with non-identical amino acids in the three species. All percentages are given in relation to the total number of aligned amino acids (non-gapped positions). Similarly, the number of indels was computed as the number of different indels (gaps affecting different positions) in the complete alignment (gapped and non-gapped sites). Therefore, an indel shared by two species was considered a single indel, while overlapping gaps were considered separately. Indel lengths were taken into account to calculate the percentage of the alignment affected by indels. In this case, all percentages are given in relation to the total length of the alignment (gapped and non-gapped positions).

We used one-way ANOVAs to test for differences between *Hox*, *Hox*-derived and non-*Hox* proteins in both parameters: the proportion of amino acid differences and the proportion of indels. We also used the Pearson correlation coefficient to test for a correlation between the two measures (e.g. to test whether proteins with a high proportion of amino acid differences also have a high proportion of indels), and a nested ANOVA [82] to test for differences in indel length among the three groups, taking into account the variation within groups.

Contribution of the homeobox and the repetitive regions to the evolutionary rates

In order to test the effect of the homeobox and the repetitive regions in our estimates of nucleotide substitutions, we repeated the previous analyses excluding one or both types of sequence. Repetitive regions were identified in three different ways. First, we searched in the UniProt Knowledgebase Release 8.6 (Swiss-Prot Release 50.6 + TrEMBL Release 33.6) [86] for annotated compositionally biased regions (defined in the feature table as COMPBIAS) in the protein sequences encoded by *Hox*, *Hox*-derived and non-*Hox* genes [see Additional file 3]. In the case of *Hox* genes, all three genes in the group contained at least one annotated repetitive region, while for *Hox*-derived and non-*Hox* genes only one entry of each group (*bcd* and *CG2520*, respectively) contained annotated repetitive regions. Note that only repeats in *D. melanogaster* are identified by using this methodology. Second, we identified simple repeats as those runs of 5 or more identical amino acids (e.g. QQQQQ), or at least 4 identical repetitions of 2 or more amino acids (e.g. GVGVGVGV), in any of the three species. By using this second approach, we extended the number of proteins with repetitive sequences in both the *Hox*-derived and non-*Hox* groups. Finally, we tried to visually annotate complex repeats as those imperfect runs of amino acid repetitions or compositionally biased regions in the protein (e.g. regions in the protein with a high content of Q, S, A, P, H, G, V, etc.). Data was analyzed using a combination of the three approaches as follows: (1) using UniProt only; (2) using UniProt + Simple repeats; and (3) using UniProt + Simple repeats + Complex repeats. Because the identification of complex repeats is somewhat subjective, we present in the main text the results obtained by identifying repeats using the second combination (UniProt + Simple repeats). However, results do not differ significantly among the three combinations [see Additional file 4].

We also calculated the proportion of amino acid differences and indels in repetitive and non-repetitive (unique) sequence in the three groups, and tested for differences between these two types of regions using a T-test for paired samples [82] on those proteins having both types of regions.

Abbreviations

t = number of nucleotide substitutions per codon; d_S = number of synonymous substitutions per synonymous site; d_N = number of nonsynonymous substitutions per nonsynonymous site; $\omega = d_N/d_S$ ratio that measures the level of functional constraint; κ = transition/transversion rate ratio; N_C = Effective Number of Codons.

Authors' contributions

SC carried out the analyses and drafted the manuscript. BN participated in obtaining the data and in the design of the analyses. AB participated in the statistical analysis. AR conceived the study, and participated in its design and coordination. All authors read and approved the final manuscript.

Additional material

Additional File 1

Parameters of gene structure, base composition and nucleotide evolution for each gene.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-6-106-S1.pdf]

Additional File 2

ANOVA and contrast analyses for all group comparisons.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-6-106-S2.pdf]

Additional File 3

Genes from D. buzzatii, D. melanogaster *and* D. pseudoobscura *used in the analyses with their accession number in Genbank or Flybase and their location on the chromosome.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-6-106-S3.pdf]

Additional File 4

Set of tables of the main text, obtained according to three different annotation criteria to define repetitive sequences.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-6-106-S4.pdf]

Additional File 5

*Structure and expression of non-*Hox *proteins.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-6-106-S5.pdf]

Acknowledgements

The authors would like to thank Natalia Petit for helpful discussions, and Kristen Panfilio and two anonymous referees for valuable comments on this manuscript. This work has been supported by grant BMC2002-01708 from the Dirección General de Enseñanza Superior e Investigación Científica (MEC, Spain) awarded to AR, a doctoral FPI fellowship from the Ministerio de Ciencia y Tecnología (BES-2003-0416) awarded to SC and a doctoral FI/DGR fellowship from the Generalitat de Catalunya awarded to BN.

References

1. Carroll SB, Grenier JK, Weatherbee SD: **From DNA to diversity: Molecular Genetics and Evolution of Animal Design.** 2nd ed. edition. Blackwell; 2005.
2. Lewis EB: **A gene complex controlling segmentation in Drosophila.** *Nature* 1978, **276(5688)**:565-570.
3. Kaufman TC, Lewis R, Wakimoto B: **Cytogenetic Analysis of Chromosome 3 in Drosophila melanogaster: the Homeotic Gene Complex in Polytene Chromosome Interval 84A-B.** *Genetics* 1980, **94(1)**:115-133.
4. Cook CE, Smith ML, Telford MJ, Bastianello A, Akam M: **Hox genes and the phylogeny of the arthropods.** *Curr Biol* 2001, **11(10)**:759-763.
5. Hughes CL, Kaufman TC: **Hox genes and the evolution of the arthropod body plan.** *Evol Dev* 2002, **4(6)**:459-499.
6. Hughes CL, Liu PZ, Kaufman TC: **Expression patterns of the rogue Hox genes Hox3/zen and fushi tarazu in the apterygote insect Thermobia domestica.** *Evol Dev* 2004, **6(6)**:393-401.
7. Lewis EB, Pfeiffer BD, Mathog DR, Celniker SE: **Evolution of the homeobox complex in the Diptera.** *Curr Biol* 2003, **13(15)**:R587-8.
8. Negre B, Ranz JM, Casals F, Caceres M, Ruiz A: **A new split of the Hox gene complex in Drosophila: relocation and evolution of the gene labial.** *Mol Biol Evol* 2003, **20(12)**:2042-2054.
9. Von Allmen G, Hogga I, Spierer A, Karch F, Bender W, Gyurkovics H, Lewis E: **Splits in fruitfly Hox gene complexes.** *Nature* 1996, **380(6570)**:116.
10. Negre B, Ruiz A: **HOM-C evolution in Drosophila: is there a need for Hox gene clustering?** *Trends Genet* 2007 in press.
11. Bonneton F: **[Extreme divergence of a homeotic gene: the bicoid case].** *Med Sci (Paris)* 2003, **19**:1265-1270.
12. Stauber M, Jackle H, Schmidt-Ott U: **The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene.** *Proc Natl Acad Sci U S A* 1999, **96(7)**:3786-3789.
13. Stauber M, Prell A, Schmidt-Ott U: **A single Hox3 gene with composite bicoid and zerknullt expression characteristics in non-Cyclorrhaphan flies.** *Proc Natl Acad Sci U S A* 2002, **99(1)**:274-279.
14. Telford MJ: **Evidence for the derivation of the Drosophila fushi tarazu gene from a Hox gene orthologous to lophotrochozoan Lox5.** *Curr Biol* 2000, **10(6)**:349-352.
15. Galant R, Carroll SB: **Evolution of a transcriptional repression domain in an insect Hox protein.** *Nature* 2002, **415(6874)**:910-913.
16. Shiga Y, Yasumoto R, Yamagata H, Hayashi S: **Evolving role of Antennapedia protein in arthropod limb patterning.** *Development* 2002, **129(15)**:3555-3561.
17. Lohr U, Yussa M, Pick L: **Drosophila fushi tarazu. a gene on the border of homeotic function.** *Curr Biol* 2001, **11(18)**:1403-1412.
18. Akam M, Averof M, Castelli-Gair J, Dawes R, Falciani F, Ferrier D: **The evolving role of Hox genes in arthropods.** *Dev Suppl* 1994:209-215.
19. Negre B, Casillas S, Suzanne M, Sanchez-Herrero E, Akam M, Nefedov M, Barbadilla A, de Jong P, Ruiz A: **Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex.** *Genome Res* 2005, **15(5)**:692-700.
20. Berleth T, Burri M, Thoma G, Bopp D, Richstein S, Frigerio G, Noll M, Nusslein-Volhard C: **The role of localization of bicoid RNA in organizing the anterior pattern of the Drosophila embryo.** *Embo J* 1988, **7(6)**:1749-1756.
21. Rushlow C, Doyle H, Hoey T, Levine M: **Molecular characterization of the zerknullt region of the Antennapedia gene complex in Drosophila.** *Genes Dev* 1987, **1(10)**:1268-1279.
22. McGinnis N, Kuziora MA, McGinnis W: **Human Hox-4.2 and Drosophila deformed encode similar regulatory specificities in Drosophila embryos and larvae.** *Cell* 1990, **63(5)**:969-976.
23. Zhao JJ, Lazzarini RA, Pick L: **The mouse Hox-1.3 gene is functionally equivalent to the Drosophila Sex combs reduced gene.** *Genes Dev* 1993, **7(3)**:343-354.
24. Bachiller D, Macias A, Duboule D, Morata G: **Conservation of a functional hierarchy between mammalian and insect Hox/HOM genes.** *Embo J* 1994, **13(8)**:1930-1941.
25. Zakany J, Gerard M, Favier B, Potter SS, Duboule D: **Functional equivalence and rescue among group 11 Hox gene products in vertebral patterning.** *Dev Biol* 1996, **176(2)**:325-328.
26. Greer JM, Puetz J, Thomas KR, Capecchi MR: **Maintenance of functional equivalence during paralogous Hox gene evolution.** *Nature* 2000, **403(6770)**:661-665.
27. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH: **Why highly expressed proteins evolve slowly.** *Proc Natl Acad Sci U S A* 2005, **102(40)**:14338-14343.
28. Pal C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158(2)**:927-931.
29. Davis JC, Brandman O, Petrov DA: **Protein evolution in the context of Drosophila development.** *J Mol Evol* 2005, **60(6)**:774-785.
30. Powell JR, Caccone A, Gleason JM, Nigro L: **Rates of DNA evolution in Drosophila depend on function and developmental stage of expression.** *Genetics* 1993, **133(2)**:291-298.
31. Riedl R: **Order in living organisms: A systems analysis of evolution.** New York , Wiley; 1978.
32. Marais G, Nouvellet P, Keightley PD, Charlesworth B: **Intron size and exon evolution in Drosophila.** *Genetics* 2005, **170(1)**:481-485.
33. Averof M: **Arthropod Hox genes: insights on the evolutionary forces that shape gene functions.** *Curr Opin Genet Dev* 2002, **12(4)**:386-392.
34. Karlin S, Burge C: **Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development.** *Proc Natl Acad Sci U S A* 1996, **93(4)**:1560-1565.
35. Ellegren H: **Microsatellites: simple sequences with complex evolution.** *Nat Rev Genet* 2004, **5(6)**:435-445.
36. Kruglyak S, Durrett RT, Schug MD, Aquadro CF: **Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations.** *Proc Natl Acad Sci U S A* 1998, **95(18)**:10774-10778.
37. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154(1)**:459-473.
38. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290(5494)**:1151-1155.
39. Long M, Betran E, Thornton K, Wang W: **The origin of new genes: glimpses from the young and old.** *Nat Rev Genet* 2003, **4(11)**:865-875.
40. Zhang J: **Evolution by gene duplication: an update.** *Trends Ecol Evol* 2003, **18(6)**:292-298.

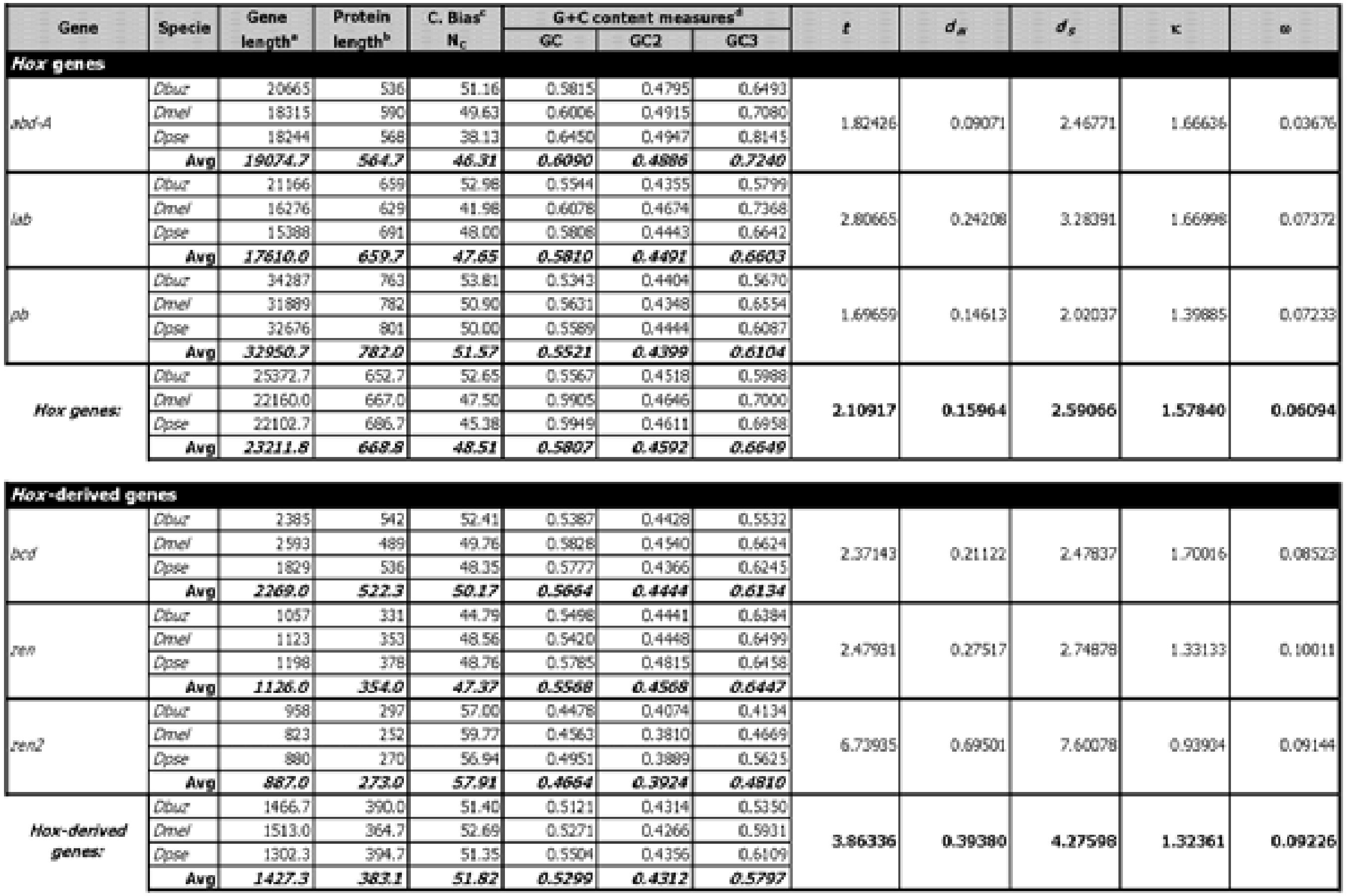
41. Baines JF, Chen Y, Das A, Stephan W: **DNA sequence variation at a duplicated gene: excess of replacement polymorphism and extensive haplotype structure in the Drosophila melanogaster bicoid region.** *Mol Biol Evol* 2002, **19(7):**989-998.
42. Barker MS, Demuth JP, Wade MJ: **Maternal Expression Relaxes Constraint on Innovation of the Anterior Determinant, bicoid.** *PLoS Genet* 2005, **1(5):**e57.
43. Akashi H: **Gene expression and molecular evolution.** *Curr Opin Genet Dev* 2001, **11(6):**660-666.
44. Wright F: **The 'effective number of codons' used in a gene.** *Gene* 1990, **87(1):**23-29.
45. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ: **Amino acid runs in eukaryotic proteomes and disease associations.** *Proc Natl Acad Sci U S A* 2002, **99(1):**333-338.
46. Fondon JW 3rd, Garner HR: **Molecular origins of rapid and continuous morphological evolution.** *Proc Natl Acad Sci U S A* 2004, **101(52):**18058-18063.
47. Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TF, Aquadro CF: **The mutation rates of di-, tri- and tetranucleotide repeats in Drosophila melanogaster.** *Mol Biol Evol* 1998, **15(12):**1751-1760.
48. Li WH: **Molecular Evolution.** Sunderland Massachusetts , Sinauer Associates, Inc.; 1997.
49. Hancock JM, Worthey EA, Santibanez-Koref MF: **A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice.** *Mol Biol Evol* 2001, **18(6):**1014-1023.
50. Karlin S, Chen C, Gentles AJ, Cleary M: **Associations between human disease genes and overlapping gene groups and multiple amino acid runs.** *Proc Natl Acad Sci U S A* 2002, **99(26):**17008-17013.
51. Albrecht A, Mundlos S: **The other trinucleotide repeat: polyalanine expansion disorders.** *Curr Opin Genet Dev* 2005, **15(3):**285-293.
52. Brown LY, Brown SA: **Alanine tracts: the expanding story of human illness and trinucleotide repeats.** *Trends Genet* 2004, **20(1):**51-58.
53. Rebers JE, Riddiford LM: **Structure and expression of a Manduca sexta larval cuticle gene homologous to Drosophila cuticle genes.** *J Mol Biol* 1988, **203(2):**411-423.
54. Rebers JE, Willis JH: **A conserved domain in arthropod cuticular proteins binds chitin.** *Insect Biochem Mol Biol* 2001, **31(11):**1083-1093.
55. Talbo G, Hojrup P, Rahbek-Nielsen H, Andersen SO, Roepstorff P: **Determination of the covalent structure of an N- and C-terminally blocked glycoprotein from endocuticle of Locusta migratoria. Combined use of plasma desorption mass spectrometry and Edman degradation to study post-translationally modified proteins.** *Eur J Biochem* 1991, **195(2):**495-504.
56. Andersen SO, Rafn K, Roepstorff P: **Sequence studies of proteins from larval and pupal cuticle of the yellow meal worm, Tenebrio molitor.** *Insect Biochem Mol Biol* 1997, **27(2):**121-131.
57. Albertini AM, Hofer M, Calos MP, Miller JH: **On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions.** *Cell* 1982, **29(2):**319-328.
58. Studier FW, Rosenberg AH, Simon MN, Dunn JJ: **Genetic and physical mapping in the early region of bacteriophage T7 DNA.** *J Mol Biol* 1979, **135(4):**917-937.
59. Pribnow D, Sigurdson DC, Gold L, Singer BS, Napoli C, Brosius J, Dull TJ, Noller HF: **rII cistrons of bacteriophage T4. DNA sequence around the intercistronic divide and positions of genetic landmarks.** *J Mol Biol* 1981, **149(3):**337-376.
60. Brake AJ, Fowler AV, Zabin I, Kania J, Muller-Hill B: **beta-Galactosidase chimeras: primary structure of a lac repressor-beta-galactosidase protein.** *Proc Natl Acad Sci U S A* 1978, **75(10):**4824-4827.
61. Fedoroff NV: **Deletion mutants of Xenopus laevis 5S ribosomal DNA.** *Cell* 1979, **16(3):**551-563.
62. Ghosal D, Saedler H: **IS2-61 and IS2-611 arise by illegitimate recombination from IS2-6.** *Mol Gen Genet* 1979, **176(2**233-238 [http://www.springerlink.com/content/u340577860368426/].
63. Post LE, Arfsten AE, Davis GR, Nomura M: **DNA sequence of the promoter region for the alpha ribosomal protein operon in Escherichia coli.** *J Biol Chem* 1980, **255(10):**4653-4659.
64. Ross DG, Swan J, Kleckner N: **Nearly precise excision: a new type of DNA alteration associated with the translocatable element Tn10.** *Cell* 1979, **16(4):**733-738.
65. Wu AM, Chapman AB, Platt T, Guarente LP, Beckwith J: **Deletions of distal sequence after termination of transcription at the end of the tryptophan operon in E. coli.** *Cell* 1980, **19(4):**829-836.
66. Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ: **The structure and evolution of the human beta-globin gene family.** *Cell* 1980, **21(3):**653-668.
67. Marotta CA, Wilson JT, Forget BG, Weissman SM: **Human beta-globin messenger RNA. III. Nucleotide sequences derived from complementary DNA.** *J Biol Chem* 1977, **252(14):**5040-5053.
68. Farabaugh PJ: **Sequence of the lacI gene.** *Nature* 1978, **274(5673):**765-769.
69. Betran E, Ashburner M: **Duplication, dicistronic transcription, and subsequent evolution of the Alcohol dehydrogenase and Alcohol dehydrogenase-related genes in Drosophila.** *Mol Biol Evol* 2000, **17(9):**1344-1352.
70. Robin GC, Russell RJ, Cutler DJ, Oakeshott JG: **The evolution of an alpha-esterase pseudogene inactivated in the Drosophila melanogaster lineage.** *Mol Biol Evol* 2000, **17(4):**563-575.
71. Puig M, Caceres M, Ruiz A: **Silencing of a gene adjacent to the breakpoint of a widespread Drosophila inversion by a transposon-induced antisense RNA.** *Proc Natl Acad Sci U S A* 2004, **101(24):**9013-9018.
72. Gonzalez J, Casals F, Ruiz A: **Duplicative and conservative transpositions of larval serum protein 1 genes in the genus Drosophila.** *Genetics* 2004, **168(1):**253-264.
73. Drysdale RA, Crosby MA, FlyBase Consortium: **FlyBase: genes and gene models.** *Nucleic Acids Res* 2005, **33:**D390-5.
74. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E, Hodgson A, George RA, Hoskins RA, Laverty T, Muzny DM, Nelson CR, Pacleb JM, Park S, Pfeiffer BD, Richards S, Sodergren EJ, Svirskas R, Tabor PE, Wan K, Stapleton M, Sutton GG, Venter C, Weinstock G, Scherer SE, Myers EW, Gibbs RA, Rubin GM: **Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence.** *Genome Biol* 2002, **3(12):**RESEARCH0079.
75. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, Couronne O, Hua S, Smith MA, Zhang P, Liu J, Bussemaker HJ, van Batenburg MF, Howells SL, Scherer SE, Sodergren E, Matthews BB, Crosby MA, Schroeder AJ, Ortiz-Barrientos D, Rives CM, Metzker ML, Muzny DM, Scott G, Steffen D, Wheeler DA, Worley KC, Havlak P, Durbin KJ, Egan A, Gill R, Hume J, Morgan MB, Miner G, Hamilton C, Huang Y, Waldron L, Verduzco D, Clerc-Blankenburg KP, Dubchak I, Noor MA, Anderson W, White KP, Clark AG, Schaeffer SW, Gelbart W, Weinstock GM, Gibbs RA: **Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution.** *Genome Res* 2005, **15(1):**1-18.
76. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2006, **34:**D16-20.
77. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I: **VISTA : visualizing global DNA sequence alignments of arbitrary length.** *Bioinformatics* 2000, **16(11):**1046-1047.
78. Berriman M, Rutherford K: **Viewing and annotating sequence data with Artemis.** *Brief Bioinform* 2003, **4(2):**124-132.
79. Hall TA: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucl Acids Symp Ser* 1999, **41:**95-98.
80. Casillas S, Barbadilla A: **PDA: a pipeline to explore and estimate polymorphism in large DNA databases.** *Nucleic Acids Res* 2004, **32(Web Server issue):**W166-9.
81. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12(10):**1611-1618.
82. Sokal RR, Rohlf FJ: **Biometry: The principles and practice of statistics in biological research.** New York , W.H. Freeman and Co.; 1995.

83. Levene H: **Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.** *Volume I*. Edited by: al. O. Stanford University Press; 1960:278-292.
84. Wernersson R, Pedersen AG: **RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences.** *Nucleic Acids Res* 2003, **31(13):**3537-3539.
85. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13(5**555-556 [http://bioinformatics.oxfordjournals.org/cgi/reprint/13/5/555].
86. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005, **33(Database issue):**D154-9.

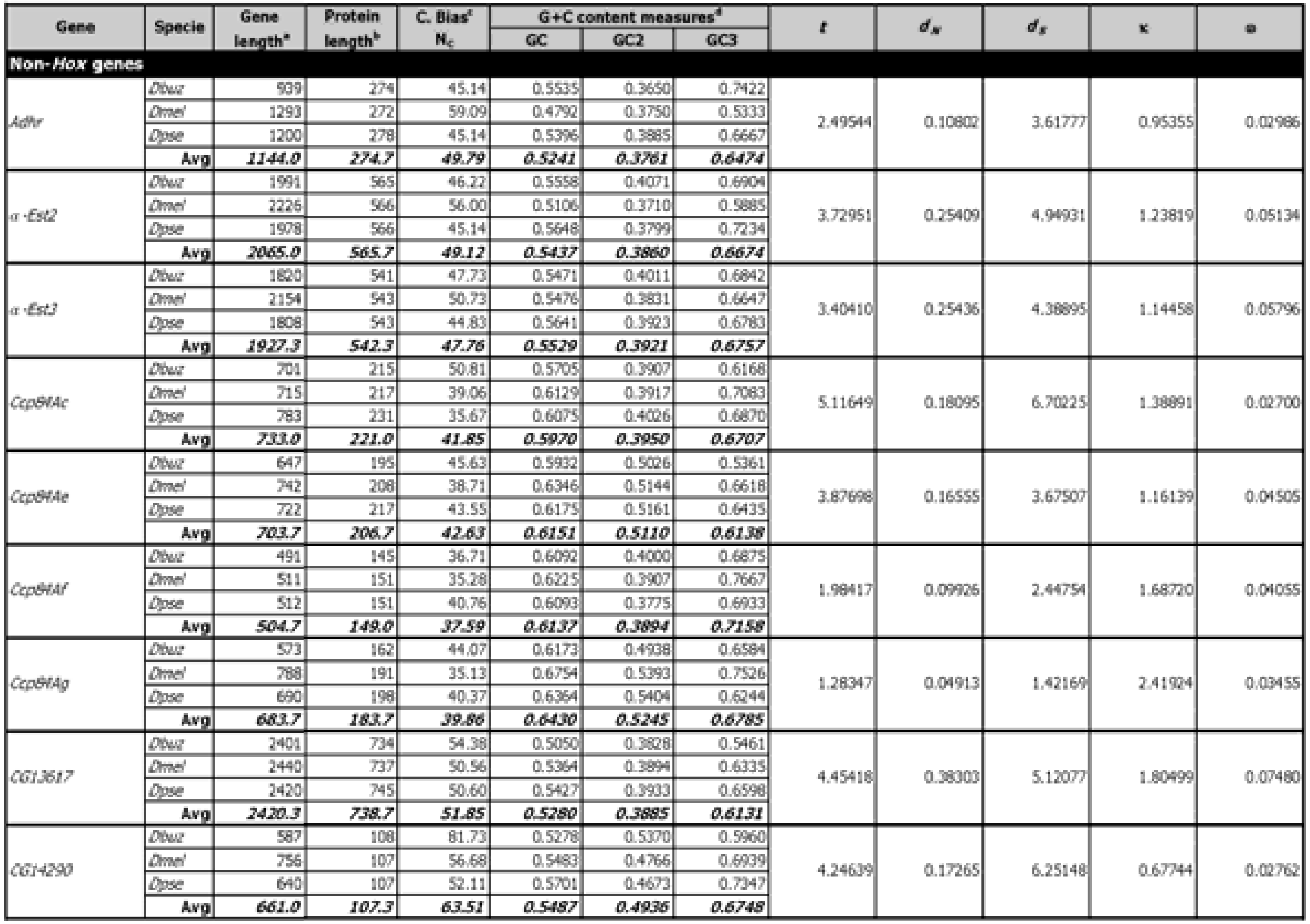
Additional Material

Additional File 1

Additional file 1. Parameters of gene structure, base composition and nucleotide evolution for each gene.



Gene	Specie	Gene length[a]	Protein length[b]	C. Bias[c]	G+C content measures[d]			*t*	d_N	d_S	κ	ω
				N_C	GC	GC2	GC3					
***Hox* genes**												
abd-A	*Dbuz*	20665	536	51.16	0.5815	0.4795	0.6493	1.82426	0.09071	2.46771	1.66636	0.03676
	Dmel	18315	590	49.63	0.6006	0.4915	0.7080					
	Dpse	18244	568	38.13	0.6450	0.4947	0.8145					
	Avg	***19074.7***	***564.7***	***46.31***	***0.6090***	***0.4886***	***0.7240***					
lab	*Dbuz*	21166	659	52.98	0.5544	0.4355	0.5799	2.80665	0.24208	3.28391	1.66998	0.07372
	Dmel	16276	629	41.98	0.6078	0.4674	0.7368					
	Dpse	15388	691	48.00	0.5808	0.4443	0.6642					
	Avg	***17610.0***	***659.7***	***47.65***	***0.5810***	***0.4491***	***0.6603***					
pb	*Dbuz*	34287	763	53.81	0.5343	0.4404	0.5670	1.69659	0.14613	2.02037	1.39885	0.07233
	Dmel	31889	782	50.90	0.5631	0.4348	0.6554					
	Dpse	32676	801	50.00	0.5589	0.4444	0.6087					
	Avg	***32950.7***	***782.0***	***51.57***	***0.5521***	***0.4399***	***0.6104***					
Hox genes:	*Dbuz*	25372.7	652.7	52.65	0.5567	0.4518	0.5988	**2.10917**	**0.15964**	**2.59066**	**1.57840**	**0.06094**
	Dmel	22160.0	667.0	47.50	0.5905	0.4646	0.7000					
	Dpse	22102.7	686.7	45.38	0.5949	0.4611	0.6958					
	Avg	***23211.8***	***668.8***	***48.51***	***0.5807***	***0.4592***	***0.6649***					
***Hox*-derived genes**												
bcd	*Dbuz*	2385	542	52.41	0.5387	0.4428	0.5532	2.37143	0.21122	2.47837	1.70016	0.08523
	Dmel	2593	489	49.76	0.5828	0.4540	0.6624					
	Dpse	1829	536	48.35	0.5777	0.4366	0.6245					
	Avg	***2269.0***	***522.3***	***50.17***	***0.5664***	***0.4444***	***0.6134***					
zen	*Dbuz*	1057	331	44.79	0.5498	0.4441	0.6384	2.47931	0.27517	2.74878	1.33133	0.10011
	Dmel	1123	353	48.56	0.5420	0.4448	0.6499					
	Dpse	1198	378	48.76	0.5785	0.4815	0.6458					
	Avg	***1126.0***	***354.0***	***47.37***	***0.5568***	***0.4568***	***0.6447***					
zen2	*Dbuz*	958	297	57.00	0.4478	0.4074	0.4134	6.73935	0.69501	7.60078	0.93904	0.09144
	Dmel	823	252	59.77	0.4563	0.3810	0.4669					
	Dpse	880	270	56.94	0.4951	0.3889	0.5625					
	Avg	***887.0***	***273.0***	***57.91***	***0.4664***	***0.3924***	***0.4810***					
Hox-derived genes:	*Dbuz*	1466.7	390.0	51.40	0.5121	0.4314	0.5350	**3.86336**	**0.39380**	**4.27598**	**1.32361**	**0.09226**
	Dmel	1513.0	364.7	52.69	0.5271	0.4266	0.5931					
	Dpse	1302.3	394.7	51.35	0.5504	0.4356	0.6109					
	Avg	***1427.3***	***383.1***	***51.82***	***0.5299***	***0.4312***	***0.5797***					



Gene	Specie	Gene length[a]	Protein length[b]	C. Bias[c]	G+C content measures[d]			*t*	d_N	d_S	κ	ω
				N_c	GC	GC2	GC3					
Non-*Hox* genes												
Adhr	*Dbuz*	939	274	45.14	0.5535	0.3650	0.7422	2.49544	0.10802	3.61777	0.95355	0.02986
	Dmel	1293	272	59.09	0.4792	0.3750	0.5333					
	Dpse	1200	278	45.14	0.5396	0.3885	0.6667					
	Avg	***1144.0***	***274.7***	***49.79***	***0.5241***	***0.3761***	***0.6474***					
α-Est2	*Dbuz*	1991	565	46.22	0.5558	0.4071	0.6904	3.72951	0.25409	4.94931	1.23819	0.05134
	Dmel	2226	566	56.00	0.5106	0.3710	0.5885					
	Dpse	1978	566	45.14	0.5648	0.3799	0.7234					
	Avg	***2065.0***	***565.7***	***49.12***	***0.5437***	***0.3860***	***0.6674***					
α-Est3	*Dbuz*	1820	541	47.73	0.5471	0.4011	0.6842	3.40410	0.25436	4.38895	1.14458	0.05796
	Dmel	2154	543	50.73	0.5476	0.3831	0.6647					
	Dpse	1808	543	44.83	0.5641	0.3923	0.6783					
	Avg	***1927.3***	***542.3***	***47.76***	***0.5529***	***0.3921***	***0.6757***					
Ccp84Ac	*Dbuz*	701	215	50.81	0.5705	0.3907	0.6168	5.11649	0.18095	6.70225	1.38891	0.02700
	Dmel	715	217	39.06	0.6129	0.3917	0.7083					
	Dpse	783	231	35.67	0.6075	0.4026	0.6870					
	Avg	***733.0***	***221.0***	***41.85***	***0.5970***	***0.3950***	***0.6707***					
Ccp84Ae	*Dbuz*	647	195	45.63	0.5932	0.5026	0.5361	3.87698	0.16555	3.67507	1.16139	0.04505
	Dmel	742	208	38.71	0.6346	0.5144	0.6618					
	Dpse	722	217	43.55	0.6175	0.5161	0.6435					
	Avg	***703.7***	***206.7***	***42.63***	***0.6151***	***0.5110***	***0.6138***					
Ccp84Af	*Dbuz*	491	145	36.71	0.6092	0.4000	0.6875	1.98417	0.09926	2.44754	1.68720	0.04055
	Dmel	511	151	35.28	0.6225	0.3907	0.7667					
	Dpse	512	151	40.76	0.6093	0.3775	0.6933					
	Avg	***504.7***	***149.0***	***37.59***	***0.6137***	***0.3894***	***0.7158***					
Ccp84Ag	*Dbuz*	573	162	44.07	0.6173	0.4938	0.6584	1.28347	0.04913	1.42169	2.41924	0.03455
	Dmel	788	191	35.13	0.6754	0.5393	0.7526					
	Dpse	690	198	40.37	0.6364	0.5404	0.6244					
	Avg	***683.7***	***183.7***	***39.86***	***0.6430***	***0.5245***	***0.6785***					
CG13617	*Dbuz*	2401	734	54.38	0.5050	0.3828	0.5461	4.45418	0.38303	5.12077	1.80499	0.07480
	Dmel	2440	737	50.56	0.5364	0.3894	0.6335					
	Dpse	2420	745	50.60	0.5427	0.3933	0.6598					
	Avg	***2420.3***	***738.7***	***51.85***	***0.5280***	***0.3885***	***0.6131***					
CG14290	*Dbuz*	587	108	81.73	0.5278	0.5370	0.5960	4.24639	0.17265	6.25148	0.67744	0.02762
	Dmel	756	107	56.68	0.5483	0.4766	0.6939					
	Dpse	640	107	52.11	0.5701	0.4673	0.7347					
	Avg	***661.0***	***107.3***	***63.51***	***0.5487***	***0.4936***	***0.6748***					

Gene	Specie	Gene length[a]	Protein length[b]	C. Bias[c] N_C	G+C content measures[d]			t	d_N	d_S	κ	ω
					GC	GC2	GC3					
CG14609	*Dbuz*	2603	599	51.51	0.4802	0.2838	0.5930	5.31870	0.21768	7.00844	1.87039	0.03106
	Dmel	2191	597	55.15	0.4662	0.2714	0.5986					
	Dpse	2126	594	51.29	0.4703	0.2677	0.5904					
	Avg	***2306.7***	***596.7***	***52.65***	***0.4722***	***0.2743***	***0.5940***					
CG14899	*Dbuz*	1023	258	52.22	0.4974	0.3721	0.5667	2.17723	0.08058	2.54854	1.71101	0.03162
	Dmel	923	261	46.42	0.5504	0.3831	0.6911					
	Dpse	982	258	46.08	0.5413	0.3915	0.6502					
	Avg	***976.0***	***259.0***	***48.24***	***0.5297***	***0.3822***	***0.6360***					
CG2520	*Dbuz*	14173	477	50.22	0.5059	0.4465	0.5098	0.91112	0.02653	1.04662	2.03996	0.02535
	Dmel	11011	468	56.50	0.5228	0.4423	0.5588					
	Dpse	11681	473	56.62	0.5159	0.4440	0.5439					
	Avg	***12288.3***	***472.7***	***54.45***	***0.5149***	***0.4443***	***0.5375***					
CG31363	*Dbuz*	8491	203	57.49	0.4663	0.4877	0.4192	1.95356	0.15912	2.32116	1.30436	0.06855
	Dmel	13033	208	41.71	0.5833	0.4904	0.7376					
	Dpse	12208	227	45.20	0.5786	0.5066	0.7227					
	Avg	***11244.0***	***212.7***	***48.13***	***0.5427***	***0.4949***	***0.6265***					
Lsp1 β	*Dbuz*	2427	788	39.80	0.4820	0.2605	0.7236	1.18839	0.09959	2.69820	1.62088	0.03691
	Dmel	2435	789	27.24	0.5560	0.2763	0.9260					
	Dpse	2432	787	28.45	0.5417	0.2630	0.8873					
	Avg	***2431.3***	***788.0***	***31.83***	***0.5266***	***0.2666***	***0.8456***					
Lsp1 γ	*Dbuz*	2376	773	38.94	0.4981	0.2717	0.7348	1.53426	0.11973	2.90247	1.41796	0.04125
	Dmel	2381	772	41.20	0.4965	0.2642	0.7480					
	Dpse	2383	773	29.41	0.5395	0.2600	0.8804					
	Avg	***2380.0***	***772.7***	***36.52***	***0.5113***	***0.2653***	***0.7877***					
Non-Hox genes:	*Dbuz*	2749.5	402.5	49.51	0.5340	0.4002	0.6203	**2.91160**	**0.15802**	**3.80668**	**1.49600**	**0.04156**
	Dmel	2906.6	405.8	45.96	0.5562	0.3973	0.6842					
	Dpse	2837.7	409.9	43.68	0.5626	0.3994	0.6924					
	Avg	***2831.3***	***406.0***	***46.38***	***0.5509***	***0.3989***	***0.6656***					
ALL CLASSES:	*Dbuz*	5798.1	436.4	50.23	0.5341	0.4120	0.6050	**2.96138**	**0.23715**	**3.55777**	**1.46600**	**0.06492**
	Dmel	5458.0	437.2	47.15	0.5569	0.4111	0.6735					
	Dpse	5370.5	447.2	45.02	0.5655	0.4134	0.6812					
	Avg	***5542.2***	***440.3***	***47.46***	***0.5522***	***0.4121***	***0.6532***					

[a] Gene length (in base pairs), excluding 5' and 3' UTRs (includes exons and introns only from the 'start' to the 'stop' codons in the CDS)

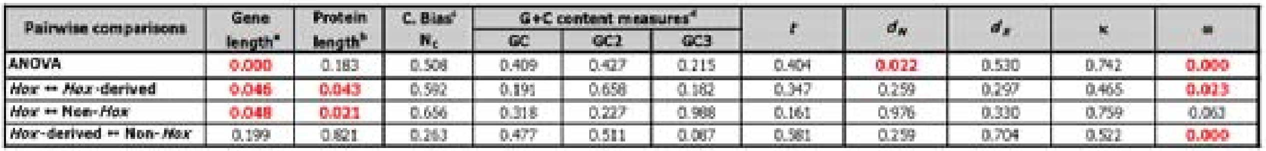
[b] Protein length (in amino acids)

[c] Codon Bias measure: *Effective Number of Codons* (Wright 1990)

[d] G+C content measures: percentage of G+C at all coding positions (GC), second coding positions (GC2) and third coding positions (GC3)

ADDITIONAL FILE 2

Additional file 2. ANOVA and contrast analyses for all group comparisons. Numbers refer to tests significances, and those in red are significant (P<0.05). Contrast tests assume homogeneity of variances for all variables except for "Gene length", "Protein length", "r" and "d_N" (which gave significant P values for the Levene test).



Pairwise comparisons	Gene length[a]	Protein length[b]	C. Bias[c]	G+C content measures[d]			r	d_N	d_S	κ	ω
			N_c	GC	GC2	GC3					
ANOVA	0.000	0.183	0.508	0.409	0.427	0.215	0.404	0.022	0.530	0.742	0.000
Hox ↔ *Hox*-derived	0.046	0.043	0.592	0.191	0.658	0.182	0.347	0.259	0.297	0.465	0.023
Hox ↔ Non-*Hox*	0.048	0.021	0.656	0.318	0.227	0.988	0.161	0.976	0.330	0.759	0.063
Hox-derived ↔ Non-*Hox*	0.199	0.821	0.263	0.477	0.511	0.087	0.381	0.259	0.704	0.522	0.000

[a] Gene length (in base pairs), excluding 5' and 3' UTRs (includes exons and introns only from the 'start' to the 'stop' codons in the CDS)

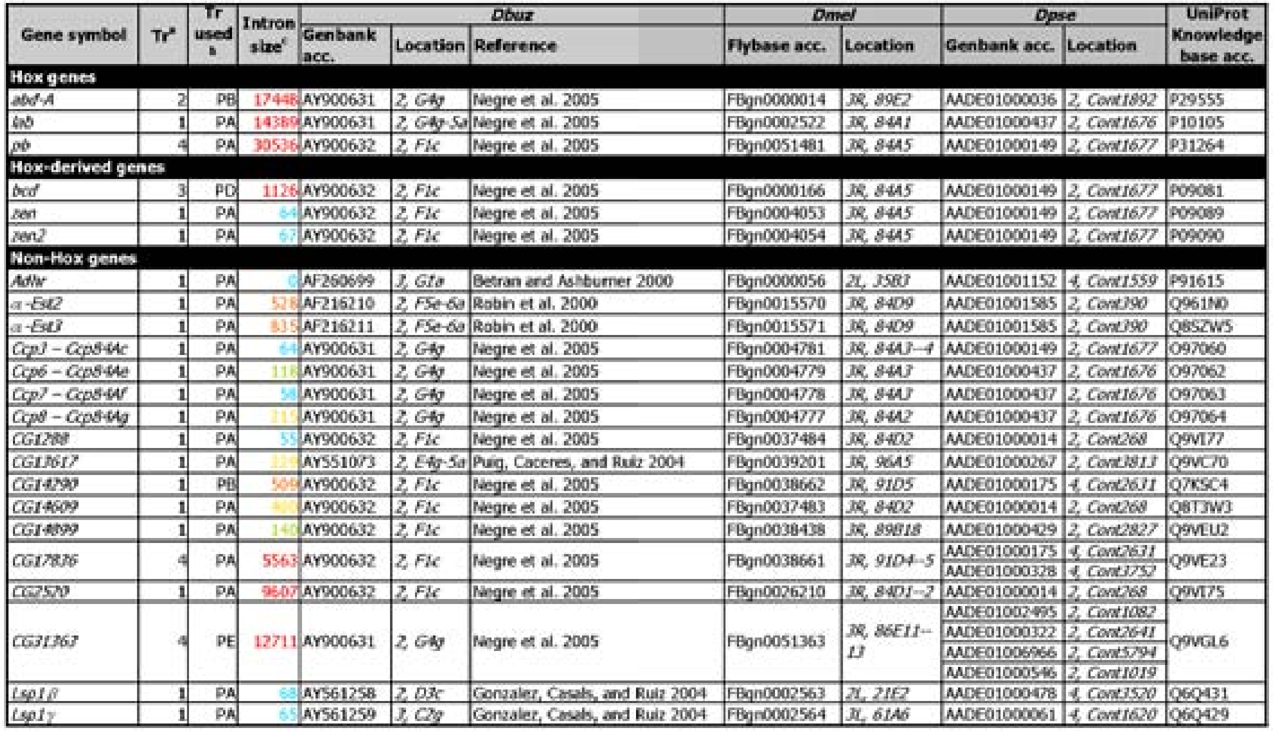
[b] Protein length (in amino acids)

[c] Codon Bias measure: *Effective Number of Codons* (Wright 1990)

[d] G+C content measures: percentage of G+C at all coding positions (GC), second coding positions (GC2) and third coding positions (GC3)

ADDITIONAL FILE 3

Additional file 3. Genes from *D. buzzatii* (*Dbuz*), *D. melanogaster* (*Dmel*) and *D. pseudoobscura* (*Dpse*) used in the analyses with their accession number in Genbank or Flybase and their location on the chromosome. In the case of *D. buzzatii*, the original references where the sequences were first published are also supplied. The last column contains the UniProt Knowledgebase accession numbers used for annotating repetitive blocks.



Gene symbol	Tr[a]	Tr used[b]	Intron size[c]	*Dbuz*			*Dmel*		*Dpse*		UniProt Knowledge base acc.
				Genbank acc.	Location	Reference	Flybase acc.	Location	Genbank acc.	Location	
Hox genes											
abd-A	2	PB	17448	AY900631	*2, G4g*	Negre et al. 2005	FBgn0000014	*3R, 89E2*	AADE01000036	*2, Cont1892*	P29555
lab	1	PA	14389	AY900631	*2, G4g-5a*	Negre et al. 2005	FBgn0002522	*3R, 84A1*	AADE01000437	*2, Cont1676*	P10105
pb	4	PA	30536	AY900632	*2, F1c*	Negre et al. 2005	FBgn0051481	*3R, 84A5*	AADE01000149	*2, Cont1677*	P31264
Hox-derived genes											
bcd	3	PD	1126	AY900632	*2, F1c*	Negre et al. 2005	FBgn0000166	*3R, 84A5*	AADE01000149	*2, Cont1677*	P09081
zen	1	PA	64	AY900632	*2, F1c*	Negre et al. 2005	FBgn0004053	*3R, 84A5*	AADE01000149	*2, Cont1677*	P09089
zen2	1	PA	67	AY900632	*2, F1c*	Negre et al. 2005	FBgn0004054	*3R, 84A5*	AADE01000149	*2, Cont1677*	P09090
Non-Hox genes											
Adhr	1	PA	0	AF260699	*3, G1a*	Betran and Ashburner 2000	FBgn0000056	*2L, 35B3*	AADE01001152	*4, Cont1559*	P91615
α-Est2	1	PA	528	AF216210	*2, F5e-6a*	Robin et al. 2000	FBgn0015570	*3R, 84D9*	AADE01001585	*2, Cont390*	Q961N0
α-Est3	1	PA	835	AF216211	*2, F5e-6a*	Robin et al. 2000	FBgn0015571	*3R, 84D9*	AADE01001585	*2, Cont390*	Q8SZW5
Ccp3 – Ccp84Ac	1	PA	64	AY900631	*2, G4g*	Negre et al. 2005	FBgn0004781	*3R, 84A3–4*	AADE01000149	*2, Cont1677*	O97060
Ccp6 – Ccp84Ae	1	PA	118	AY900631	*2, G4g*	Negre et al. 2005	FBgn0004779	*3R, 84A3*	AADE01000437	*2, Cont1676*	O97062
Ccp7 – Ccp84Af	1	PA	58	AY900631	*2, G4g*	Negre et al. 2005	FBgn0004778	*3R, 84A3*	AADE01000437	*2, Cont1676*	O97063
Ccp8 – Ccp84Ag	1	PA	315	AY900631	*2, G4g*	Negre et al. 2005	FBgn0004777	*3R, 84A2*	AADE01000437	*2, Cont1676*	O97064
CG1288	1	PA	55	AY900632	*2, F1c*	Negre et al. 2005	FBgn0037484	*3R, 84D2*	AADE01000014	*2, Cont268*	Q9VI77
CG13617	1	PA	329	AY551073	*2, E4g-5a*	Puig, Caceres, and Ruiz 2004	FBgn0039201	*3R, 96A5*	AADE01000267	*2, Cont3813*	Q9VC70
CG14290	1	PB	509	AY900632	*2, F1c*	Negre et al. 2005	FBgn0038662	*3R, 91D5*	AADE01000175	*4, Cont2631*	Q7KSC4
CG14609	1	PA	400	AY900632	*2, F1c*	Negre et al. 2005	FBgn0037483	*3R, 84D2*	AADE01000014	*2, Cont268*	Q8T3W3
CG14899	1	PA	140	AY900632	*2, F1c*	Negre et al. 2005	FBgn0038438	*3R, 89B18*	AADE01000429	*2, Cont2827*	Q9VEU2
CG17836	4	PA	5563	AY900632	*2, F1c*	Negre et al. 2005	FBgn0038661	*3R, 91D4–5*	AADE01000175 AADE01000328	*4, Cont2631* *4, Cont3752*	Q9VE23
CG2520	1	PA	9607	AY900632	*2, F1c*	Negre et al. 2005	FBgn0026210	*3R, 84D1–2*	AADE01000014	*2, Cont268*	Q9VI75
CG31363	4	PE	12711	AY900631	*2, G4g*	Negre et al. 2005	FBgn0051363	*3R, 86E11–13*	AADE01002495 AADE01000322 AADE01006966 AADE01000546	*2, Cont1082* *2, Cont2641* *2, Cont5794* *2, Cont1019*	Q9VGL6
Lsp1β	1	PA	68	AY561258	*2, D3c*	Gonzalez, Casals, and Ruiz 2004	FBgn0002563	*2L, 21E2*	AADE01000478	*4, Cont3520*	Q6Q431
Lsp1γ	1	PA	65	AY561259	*3, C2g*	Gonzalez, Casals, and Ruiz 2004	FBgn0002564	*3L, 61A6*	AADE01000061	*4, Cont1620*	Q6Q429

[a] Number of transcripts that exist in all the three species.

[b] The longest transcript shared by all the three species and that was used for the analyses.

[c] Total intron size in base pairs (sum of all introns in the gene, including those outside the limits of the coding sequence, in *D. melanogaster*).

<100 <200 <500 <1000 >1000

ADDITIONAL FILE 4

This file contains Tables 1, 2 and 3 of the main text, obtained according to three different annotation criteria to define repetitive sequences (see ***Methods*** in the main text):

- **Section 1 – Annotation criterion = Uniprot**
- **Section 2 – Annotation criterion = Uniprot + Simple repeats** (note that this is the annotation criterion selected for the main text; tables in this section are those shown in the main text)
- **Section 3 – Annotation criterion = Uniprot + Simple repeats + Complex repeats**

SECTION 1. ANNOTATION CRITERION = Uniprot

Table 1. Mean nucleotide substitution parameters, ANOVAs and contrast tests for the three groups of genes.

		t		d_N	d_S		ω		
Complete coding sequences	*Hox*	2.10917		0.15964	2.59066		0.06094	*	
	Hox-derived	3.86337		0.39380	4.27598		0.09226	***	
	Non-*Hox*	2.91160		0.15802	3.80668		0.04156		
	ANOVA	n.s.		*	n.s.		***		
Coding sequences excluding the homeobox	*Hox*	2.27653		0.18257	2.65921		0.06673	**	
	Hox-derived	5.04914		0.54809	5.26666		0.11320		*
	Non-*Hox*	2.91160		0.15802	3.80668		0.04156	***	
	ANOVA	n.s.		**	n.s.		***		
Coding sequences excluding repetitive regions	*Hox*	1.88711		0.13826	2.36833		0.05917	*	
	Hox-derived	3.80869	*	0.39291	4.21256		0.09423		
	Non-*Hox*	2.91375		0.15726	3.81781		0.04071	***	
	ANOVA	n.s.		*	n.s.		***		
Coding sequences excluding the homeobox and repetitive regions	*Hox*	2.01610		0.16274	2.35658		0.06782	**	
	Hox-derived	4.99467		0.54854	5.20047	*	0.11617		
	Non-*Hox*	2.91375		0.15726	3.81781		0.04071	***	
	ANOVA	n.s.		**	n.s.		***		

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001). For pairwise comparisons, only significant values are shown. Contrast tests assume homogeneity of variances where the Levene test does not give a significant P value.

Table 2. Percentage of amino acid differences in the alignment (±SD) in the three groups of proteins.

	TOTAL	**UNIQUE**	**REPETITIVE**	**T-test§**
Hox	22.80 ± *10.44*	19.88 ± *10.84*	33.94 ± *11.71*	**
***Hox*-derived**	40.43 ± *18.26*	40.17 ± *18.52*	33.93	
Non-*Hox*	23.77 ± *10.81*	23.66 ± *11.02*	13.51	
ANOVA	n.s.	n.s.	n.s.	

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001)
§ T-test for paired samples (unique *versus* repetitive) on proteins having both types of regions [ABD-A, LAB, PB, BCD and LAP (product of *CG2520*)].

Table 3. Percentage of indels in the alignment (±SD) in the three groups of proteins.

	TOTAL	**UNIQUE**	**REPETITIVE**	**T-test§**
Hox	25.77 ± *4.31*	21.61 ± *6.67*	35.97 ± *5.52*	*
***Hox*-derived**	37.53 ± *9.63*	35.36 ± *12.80*	60.61	
Non-*Hox*	8.73 ± *10.24*	8.50 ± *10.34*	20.88	
ANOVA	***	**	n.s.	

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001)
§ T-test for paired samples (unique *versus* repetitive) on proteins having both types of regions [ABD-A, LAB, PB, BCD and LAP (product of *CG2520*)].

SECTION 2. ANNOTATION CRITERION = Uniprot + Simple repeats

Table 1. Mean nucleotide substitution parameters, ANOVAs and contrast tests for the three groups of genes.

		t		d_N	d_S		*ω*		
Complete coding sequences	*Hox*	2.10917		0.15964	2.59066		0.06094	*	
	Hox-derived	3.86336		0.39380	4.27598		0.09226		
	Non-*Hox*	2.91160		0.15802	3.80668		0.04156	***	
	ANOVA	n.s.		*	n.s.		***		
Coding sequences excluding the homeobox	*Hox*	2.27653		0.18257	2.65921		0.06673	**	
	Hox-derived	5.04914		0.54809	5.26666		0.11320		*
	Non-*Hox*	2.91160		0.15802	3.80668		0.04156	***	
	ANOVA	n.s.		**	n.s.		***		
Coding sequences excluding repetitive regions	*Hox*	1.81997		0.12399	2.35029		0.05310	*	
	Hox-derived	3.71981	*	0.37759	4.14242	*	0.09042		
	Non-*Hox*	2.85593		0.15444	3.76458		0.04035	***	
	ANOVA	n.s.		*	n.s.		***		
Coding sequences excluding the homeobox and repetitive regions	*Hox*	1.94286		0.14684	2.33783		0.06146	**	
	Hox-derived	4.88928		0.53011	5.12014	**	0.11245	***	
	Non-*Hox*	2.85593		0.15444	3.76458		0.04035		
	ANOVA	n.s.		**	n.s.		***		

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001). For pairwise comparisons, only significant values are shown. Contrast tests assume homogeneity of variances where the Levene test does not give a significant P value.

Table 2. Percentage of amino acid differences in the alignment (±SD) in the three groups of proteins.

	TOTAL	**UNIQUE**	**REPETITIVE**	**T-test§**
Hox	22.80 ± *10.44*	18.22 ± *10.50*	37.11 ± *12.33*	
***Hox*-derived**	40.43 ± *18.26*	39.00 ± *19.64*	62.97 ± *24.08*	***
Non-*Hox*	23.77 ± *10.81*	23.38 ± *10.93*	55.46 ± *31.35*	
ANOVA	n.s.	n.s.	n.s.	

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001)

§ T-test for paired samples (unique *versus* repetitive) on proteins having both types of regions [ABD-A, LAB, PB, BCD, ZEN, Ccp84Ac, CG13617, CG14290 and LAP (product of *CG2520*)].

Table 3. Percentage of indels in the alignment (±SD) in the three groups of proteins.

	TOTAL	UNIQUE	REPETITIVE	T-test§
Hox	25.77 ± *4.31*	16.21 ± *8.40*	44.82 ± *2.38*	
***Hox*-derived**	37.53 ± *9.63*	34.88 ± *12.40*	75.64 ± *34.45*	**
Non-*Hox*	8.73 ± *10.24*	8.46 ± *10.28*	23.79 ± *25.66*	
ANOVA	***	**	n.s.	

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001)

§ T-test for paired samples (unique *versus* repetitive) on proteins having both types of regions [ABD-A, LAB, PB, BCD, ZEN, Ccp84Ac, CG13617, CG14290 and LAP (product of *CG2520*)].

SECTION 3. ANNOTATION CRITERION = Uniprot + Simple repeats + Complex repeats

Table 1. Mean nucleotide substitution parameters, ANOVAs and contrast tests for the three groups of genes.

		t		d_N	d_S		ω		
Complete coding sequences	*Hox*	2.10917		0.15964	2.59066		0.06094	*	
	Hox-derived	3.86337		0.39380	4.27598		0.09226		
	Non-*Hox*	2.91160		0.15802	3.80668		0.04156	***	
	ANOVA	n.s.		*	n.s.		***		
Coding sequences excluding the homeobox	*Hox*	2.27653		0.18257	2.65921		0.06673	**	
	Hox-derived	5.04914		0.54809	5.26666		0.11320		*
	Non-*Hox*	2.91160		0.15802	3.80668		0.04156	***	
	ANOVA	n.s.		**	n.s.		***		
Coding sequences excluding repetitive regions	*Hox*	1.47343		0.08116	2.09576		0.03615	*	
	Hox-derived	3.57569	*	0.33831	4.17038	*	0.07474		
	Non-*Hox*	2.69427		0.14241	3.72116		0.03739	**	
	ANOVA	n.s.		n.s.	n.s.		*		
Coding sequences excluding the homeobox and repetitive regions	*Hox*	1.52473		0.09941	2.00530		0.04554		
	Hox-derived	4.70398	*	0.48080	5.11828	**	0.09217	***	
	Non-*Hox*	2.69427		0.14241	3.72116		0.03739		
	ANOVA	n.s.		*	n.s.		**		

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001). For pairwise comparisons, only significant values are shown. Contrast tests assume homogeneity of variances where the Levene test does not give a significant P value.

Table 2. Percentage of amino acid differences in the alignment (±SD) in the three groups of proteins.

	TOTAL	UNIQUE	REPETITIVE	T-test§
Hox	22.80 ± *10.44*	13.64 ± *11.65*	37.00 ± *10.69*	
***Hox*-derived**	40.43 ± *18.26*	35.87 ± *22.13*	57.50 ± *9.77*	**
Non-*Hox*	23.77 ± *10.81*	21.75 ± *11.57*	34.16 ± *31.78*	
ANOVA	n.s.	n.s.	n.s.	

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001)

§ T-test for paired samples (unique *versus* repetitive) on proteins having both types of regions [ABD-A, LAB, PB, BCD, ZEN, Ccp84Ac, Ccp84Ae, Ccp84Af, Ccp84Ag, CG13617, CG14290, LAP (product of *CG2520*) and CG31363].

Table 3. Percentage of indels in the alignment (±SD) in the three groups of proteins.

	TOTAL	UNIQUE	REPETITIVE	T-test§
Hox	25.77 ± *4.31*	9.72 ± *5.29*	43.60 ± *6.39*	
***Hox*-derived**	37.53 ± *9.63*	27.78 ± *17.38*	66.43 ± *7.83*	***
Non-*Hox*	8.73 ± *10.24*	3.45 ± *5.29*	23.41 ± *20.84*	
ANOVA	***	***	*	

n.s. (P>0.05), * (P<0.05), ** (P<0.01), *** (P<0.001)

§ T-test for paired samples (unique *versus* repetitive) on proteins having both types of regions [ABD-A, LAB, PB, BCD, ZEN, Ccp84Ac, Ccp84Ae, Ccp84Af, Ccp84Ag, CG13617, CG14290, LAP (product of *CG2520*) and CG31363].

ADDITIONAL FILE 5

Additional file 5. Structure and expression of non-*Hox* proteins.

Gene symbol	Protein expression	Protein structure
Adhr	Enzyme with oxidoreductase activity that is expressed always with ADH in the embryo gastric caecae, larvae and the adult fat body and gut *[1]*.	It contains a short chain dehydrogenase domain (adh-short; PF00106).
α-Est2	Enzymes that act on carboxylic esters during embryogenesis, larvae and pupae *[2]*.	The catalytic apparatus (Carboxylesterase; PF00135) involves three residues (the catalytic triad): a serine, a glutamate or aspartate, and a histidine.
α-Est3		
Ccp3 – Ccp84Ac	Structural constituent of the larval cuticle.	Conserved C-terminal section *[3]* and include a 35-36 amino acid motif known as the R&R consensus, present in many insect cuticle proteins, an extended form of which has been shown to bind chitin (chitin-bind 4; PF00379) *[4]*. Outside these conserved domains, cuticular proteins share hydrophobic regions dominated by tetrapeptide repeats (A-A-P-A/V), which are presumed to be functionally important *[5,6]* and are responsible for the high percentage of indels found in these proteins.
Ccp6 – Ccp84Ae		
Ccp7 – Ccp84Af		
Ccp8 – Ccp84Ag		
CG13617	Nucleic acid/zinc ion binding protein.	It contains a classical zinc finger domain (zf-C2H2; PF00096). The C2H2 zinc finger is composed of 25 to 30 amino acid residues including 2 conserved Cys and 2 conserved His residues in a C-2-C-12-H-3-H type motif. The 12 residues separating the second Cys and the first His are mainly polar and basic, implicating this region in particular in nucleic acid binding. The zinc finger motif is an unusually small, self-folding domain in which Zn is a crucial component of its tertiary structure binding to the conserved Cys and His residues. Fingers have been found to bind to about 5 base pairs of nucleic acid containing short runs of guanine residues.
CG14290	Unknown	No domains identified.
CG14609	Unknown	No domains identified.
CG14899	Could be a membrane protein.	Putative membrane domains predicted.
CG2520	Binding protein that is expressed in the embryonic nervous system and garland cell, and in the larval neuromuscular synapses.	It contains three defined domains: an ENTH domain of unknown function (PF01417) located at the N-termini and composed of 9 alpha-helices connected by loops; a phosphoinositide-binding clathrin adaptor (IPR008943) involved in clathrin-mediated endocytosis; and an ANTH domain (PF07651) involved in phosphatidylinositol 4,5-bisphosphate binding. It contains annotated repetitive sequences in UniProt.
CG31363	Soluble unfolded molecule associated with microtubules through the cell cycle *[7]*. It is expressed in the young embryo, larval nervous system, precursors of eye photoreceptors and adult ovary.	No domains identified. It contains 2 degenerated repeats around the sequence PPGG, separated by a Serine-rich region *[7]*.
Lsp1β	Proteins from the hemolymph of insects, which are expressed in larvae and may serve as a store of amino acids for synthesis of adult proteins.	Structurally related to arthropod hemocyanins. They contain an N-terminal domain (Hemocyanin-N; PF03722), a copper-containing domain (Hemocyanin-M; PF00372) and a C-terminal ig-like domain (Hemocyanin-C; PF03723).
Lsp1γ		

References:

1. Betran E, Ashburner M: **Duplication, dicistronic transcription, and subsequent evolution of the Alcohol dehydrogenase and Alcohol dehydrogenase-related genes in Drosophila.** *Mol Biol Evol* 2000, **17:**1344-1352.

2. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP: **Gene expression during the life cycle of Drosophila melanogaster.** *Science* 2002, **297:**2270-2275.

3. Rebers JE, Riddiford LM: **Structure and expression of a Manduca sexta larval cuticle gene homologous to Drosophila cuticle genes.** *J Mol Biol* 1988, **203:**411-423.

4. Rebers JE, Willis JH: **A conserved domain in arthropod cuticular proteins binds chitin.** *Insect Biochem Mol Biol* 2001, **31:**1083-1093.

5. Talbo G, Hojrup P, Rahbek-Nielsen H, Andersen SO, Roepstorff P: **Determination of the covalent structure of an N- and C-terminally blocked glycoprotein from endocuticle of Locusta migratoria. Combined use of plasma desorption mass spectrometry and Edman degradation to study post-translationally modified proteins.** *Eur J Biochem* 1991, **195:**495-504.

6. Andersen SO, Rafn K, Roepstorff P: **Sequence studies of proteins from larval and pupal cuticle of the yellow meal worm, Tenebrio molitor.** *Insect Biochem Mol Biol* 1997, **27:**121-131.

7. Karpova N, Bobinnec Y, Fouix S, Huitorel P, Debec A: **Jupiter, a new Drosophila protein associated with microtubules.** *Cell Motil Cytoskeleton* 2006, **63:**301-312.

Part 3

Discussion

3

Discussion

3.1. Bioinformatics of genetic diversity

Population genetics theory has been inspired and tested on empirical data coming from a limited number of species and genes. Re-sequencing from one or more populations or species was laborious and expensive until the arrival of large-scale sequencing methods, and this was the main reason for the data shortage in the field. The present scenario has changed dramatically. As sequence data is growing exponentially, the development of efficient software to deal and store this huge avalanche of information has become a high priority in this massive information era (Collins *et al.* 2003). Even though many programs have been developed that successfully analyze local data in terms of nucleotide variability, they usually require the previous alignment of input sequences, which implies that sequences are known to be polymorphic. Thus, mining tools fitting the population genetics standards for managing and analyzing large datasets without human inspection have long been lacking. As a consequence, very few comparative studies of polymorphism and divergence patterns across several species are reported in the literature. Polymorphix (Bazin *et al.* 2005) should be noted here as a database of intraspecific sequence polymorphism that allows one to select sets of within-species eukaryotic sequence families and to visualize multiple alignments and phylogenetic trees. Polymorphix is useful for meta-analyses of population genomics (Bazin *et al.* 2006) and as a bibliographic tool in population genetics, in the sense that it retrieves and groups nucleotide sequences from public databases based on bibliographic and similarity criteria. However, the method to cluster sequences in families used in Polymorphix is quite rough (they perform similarity searches using Mega BLAST (Zhang *et al.* 2000) for all sequences against themselves), and remove paralogous sequences by sequence similarity and bibliographic criteria only. Also, they do not provide any quality measure to assess the confidence on the families, or any diversity measure to compare diversity values across genes or species.

In the first step of this work, an elaborated bioinformatic system (PDA —Pipeline Diversity Analysis—) has been created to explore and estimate polymorphism, as well as detect SNPs, in large DNA databases at any gene or species from which two or more sequences have been determined. Later, the efficacy of this system has been tested on the dataset corresponding to the *Drosophila* genus, and all the estimations have been made available through a comprehensive website (DPDB —Drosophila Polymorphism Database—) that includes various options for searching the database, specific additional data of interest for each polymorphic set and different analytic tools to re-analyze the data. Both the pipeline PDA and database DPDB will be discussed together in this section.

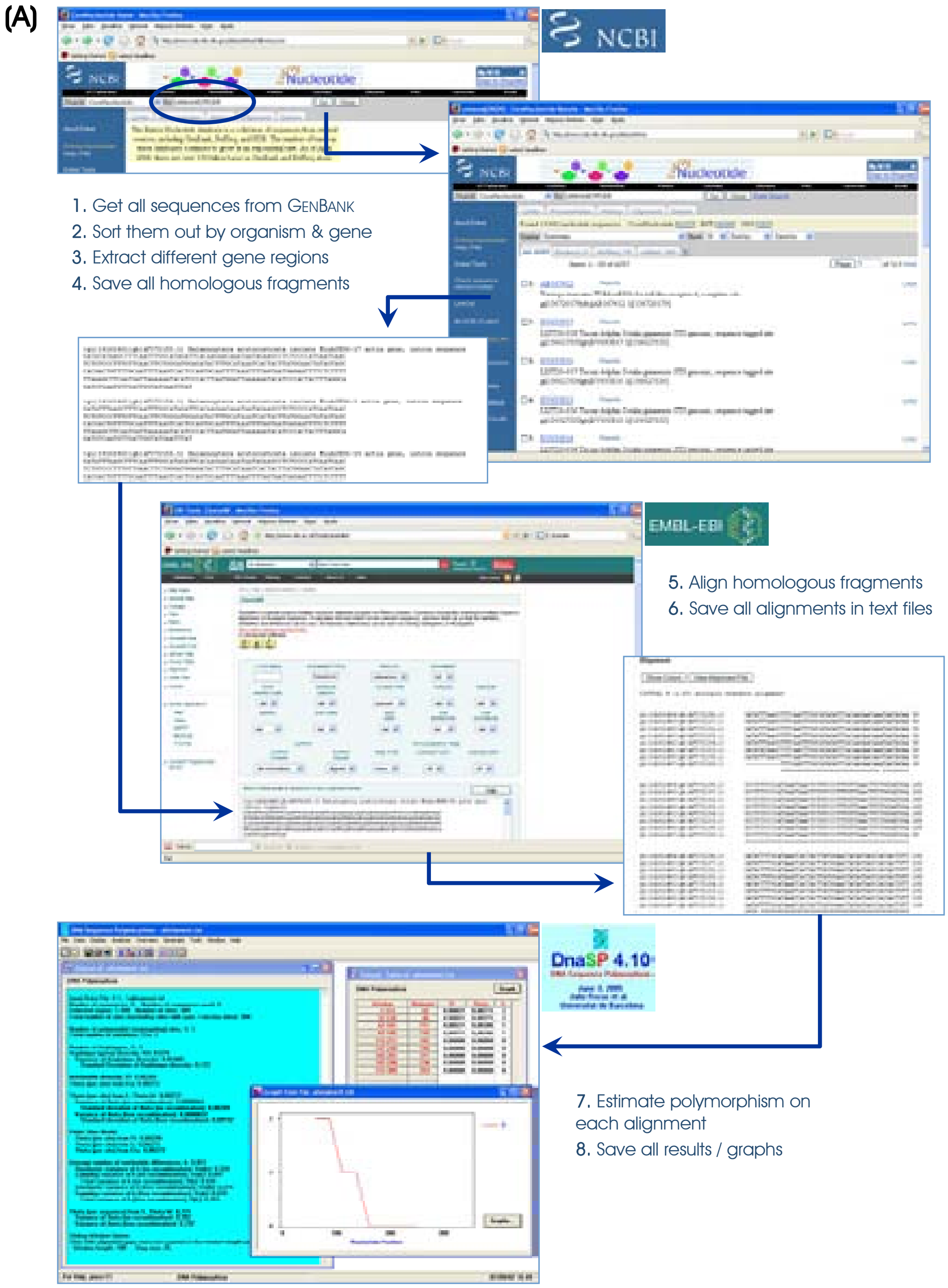
3.1.1. THE CHALLENGE: AUTOMATING THE ESTIMATION OF GENETIC DIVERSITY

The large-scale estimation of genetic diversity from sources of heterogeneous sequences requires the development of elaborate modules for data mining and analysis which operate together to automatically extract the available sequences from public databases, align them and compute diversity measures. *A priori*, the automation of this process seems doomed to failure, since variation estimates usually require a careful manual inspection. The main limitation of this process is undoubtedly the heterogeneous nature of the sequences, because such an automatic process can lump together sequences that are fragmented, paralogous, from different populations or chromosome arrangements, or simply incorrectly annotated sequences. Also critical is the multiple alignment of sequences, which is sensitive to the choice of algorithm, the input parameters and the intrinsic characteristics of the sequences. However, millions of haplotypic sequences (including those of complete chromosomes) that are today stored in public databases are an outstanding resource for the estimation of genetic diversity that cannot be neglected. Therefore, while aware of the limitations, we have tackled the bioinformatic automation of genetic diversity and developed both appropriate methods for data grouping and analysis, and rigorous controls for data quality assessment.

PDA is an exploratory tool that can typically be used to explore how many polymorphic sequences are available in GenBank for one or several species of interest, or how much variation there is in such sequences. The manual process of getting this data would involve going to GenBank and searching for all the sequences belonging to the species of interest, grouping them manually by organism and gene and extracting from them the fragments corresponding to the coding regions or any functional region we want to analyze. Then we would need to use any aligning software such as ClustalW (Chenna *et al.* 2003) to align the different subgroups one by one. Finally, we would need to take all the resulting alignments, revise their quality and use a different program, typically DnaSP (Rozas *et al.* 2003), to calculate the diversity values on them, also one alignment by one, and store the results manually in a file or database (Figure 16A). PDA incredibly speeds this process as we only need to go to the PDA main page and submit one single job specifying which species we want to analyze. Then, all the process above is done automatically by the PDA pipeline (Figure 16B). The program gives back a database containing all the sequences and measures of DNA diversity, as well as a set of HTML pages for the interactive exploration and reanalysis of the results, including a histogram maker tool to create graphical displays. Therefore, PDA is especially suitable for creating on-line polymorphism databases —such as DPDB— providing searchable collections of polymorphic sequences together with their associated diversity measures. It is thus a valuable contribution to bioinformatics and population genetics.

3.1.2. Data model

A key step in the process of large-scale management of sequence data is to define appropriate bioinformatic data objects that facilitate the storage, representation and analysis of genetic diversity from raw data. PDA introduces two novel data objects based on two basic storing units: the 'polymorphic set' and the 'analysis unit'. The polymorphic set is a group of homologous sequences for a given gene and species obtained from the public databases. From sequence annotations, homologous subgroups are created for each polymorphic set corresponding to different functional regions (e.g. CDS, exon, intron, UTR, promoter, etc.). Every group is then aligned and selected according to different quality criteria to form the analysis units on which the commonly used diversity parameters are estimated. All the data (both gathered and generated by PDA) is finally stored in a relational MySQL database which was designed according to this data model (Figure 17).



1. Get all sequences from GenBank
2. Sort them out by organism & gene
3. Extract different gene regions
4. Save all homologous fragments

5. Align homologous fragments
6. Save all alignments in text files

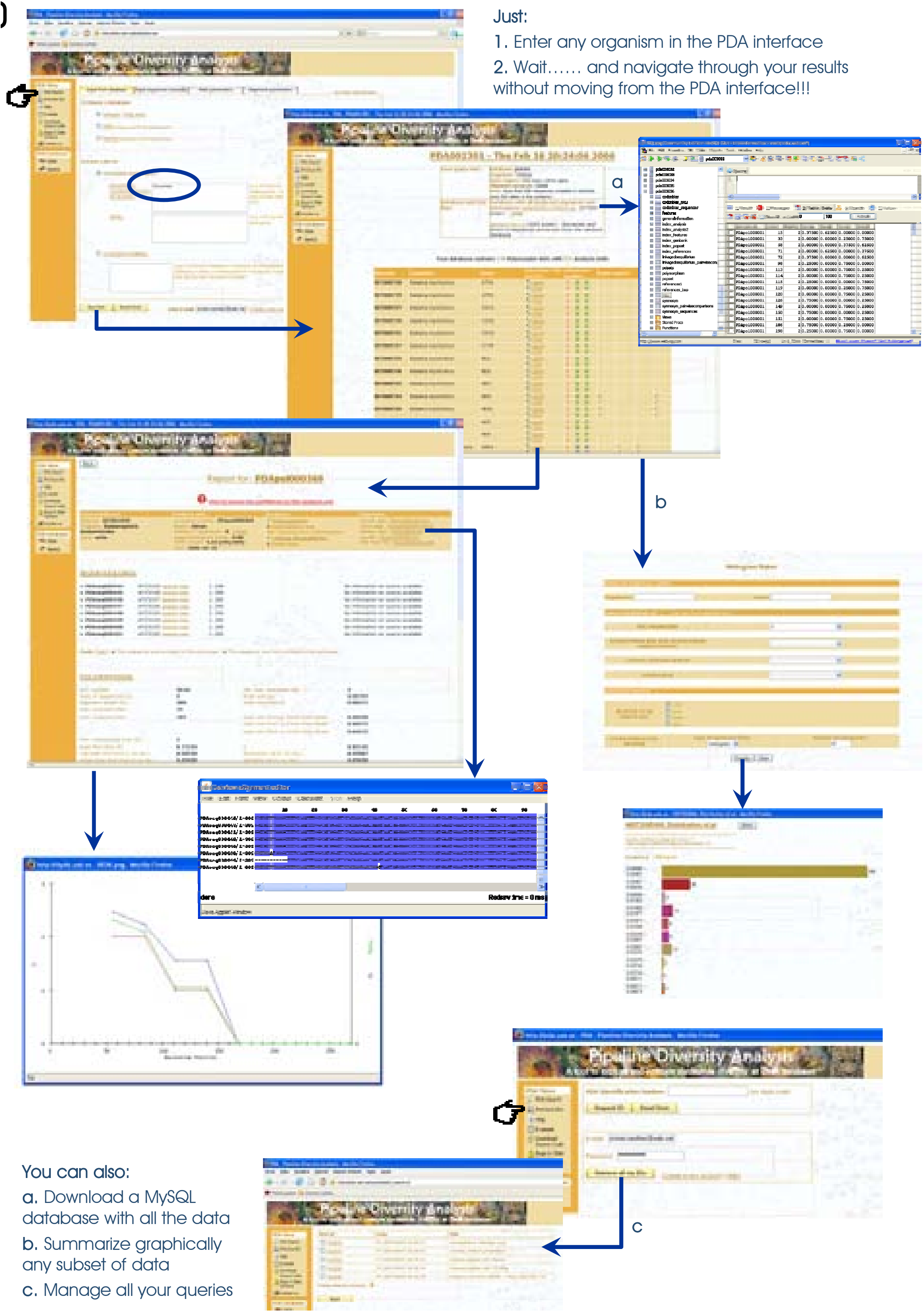
7. Estimate polymorphism on each alignment
8. Save all results / graphs

Figure 16

Large-scale exploration and estimation of genetic diversity: manual procedure *versus* PDA

In this example, polymorphism is estimated for all haplotypic sequences of Cetaceans stored in GenBank, using two different procedures: (A) manual gathering and analysis of sequences, and (B) automatic mining and analysis using PDA.

(B)



Just:

1. Enter any organism in the PDA interface

2. Wait…… and navigate through your results without moving from the PDA interface!!!

You can also:

a. Download a MySQL database with all the data

b. Summarize graphically any subset of data

c. Manage all your queries

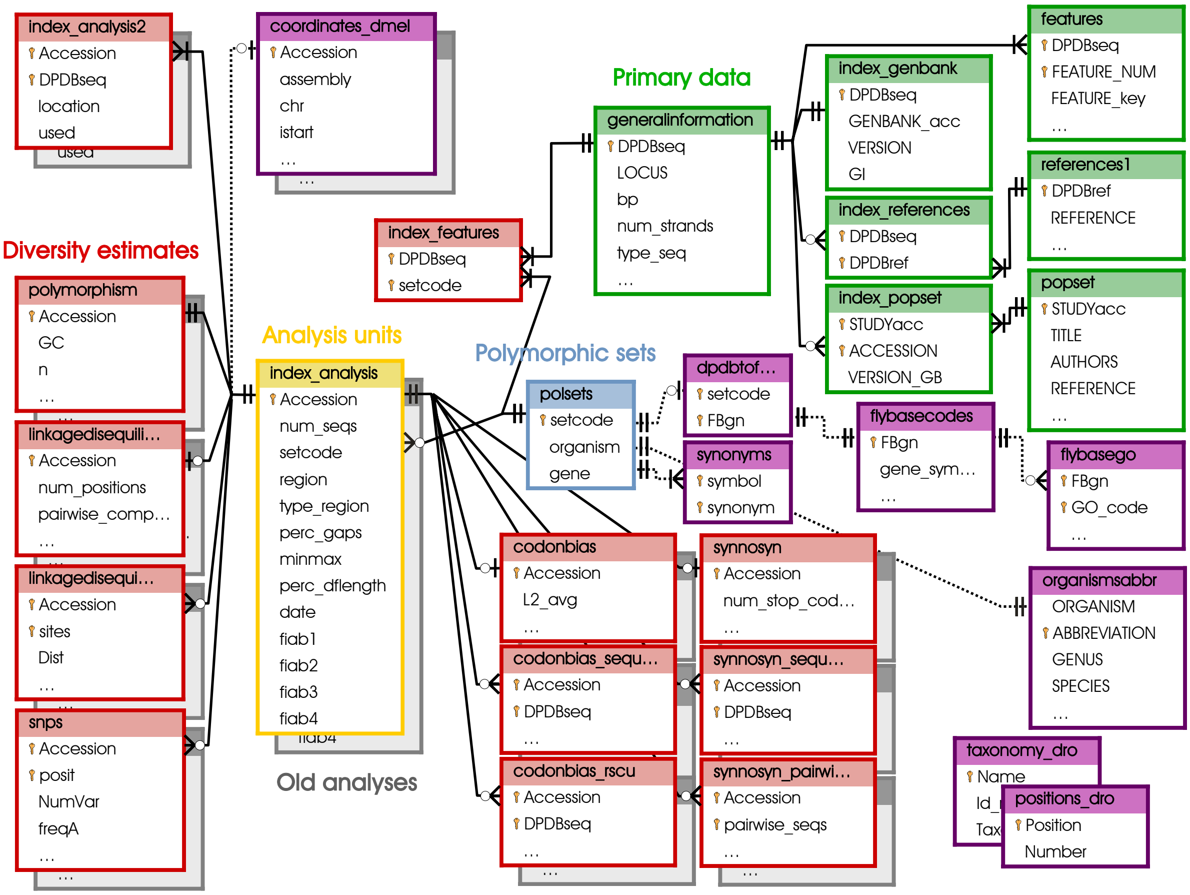


Figure 17
Structure of the PDA-DPDB databases

Entity-Relationship Diagram for the PDA-DPDB database. Color code: green, primary data gathered from external sources; blue, polymorphic sets; yellow, analysis units; red, diversity estimates obtained by PDA; grey, old analyses; violet, DPDB-specific tables that extend the information for *Drosophila* data.

3.1.3. Data gathering and processing

The PDA pipeline is made up of a set of PERL modules that automate the mining and analysis of sequences stored in GENBANK (Figure 18). Using PDA we get all the publicly available nucleotide sequences for any given organism, gene or accession number which are well annotated (we exclude sequences from sections CON, EST, GSS, HTC, HTG, PAT, STS, SYN, TPA, and WGS of GENBANK, as well as sequences without gene

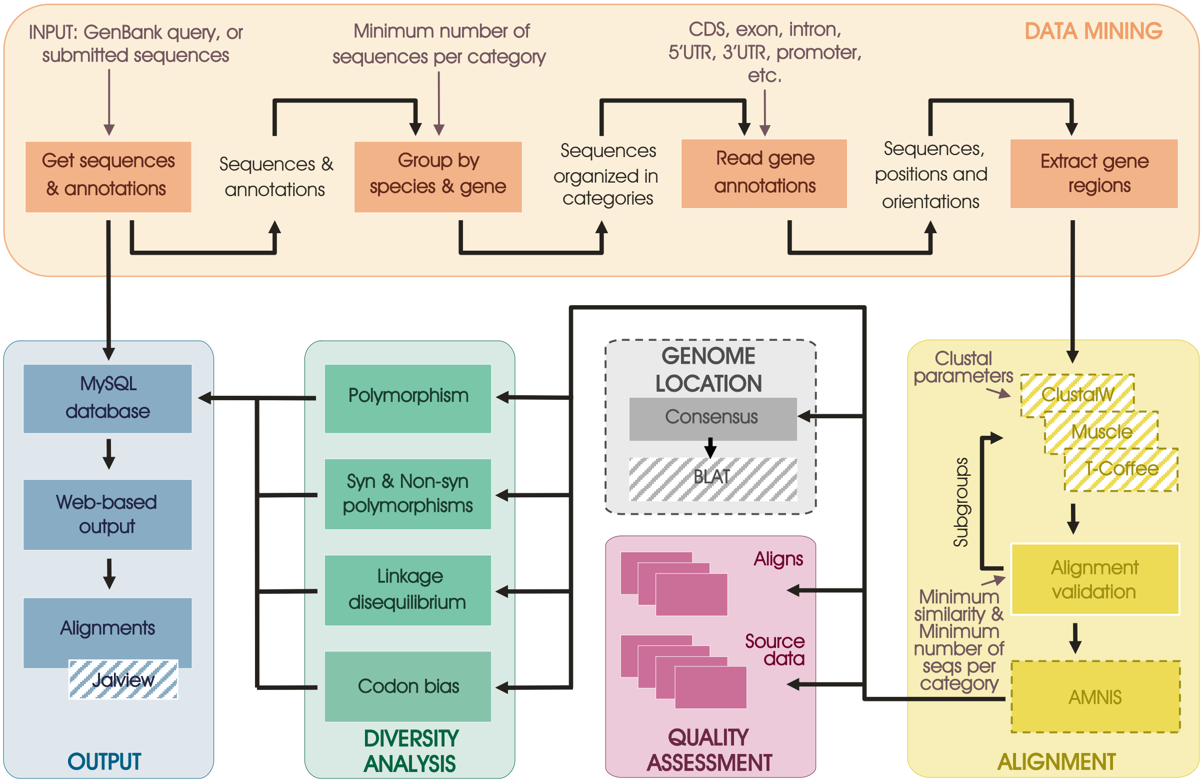


Figure 18
The PDA workflow

Dashed lines indicate optional modules. Striped colors represent external programs implemented in PDA. AMNIS, Algorithm for Maximization of the Number of Informative Sites (CASILLAS and BARBADILLA 2006). The Genome Location module is currently implemented for DPDB only.

annotations), together with their cross-references to the POPSET database. For thespecific case of DPDB, PDA extracts additional information from FLYBASE (CROSBY *et al.* 2007) including annotated GENE ONTOLOGY (GO) terms (ASHBURNER *et al.* 2000). Also in the case of DPDB, the annotated sequences of the complete chromosomes of *D. melanogaster* (ADAMS *et al.* 2000) are used for the estimation of genetic diversity. As a result, the number of analysis units in this species increases by ~50%, since many genes with a single sequence in GENBANK in addition to the genome sequence that were previously discarded can now be analyzed together with its corresponding allele in the genomic sequence.

One serious problem in large-scale studies of genetic diversity is the automatic detection of homologous DNA regions. PDA determines homologous sequences based on gene name. However, sequences stored in GenBank use sometimes different names for the same gene, and thus homologous sequences could eventually be grouped into different polymorphic sets. Given the importance and long way of *Drosophila* as a model organism, this species has a well-curated list of accepted gene symbols with several recorded gene synonyms. For the creation of DPDB, this list has been downloaded from FlyBase and gene names from GenBank are replaced by their accepted gene symbol before being introduced into the DPDB database. Following this procedure, the fraction of redundant polymorphic sets in DPDB is expected to be low (~98% of the *D. melanogaster* genes that are currently analyzed in DPDB match an accepted gene symbol in FlyBase).

Once the homologous sequences are determined, sequences are aligned. PDA implements three different aligning programs: ClustalW, Muscle (Edgar 2004) and T-Coffee (Notredame *et al.* 2000). Muscle has been shown to achieve a better accuracy (see Section 3.1.5), especially in alignments with a high proportion of gaps, and it is thus used to align sequences in DPDB. PDA deals with the problem of non-homology in alignments by grouping sequences by similarity (a customizable minimum identity must exist between each pair of sequences within an alignment, e.g. 95% in DPDB). On the other hand, given that sites with gaps are not used for the estimation of single nucleotide polymorphism, inclusion of short sequences tends to reduce the amount of informative content in an alignment. As a result, we have developed an Algorithm for the Maximization of the Number of Informative Sites (AMNIS) of an alignment in which sequences are grouped by length in order to score the largest number of informative sites for the calculation of diversity measures (Casillas and Barbadilla 2006).

Finally, each alignment is mapped to the genome sequence of *D. melanogaster* in DPDB. First, a consensus sequence is obtained from the multiple sequence alignment. The consensus is then aligned to the *D. melanogaster* genome using BLAT (Kent 2002) and the corresponding coordinates are obtained and provided with the alignments. These coordinates are used to link each analysis unit to the genome browsers in FlyBase and UCSC (Kuhn *et al.* 2007). This allows users to integrate analysis of polymorphism within species with other comparative or functional genomic resources that are aligned to the reference genome sequence. Additional links to FlyBase based on gene name, and related GO terms are also provided.

3.1.4. Confidence assessment for each polymorphic set

PDA provides several measures to assess the confidence on each polymorphic set, according to both the data source and the quality of the alignment. For the data source, we provide four criteria to help determining if the sequences were initially reported as part of a polymorphism study: (i) one or more sequences from the alignment are stored in the PopSet database, (ii) all the sequences have consecutive GenBank accession numbers, (iii) all the sequences share at least one reference, and (iv) one or more references are from journals that typically publish polymorphism studies (i.e. *Genetics*, *Journal of Molecular Evolution*, *Molecular Biology and Evolution*, *Molecular Phylogenetics and Evolution* and *Molecular Ecology*). To assess the quality of an alignment we use three other criteria: (i) the number of analyzed sequences, (ii) the percentage of gaps or ambiguous bases within the alignment, and (iii) the percentage of difference in length between the shortest and the longest sequences. Users of PDA are advised to: (i) revise all the previous parameters, (ii) check the alignments and phylogenetic trees provided for each alignment, (iii) revise the origin of the sequences, (iv) pay special attention to estimates of polymorphism giving extreme values, and (v) reanalyze the data changing the parameters or the input sequences when needed.

3.1.5. Testing the quality of the PDA output

Given the high heterogeneity of the source data and the fact that all the process —including the multiple alignment of sequences— is done automatically, a main caveat of PDA is whether its estimations are reliable. In order to test the quality of the PDA output, we automatically compiled and analyzed with PDA a controlled set of data including ~100-250 noncoding regions in three different populations of *D. melanogaster* (Glinka *et al.* 2003; Orengo and Aguade 2004; Ometto *et al.* 2005a). We compared the π values given by the authors (they aligned the sequences manually) with those obtained after aligning the sequences using 5 different aligning software: ClustalW, Dialign (Morgenstern 2004), Mafft (Katoh *et al.* 2005), Muscle and T-

COFFEE (Figure 19). Our alignments contained one additional *D. melanogaster* sequence (that from the genome sequencing project), although it was not used to compute π values because of its different origin. All correlation coefficients were >0.95 and the best results were obtained with the software MUSCLE and T-COFFEE (both implemented in PDA). Therefore, the automatic compilation and alignment of sequences with PDA proves to be very efficient, as we can recover π values from manually-performed studies.

In the previous correlation graphs (Figure 19), dots accumulate more often above the expected line than they do below, indicating that PDA tends to give higher π values than those by the authors. This is affected by the default input parameters of the aligning programs and it is expected to change depending on whether they penalize more nucleotide substitutions or gaps. If nucleotide substitutions are more penalized and gaps are opened more frequently in order to avoid nucleotide substitutions, π estimates will be lowered (as gaps are not used for computing diversity estimates); rather, if gaps are more penalized and nucleotide substitutions are more frequently allowed, π estimates will tend to increase. Thus, even though we show that PDA estimates are generally robust and suitable for large-scale studies of polymorphism, additional testing should be done under different conditions, e.g. changing default program parameters, using other aligning programs, or analyzing sequences from other species.

In a second test, we aligned outgroup sequences to our previous alignments: (i) the *D. simulans* sequence, (ii) the *D. yakuba* sequence, and (iii) both the *D. simulans* and the *D. yakuba* sequences. We realigned with MUSCLE (that performed best on the previous test) and recomputed the π values (Figure 20). We again got very high correlation coefficients (all >0.96). Therefore, the incorporation of outgroup sequences to the alignments does not distort the estimates, and thus PDA could ideally be used to combine polymorphism and divergence data for a wide range of tests for selection.

3.1.6. DPDB CONTENTS AND QUALITY OF THE DATA

In order to test the efficacy of the PDA system in generating on-line polymorphism databases, we chose the dataset corresponding to the *Drosophila* genus as a pilot test and generated the DPDB database. At the time of writing, the DPDB database contained >40,000 sequences from GENBANK, corresponding to 392 species and 15,177 different genes (see Figure 2 in 'Article 4'). When these sequences were filtered and analyzed, DPDB could gather informative data for 1,898 polymorphic sets (from 145 species and 1,184 different genes), and estimations were calculated on 3,741 analysis units, mostly corresponding to the functional regions CDS, exon, and intron. The best-represented species was *D. melanogaster* (53.2% of all analysis units), and the gene with the highest number of alignments was *Adh* (5.3% of all analysis units), which Drosophilists should be proud to note is the first gene in any species whose population genetics was studied using re-sequencing methods (KREITMAN 1983). In terms of quality of the alignments, many estimates were performed on alignments with <6 sequences (45.2%), but most of the alignments contain <10% of gaps or ambiguous bases (95.5%) and small differences in sequences length (84.8%). In terms of quality of the data source, only 26% of the analysis units contain sequences from POPSET, which means that DPDB contains an additional 3-fold more genomic regions that would otherwise be overlooked if only sequences from polymorphism studies were searched. The PDA retrieval system used in the construction of the DPDB database has thus provided a notable enrichment of the available diversity data. Daily-updated statistics of the DPDB database can be monitored at the Statistics section of the DPDB website.

3.1.7. USING PDA AND DPDB FOR LARGE-SCALE ANALYSES OF GENETIC DIVERSITY

As can be gathered from the text above, PDA is a powerful analytic pipeline to obtain and synthesize the existing empirical evidence of genetic diversity at any species or gene. Thus, the possibilities of interesting hypotheses that can be tested with PDA are endless. For example, 'Article 5' proves the action of purifying selection maintaining highly conserved noncoding sequences by combining genomic data from recently completed insect genome projects with population genetic data in *D. melanogaster*. For this study, a selected set of noncoding data from genome scans was gathered and analyzed both for point mutations and indels using a modified version of PDA.

On the other hand, DPDB provides pre-computed estimates of nucleotide diversity for a large number of genes and species of *Drosophila* which, in conjunction with the web interface, greatly facilitate both multi-species and multi-locus genetic diversity analyses. For example, obtaining average estimates of polymorphism in different gene

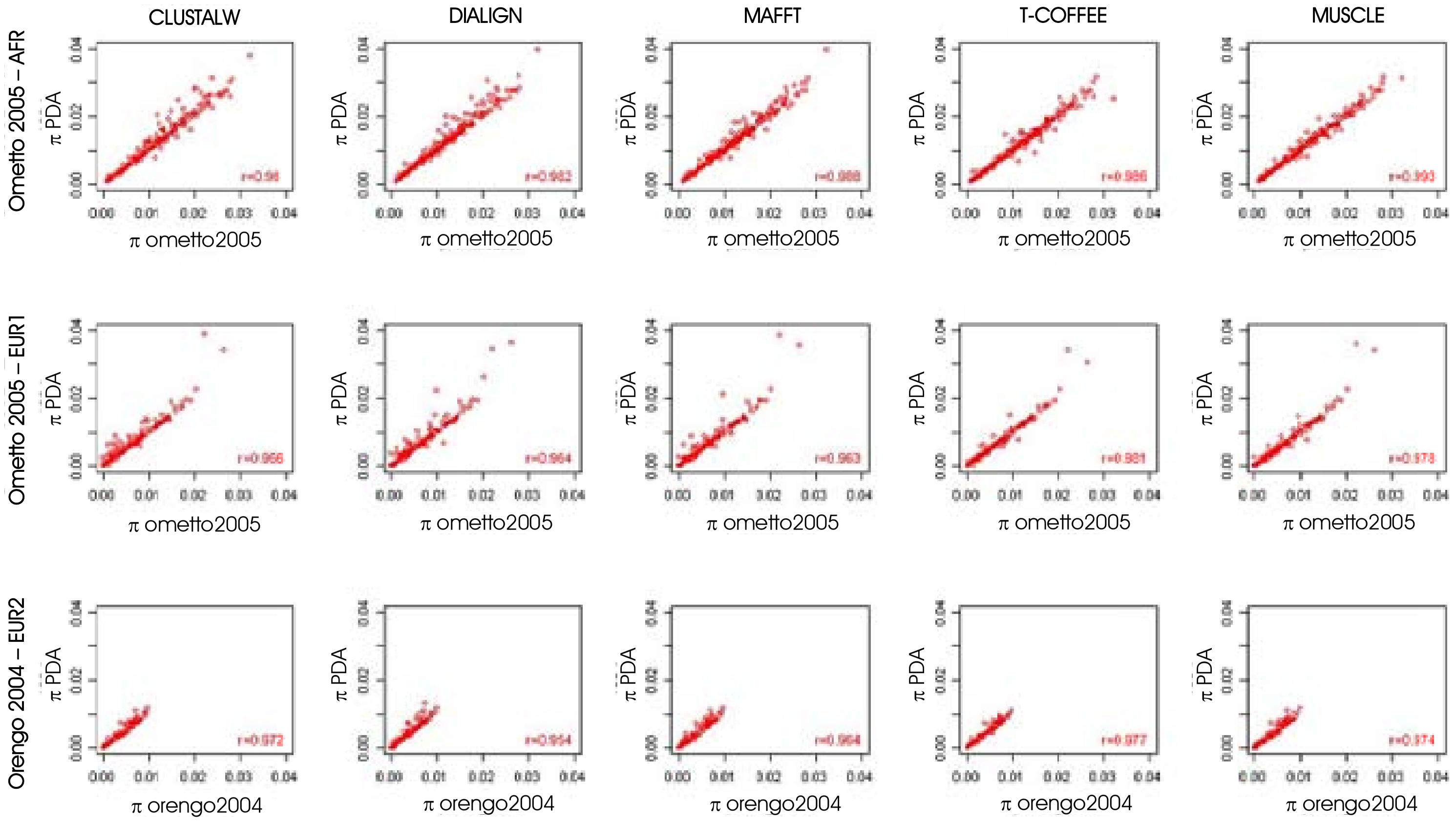


Figure 19

Correlations for π between the values obtained from published manually-curated alignments *versus* those obtained automatically with PDA

Data is from noncoding regions of the X chromosome in three different populations of *D. melanogaster* (rows). Automatic alignments were generated by PDA using one of the following aligning software: CLUSTALW, DIALIGN, MAFFT, T-COFFEE, or MUSCLE (columns).

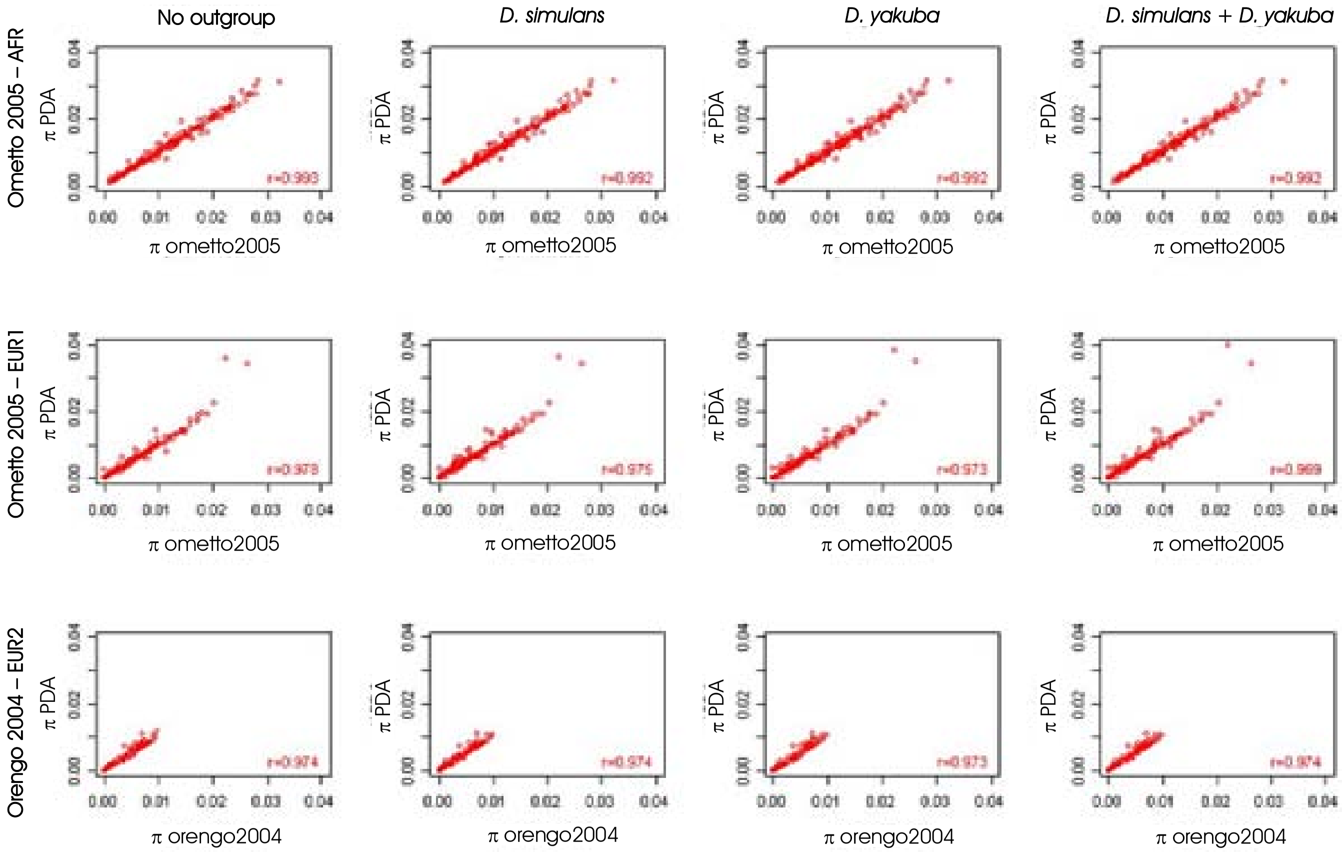


Figure 20

Correlations for π between the values obtained from published manually-curated alignments *versus* those obtained automatically with PDA when outgroup species are added

Data is from noncoding regions of the X chromosome in three different populations of *D. melanogaster* (rows). Automatic alignments were generated by PDA using the aligning software MUSCLE. PDA alignments include only sequences from *D. melanogaster* (No outgroup), sequences from *D. melanogaster* + one sequence from *D. simulans* (*D. simulans*), sequences from *D. melanogaster* + one sequence from *D. yakuba* (*D. yakuba*), or sequences from *D. melanogaster* + one sequence from *D. simulans* + one sequence from *D. yakuba* (*D. simulans* + *D. yakuba*) (columns).

regions and across different groups of *Drosophila* is easy by using the DPDB query interface (see Figure 1 in 'Article 4'). More interestingly, subsets of data from DPDB have already been used in large-scale analyses of nucleotide diversity. For example, our group has recently studied the association between coding polymorphism levels, intron content and expression patterns in *D. melanogaster* using DPDB (PETIT *et al.* 2007) (see Appendix II). This study reports that genes with low nonsynonymous polymorphism contain long introns with a high content of CNS sequences, and that genes with CNSs in their introns have more complex regulation. All these examples illustrate the power of PDA and DPDB to reveal new knowledge about the evolutionary process without the need for labor-intensive sequence retrieval or data processing on the part of the user.

3.1.8. WHAT IS NEXT?

The perspectives for PDA and DPDB are rather optimistic. Apart from their usage in large-scale studies of genetic diversity, we plan to extend the amount of taxa having pre-computed diversity estimates compiled in an on-line database, and the variety of tests provided today. On the one hand, after the implementation of PDA to generate the first on-line compilation of genetic diversity estimates (DPDB), our group is extending the availability of polymorphism databases to other taxa outside *Drosophila*. The MAMPOL database is the next database built using the PDA system and comprises diversity estimates for species of the Mammalia class (EGEA *et al.* 2007) (see Appendix I). In this database, polymorphic sets are categorized according to the species group (i.e. primates, rodents and other mammals) and gene location (i.e. nuclear or mitochondrial). In the future, databases created with PDA will be extended to all metazoans.

On the other hand, we expect to improve the PDA system by providing divergence data (i.e. outgroup sequences to each polymorphic set), additional tests of neutrality such as the McDonald-Kreitman test (MCDONALD and KREITMAN 1991) and derived allele frequency distributions, as well as the estimation of indel polymorphism. We will also create a specific section within DPDB that will include all the SNPs discovered using the PDA system. Finally, we will develop new methods to deal with unannotated noncoding sequences from genome scans (GLINKA *et al.* 2003; ORENGO and AGUADE 2004; OMETTO *et al.* 2005a) and data coming from SNP mapping studies (e.g.

the FlySNP Project) (Berger *et al.* 2001), whole genome shotgun and tiling array re-sequencing (e.g. the Drosophila Population Genomics Project, DPGP). It is thus our goal that the PDA system and its associated on-line polymorphism databases become comprehensive and reference resources for genetic variation, describing different types of genetic variation (e.g. SNPs and indels), distinct functional regions (e.g. coding and noncoding) and accepting diverse sources of data (e.g. re-sequencing data, SNP typing and whole genome sequencing).

3.2. Using patterns of sequence evolution to infer constraint and adaptation in *Drosophila* noncoding DNA

The euchromatic portion of the *Drosophila* genome sequence is mainly occupied by intronic and intergenic regions that do not code for proteins and that are poorly understood (Misra *et al.* 2002). As a consequence, >75% of the *Drosophila* genome has not yet been assigned a function and is still unannotated. This makes its study much more difficult, especially because bioinformatic tools such as PDA depend on well annotated sequences. The recent availability of many complete genome sequences, however, is providing an invaluable resource to try to determine which fraction of the ncDNA is functional and thus provide a better understanding of the mechanisms of transcriptional regulation and the evolution of development (Bergman *et al.* 2002; Enright *et al.* 2003; Lai *et al.* 2003; Costas *et al.* 2004; Negre *et al.* 2005; Siepel *et al.* 2005). *Drosophila* is indeed a desirable model system for the purpose of these studies. It has a relatively small genome sequence and many sequence data is currently available, including both high-quality re-sequencing sequences in *D. melanogaster* ncDNA and a wealth of comparative genomic data in many other species throughout the *Drosophila* genus (Figure 15).

3.2.1. *Drosophila* noncoding DNA

Unlike large mammalian genomes, most of the ncDNA in *Drosophila* is unique, with <6% confidently identified as repetitive transposable element sequences (Quesneville *et al.* 2005; Bergman *et al.* 2006). Thus, the *Drosophila* compact genome may imply that most of its noncoding sequence is associated with a regulatory function, such as *cis*-regulatory elements (BERGMAN *et al.* 2002; COSTAS *et al.* 2004; NEGRE *et al.* 2005) or noncoding RNAs (ENRIGHT *et al.* 2003; LAI *et al.* 2003). Indeed, this is supported by several findings reported recently in *Drosophila*. First, PETROV *et al.* have shown that unconstrained ncDNA is quickly removed from the *Drosophila* genome (PETROV *et al.* 1996; PETROV and HARTL 1998). This process may purge the fly genome of 'junk' DNA (e.g. pseudogenes) and enrich ncDNA for functional elements. Second, it has been shown that genes with complex expression (e.g. specific TFs or genes involved in pattern specification, embryonic development, cell differentiation, or receptor activity) tend to have longer associated intergenic regions (NELSON *et al.* 2004) and a higher abundance of CNSs within introns or nearby intergenic regions (NEGRE *et al.* 2005; PETIT *et al.* 2007). These findings suggest that the mere presence of ncDNA in *Drosophila* may not involve lack of function. Third, long introns and intergenic regions evolve slower in *Drosophila* (HADDRILL *et al.* 2005; HALLIGAN and KEIGHTLEY 2006), consistent with the interpretation that long noncoding regions may harbor most of the functional noncoding elements. Furthermore, intron length is also negatively correlated with its corresponding coding sequence evolution (MARAIS *et al.* 2005; PETIT *et al.* 2007), resulting in sequence evolution being coupled between *cis*-regulatory sequences and its associated coding DNA (CASTILLO-DAVIS *et al.* 2004; PETIT *et al.* 2007). And finally, adaptive substitutions may be common in both intronic and intergenic regions in *Drosophila* (ANDOLFATTO 2005), which can only occur if functional nucleotide sites in ncDNA are abundant.

3.2.2. DETECTION OF CNSS BY COMPARATIVE GENOMICS

Sequence conservation across distantly related species is typically interpreted as the signature of functional elements that are maintained over the time by the action of purifying selection (BERGMAN *et al.* 2002). The general approach of detecting functional sequences by sequence comparison is known as phylogenetic footprinting (or DNA footprinting). For example, CNSs detected in ncDNA have been successfully used to guide the prediction of *cis*-regulatory sequences (BERGMAN *et al.* 2002; COSTAS *et al.* 2004) and functional noncoding RNAs (ENRIGHT *et al.* 2003; LAI *et al.* 2003). However, the ability to detect functional sequences by comparative genomics extremely depends on the phylogenetic distance of the species compared and the type of regions sampled (NEGRE

2005; Negre *et al.* 2005; Richards *et al.* 2005). In general, the comparison of species covering large evolutionary distances fails to detect lineage-specific regulatory elements, while the comparison of too closely-related species does not allow discriminating neutral *versus* constrained DNA. Also, genes with a complex expression, which harbor long conserved intergenic regions, may target most of the *cis*-regulatory sequences predicted by genome comparisons, while house-keeping genes or general TFs may require subtler methods or the comparison of sibling species to detect conservation (Negre 2005).

After Dermitzakis and Clark (2002) showed that ~32-40% of the functional binding sites for TFs in human regulatory regions are not conserved in rodents, novel computational methods for finding regulatory elements were required. The comparison of multiple species at different phylogenetic distances (including sibling species) with the inclusion of this divergence data in the method has been shown to be the most efficient method to detect functional elements at different groups of species (Boffelli *et al.* 2003). This method is known as phylogenetic shadowing. A sophisticated phylogenetic shadowing method today is that by Siepel *et al.* (2005), which uses a phylogenetic Hidden Markov Model on different sets of species (e.g. conserved elements in *Drosophila* are being defined using genome sequences from 12 flies, mosquito, honeybee and beetle) (see Figure 21). The phastCons predictions by Siepel *et al.* (2005) on *Drosophila* estimate that >30% of the ncDNA in this species is within conserved blocks of several tens of base pairs. Note that phastCons percent estimates are up to 2 times higher than those previously given by other studies that used pairwise alignments and simple percent identity-based methods for identifying conserved elements (Bergman and Kreitman 2001; Negre *et al.* 2005). However, even the most rigorous definition of CNS is unlikely to capture all functionally important elements in the genome, especially those which arise through lineage-specific gain-of-function events. Therefore, the fraction of functional ncDNA may easily exceed these predictions of CNSs.

3.2.3. Evolutionary forces governing patterns of noncoding DNA in *Drosophila*

Which are the evolutionary forces governing the evolution of CNSs in *Drosophila*? Are they functionally constrained? Despite the widespread assumption that CNSs are

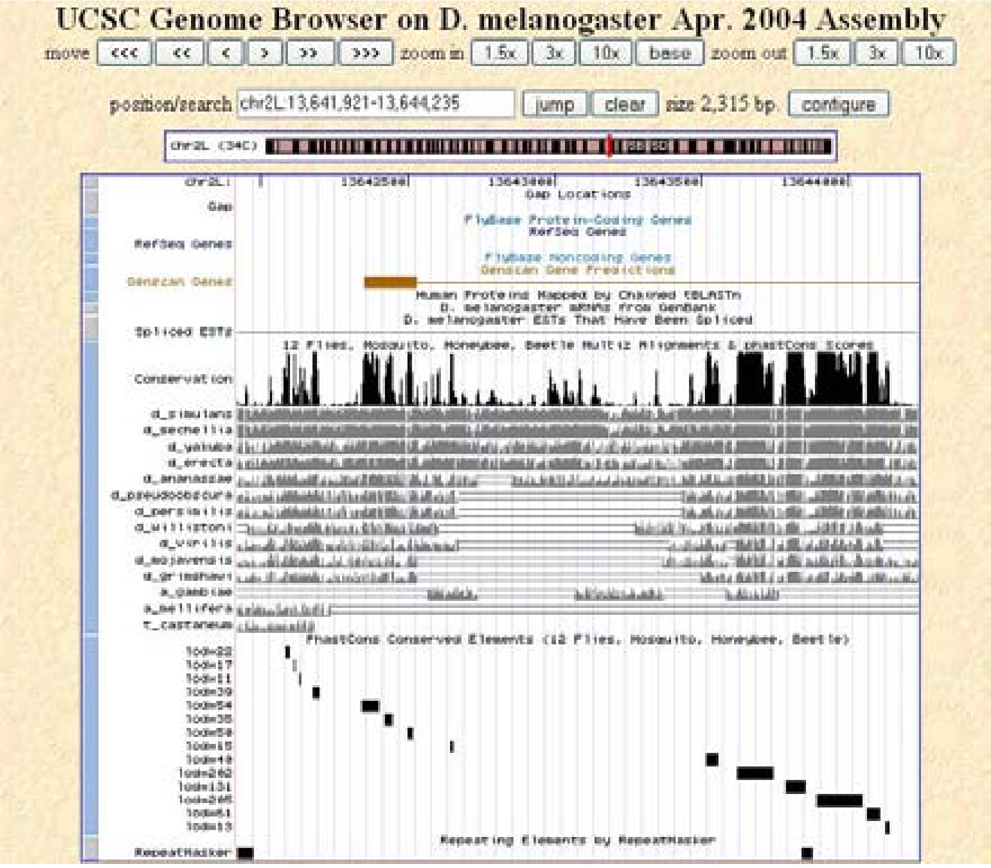

UCSC Genome Browser on D. melanogaster Apr. 2004 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr2L:13,641,921-13,644,235 jump clear size 2,315 bp. configure
chr2L (34C)
chr2L:
13642500
13643000
13643500
13644000
Gap Locations
Gap
FlyBase Protein-Coding Genes
RefSeq Genes
RefSeq Genes
FlyBase Noncoding Genes
Genscan Gene Predictions
Genscan Genes
Human Proteins Mapped by Chained tBLASTn
D. melanogaster mRNAs from GenBank
D. melanogaster ESTs That Have Been Spliced
Spliced ESTs
12 Flies, Mosquito, Honeybee, Beetle Multiz Alignments & phastCons Scores
Conservation
d_simulans
d_sechellia
d_yakuba
d_erecta
d_ananassae
d_pseudoobscura
d_persimilis
d_willistoni
d_virilis
d_mojavensis
d_grimshawi
a_gambiae
a_mellifera
t_castaneum
PhastCons Conserved Elements (12 Flies, Mosquito, Honeybee, Beetle)
lod=22
lod=17
lod=11
lod=39
lod=54
lod=35
lod=58
lod=15
lod=48
lod=262
lod=131
lod=245
lod=61
lod=13
Repeating Elements by RepeatMasker
RepeatMasker

Figure 21

PHASTCONS conserved elements displayed in the UCSC Genome Browser

UCSC Genome Browser displaying sequence annotations for a specific region of *D. melanogaster* chromosome 2. Note that sequence conservation between *D. melanogaster* and other insect genomes is shown in different tracks, as well as all conserved elements detected by PHASTCONS based on this conservation.

functionally constrained, they could potentially arise from a non-selective mode of evolution that invokes mutation and genetic drift only (CLARK 2001). This hypothesis, however, assumes extremely low mutation rates that vary over small spatial scales and no molecular mechanism has yet been proposed to produce such localized mutation cold spots. These two hypotheses can be distinguished from the patterns of noncoding sequence evolution. Such patterns in *Drosophila* include: (i) single nucleotide polymorphisms and fixed differences are reduced in CNSs compared to flanking regions,

(ii) CNSs have an excess of point mutations within species relative to divergence between species, and (iii) CNSs have an excess of rare SNPs relative to flanking regions. Even though both the mutational cold-spot hypothesis and the functional constraint hypothesis predict that levels of within-species polymorphism and between-species divergence should be reduced in CNSs compared to non-CNS flanking sequences, they do make different predictions concerning their relative reductions. While either low mutation rates or the elimination of deleterious alleles by purifying selection reduce the number of segregating polymorphisms and fixed differences within CNSs, only the elimination of deleterious alleles by purifying selection can explain a greater reduction in fixed differences relative to within-species polymorphisms. The reason is that, assuming that CNSs are maintained under 'weak' purifying selection, deleterious alleles will rarely reach fixation in these regions, but they can be maintained as low-frequency polymorphisms for some time before being eliminated. On the contrary, mutation rates are expected to affect similarly within and between species, and thus the mutational cold-spot hypothesis predicts similar reductions of polymorphisms and fixed differences. As a consequence, while the mutational cold-spot hypothesis is not expected to alter the allele frequency spectrum predicted under the neutral theory, the functional constraint hypothesis predicts a shift in the derived allele frequency (DAF) spectrum towards rare alleles. Then, the second and third patterns of sequence evolution found on *Drosophila* CNSs are only compatible with our initial prediction that CNSs are maintained by the action of purifying selection. Specifically, we have found that ~80-85% of CNS sites are functionally constrained, with an average strength of selection on the order of ~10-100-fold greater than the reciprocal of the effective population size. Finally, no evidence of positive selection has been found on *Drosophila* CNS sites.

The conclusion that *Drosophila* CNSs are maintained by purifying selection supports previous analyses that have made this assumption based on reduced rates of molecular evolution (BERGMAN and KREITMAN 2001; SIEPEL *et al.* 2005; HALLIGAN and KEIGHTLEY 2006). Specifically, our results support the UCSC PHASTCONS highly conserved track (SIEPEL *et al.* 2005) as being able to identify selectively constrained regions of the *D. melanogaster* genome. The characteristics of CNSs detected in our sampled regions show differences between introns and intergenic sequences. The intergenic regions sampled contain more CNSs than introns (~3.40-5.09 *versus* ~2.73-4.65

CNSs per region of ~800 bp, respectively), and these CNSs are also longer on average (~53.3-60.0 *versus* ~34.2-55.4 bp). As a consequence, the percentage of sequence covered by CNSs is higher in intergenic regions than in introns (~36.1-39.2% *versus* ~20.6-31.2%). And finally, CNSs in intergenic regions are more constrained than in introns. As a result, intergenic regions in *Drosophila* probably play a more important role in gene regulation than introns.

It is important noticing that our definition of CNSs by the PHASTCONS method allows indels within these regions. Therefore, we can also determine the patterns of indel evolution in CNSs compared to non-CNS flanking regions. We expect CNSs to be highly constrained for indels, possibly more than for point mutations. However, our results showed rather surprising patterns of indel evolution in CNSs: (i) both polymorphic and fixed indels are reduced in CNSs compared to flanking regions, (ii) polymorphic indels are only slightly overrepresented than fixed indels in CNSs compared to non-CNSs, and (iii) CNSs do not show an excess of rare indels relative to non-conserved flanking regions. After discarding that low power was the reason for the patterns found, only two possible explanations remained: (i) CNSs are not constrained for indels, or (ii) both CNSs and non-CNSs are similarly constrained for indels. The hypothesis that CNSs are cold spots for indels seems very unlikely, since CNSs are indeed selectively constrained for point mutations and are expected to be even more constrained for indels. Thus, the lack of strong differences in the evolutionary patterns between CNSs and non-CNSs for indels could be the result of the high constraint for indels in the spacers between conserved regions.

Evolutionary patterns for single nucleotide changes in the spacer regions between CNSs were determined by comparison with 4-fold degenerate coding sites, finding that these regions are not exclusively evolving under the influence of mutation and random drift: (i) ~18% of non-CNS sites are purged by the action of purifying selection, and (ii) ~18% of substitutions are fixed by positive selection. These results imply that the proportion of functionally relevant nucleotides in non-CNS regions is FRN = C + (1 – C)α ≈ 33%, ~2.5-fold less than the proportion of functional sites in CNS regions (~80-85%). Given that the majority of *Drosophila* ncDNA is found in non-CNS regions (~65-80%), the number of functional sites in CNS and non-CNS regions is approximately equivalent. Although less densely packed than CNSs, non-CNS regions also harbor some functionally important sites under purifying selection that were missed out by PHASTCONS due to their low sequence conservation across species. These regions probably represent species-specific regulatory elements that cannot be detected by genome comparison methods. Interestingly, non-CNS regions are also the preferred targets of positive selection. The overall relaxed selective constraints in non-CNS regions may facilitate the emergence of new regulatory elements that are beneficial and evolve adaptively until their fixation. Selective constraints on CNSs coupled with the signature of adaptive evolution in non-CNSs might be expected under the model of stabilizing selection proposed by LUDWIG *et al.* (2000), whereby loss of ancestral transcription factor binding sites (TFBSs) in CNS regions by mutations or small indels may lead to compensatory adaptive fixation of lineage-specific binding sites in non-CNS regions that restore the *cis*-regulatory function. Intriguingly, if this type of selection is common, the identification of *cis*-regulatory regions by sequence comparison across species (e.g. PHASTCONS predictions) may be missing out many of the functionally important sites in the genomes, including most positively-selected sites in non-CNS regions. Thus, additional methods should be used to complete the catalog of *cis*-regulatory regions.

CAMERON *et al.* (2005) studied five functionally characterized *cis*-regulatory regions in sea urchin and determined that, while single-nucleotide substitutions and small indels occurred freely at many positions within these regions, large indels (>20 bp) tended to appear only in the flanking sequences. They foresaw that a computational search for domains of large indel suppression would permit formulating a library of putative *cis*-regulatory sequences around any given gene. LUNTER *et al.* (2006) have recently published a ground-breaking methodology that, instead of nucleotide substitutions, uses the evolutionary imprinting of indels to infer selection. They have defined a neutral model of indel evolution fitting the human-mouse ancestral repeat data and predict functional DNA as those unusually ungapped regions that cannot be explained under the neutral model. The method is surprisingly powerful using only two or three mammalian genomes and it is able to identify protein-coding genes, micro-RNAs and unannotated material under indel-purifying selection at a predicted 10% false-discovery rate and 75% sensitivity (when human, mouse and dog genomes are used). Furthermore, the method allows identifying sequence that is subject to heterogeneous selection: positive selection with respect to nucleotide substitutions and purifying selection with respect to indels, enabling the genome-wide investigation of positive selection on functional elements other than protein-coding genes. It would be interesting to apply the method by LUNTER *et al.* (2006) and other alternative CNS definitions to the set of noncoding data compiled for this thesis and see how they affect our estimates of constraint and adaptation in CNS and non-CNS regions, as well as identify non-CNS regions that are subject to heterogeneous selection.

3.2.4. AN INTEGRATIVE MODEL OF *CIS*-REGULATORY EVOLUTION FOR *DROSOPHILA*

Both the protein machinery involved in transcription and the structure of genes at the DNA level are more complex in metazoans than are in prokaryotes (Box 4) (CARROLL *et al.* 2001). These differences result from the need of multicellular organisms to fine-tuning individual gene expression in specific cells at particular times during development. The regulation machinery in these organisms involves sequence-specific protein-DNA interactions that occur between particular TF residues and short DNA sequence motifs (i.e. TFBSs) of ~6-30 bp in length. These binding sites have been typically identified computationally as ≥6 nucleotide motifs of 100% between-species conservation using methods of phylogenetic footprinting. The brevity of these motifs implies, however, that

Box 4 Gene regulation in metazoans

Initiating transcription in complex multicellular organisms requires several dozen different proteins which interact with each other in specific ways (WRAY *et al.* 2003) (Figure 22). These include at least: (i) the RNA polymerase II with its associated transcription factors, which operate near the transcription start site, (ii) cell/tissue-specific activators and repressors that affect the transcriptional activity by binding to other DNA sequences called enhancers, and (iii) co-activators/co-repressors that join different activators/repressors and the general transcription machinery and influence the local state of the chromatin. Enhancers may be located in the vicinity of the promoter regions, or many thousands of base pairs away from the promoter, and even in the opposite orientation. Many of the genes in the toolkit for animal development (e.g. *Hox* genes) are transcriptional activators or repressors that interact with the regulatory sequences of other genes and modulate their transcriptional activity.

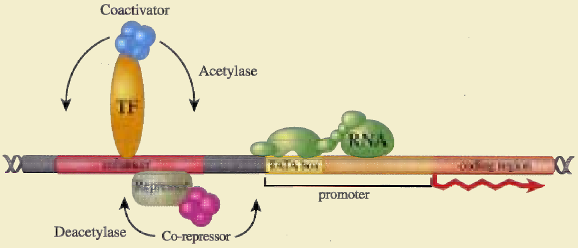


Figure 22
The regulation machinery in metazoans
[Figure from CARROL *et al. (2001).]*

they will occur individually at random many times in the huge noncoding sequence of the genome. Therefore, their regulatory specificity is not defined by their occurrence in the genome exclusively, but also by their nonrandom functional distribution. Indeed, verified functional regulatory elements always consist of relatively dense clusters of distinct sites, which are recognized by many DNA binding proteins.

CNSs detected in noncoding regions of *Drosophila* by PHASTCONS are usually several tens of base pairs. Typical CNSs are thus aggregations of many binding sites configuring what are known as *cis*-regulatory modules (CRMs, previously named 'enhancer elements') (Figure 23) (ONDEK *et al.* 1988; BERGMAN 2001). Because CNSs are rich in sequence-specific binding sites, the sequence is highly constrained in evolution and thus conserved among distant species. Furthermore, these modules hold spatial interactions among different TFs that recognize neighboring binding sites and thus they bear strong constraints in length variations. Observed indels within modules (e.g. within CNSs) are typically of few base pairs (in our sampled regions, >95% of fixed indels within CNSs are <15 bp, or <12 bp for polymorphic indels). Single binding sites can possibly act as modules as well (FROMENTAL *et al.* 1988). In turn, CNSs have been shown to be clustered in the chromosome (BERGMAN *et al.* 2002). Aggregations of various modules form the enhancers (ONDEK *et al.* 1988; BERGMAN 2001). Even though spacer regions between modules are sequence-independent, they interact at intermediate distances and thus they carry moderate spatial constraints. Spacing differences allowed between modules (e.g. spacer regions between CNSs) are on the order of ~100 bp. In our sampled regions, >95% of fixed indels in spacer regions are <25 bp, or <21 bp for polymorphic indels; note that these lengths are underestimates of typical permitted lengths in the genome since we have not analyzed complete chromosomes but pieces of ncDNA of ~800 bp only. Enhancers are positionally independent by definition and operate on promoters that may reside several kbs away by looping; therefore their 3' or 5' orientation with respect to the promoter seems not to be essential.

Spatial constraints within regulatory regions have been argued previously in *Drosophila* ncDNA based on the non-random distribution of CNSs and the strong correlation in the length between neighboring CNSs across divergent species (BERGMAN *et al.* 2002). More recently, OMETTO *et al.* (2005b) argued for spatial constraints acting within *Drosophila* ncDNA based on the ratio of insertions to deletions and the size

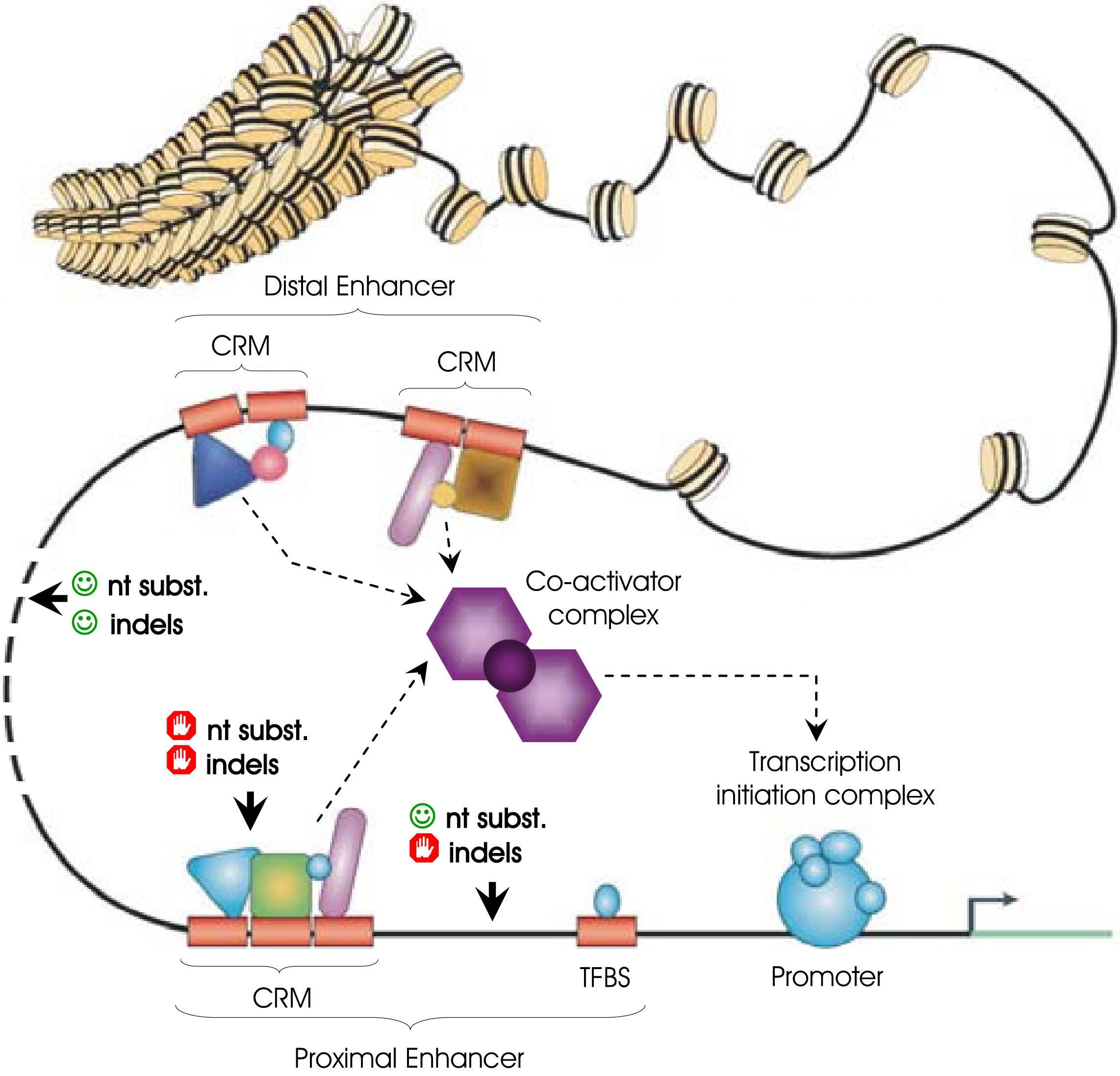


Figure 23

Components of transcriptional regulation and sequence/spatial constraints in *cis*-regulatory regions

TFs bind to specific TFBSs that lie either proximal or distal to a transcription start site. Sets of TFs can operate together in functional CRMs to achieve specific regulatory properties. Interactions between bound TFs and cofactors stabilize the transcription-initiation machinery to enable gene expression. The regulation that is achieved by sequence-specific binding TFs is highly dependent on the three-dimensional structure of the chromatin. These different functional regions are subject to different sequence and spatial constraints: (i) CRMs (i.e. CNSs) are subject to strong sequence/spatial constraints, (ii) spacers between CRMs within single enhancers are subject to moderate spatial constraints but are free to nucleotide changes, and (iii) DNA between separate enhancers seems to be free to either nucleotide changes or indels. See text for details. *[Figure modified from WASSERMAN and SANDELIN (2004).]*

distribution of deletions segregating in natural populations. In this thesis we also show that length constraints within CNSs and in the spacer regions between them seem to be similarly strong, thus reinforcing the hierarchical model of *cis*-regulatory regions firstly described by ONDEK *et al.* (1988) and later elaborated by BERGMAN (2001), here represented schematically in Figure 23. These spatial constraints extend to the level of enhancers. Enhancers lie in vast amounts of freely-evolving apparently nonfunctional genomic DNA. However, these sequences may also play an important role in transcriptional regulation: they provide genomic space (DAVIDSON 2006). This space allows distant sequence-specific regulatory elements to loop and associate with one another, and with the basal transcriptional apparatus at the promoter region of the gene. Thus, even though the space is functionally essential, its specific sequence or even the exact length are rather irrelevant.

3.2.5. DOES THE SAME MODEL APPLY TO OTHER SPECIES?

These findings in *Drosophila* closely parallel those recently found for mammalian CNSs and predicted micro-RNA binding sites using population genetic data from human SNP studies (KEIGHTLEY *et al.* 2005a; KRYUKOV *et al.* 2005; CHEN and RAJEWSKY 2006; DRAKE *et al.* 2006). Since there is no population evidence or molecular mechanism that support the mutation cold-spot hypothesis, similar results in disparate organisms such as flies and mammals, together with the non-random spacing of CNSs in flies and worms (BERGMAN *et al.* 2002; WEBB *et al.* 2002), argue against the interpretation of CNSs as mutational cold spots in any organism. Thus, purifying selection seems to be a general force acting to maintain highly conserved noncoding sequences in every metazoan genome. However, the strength of this selection may vary considerably from species to species. KEIGHTLY *et al.* (2005a) have shown that selective constraints in CNSs and their flanking regions are about one-half as strong in hominids as in murids, implying that hominids may have accumulated many slightly deleterious mutations in functionally important noncoding regions. In turn, our results in *Drosophila* show a stronger signature of negative selection than that previously found in mammals (KEIGHTLEY *et al.* 2005a; KRYUKOV *et al.* 2005; DRAKE *et al.* 2006; CHEN *et al.* 2007a), even though each study uses different methods and data types. These differences are likely to be a consequence of the lower effective population size of hominids than murids, and murids than drosophilids, leading to a reduced efficacy of selection (hominids < murids < dosophilids). For the same reason, evidence of positive selection is also more common in drosophilids than in mammals (Table 6). However, further studies in disparate taxa will be necessary to confirm the generality of this conclusion.

Mammals differ from other species in the proportion of ncDNA which is conserved and thus potentially functional (Figure 10). According to PHASTCONS, the estimated proportion of conserved ncDNA in drosophilids is >30%, but drop to <5% in hominids. The low effective population sizes of mammalian species are probably unable to keep 'junk' DNA (e.g. TEs and repetitive DNA) at a low proportion (KONDRASHOV 2005; LYNCH 2006), and thus the overall fraction of functionally important sites in ncDNA is smaller in these species. The structure and hierarchy of *cis*-regulatory regions proposed for *Drosophila* (Figure 23), however, should be applicable to any metazoan genome, albeit the amount of genomic space flanking independent regulatory elements varies widely among species. For example, the method by LUNTER *et al.* (2006) estimates that the fraction of functional ncDNA in the human euchromatin is ~2.56-3.25%. Note that these estimates closely match those obtained by the PHASTCONS method (SIEPEL *et al.* 2005) and report that more than half of the functional complement of the human genome is non-protein-coding (coding sequences represent only ~1.2% of the human euchromatin). The current availability of many complete genome sequences will provide the definite resource on which to provide unequivocal evidence for spatial constraints and quantify the mode and strength of selection acting on both indels and single nucleotide mutations in the ncDNA of several species. Altogether, this will shed light on the functions encoded in this abundant but still unknown fraction of metazoan genomes.

3.3. CODING EVOLUTION OF *HOX* GENES: FAST DIVERGENCE *VERSUS* A PARADIGM OF FUNCTIONAL CONSERVATION

Although it is commonly accepted that most of the genetic and morphological complexity resides on the noncoding fraction of the genome, the coding sequence of TFs that interact with the corresponding TFBSs in *cis*-regulatory regions may play an important role as well. For example, *Hox* genes encode TFs that interact with specific

TFBSs of other genes downstream in the regulatory cascade of development and modulate their expression. Mutations in *Hox* genes will on average have more deleterious fitness consequences than mutations occurring in genes expressed later on, because they may have cascading consequences for the later steps in development and thus may broadly alter the adult phenotype (RIEDL 1978; POWELL *et al.* 1993; CARROLL 2005; DAVIS *et al.* 2005). In fact, the large effects of *Hox* genes on morphology suggest that they regulate, directly or indirectly, a large number of genes. Estimates of evolution for different regions of the coding sequences of *Hox* genes may be a first step in predicting to what degree changes in such sequences are functionally significant.

3.3.1. PREDICTING THE RATE OF EVOLUTION OF *HOX* GENES

Different studies so far predict a slow rate of sequence evolution for *Hox* genes. First, DAVIS, BRANDMAN and PETROV (2005) find a highly significant relationship between the developmental timing of gene expression and their nonsynonymous evolutionary rate: genes expressed early in development are likely to have a slower rate of evolution at the protein level than those expressed later. However, the strongest negative relationship between expression and evolutionary rate occurs only after the main burst of expression of segment polarity and *Hox* genes in embryonic development. This observation may mislead about the expected rate of evolution for these genes. Besides, *Hox* genes are known to contain a long intron and nearby intergenic regions replete with regulatory elements (NEGRE *et al.* 2005). As previously pointed out, MARAIS *et al.* (2005) find a negative correlation between intron size and divergence rate at the protein level in *Drosophila*, likely due to a higher abundance of *cis*-regulatory elements in introns (especially first introns) in genes under strong selective constraints. These findings sustain *Hox* genes as firm candidates to be strongly constrained at the protein-coding level.

However, previous observations of protein sequences by KARLIN and BURGE (1996) determined that many essential developmental genes, including *Hox* genes, contain long microsatellites within their coding sequence (e.g. trinucleotide repeats that do not disrupt the open reading frame). These microsatellites promote frequent insertions and deletions and may be responsible for a higher than expected evolutionary rate of *Hox* genes. Also, AVEROF (2002) showed that a large portion of the sequence of *Hox* proteins diverges so fast that it is difficult to align homologous genes from different arthropod classes. Thus, these contradictory evidences question the veritable rate of evolution of the coding sequences of *Hox* genes.

3.3.2. Estimating the rate of evolution of *Hox* genes: nucleotides *versus* indels

We have studied the rate of evolution of *Hox* genes and shown that they evolve differently from other essential genes expressed in early development, with complex expression patterns or with long introns rich in *cis*-regulatory elements. On the one hand, both the number of nonsynonymous substitutions —or amino acid replacements— and the degree of functional constraint are not significantly different between *Hox* and non-*Hox* genes. Therefore, if we take into account aligned sites only, *Hox* genes seem to evolve at a similar rate than other non-*Hox* genes, even though they would be expected to evolve slower. *Hox*-derived genes evolve at higher rates and have lower levels of functional constraint than the other two groups of genes. On the other hand, the results for indels are even more surprising. The percentage of indels in *Hox* and *Hox*-derived proteins is much higher than that in non-*Hox* proteins (25.77%, 37.53% and 8.73%,

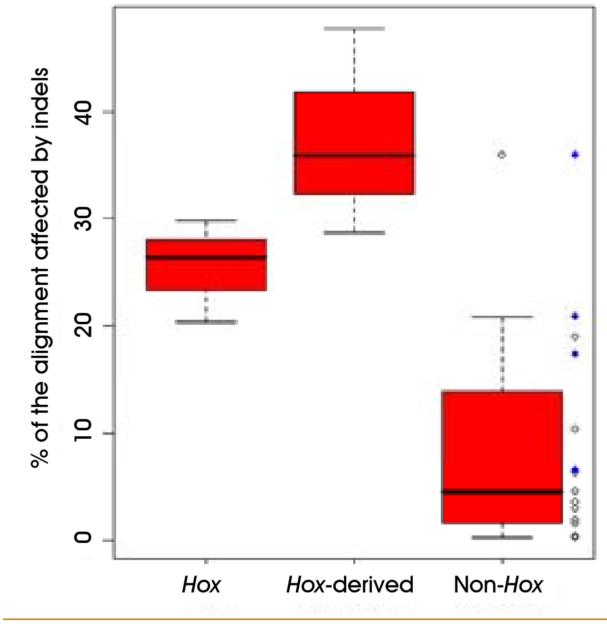


Figure 24
Proportion of indels in *Hox*, *Hox*-derived and non-*Hox* proteins

The proportion of indels in *Hox* and *Hox*-derived proteins is much higher than that in non-*Hox* proteins. Among non-*Hox* proteins, those belonging to the cluster of cuticular genes tend to have the highest values in their group; these proteins contain hydrophobic regions dominated by tetrapeptide repeats within which most indels occur. Horizontal bar: median. Box: interquartile range. Whiskers: range of the data up to 1.5 times the interquartile range. Open dots: extreme values. Dots at the right of the figure: values for each non-Hox protein; those in blue correspond to proteins of the cluster of cuticular genes (*Ccp84Ac*, *Ccp84Ae*, *Ccp84Af*, *Ccp84Ag*).

respectively) (Figure 24). In conclusion, *Hox* proteins are as divergent as non-*Hox* proteins in terms of amino acid changes, but they are much more divergent in terms of indels. A lack of correlation between the proportion of indels and amino acid differences in the set of genes used in this study highlights the different evolutionary mechanisms that regulate both types of changes.

The two last studies in this thesis emphasize the importance of studying not only nucleotide substitutions but also indels when determining rates of evolution, either in coding or in noncoding regions. Both studies have demonstrated a non-correlated distribution of single nucleotide substitutions and indels, which implies that nucleotides and indels behave differently and must be studied apart. However, most population genetics studies so far have focused on the study of aligned nucleotides only, thus overlooking an important part of the story. A critical consequence of this fact is the lack of advanced models and tools to understand and quantify the evolution of indels. Novel methods incorporating data of indels in sequences will be essential to fill this vacuum (CHEN *et al.* 2007b; CHEN *et al.* 2007c; LUNTER 2007).

3.3.3. THE IMPACT OF HOMOPEPTIDES AND OTHER SHORT REPETITIVE MOTIFS IN THE CODING EVOLUTION OF *HOX* PROTEINS

Multiple long homopeptides are found in 7% of *Drosophila* proteins, most of which are essential developmental proteins expressed in the nervous system and involved in transcriptional regulation (KARLIN and BURGE 1996; KARLIN *et al.* 2002a). These homopeptides could be tolerated insertions that may play a role as transcriptional activity modulators. Some examples have been described in *Hox* and *Hox*-derived proteins that illustrate the acquisition of new functions in the insect lineage while maintaining their homeotic role (GALANT and CARROLL 2002). In these examples, selection against coding changes might have been relaxed because of functional redundancy among *Hox* paralogs. This may in turn have facilitated the functional divergence of *Hox* proteins. Moreover, it has been shown that homopeptide sequences in developmental proteins are a source of variation in natural populations, affecting visible traits by expanding or contracting at very high rates (FONDON and GARNER 2004) (Box 5). One potential role for these homopeptides is to serve as spacer elements between functional domains, to provide flexibility to the three-dimensional conformation and fine-tuning domain orientation of the protein in its interactions with DNA and other proteins. Excitingly, CHAN *et al.* (2007) have recently discovered a significant association between indel frequencies and protein essentiality in three different species (*Bacillus subtilis*, *Escherichia coli* and *Saccharomyces cerevisiae*), such that indels occur more often in essential proteins and those that are highly connected. This supports a possible role of these insertions and deletions in the regulation and modification of protein-protein interactions. In the other side of the coin, excessive expansions of homopeptides have often been associated with disease in humans

Box 5 Evolution of microsatellites by replication slippage (slipped-strand mispairing)

Microsatellites (i.e. tandem repetitions of short DNA sequences) undergo frequent increases and decreases in copy number, usually in small changes (see Table 1 for typical mutation rates). Their mutation mechanism results from a balance between slippage events and point mutations in a sort of stepwise mutation process (KRUGLYAK *et al.* 1998; ELLEGREN 2004) (Figure 25).

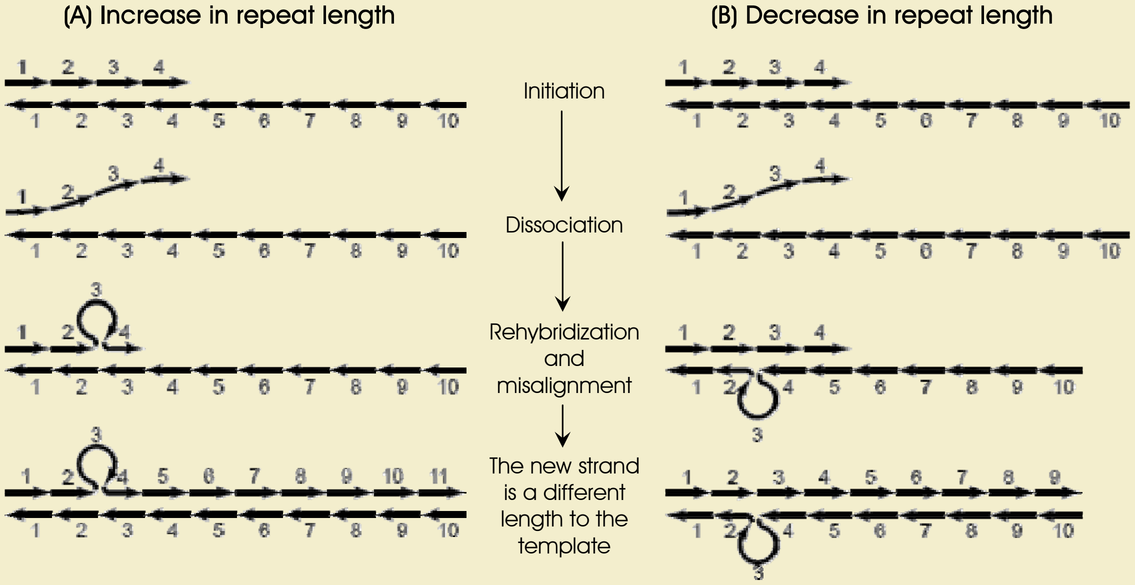

(A) Increase in repeat length
(B) Decrease in repeat length
Initiation
Dissociation
Rehybridization and misalignment
The new strand is a different length to the template
1 2 3 4 5 6 7 8 9 10 11

Figure 25
Model of microsatellite mutation by replication slippage

Repeat units are denoted by arrows. When the repetitive region is being replicated, the two strands can dissociate and be misaligned when re-associated: (A) if the nascent strand is incorrectly realigned e.g. one unit downstream of the template strand, a loop formed on the nascent strand results in this nascent strand being one unit longer than the template strand; (B) if the incorrect alignment occurs upstream of the template strand, the nascent strand will become one repeat unit shorter than the template strand. *[Figure from* ELLERGEN *(2004).]*

(HANCOCK *et al.* 2001; KARLIN *et al.* 2002b; BROWN and BROWN 2004; ALBRECHT and MUNDLOS 2005). Amazingly, essential developmental proteins —like homeotic proteins— that apparently need such homopeptides for their correct functioning have to suffer the consequences of the quick and apparently unpredictable evolution of this kind of repetitive sequences. This is a beautiful example of cost/benefit trade-off at the molecular evolutionary level.

Hox and *Hox*-derived proteins contain many homopeptides and other types of repetitive regions, which we studied separately. These repetitive regions are richer in both amino acid differences and indels than unique sequence. However, they are not enough to explain all the differences found between *Hox* and *Hox*-derived genes, and non-*Hox* genes. *Hox* and *Hox*-derived genes have a tendency to accumulate indels outside these repetitive regions that is not observed in non-*Hox* genes. The excess of indels in these genes may be the result of spontaneous deletions between short repeated sequences, which have been described in different organisms from phages (STUDIER *et al.* 1979; PRIBNOW *et al.* 1981) to humans (MAROTTA *et al.* 1977; EFSTRATIADIS *et al.* 1980). Two different models can explain the generation of spontaneous deletions: slipped mispairing during DNA synthesis, and recombination events mediated by enzymes that recognize short sequence similarities of as few as 5-8 bp. In either case, the repetitive and compositionally biased nature of several regions within *Hox* and *Hox*-derived sequences might explain the major incidence of indels in these two groups of genes. This would also explain the large differences in protein lengths among species that have been observed in some *Hox* proteins. This higher probability of mutation would presumably be accompanied by a higher tolerance towards indels of *Hox* and *Hox*-derived proteins outside their binding domains. The main implication of this result is that the estimation of evolutionary rates at any DNA sequence may require its partition into regions of different nature (e.g. unique *versus* repetitive) in order to understand the results completely.

3.3.4. AN INTEGRATIVE VIEW OF THE CO-EVOLUTION BETWEEN TRANSCRIPTION FACTOR CODING REGIONS AND *CIS*-REGULATORY DNA

Changes in gene regulation have been typically attributed to either changes in *cis*-regulatory sequences affecting one gene, or changes in *trans*-acting regulators affecting many genes (Brem *et al.* 2002; Wittkopp *et al.* 2004; Carroll 2005). Mutations in *cis*-regulatory sequences generally imply the gain or loss of TFBSs in a lineage-specific manner. For example, 30-50% of experimentally identified TFBSs in *Drosophila* lie outside of conserved blocks (Emberly *et al.* 2003), and 40% of human and mouse TFBSs are species-specific (Dermitzakis *et al.* 2003). Also, more than half of the binding sites extracted from the Yeastract database are not conserved among three closely-related yeast species (Doniger and Fay 2007). Binding site loss and gain have been typically explained by turnover, where either: (i) the concurrent gain of TFBSs generates redundancy and lowers constraint, which permit the degradation of pre-existing TFBSs, or (ii) the loss of ancestral TFBSs in CNS regions by mutations or small indels leads to compensatory adaptive fixation of lineage-specific TFBSs in non-CNS regions that restore the *cis*-regulatory function (Ludwig *et al.* 2000). The turnover of TFBSs has been demonstrated empirically and provides a neat explanation for the divergence in *cis*-regulatory sequences without any change in regulatory function. In the absence of binding site turnover, binding site loss results in species-specific changes in gene regulation, which seems to apply to nearly half of all loss events in yeast (Doniger and Fay 2007).

The frequent gain and loss of TFBSs implies that *cis*-regulatory sequences are labile and, as a consequence, the coding sequence of TFs must be labile too to adapt to the frequent changes in the composition of active TFBSs. The outcoming results of this thesis support a possible co-evolution between the coding sequence of TFs and their target binding sites, such that changes in nucleotide distances between TFBSs or substitution of active TFBSs may be accompanied by complementary changes in the sequences spacing the binding domains in the corresponding TFs. First, the fact that both CNSs and non-CNSs tolerate indels within their sequences corroborates the conclusion that *cis*-regulatory regions are labile. Second, a non-negligible fraction of the sequence spacing conserved blocks in ncDNA shows evidences of selection (both negative and positive). As previously mentioned, this supports the hypothesis that non-CNSs are a frequent source of new, lineage-specific TFBSs. And finally, the unstable sequence of *Hox* proteins by means of frequent indel changes owing to their characteristic repetitive composition might facilitate their adaptation to modifications in the active TFBSs. For instance, while the homeodomain is highly conserved among disparate taxa, the rest of the *Hox* protein sequences varies largely even among sibling species, specially through indel changes. Indels are especially rare among protein-coding sequences, since they easily interrupt the proper reading frame (i.e. by indels not multiple of three) and create a non-functional protein. The fact that indels are frequent among *Hox* coding sequences suggests that they may be functional, perhaps by providing flexibility to *Hox* proteins to adapt to their corresponding *cis*-regulatory regions. However, further studies associating changes in *cis* (i.e. TFBSs) and *trans* (i.e. TF coding sequences) will be necessary to confirm this statement (DAVIES *et al.* 2007). Also, studies of polymorphism in *Hox* genes will be needed to determine how frequently indels within species occur. Finally, more TFs will have to be studied in order to extend these findings outside of *Hox* genes.

Part 4

Conclusions

4
Conclusions

1. The automation of the estimation of genetic diversity from large sources of heterogeneous sequences can be successfully achieved if the appropriate tools for data gathering, processing, filtering and quality checking are developed. Thus, large-scale analyses based on automated estimates are reliable and can be used to answer hypotheses formulated within a population genetics framework.

2. PDA is a powerful analytic pipeline to obtain and synthesize the existing empirical evidence of genetic diversity at any species or gene. Importantly, it includes several filters and quality parameters that overcome the intrinsic difficulties to the automation of the process of the large-scale estimation of genetic diversity. Furthermore, PDA is a useful tool to generate databases of knowledge from raw data for any species or group of species.

3. DPDB is the first database that allows the search of DNA sequences according to different parameter values of nucleotide diversity, the degree of linkage disequilibrium or codon bias, and it allows filtering the results according to different confidence criteria. It also allows comparing diversity values across different species or taxonomic groups, and generating graphical distributions of any diversity measure. DPDB is thus a comprehensive resource for population geneticists working on the *Drosophila* model system.

4. Both PDA and DPDB have been successfully used to resolve different population genetics questions, some of which contribute to this thesis, and their possibilities of application in large-scale analyses of genetic diversity are endless.

5. Patterns of nucleotide sequence evolution in *Drosophila* CNSs are incompatible with the mutational cold-spot hypothesis to explain their existence and support the hypothesis that they are maintained by the action of purifying selection. Specifically, we estimate that ~85% of CNS sites in *Drosophila* are functionally constrained, with an average strength of selection on the order of ~10-100-fold greater than the reciprocal of the effective population size. Compared to similar studies in mammals, the estimated strength and number of sites under purifying selection is greater for *Drosophila* CNSs than human CNSs, as is expected given the higher effective population size of flies.

6. Patterns of sequence evolution for indels suggest that length constraints within CNSs and in the spacer regions between them are similarly strong, thus reinforcing a previously described hierarchical model of *cis*-regulatory regions.

7. We find no evidence of positive selection acting on *Drosophila* CNSs, although we do find evidence for the action of recurrent positive selection in the spacer regions between CNSs. Selective constraints on CNSs coupled with the signature of adaptive evolution in non-CNSs support a model of stabilizing selection whereby loss of ancestral TFBSs in CNS regions may lead to compensatory adaptive fixation of lineage-specific binding sites in non-CNS regions that restore the *cis*-regulatory function. If this type of selection is common, the identification of *cis*-regulatory regions by sequence comparison across species may be missing out many of the functionally important sites in the genomes, and thus additional methods should be used to complete the catalog of *cis*-regulatory regions.

8. Both the number of nonsynonymous substitutions and the degree of functional constraint are not significantly different between *Hox* and non-*Hox* genes, while *Hox*-derived genes exhibit higher rates of substitution and lower levels of constraint compared to the other two groups of genes. In terms of indels, *Hox* and *Hox*-

derived genes contain significantly more indels than non-*Hox* genes in their coding sequences. Thus, *Hox* genes evolve faster than would be predicted given their important function in early development.

9. *Hox* genes have a tendency to accumulate indels within their coding sequences that may result from a combination of two factors: (i) the repetitive and compositionally biased nature of several regions within these genes, which would promote indel mutations, and (ii) a high tolerance of indels of these genes outside their binding domains. This high rate of indels may allow *Hox* genes (and possibly other TFs) to adapt to the frequent changes in the composition of active TFBSs, and thus these results support a possible co-evolution between the coding sequence of TFs and their target binding sites, such that changes in nucleotide distances between TFBSs or substitution of active TFBSs may be accompanied by complementary changes in the sequences spacing the binding domains in the corresponding TFs.

10. A non-correlated distribution of single nucleotide substitutions and indels both in noncoding and in coding sequences implies that nucleotides and indels behave differently and must be studied apart. This observation, together with the fact that most population genetics studies so far have focused on the study of aligned nucleotides only, emphasize the importance of developing novel methods that incorporate data of indels in studying sequence evolution.

Conclusions

1. L’automatització de l’estimació de la diversitat genètica a partir de fonts de dades heterogènies és factible si es desenvolupen les eines adequades d’extracció, processat, filtrat i control de qualitat de les dades. Per tant, les anàlisis a gran escala basades en estimes automatitzades són fiables i es poden utilitzar per a respondre hipòtesis formulades dins el marc de la genètica de poblacions.

2. PDA és una eina analítica potent per a obtenir i sintetitzar l’evidència empírica existent de la diversitat genètica a qualsevol espècie i gen. Cal destacar que inclou varis filtres i paràmetres de qualitat que superen les dificultats intrínseques de l’automatització del procés d’estimació de la diversitat genètica a gran escala. A més, PDA és una eina útil per a generar bases de dades de coneixement a partir de dades brutes de qualsevol espècie o grup d’espècies.

3. DPDB és la primera base de dades que permet la cerca de seqüències de DNA segons diferents paràmetres de diversitat nucleotídica, el grau de desequilibri de lligament o el biaix en l’ús de codons, i permet filtrar els resultats en funció de diferents criteris de confiança. També permet comparar valors de diversitat a diferents espècies o grups taxonòmics, així com generar distribucions gràfiques de qualsevol mesura de diversitat. DPDB és doncs un recurs integral, de gran utilitat per als genetistes de poblacions que treballen en *Drosophila* com a model.

4. Tant PDA com DPDB han estat utilitzats amb èxit per a resoldre diferents qüestions de la genètica de poblacions, algunes de les quals contribueixen a aquesta tesi, i les seves possibilitats d’aplicació a anàlisis de diversitat genètica a gran escala són il·limitades.

5. Els patrons d'evolució de seqüències nucleotídiques a regions conservades no codificadores (CNSs) de *Drosophila* són incompatibles amb la hipòtesi de regions fredes de mutació per a explicar la seva existència, donant suport doncs a la hipòtesi que les CNSs són mantingudes per l'acció de la selecció purificadora. En concret, estimem que el ~85% dels llocs a CNSs a *Drosophila* estan constrenyits funcionalment, amb una força de selecció de l'ordre de ~10-100 vegades més gran que la inversa de la mida efectiva de la població. En comparació amb estudis similars a mamífers, la força de selecció estimada i el número de llocs afectats per selecció purificadora són majors a les CNSs de *Drosophila* que a les d'humans, la qual cosa és esperada donada la major mida efectiva de la població a les mosques.

6. Els patrons d'evolució de les insercions i delecions (indels) suggereixen que els constrenyiments funcionals sobre la longitud són tant importants a les CNSs com a les regions espaiadores entre elles, reforçant doncs un model jeràrquic de regulació prèviament descrit, format per regions *cis*-reguladores.

7. No trobem evidència de selecció positiva actuant a les CNSs de *Drosophila*, tot i que sí que trobem evidència de l'acció de selecció positiva recurrent a les regions espaiadores entre CNSs. Els constrenyiments selectius a les CNSs junt amb l'evidència d'evolució adaptativa a les regions no CNSs donen suport a un model de selecció estabilitzadora on la pèrdua de llocs d'unió a factors de transcripció (TFBSs) de les regions CNSs podria ser compensada per la fixació adaptativa de llocs d'unió específics de llinatge a les regions no CNSs, restaurant així la funció *cis*-reguladora. Si aquest tipus de selecció és freqüent, la identificació de regions *cis*-reguladores mitjançant la comparació de seqüències entre espècies no detectaria molts dels llocs funcionalment importants dels genomes i, per tant, s'haurien d'utilitzar mètodes addicionals per a completar el catàleg de regions *cis*-reguladores.

8. Tan el número de substitucions no sinònimes com el grau de constrenyiment funcional no són significativament diferents entre els gens *Hox* i els gens no *Hox*, mentre que els gens derivats de *Hox* presenten nivells més alts de substitució i un

menor constrenyiment comparat amb els altres dos grups de gens. En quan als indels, els gens *Hox* i els derivats de *Hox* contenen significativament més indels que els gens no *Hox* a les seves seqüències codificadores. Per tant, els gens *Hox* evolucionen més ràpidament del que s'esperaria donada la seva important funció en el desenvolupament primerenc.

9. Els gens *Hox* tenen una tendència a acumular indels dins les seves seqüències codificadores que podria resultar de la combinació de dos factors: (i) la composició repetitiva i composicionalment esbiaixada de moltes de les seves regions codificadores, la qual promouria mutacions del tipus indels, i (ii) una major tolerància d'aquests gens als indels fora dels seus dominis d'unió. Aquesta elevada taxa d'indels permetria als gens *Hox* (i possiblement a altres factors de transcripció, TFs) adaptar-se als canvis freqüents en la composició dels TFBSs actius, i per tant aquests resultats donen suport a una possible coevolució entre la seqüència codificadora dels TFs i els seus llocs d'unió diana, de tal manera que canvis en les distàncies nucleotídiques entre TFBSs o la substitució de TFBSs actius podrien estar acompanyats per canvis complementaris a les seqüències que separen els dominis d'unió dels TFs corresponents.

10. Les distribucions de substitucions nucleotídiques i d'indels no estan correlacionades, ni a les seqüències no codificadores ni a les codificadores, la qual cosa implica que els nucleòtids i els indels es comporten de manera diferent i s'han d'estudiar per separat. Aquesta observació, junt amb el fet que la majoria d'estudis de genètica de poblacions fins al moment es centren només en l'estudi de nucleòtids alineats, emfatitzen la necessitat de desenvolupar nous mètodes que incorporin dades d'indels per a l'estudi de l'evolució de les seqüències.

Conclusiones

1. La automatización de la estimación de la diversidad genética a partir de fuentes de datos heterogéneos es factible si se desarrollan las herramientas adecuadas de extracción, procesado, filtrado y control de calidad de los datos. Por lo tanto, los análisis a gran escala basados en estimas automatizadas son fiables y se pueden utilizar para responder hipótesis formuladas dentro del marco de la genética de poblaciones.

2. PDA es una herramienta analítica potente para obtener y sintetizar la evidencia empírica existente de la diversidad genética en cualquier especie y gen. Cabe destacar que incluye varios filtros y parámetros de calidad que superan las dificultades intrínsecas de la automatización del proceso de estimación de la diversidad genética a gran escala. Además, PDA es una herramienta útil para generar bases de datos de conocimiento a partir de datos brutos de cualquier especie o grupo de especies.

3. DPDB es la primera base de datos que permite la búsqueda de secuencias de DNA según distintos parámetros de diversidad nucleotídica, el grado de desequilibrio de ligamiento o el sesgo en el uso de codones, y permite filtrar los resultados en función de distintos criterios de confianza. También permite comparar valores de diversidad en distintas especies o grupos taxonómicos, así como generar distribuciones gráficas de cualquier medida de diversidad. DPDB es pues un recurso integral, de gran utilidad para los genetistas de poblaciones que trabajan en *Drosophila* como modelo.

4. Tanto PDA como DPDB has sido utilizados con éxito para resolver distintas cuestiones de la genética de poblaciones, algunas de las cuales contribuyen a esta tesis, y sus posibilidades de aplicación a análisis de diversidad genética a gran escala son il·limitadas.

5. Los patrones de evolución de secuencias nucleotídicas en regiones conservadas no codificadoras (CNSs) de *Drosophila* son incompatibles con la hipótesis de las regiones frías de mutación para explicar su existencia, apoyando pues la hipótesis que las CNSs son mantenidas por la acción de la selección purificadora. En concreto, estimamos que el ~85% de los sitios CNSs en *Drosophila* están constreñidos funcionalmente, con una fuerza de selección del orden de ~10-100 veces mayor que la inversa del tamaño efectivo de la población. En comparación con estudios similares en mamíferos, la fuerza de selección estimada y el número de sitios afectados por selección purificadora son mayores en las CNSs de *Drosophila* que en las de humanos, lo cual es esperado dado el mayor tamaño efectivo de la población en las moscas.

6. Los patrones de evolución de las inserciones y deleciones (indels) sugieren que los constreñimientos funcionales sobre la longitud son tan importantes en las CNSs como en las regiones espaciadoras entre ellas, reforzando pues un modelo jerárquico de regulación previamente descrito, que está constituido por regiones *cis*-reguladoras.

7. No encontramos evidencia de selección positiva actuando en las CNSs de *Drosophila*, aunque si que encontramos evidencia de la acción de selección positiva recurrente en las regiones espaciadoras entre CNSs. Los constreñimientos selectivos en las CNSs junto con la evidencia de evolución adaptativa en las regiones no CNSs apoyan un modelo de selección estabilizadora donde la pérdida de sitios de unión a factores de transcripción (TFBSs) de las regiones CNSs podría ser compensada por la fijación adaptativa de sitios de unión específicos de linaje en las regiones no CNSs, restaurando así la función *cis*-reguladora. Si este tipo de selección es frecuente, la identificación de regiones *cis*-reguladoras mediante la comparación de secuencias entre especies podría estar pasando por alto muchos de los sitios funcionalmente importantes de los genomas y, por lo tanto, se deberían utilizar métodos adicionales para completar el catálogo de regiones *cis*-reguladoras.

8. Tanto el número de sustituciones no sinónimas como el grado de constreñimiento funcional no son significativamente distintos entre los genes *Hox* y los genes no *Hox*, mientras que los genes derivados de *Hox* presentan niveles más altos de sustitución y un menor constreñimiento comparado con los otros dos grupos de genes. En cuanto a los indels, los genes *Hox* y los derivados de *Hox* contienen significativamente más indels que los genes no *Hox* en sus secuencias codificadoras. Por lo tanto, los genes *Hox* evolucionan más rápidamente de lo que se esperaría dada su importante función en el desarrollo temprano.

9. Los genes *Hox* tienen una tendencia a acumular indels dentro de sus secuencias codificadoras que podría resultar de la combinación de dos factores: (i) la composición repetitiva y composicionalmente sesgada de muchas de sus regiones codificadoras, la que promovería mutaciones del tipo indels, y (ii) una mayor tolerancia de estos genes a los indels fuera de sus dominios de unión. Esta elevada tasa de indels permitiría a los genes *Hox* (y posiblemente a otros factores de transcripción, TFs) adaptarse a los cambios frecuentes en la composición de TFBSs activos, y por lo tanto estos resultados apoyan una posible coevolución entre la secuencia codificadora de los TFs y sus sitios de unión diana, de forma que cambios en las distancias nucleotídicas entre TFBSs o la sustitución de TFBSs activos podrían ir acompañados de cambios complementarios en las secuencias que separan los dominios de unión de los TFs correspondientes.

10. Las distribuciones de sustituciones nucleotídicas y de indels no están correlacionadas, ni en las secuencias no codificadoras ni en las codificadoras, lo que implica que los nucleótidos y los indels se comportan de forma distinta y se deben estudiar por separado. Esta observación, junto con el hecho que la mayoría de estudios de genética de poblaciones hasta el momento se centran solo en el estudio de nucleótidos alineados, enfatizan la necesidad de desarrollar nuevos métodos que incorporen datos de indels para el estudio de la evolución de las secuencias.

Appendix

I

D624–D629 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 16 November 2006*
doi:10.1093/nar/gkl833

MamPol: a database of nucleotide polymorphism in the Mammalia class

Raquel Egea, Sònia Casillas, Enol Fernández[1], Miquel Àngel Senar[1] and Antonio Barbadilla*

Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain and [1]Departament d'Arquitectura d'Ordinadors i Sistemes Operatius, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

Received July 31, 2006; Revised and Accepted October 6, 2006

ABSTRACT

Multi-locus and multi-species nucleotide diversity studies would benefit enormously from a public database encompassing high-quality haplotypic sequences with their associated genetic diversity measures. MamPol, 'Mammalia Polymorphism Database', is a website containing all the well-annotated polymorphic sequences available in GenBank for the Mammalia class grouped by name of organism and gene. Diversity measures of single nucleotide polymorphisms are provided for each set of haplotypic homologous sequences, including polymorphism at synonymous and non-synonymous sites, linkage disequilibrium and codon bias. Data gathering, calculation of diversity measures and daily updates are automatically performed using PDA software. The MamPol website includes several interfaces for browsing the contents of the database and making customizable comparative searches of different species or taxonomic groups. It also contains a set of tools for simple re-analysis of the available data and a statistics section that is updated daily and summarizes the contents of the database. MamPol is available at http://mampol.uab.es/ and can be downloaded via FTP.

INTRODUCTION

Nucleotide sequences available in public databases for different organisms can be used to describe the general patterns of genetic diversity in natural populations across a wide spectrum of different taxa (1) and to infer the molecular evolutionary forces that shape the observed patterns (2,3). For this endeavor, a secondary database that provides searchable collections of polymorphic sequences with their associated genetic diversity measures would greatly facilitate both multi-locus and multi-species diversity studies. However, population geneticists still lack this basic resource.

Databases of genetic polymorphisms such as Popset (4), ALFRED (5) and dbSNP (4) rely on author submissions and contain little additional data analysis. On the contrary, Polymorphix (6) is a database that collects eukaryotic genomic DNA sequences available in EMBL/GenBank and groups them by similarity and bibliographic criteria, but does not provide any measure of sequence diversity. The only database that provides genetic diversity estimates and also permits queries about polymorphic sequences by such estimates is the *Drosophila* Polymorphism Database, DPDB (7). DPDB stores all the well-annotated nuclear sequences of the *Drosophila* genus available in GenBank, grouped by organism, gene and degree of similarity in polymorphic sets, and provides the commonly used measures of diversity. Database building and updating is totally automated using PDA software (8).

The Mammalia class is the taxonomic group with the largest amount of nucleotide information. Most intraspecies nucleotide variation in this taxon comes from the analyses of haplotypic sequences for one or more genes in a given species, but no database permits searches for polymorphic sets in accordance with different parameter values of nucleotide diversity, linkage disequilibrium or codon bias. Here we present a new database containing polymorphism data for the Mammalia class, including both nucleotide sequences and their associated diversity estimates, which was built using the DPDB database as a reference. Human data have not been included, because an extensive SNP database for human polymorphism already exists (HapMap) with more than 11 million SNPs positioned in the genome (4,9).

The MamPol database provides estimates of both one-dimensional and multi-dimensional measures of nucleotide diversity in polymorphic sets. One-dimensional measures, such as the distribution of Nei's diversity values (10) along sliding windows, permit the detection of differently constrained regions (11). Multi-dimensional measures of diversity permit searches for association among variable sites, as summarized by linkage disequilibrium estimators, providing

*To whom correspondence should be addressed. Tel: +34 93 581 2730; Fax: +34 93 581 2387; Email: antonio.barbadilla@uab.es

key information on the effective recombination and evolution of a DNA region (12).

The MamPol database was built using an optimized version of PDA v. 2 (8) that runs on a computing grid. We have also included a manually curated list of synonyms for mammalian gene names in order to detect and collect together sequences of the same gene that have been annotated differently. The database includes both nuclear and mitochondrial nucleotide sequences that can be queried independently in order to emphasize differences in their evolution due to their different origins (1). Another major improvement with respect to DPDB is the comparative search module, in which different taxa can be compared for diversity levels. All the data and results are stored in various MySQL databases that can be freely downloaded via FTP.

DATABASE BUILDING

Data retrieving

Data retrieving, calculation of the diversity measures and updating are performed by PDA (8), a pipeline made up of a set of Perl modules that automates the mining and analysis of data. PDA provides all the well-annotated genomic DNA sequences available in GenBank for mammals except for the genus *Homo*, as well as their associated information, and cross-references to Popset. To ensure that the sequences obtained are reliable, those coming from the CON, EST, GSS, HTC, HTG, PAT, STS, SYN and TPA sections are excluded before being downloaded. Entries matching the keyword 'geneID' are also excluded.

Gene synonyms and creation of polymorphic sets

Sequences are grouped by name of species and gene to create 'polymorphic sets' (7). As this process is totally automated, sequences corresponding to the same gene but with different gene names are placed in different groups. To avoid this, a list of synonymous gene names was manually created. The product of each gene name was searched in GenBank to facilitate the identification of synonymous gene names. Most of the synonymous gene names found differed only very slightly in terms of punctuation (e.g. beta casein versus beta-casein) or the order and/or length of the components (e.g. beta2 adrenergic receptor versus adrenergic receptor beta 2). Totally different gene names with the same product were putative synonymous gene names. To confirm that they correspond to the same gene, the names were consulted in the Entrez Gene section of GenBank (4) or in the BioThesaurus website (13). Both databases give species-specific gene aliases and because of that, gene aliases for one species might not be shared with others. To ensure that gene aliases comprised sequences from a single gene, a similarity search among the sequences was performed. Sequences for the same organism with 95% similarity or higher were considered as synonymous gene names and up to the 5% difference was attributed to population polymorphism. The final list available in the search pages of the website contains synonymous gene names taken from our data (except those that differ only in terms of punctuation or other small differences), the aim is to manage the database content properly without creating an extensive list.

Grouping by similarity and length

For each polymorphic set, subgroups of homologous sequences are created corresponding to the different functional regions (genes, CDSs, exons, introns, UTRs and promoters) found in the sequence annotations. Note that only sequences with functional regions in their annotations will be downloaded and grouped. Subgroups are aligned with ClustalW (14). ClustalW uses a fast and reliable multiple alignment algorithm to align sequences that supposedly are not very distant, as is the case with polymorphic sequences. ClustalW parameter values were optimized for alignments of polymorphic sequences. A 95% similarity between each pair of sequences in the alignment was fixed as the minimum percentage score (15). If the score assigned to a sequence is lower in comparison to any of the others, the sequence is extracted from the alignment. Sequences can also be substructured in different subsets. In this case, subsets are first made by considering the highest scores among the pairs of sequences, and second, their length in base pairs. Extraction from the group is random only when two sequences have the same score and length, since they supposedly contain the same amount of information.

By using this filter, most data heterogeneity can be avoided, e.g. two sequences corresponding to different genes but with the same gene alias are separated or two sequences corresponding to different parts of the same gene. However, paralogous genes, such as pseudogenes, with the same gene name annotation and those that have diverged by <5% will still be grouped together. On the other hand, highly polymorphic genes such as genes from the MHC will be grouped separately since their similarity is <95%. The reanalysis option (see Analysis Tools below) is useful for these special cases.

In order to increase the quantity of informative sites in the alignments, short sequences might be excluded from the alignments following the estimate optimization method (8). On these grounds, two or more subsets of sequences can be obtained from a given polymorphic set if sequences differ considerably in similarity and/or length. The final sets of sequences, on which estimates are performed, are called 'analysis units'.

Diversity measures and data storage

Commonly used diversity measures are calculated on these analysis units, including polymorphism at synonymous and non-synonymous sites, linkage disequilibrium and codon bias [see Table 1 in Ref. (15) for a detailed description of all the estimates].

Both primary and secondary information is stored in relational MySQL databases (for structure see the Help section of the website). Sequences, polymorphic sets and analysis units are given a single identification number to facilitate cross-database referencing. The information is divided into three databases: (i) for primates, (ii) for rodents and (iii) for all other mammals. This division is the same as that made by GenBank to store the CoreNucleotide information and the sole intention is to make searches faster and at the same time totally transparent to the user.

The databases are updated daily, searching for new sequences in GenBank and reanalyzing only the affected polymorphic sets.

Computing grid

Creation of the complete database is a fully automated and highly time-consuming process, due to the large amount of sequences that must be retrieved and analyzed from GenBank. To cope with the computational burden, a pipeline has been implemented that is able to take advantage of the multiple computational resources available in the University's campus grid. These resources consist of more than 250 laboratory computers managed by Condor software (16), a high-throughput batch queuing system.

MamPol runs multiple instances of its internal modules on this computing pool during two specific stages of the creation of the database. These two stages were selected because they were suitable for a parallel decomposition that could achieve a significant reduction in computing time. First, concurrent access to sequences in GenBank is carried out by dividing the total set into small sequence subsets. Second, alignments in each subgroup of homologous sequences are executed concurrently by running multiple instances of ClustalW on different machines.

Significant improvements were achieved in the overall performance of our Condor-enabled version of MamPol when it was used to create the database for the Mammalia class. Over the course of the computation, up to 50 machines were used during the GenBank access stage (this took 1.5 h, in contrast to the 22 h taken by the sequential case) and up to 10 machines were used during sequence alignments (∼40 h, in contrast to 245 h in the sequential case).

While updating, if the number of sequences to retrieve and/or the number of polymorphic sets to reanalyze is high, the Condor-enabled version is used.

THE MamPol WEBSITE

The MamPol website (http://mampol.uab.es/) integrates the information from the databases and offers several interfaces for browsing the contents of the database in different ways. It also includes tools for the reanalysis of polymorphic sets, a website Help section, a Statistics section in which the contents of the database are summarized and a series of links of interest classified by different categories. The database contents can also be downloaded via FTP.

Database queries and output

Queries about the contents of the database can be made using a web interface implemented as Perl CGI scripts based on SQL searches. The user can directly select the species of interest from the list of species or select a group in a higher taxonomic level in the taxonomic list. The latter is an expandable list, which includes all the taxonomic levels for the mammalian class and permits selection on any level. Gene names can also be selected from the list of genes or in the list of gene name aliases. In all these lists, mitochondrial and nuclear data are separated, as well as data for rodents, primates and all other mammals. These subdivisions are made to make searches faster and to facilitate searches of a particular subdivision, although combined queries can also be made.

General search

When selecting a polymorphic set, the user can also use filters for the diversity values and/or for the degree of confidence in the polymorphic set (see the Help section). The first output page lists all the polymorphic sets by organism, gene and analysis unit, showing additional information about the quality of the alignment, the confidence in the data source and the date of the last update. A complete report for each analysis unit can then be obtained through the corresponding link as well as access to the primary database (individual sequences, references and polymorphic studies in the Popset database). It is also possible to easily reanalyze any polymorphic set with PDA, including or excluding sequences or changing the default parameters. Furthermore, sequences can be directly downloaded in the FASTA format.

Graphical search

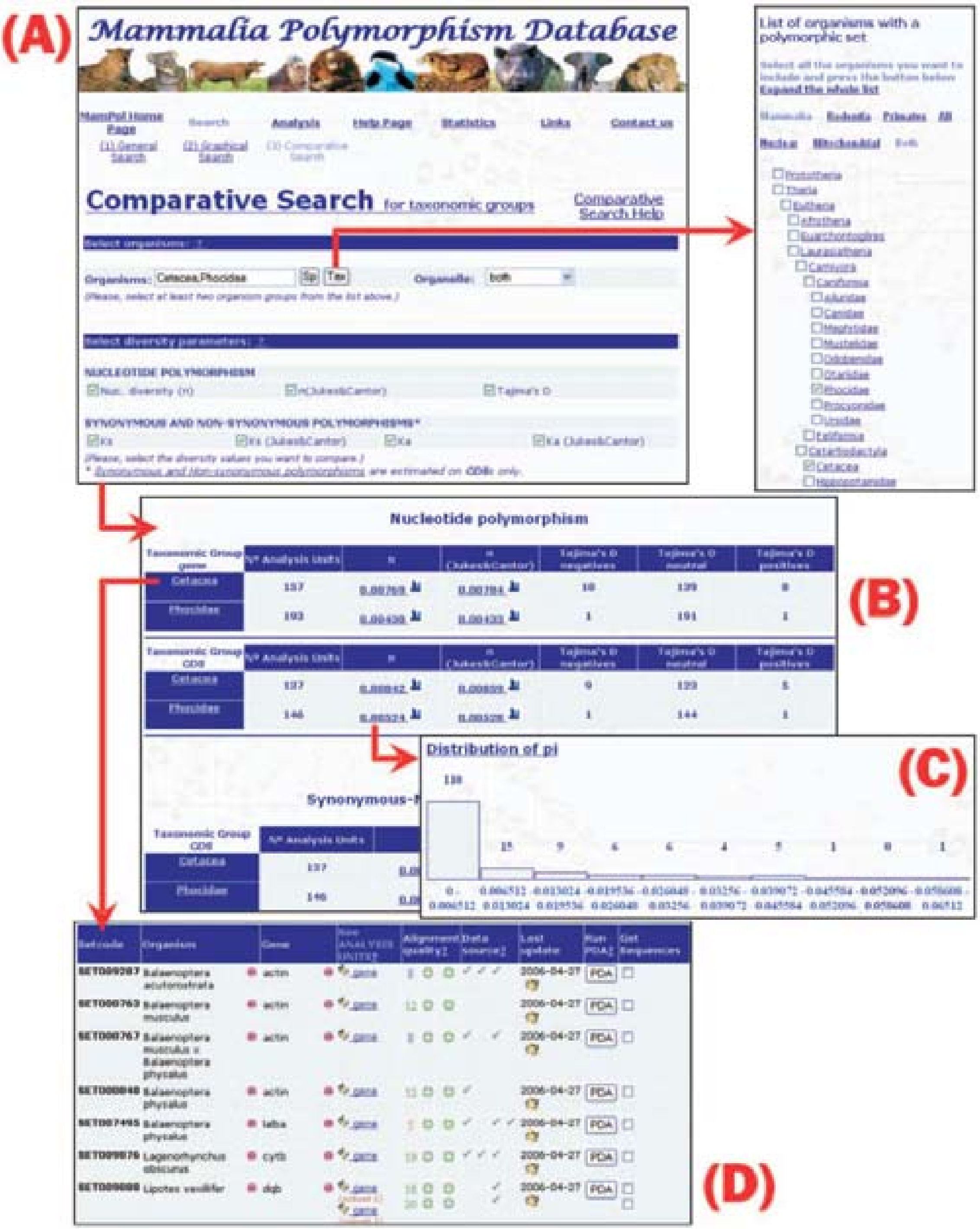
There is a graphical interface in which the user can select the graphical distribution of any of the diversity parameters estimated. Selection of the polymorphic sets and filters is the same as for the general search described above. The first output page shows the distribution of the selected parameter, which can be ordered by rank or by classes. Each class has a link for viewing the corresponding polymorphic sets as in the general search.

Comparative search

There is a totally new interface for making comparative searches among taxa (Figure 1). The user can select two or more species or taxonomic groups and compare the polymorphism levels at synonymous and non-synonymous sites, place filters for the quality of the alignments and select any functional region to be included in the search. The first output page gives the number of analysis units for each group and the mean values of the selected diversity values. Tajima's D estimates are divided into negative, zero and positive values. When the diversity mean value is different from zero and there are more than two analysis units, there is a link that displays the graphical distribution of the diversity parameter. There is also a link to the general search results page for each taxonomic group shown. Different functional regions are compared separately in order to avoid any overrepresentation of the same sequences (from different functional regions).

Analysis tools

The website includes a set of common analysis tools running on our server, therefore avoiding the need for connection to other servers. These tools are divided into different modules for sequence comparison and the estimation of nucleotide diversity. The first module includes: (i) ClustalW software (v. 1.83), with the default parameters used to create the database; and (ii) Jalview (17), which makes it possible to display and edit sequence alignments. The second module includes two other tools: (i) SNPs-Graphic, which makes it possible to perform variation analysis using the sliding window method, obtaining both estimates in different regions of the alignment and graphic representations; and (ii) the PDA pipeline, whereby the user can reanalyze the polymorphic sets by adding or deleting sequences or changing the default parameters. This is a very useful tool, especially in cases



Mammalia Polymorphism Database

MamPol Home Page | Search | Analysis | Help Page | Statistics | Links | Contact us
(1) General Search | (2) Graphical Search | (3) Comparative Search

Comparative Search for taxonomic groups

Comparative Search Help

Select organisms: ?

Organisms: Cetacea,Phocidae [Sp] [Tax] Organelle: both

(Please, select at least two organism groups from the list above.)

Select diversity parameters: ?

NUCLEOTIDE POLYMORPHISM

☑ Nuc. diversity (π) ☑ π(Jukes&Cantor) ☑ Tajima's D

SYNONYMOUS AND NON-SYNONYMOUS POLYMORPHISMS*

☑ Ks ☑ Ks (Jukes&Cantor) ☑ Ka ☑ Ka (Jukes&Cantor)

(Please, select the diversity values you want to compare.)

* Synonymous and Non-synonymous polymorphisms are estimated on CDS only.

List of organisms with a polymorphic set

Select all the organisms you want to include and press the button below

Expand the whole list

Mammalia Rodentia Primates All

Nuclear Mitochondrial Both

- □ Prototheria
- □ Theria
  - □ Eutheria
    - □ Afrotheria
    - □ Euarchontoglires
    - □ Laurasiatheria
      - □ Carnivora
        - □ Caniformia
          - □ Ailuridae
          - □ Canidae
          - □ Mephitidae
          - □ Mustelidae
          - □ Odobenidae
          - □ Otariidae
          - ☑ Phocidae
          - □ Procyonidae
          - □ Ursidae
        - □ Feliformia
      - □ Cetartiodactyla
        - ☑ Cetacea
        - □ Hippopotamidae

Nucleotide polymorphism

Taxonomic Group gene	Nº Analysis Units	π	π (Jukes&Cantor)	Tajima's D negatives	Tajima's D neutral	Tajima's D positives
Cetacea	157	0.00769	0.00784	10	139	8
Phocidae	193	0.00430	0.00433	1	191	1

Taxonomic Group CDS	Nº Analysis Units	π	π (Jukes&Cantor)	Tajima's D negatives	Tajima's D neutral	Tajima's D positives
Cetacea	137	0.00842	0.00859	9	123	5
Phocidae	146	0.00524	0.00528	1	144	1

Synonymous-N

Taxonomic Group CDS	Nº Analysis Units
Cetacea	137
Phocidae	146

Distribution of pi

Setcode	Organism	Gene	Nº ANALYSIS UNITS?	Alignment quality?	Data source?	Last update	Run PDA?	Get Sequences
SET009287	Balaenoptera acutorostrata	● actin	● gene	3 ○ ○	✓ ✓ ✓	2006-04-27	PDA	□
SET000763	Balaenoptera musculus	● actin	● gene	12 ○ ○		2006-04-27	PDA	□
SET000767	Balaenoptera musculus x Balaenoptera physalus	● actin	● gene	3 ○ ○	✓ ✓	2006-04-27	PDA	□
SET000848	Balaenoptera physalus	● actin	● gene	11 ○ ○	✓	2006-04-27	PDA	□
SET007495	Balaenoptera physalus	● lalba	● gene	1 ○ ○	✓ ✓ ✓	2006-04-27	PDA	□
SET009876	Lagenorhynchus obscurus	● cytb	● gene	13 ○ ○	✓ ✓ ✓	2006-04-27	PDA	□
SET009888	Lipotes vexillifer	● dqb	● gene gene	18 ○ ○ 20 ○ ○	✓ ✓	2006-04-27	PDA	□ □

Figure 1. Example of the MamPol interface. We illustrate a comparative search comparing two distinct taxonomic groups: Phocidae and Cetacea. (**A**) Comparative Search page (with the taxon checking list window where the two taxa are selected). (**B**) First output page of the comparative query reporting all the analysis units and estimates of diversity for each taxon. (**C**) Distribution of *Pi* values for a taxonomic group and gene region. (**D**) Partial list of all polymorphic sets in the clicked Cetacea group, with its different analysis units.

where one polymorphic set erroneously includes different paralogous genes, or one polymorphic set should be split into different analysis units in accordance with the different origins of the source sequences.

Statistics

The Statistics section summarizes the contents of both the primary and secondary databases. It is updated daily, and includes tabular and graphic information. The information is divided for rodents, primates and all other mammals, and between mitochondrial and nuclear data.

The distributions of polymorphic sets in accordance with different parameters, such as species and genes, are shown. The number of analysis units per taxon can be viewed in the 'Phylogeny of the Mammalia class' graph. Analysis units are then classified by gene region, the quality of the alignments and the confidence in the data source. Average diversity estimates by gene regions are also given. Finally, certain important statistics on the primary database are displayed, such as the number of sequences, species, genes and references, in different classifications.

At the time of writing, MamPol contained 5021 polymorphic sets corresponding to 1555 different species and 1633 different genes. A total of 15 746 analysis units were analyzed, mostly corresponding to the gene (6855), CDS (5424) and exon (2526) regions.

The statistics on the quality of the alignments show that a high percentage of analysis units have less than six sequences (86%), but most contain few gaps within the alignment (98.8%), and sequences are generally of similar length (80%). Finally, according to the confidence in the data source, only ~30% of the analysis units come from sequences whose primary focus is the study of polymorphism. Therefore, PDA has gathered the other 70% of the analysis units from sequences that would otherwise be overlooked if searching among polymorphism studies only, and it has therefore provided a notable increase in the amount of diversity data. Overall, these statistics highlight the amount and quality of the data used to estimate polymorphism in the MamPol database.

MamPol IN ACTION

The MamPol database provides estimates of nucleotide diversity for a large number of genes and species of mammals, and the website interface makes it possible to perform totally customizable queries in a single step. This greatly facilitates a wide range of large-scale analyses. For example, multi-locus polymorphic data can be used to detect adaptation on the population level and to discriminate between selection and demographic effects (18,19). On the other hand, multi-species polymorphic data make it possible to describe and compare the patterns of nucleotide diversity in organisms with different biologies, both for nuclear and mitochondrial genes. Both types of analyses may help, for example, to find a covariation between the coding and the non-coding regions of a gene, depending on different factors such as the complexity of expression [(20), Natalia Petit *et al.*, personal communication].

As an example for the use of MamPol, we show a simple study searching for evidence relating nucleotide diversity and the risk of threatened species becoming extinct. It was made using comparative queries on the website interface. Independent evidence from stochastic computer projections has demonstrated that inbreeding depression increases the risk of threatened species becoming extinct in natural habitats when all other threatening processes are included in the models. Therefore, most taxa are not driven to extinction before genetic factors have an adverse effect upon them. Spielman *et al.* (21) have recently shown that threatened species exhibit lower levels of allozyme heterozygosity in comparison with taxonomically related non-threatened taxa. By using MamPol, we have compared the corresponding levels of nucleotide diversity for these two groups and found that threatened taxa have significantly less genetic diversity than comparable non-threatened taxa (Wilcoxon's signed rank test, one-tailed $P = 0.0174$, $n = 55$) (Supplementary Table S1 and Figure S1). Specifically, nucleotide diversity was lower in threatened taxa in 70.9% of all comparisons. These differences in both heterozygosity and nucleotide diversity indicate lower evolutionary potential, higher compromised reproductive fitness and a higher risk of extinction in the wild.

This example illustrates the power of MamPol. The wide range of potential queries that can be performed on nucleotide diversity greatly facilitate comprehensive metaanalyses involving both multi-locus and multi-species polymorphic data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Natalia Petit, Elisa Heymann and two anonymous referees for valuable discussions, and David Owen for his critical reading of this manuscript. Computational resources used in this work were kindly provided by the Oliba Project of the Universitat Autonoma de Barcelona. R.E. was supported by the Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya i el Fons Social Europeu (Grant 2005FI 00328). S.C. was supported by the Ministerio de Ciencia y Tecnología (Grant BES-2003-0416). This work was supported by the Ministerio de Educación y Ciencia (Grants BFU2006-08640/BMC and TIN 2004-03388). Funding to pay the Open Access publication charges for this article was provided by the Ministerio de Educación y Ciencia (Grants BFU2006-08640/BMC awarded to Antonio Barbadilla and TIN 2004-03388 awarded to Miquel Ángel Senar).

Conflict of interest statement. None declared.

REFERENCES

1. Bazin,E., Glemin,S. and Galtier,N. (2006) Population size does not influence mitochondrial genetic diversity in animals. *Science*, **312**, 570–572.
2. McVean,G.A. and Vieira,J. (2001) Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics*, **157**, 245–257.
3. Orengo,D.J. and Aguade,M. (2004) Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. *Genetics*, **167**, 1759–1766.
4. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **34**, D173–D180.
5. Rajeevan,H., Osier,M.V., Cheung,K.H., Deng,H., Druskin,L., Heinzen,R., Kidd,J.R., Stein,S., Pakstis,A.J., Tosches,N.P. *et al.* (2003)

ALFRED: the ALelle FREquency Database. Update. *Nucleic Acids Res.*, **31**, 270–271.

6. Bazin,E., Duret,L., Penel,S. and Galtier,N. (2005) Polymorphix: a sequence polymorphism database. *Nucleic Acids Res.*, **33**, D481–D484.
7. Casillas,S., Petit,N. and Barbadilla,A. (2005) DPDB: a database for the storage, representation and analysis of polymorphism in the *Drosophila* genus. *Bioinformatics*, **21** (Suppl. 2), ii26–ii30.
8. Casillas,S. and Barbadilla,A. (2006) PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. *Nucleic Acids Res.*, **34**, W632–W634.
9. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
10. Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
11. Vilella,A.J., Blanco-Garcia,A., Hutter,S. and Rozas,J. (2005) VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*, **21**, 2791–2793.
12. McVean,G.A., Myers,S.R., Hunt,S., Deloukas,P., Bentley,D.R. and Donnelly,P. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.
13. Liu,H., Hu,Z.Z., Zhang,J. and Wu,C. (2006) BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103–105.
14. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
15. Casillas,S. and Barbadilla,A. (2004) PDA: a pipeline to explore and estimate polymorphism in large DNA databases. *Nucleic Acids Res.*, **32**, W166–W169.
16. Litzkow,M.J., Livny,M. and Mutka,M.W. (1988) *Condor—a hunter of idle workstations. Proc. of 8th International Conference on Distributed Computing Systems*, San Jose, CA, USA, pp. 104–111.
17. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
18. Hudson,R.R., Kreitman,M. and Aguade,M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.
19. Storz,J.F. and Nachman,M.W. (2003) Natural selection on protein polymorphism in the rodent genus *Peromyscus*: evidence from interlocus contrasts. *Evol. Int. J. Org. Evol.*, **57**, 2628–2635.
20. Marais,G., Nouvellet,P., Keightley,P.D. and Charlesworth,B. (2005) Intron size and exon evolution in *Drosophila*. *Genetics*, **170**, 481–485.
21. Spielman,D., Brook,B.W. and Frankham,R. (2004) Most species are not driven to extinction before genetic factors impact them. *Proc. Natl Acad. Sci. USA*, **101**, 15261–15264.

Appendix II

J Mol Evol (2007) 64:511–518
DOI: 10.1007/s00239-006-0047-5

JOURNAL OF MOLECULAR EVOLUTION

Protein Polymorphism Is Negatively Correlated with Conservation of Intronic Sequences and Complexity of Expression Patterns in *Drosophila melanogaster*

Natalia Petit, Sònia Casillas, Alfredo Ruiz, Antonio Barbadilla

Departament de Genètica i Microbiologia, Facultat de Biociències, Universitat Autònoma de Barcelona, 08193, Bellaterra, Barcelona, Spain

Received: 27 February 2006 / Accepted: 17 January 2007 [*Reviewing Editor:* Dr. Dmitri Petrov]

Abstract. We report a significant negative correlation between nonsynonymous polymorphism and intron length in *Drosophila melanogaster*. This correlation is similar to that between protein divergence and intron length previously reported in *Drosophila*. We show that the relationship can be explained by the content of conserved noncoding sequences (CNS) within introns. In addition, genes with a high regulatory complexity and many genetic interactions also exhibit larger amounts of CNS within their introns and lower values of nonsynonymous polymorphism. The present study provides relevant evidence on the importance of intron content and expression patterns on the levels of coding polymorphism.

Key words: Coding nucleotide polymorphism — Gene expression pattern — Protein evolution — Conserved noncoding sequences — *Drosophila melanogaster*

Introduction

The growing amount of annotated genomic sequences allows the study of patterns of covariation between different functional regions of the genome. Recent work is addressing the evolutionary and functional relationship between noncoding and coding sequences within a gene and the mediator role that gene expression could play in this relationship. Castillo-Davis et al. (2004) analyzed a set of orthologous genes of *Caenorhabditis elegans* and *C. briggsae* and found that nucleotide divergence is coupled between coding and cis-regulatory sequences, that is, less divergent proteins exhibit lower rates of cis-regulatory evolution. Cis-regulatory sequences were identified by looking for shared motifs (regions of high local similarity) in regions upstream of homologous genes. Likewise, Marais et al. (2005) observed a negative correlation between protein evolution, measured as nonsynonymous divergence, and intron length when comparing orthologous genes of *Drosophila melanogaster* and *D. yakuba*. They suggest that genes coding for proteins under strong selective constraint also have more cis-regulatory elements (within introns). This relationship between coding evolution and intron length might be mediated by gene expression level or regulatory complexity. On the one hand, protein evolution is associated with gene expression: negatively with the breadth in mammals (Duret and Mouchiroud 2000) and the levels in several species (Akashi 2001; Pál et al. 2001; Zhang and Li 2004; Rocha and Danchin 2004; Marais et al. 2004; Drummond et al. 2006) and positively with differences in expression levels between species in *Drosophila* and humans (Nuzhdin et al. 2004; Khaitovich et al. 2005). On the other hand, intron length and gene expression levels are negatively correlated (Castillo-Davis et al. 2002; Vinogradov 2004; Seoighe et al. 2005). Selection for economy in trasnscription had been proposed to explain the latter correlation (Akashi 2001; Castillo-Davis et al. 2002;

Correspondence to: Natalia Petit; *email:* natalia.petit@uab.es

Seoighe et al. 2005). As transcription is costly, shorter introns are selected in highly expressed genes, increasing the efficiency of the transcriptional process. However, another view (the genomic design hypothesis) points out that the shorter introns of highly expressed genes reflect the low levels of epigenetic regulation in housekeeping genes. Tissue specific genes exhibit lower levels of expression and require greater levels of epigenetic regulation (Vinogradov 2004). The genomic design hypothesis is supported by the fact that intergenic distances are also shorter in housekeeping genes than the tissue specific ones (Nelson et al. 2004; Vinogradov 2004), and this observation is not explained by the efficiency transcription model (Seoighe et al. 2005). Altogether, these observations suggest that cis-regulatory elements may play an important role in the variation dynamics of the coding sequence. Recent genomic studies are indicating the functional importance of noncoding sequences (Siepel et al. 2005; Keightley et al. 2005; Andolfatto 2005). In an evolutionary study on noncoding DNA in *Drosophila*, Andolfatto (2005) has shown that ~60% of noncoding sequences are under purifying selection, and that a significant fraction of nucleotide substitutions, nearly 20%, is due to positive selection. Thus, noncoding evolution seems to be as important, or more, for organismic evolution than that of coding sequences.

Here, we integrate genomic, polymorphism and regulatory complexity data to test the following hypotheses: (i) the amount of conserved noncoding sequences (CNS) within introns is negatively correlated with the protein polymorphism of a gene, accounting for the correlation found between protein evolution and intron length, and (ii) the previous correlation is explained by the regulatory role of conserved sequences. We use polymorphism instead of divergence data, because polymorphism is the variation stage prior to divergence, and much of nonsynonymous polymorphism is thought to be slightly deleterious and hence constrained by purifying selection (Fay and Wu 2003). Strongly deleterious and adaptive mutations are eliminated or fixed rapidly, and contrarily to divergence, they do not contribute perceptibly to polymorphism. We followed a three-step approach in this study: (1) we estimated the correlation between nonsynonymous polymorphism and intron length; (2) intronic sequences were then split into conserved and nonconserved portions to test whether the correlation estimated in step 1 can be attributed to the conserved (putative cis-regulatory) sequences; and (3) data from gene expression patterns and regulation were integrated in the analyses. This work provides relevant new evidence on the emerging view that the amount of conserved cis-regulatory elements within introns and the degree of constraint of coding sequences are coupled.

Methods

All available polymorphism coding sequences in *Drosophila melanogaster* were collected from DPDB (Drosophila Polymorphism Data Base, http//www.dpdb.uab.es; Casillas et al. 2005). After careful data filtering, 107 polymorphic genes were selected for this analysis, from which 88 contained introns (see Supplementary Table). During the data mining and analyses, manual inspection and data filtering were done to improve the suitability and confidence in the raw data: (1) the name (or alias) of each gene must be described in FlyBase (http://www.flybase.org; Drysdale et al. 2005) to be included in the analysis; (2) the reference sequences from the *D. melanogaster* sequencing project were discarded; (3) only those genes with at least five sequences were selected for analyses; (4) each alignment was compared with its corresponding reference sequence from GenBank (http://www.ncbi.nlm.nih.gov/mapview; Release 4.1) to check the correct homology of aligned sequences; and (5) for dubious alignments (for example, alignments with >10% excluded sites due to gaps or for alignments with extreme values of polymorphism), the origin of each sequence was traced. If the strain or the geographical origin of a sequence was not recorded, then the sequence was excluded from the analysis, and the remaining sequences reanalyzed again. (6) To check the suitability of the methodology, diversity estimates of genes analyzed by Moriyama and Powell (1996) were compared with our estimates, obtaining almost-coincident diversity values when the sequences were the same and close agreement when additional sequences were included.

Nei's nucleotide diversity parameter, π, was estimated for each gene in synonymous (π_s) and nonsynonymous (π_n) sites (Nei and Gojobori 1986). Polymorphism analyses were performed using PDA (Pipeline Diversity Analysis; http://www.pda.uab.es; Casillas and Barbadilla 2004). Nucleotide diversity data for chromosome X was multiplied by 4/3 to compensate for effective population size differences (Charlesworth et al. 1987).

Both total intron and transcript length were estimated from Release 4.1 of the annotated genome of *D. melanogaster*. Sequences were considered intronic if they were located between exons, between UTRs, or between UTRs and exons. When more than a transcript was annotated, the longest transcript was considered for the estimation of the intron size.

The searching of conserved noncoding sequences within introns was carried out using the Vista Genome Browser (http://www.genome.lbl.gov/vista; Couronne et al. 2003), which is a very useful and widely used tool in comparative genomics. We compared the *D. melanogaster* genome with the available genome data from six other species (*D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. virilis*, and *D. mojavensis*). For the comparison between *D. melanogaster* and *D. yakuba* species, the size of the sliding window used to calculate conservation scores of each base pair (Calc windows) was 100 bp; the minimum width of a conserved region (Min Cons Width) was 100 bp, and the minimum percentage conservation identity that must be maintained over the window (Cons Identity) was 70%. A CNS length measure was estimated per gene by summing up the length of all conserved blocks within the introns of the gene, discarding regions with repeated sequences as determined in Vista Genome Browser. For comparisons between *D. melanogaster* and the other *Drosophila* species, the default parameter values were modified to detect smaller conserved blocks between more phylogenetically distant species: Calc windows = 50 bp, Min Cons Width = 25 bp, and Cons Identity = 90%. Three different groups of genes were defined: (1) genes without introns, (2) genes without CNS within introns, and (3) genes with CNS within introns, for all comparisons of *D. melanogaster* to the other six non-*melanogaster* species.

Data on expression patterns were obtained from Flybase (Drysdale et al. 2005). Nelson et al. (2004) devised an index of expression pattern (FBx index) by counting the number of mutant phenotypes of embryos, larvae, and adults affecting different expression domains and tissues. Data were obtained from the section "Expression and Phenotype" of Flybase. This index is a surrogate measure of regulatory complexity, because genes expressed in a greater number of tissues and domains tend to require a greater number of regulatory elements to conduct their expression. To correct for potential bias of overrepresentation of some genes, Nelson et al. (2004) used the Fbxbin index that groups the genes in 10 categories (Bin 1 = 1 record; Bin 2 = 2 records, Bin 3 = 3 records, Bin 4 = 4–5 records, Bin 5 = 6–8 records, Bin 6 = 9–13 records, Bin 7 = 14–18 records, Bin 8 = 19–29 records, Bin 9 = 30–49 records, and Bin 10 = > 50 records). They found that genes with complex functions (e.g., developmental genes), and expressed in a wide variety of specific tissues, have a higher FBxbin index and are flanked by longer noncoding DNA than genes with simple or housekeeping functions. Therefore, we used Fbxbin as an indicator of regulatory complexity in our gene set and tested for a correlation with nonsynonymous polymorphism.

To test for differences in biological functions between the groups of genes with or without CNS within introns, Gene Ontology terms (Ashburner et al. 2000) were got using FatiGo (http://www.fatigo.bioinfo.cipf.es; Al-Shahrour et al. 2005). A test for differences of unequal distribution of terms between the two groups of genes was performed. FatiGo uses Fisher's exact test for 2 × 2 contingency tables and calculates the significant differences in gene ontology term distribution between two groups of genes using methods adjusted for multiple tests.

For the estimation of recombination rates we used the software Recomb-Rate (Comeron et al. 1999), which takes into account the cytological localization of the genes. The program considers that recombination rate is proportional to the amount of DNA in each division along the chromosome versus the change of position of the genetic map (Kliman and Hey 1993).

Nonparametric tests were usually performed on the data because of the deviation from normal distribution of the variables.

Results and Discussion

Relationship Among Intron Length, Conserved Sequence Within Introns, and Nonsynonymous Polymorphism

An association between protein evolution and length of introns was reported by Marais et al. (2005) for *D. melanogaster*. Besides this correlation, intron length had been found to be associated with recombination rates (Carvalho and Clark 1999; Comeron and Kreitman 2000) and levels of expression (Castillo-Davis et al. 2002; Vinogradov 2004; Seoighe et al. 2005; Marais et al. 2005). We have tested the association between intron length and protein polymorphism, and assessed different possible explanations for this association. The correlation between nonsynonymous polymorphism (π_n) and total intron length was highly significant ($r_{\text{Spearman}} = -0.412$, $p < 10^{-4}$, $N = 107$; Fig. 1) and remained so even when intronless genes were excluded from the analysis ($r_{\text{Spearman}} = -0.527$, $p < 10^{-4}$, $N = 85$). Hence, genes with longer introns show lower levels of nonsynonymous polymorphism. This is in agreement with the negative correlation found by Marais et al. (2005) between nonsynonymous divergence (d_n) and intron length. However, the absolute value of our correlation was more than twice theirs (–0.412 vs –0.19). This difference could be due to sporadic adaptive mutations contributing to divergence which loosen the correlation. For comparison, we estimated the correlation between intron length and d_n between *D. melanogaster* and *D. yakuba* for the genes analyzed here. The value of this correlation was not significantly different from that of Marais et al. (2005) ($r_{\text{Spearman}} = -0.237$, $p = 0.03$, $N = 84$, vs $r_{\text{Spearman}} = -0.190$, $p < 10^{-4}$, $N = 570$; χ^2 test for homogeneity of two correlation coefficients, $p = 0.549$ [Sokal and Rholf 1981]), indicating that the length of introns in our dataset is not biased.

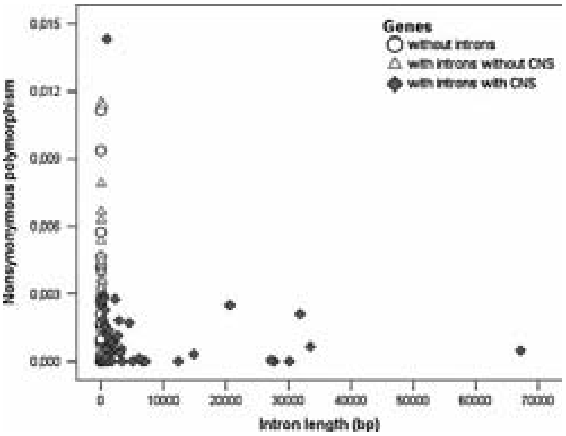


Fig. 1. Relationship between the level of nonsynonymous polymorphism and intron length for the 107 genes analyzed in *Drosophila melanogaster* ($r_{\text{Spearman}} = -0.412$, $p < 10^{-4}$, $N = 107$). Means and standard deviations are given in Table 1. Symbols refer to the different gene groups: circles, genes without introns; triangles, genes with introns without CNS; and diamonds, genes with introns with CNS.

A possible explanation for the negative correlation between nonsynonymous polymorphism and intron length is the presence of regulatory sequences within introns, which would increase the length of introns in genes with complex patterns of expression (Marais et al. 2005; Vinogradov 2004). CNS are though to be cis-regulatory elements of the expression of a gene (Hardison 2000; Bergman et al. 2002; Negre et al. 2005). Thus, to test for the hypothesis that the correlation between intron length and protein evolution is due to the presence of regulatory elements within introns, we searched for CNS within introns by comparing orthologous gene sequences between *D. melanogaster* and *D. yakuba*. The CNS total length within introns was calculated for each gene by summing up the length of all conserved blocks detected within introns (see Methods). Nonsynonymous polymorphism (π_n) was negatively correlated with CNS length ($r_{\text{Spearman}} = –0.346$, $p = 0.001$, $N = 83$). We also found a high correlation value between the length of introns and the CNS length ($r_{\text{Pearson}} = 0.946$, $p < 10^{-6}$, $N = 83$), which is consistent with the observation of Haddrill et al. (2005)

Table 1. Test for differences among the means of the three groups of genes for the different analyzed variables

	Gene group: mean ± SD (N)			p-value	
	Without introns (1)	With introns without CNS (2)	With introns with CNS (3)	ANOVA Kruskall-Wallis (gene groups 1–3)	Kolmogorov-Smirnov (Gene groups 2 vs 3)
π_{n}[a]	0.0027 ± 0.0031 (20)	0.0024 ± 0.0029 (62)	0.0006 ± 0.0008 (23)	**0.010**	**< 0.01**
π_{s}[a]	0.0137 ± 0.0113 (20)	0.0180 ± 0.0167 (62)	0.0113 ± 0.0147 (23)	0.202	> 0.1
$\pi_{\mathrm{n}}/\pi_{\mathrm{s}}$[a]	0.221 ± 0.248 (18)	0.283 ± 0.609 (54)	0.068 ± 0.107 (20)	**0.037**	< 0.1
Total intron length (bp)[b]	—	912 ± 1975 (62)	11505 ± 16076 (23)	—	**< 0.001**
Expression pattern index[c]	4.67 ± 3.88 (6)	3.89 ± 2.96 (40)	7.58 ± 2.47 (16)	**0.013**	**< 0.005**
Recombination rate[d]	0.0019 ± 0.0014 (20)	0.0024 ± 0.0016 (62)	0.0027 ± 0.0025 (23)	0.134	> 0.1

Note. Gene groups were defined according to the presence of introns and CNS in all comparisons between the *D. melanogaster* genome and six other *Drosophila* species using the Vista genome browser (see Methods). Significant p-values are in boldface.

[a] π_{n}, nonsynosnymous polymorphism; π_{s}, synosnymous polymorphism; $\pi_{\mathrm{n}}/\pi_{\mathrm{s}}$, relationship between the two types of polymorphism sites (selective constraints).

[b] Total intron lengths (base pairs) estimated from Release 4.1 of the annotated genome of *D. melanogaster* (see Methods).

[c] FBxbin index as defined by Nelson et al. (2004) (see Methods).

[d] Recombination rate estimated using Recomb-rate program of Comeron et al. (1999).

that long introns (> 86 bp) are more conserved than short introns. To elucidate the importance of conserved and nonconserved intronic regions in the correlation between π_{n} and total intron length, we estimated the correlation of π_{n} with the residuals of the regression line that predicts total intron length from CNS content. When this was done, the correlation between π_{n} and (corrected) total intron length vanished ($r_{\mathrm{Spearman}} = 0.014$, $p = 0.890$, $N = 83$). Therefore, the original correlation between π_{n} and total intron length could be ascribed to the CNS content of introns. In Fig. 1 we show that genes with longer introns are those with CNS within introns.

We tested the possibility that proteins with alternative splicing in the dataset ($N = 12$) could be biasing our results, because different selective pressures in alternative introns and exons could influence the nonsynonymous polymorphism detected. However, when only genes without known alternative splicing were analyzed ($N = 73$), the correlation between CNS length and π_{n} remained significant ($r_{\mathrm{Spearman}} = -0.347$, $p = 0.003$).

The divergence time between *D. melanogaster* and *D. yakuba* is nearly 6 million years (Smith and Eyre-Walter 2002), and most of the CNS detected in long introns could be contingent on the short divergence time. To avoid this potential bias, we searched for intronic CNS between *D. melanogaster* and other *Drosophila* species with different times of divergence, whose genomes are sequenced and aligned with *D. melanogaster* in the Vista Genome Browser (http://www.genome.lbl.gov/vista; Couronne et al. 2003). As our hypothesis assumes that the correlation between π_{n} and total intron length can be explained by the regulatory nature of CNS, we diminished the window size and increased the percentage of identity in the Vista Genome Browser for these comparisons (see Methods). As expected, the number of genes with introns bearing CNS decreases with phylogenetic distance: 61 with *D. yakuba*, 58 with *D. erecta*, 35 with *D. ananassae*, 25 with *D. pseudoobscura*, and 24 with *D. virilis* and *D. mojavensis*. The diminution could indicate either that much of the sequence is not functional and is neutrally diverging or that the evolution of these sequences is linage specific. The latter case would imply a rapid divergence of regulatory elements, in agreement with the recent estimates of Andolfatto (2005) that about 20% of intron substitutions between *D. melanogaster* and *D. simulans* are adaptive, even though we still detect 23 genes (26% of our data set) containing intronic CNS in all the analyzed species. The π_{n} value of this gene group is significantly lower than both the group of genes without introns and the group of genes with introns but without CNS (ANOVA Kruskall-Wallis, $p = 0.010$; Table 1, Fig. 2a). The average length of introns also differs between genes with and genes without CNS (Kolmorogov-Smirnov test, $p < 0.001$; Table 1). We conclude that genes with intronic conserved sequences that are putatively functional have long introns and low values of nonsynonymous polymorphism.

Evaluation of the Differences in Background Mutation Rate in Genes With and Without CNS

Heterogeneity in mutation rates between different gene regions might also explain our results (Clark 2001). A low mutation rate along a gene could produce a correlation between CNS amount and nonsynonymous polymorphism when analyzed with other genes with higher background mutation rate. To test this possibility, we compared synonymous polymorphism (π_{s}, which can be considered almost neutral) and the ratio $\pi_{\mathrm{n}}/\pi_{\mathrm{s}}$ (which can be taken as a measure of selective constraint) among the three gene groups (genes without introns, genes with introns without CNS, and genes with introns with CNS). The results showed that synonymous variation does not differ significantly among the three gene groups (ANOVA Kruskall-Wallis, p = 0.180; Table 1). On the contrary, genes with intronic CNS had significantly lower values of π_n/π_s than the other two gene groups (ANOVA Kruskall-Wallis, p = 0.037; Table 1). This group effect would not be expected if genes with CNS had a lower background mutation, and it is consistent with the hypothesis that genes with CNS are more constrained.

Evaluation of the Effect of the Recombination Rate in the Association Between CNS Content Within Introns and Nonsynonymous Polymorphism

Recombination rate has been found to be positively correlated with levels of nucleotide polymorphism (Begun and Aquadro 1992; Moriyama and Powell 1996) and negatively associated with intron length (Carvalho and Clark 1999; Comeron and Kreitman 2000) in *D. melanogaster*. Likewise, differences in evolution rates have been detected between genes with low and genes with high recombination rates (Presgraves 2005). Thus, a correlation between nucleotide polymorphism and intron length might be due to recombination rate variation. To test if the recombination rate can account for our results, we have estimated the correlation between recombination rate and coding polymorphism and between recombination rate and intron length in our data set. Recombination rate was positively correlated with synonymous polymorphism (π_s) (r_{Spearman} = 0.345, p = < 10^{-4}, n = 107) but not with nonsynonymous polymorphism π_n (r_{Spearman} = 0.16, p = 0.099, n = 107). The correlation between recombination rate and total intron length was also not significant (r_{Spearman} = −0.110, p=0.307, n=85). Furthermore, significant differences were not found in recombination rate between genes with and genes without intronic CNS (ANOVA Kruskall-Wallis, p = 0.134; Table 1). Therefore, recombination rate does not seem to influence the association found between π_n and total intron length or CNS content.

Relationship Among Expression Patterns, Regulation, and Nonsynonymous Polymorphism

Our results confirm two previous observations: (1) longer introns are less variable (Haddrill et al. 2005), and (2) longer introns are associated with lower values of coding variation (Marais et al. 2005). Our second observation extends the evidence from interspecific divergence to intraspecific polymorphism. Marais et al. (2005) proposed that longer introns contain a larger number of elements that regulate the expression of genes bearing them. This agrees with our analysis of CNS as indicators of regulatory elements (Hardison 2000; Bergman et al. 2002; Negre et al. 2005). To further support this assumption, we have incorporated in our analyses data on the gene expression pattern of *D. melanogaster* obtained from Flybase (Drysdale et al. 2005), following the approach of Nelson et al. (2004), to measure the regulatory complexity of genes. In these data, higher values of the expression pattern index (FBxbin index; see Methods) denote higher regulatory complexity. Table 1 and Fig. 2b show the comparison of expression patterns calculated for the three gene groups. Genes with CNS in their introns exhibit a significantly higher complexity in their expression pattern than genes without CNS or without introns (ANOVA Kruskall-Wallis, p = 0.013). Figure 3 shows the association among the three analyzed variables: nonsynonymous polymorphism, expression pattern index, and CNS content within introns. The corre-

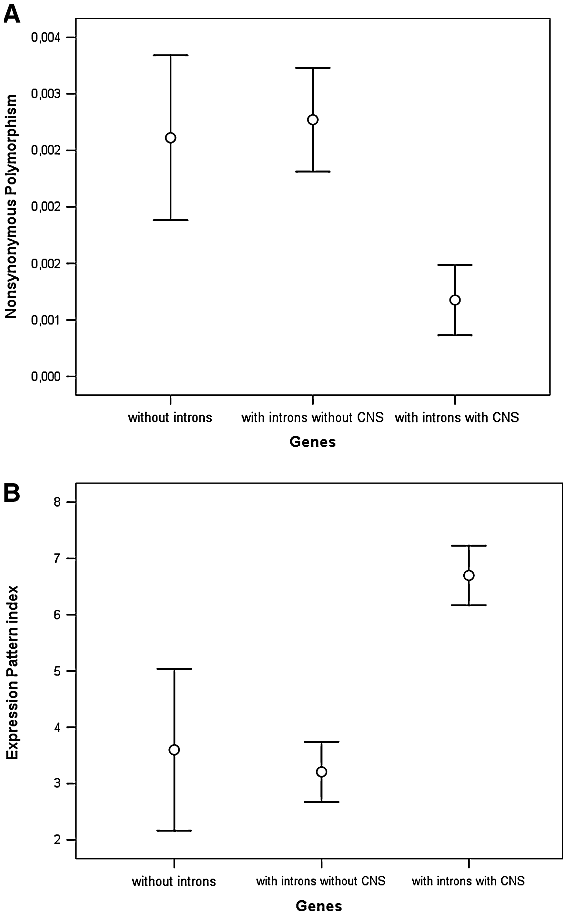


Fig. 2. Mean and standard errors of (**a**) nonsynonymous polymorphism and (**b**) expression pattern index (FBxbin) of the three groups of genes. Bars indicate two standard errors. The mean, standard deviation, sample size, and p-values are given in Table 1.

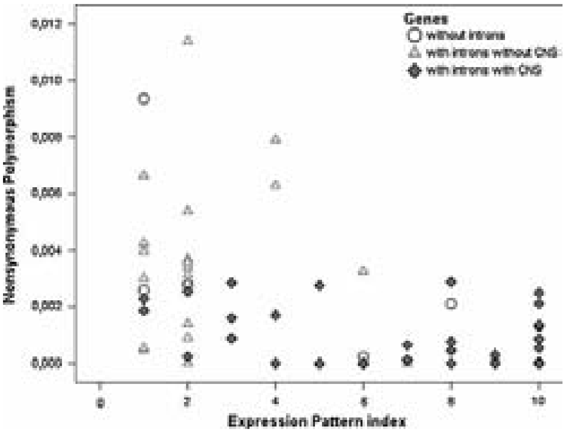


Fig. 3. Nonsynonymous polymorphism is negatively associated with expression pattern index ($r_{\mathrm{Spearman}} = -0.525$, $p < 10^{-4}$). Genes with high regulatory complexity have CNS and low nonsynonymous polymorphism. Circles, genes without introns; triangles, genes with introns without CNS; and diamonds, genes with introns with CNS.

lation between nonsynonymous polymorphism and the expression pattern index is negative and significant ($r_{\mathrm{Spearman}} = -0.525$, $p < 10^{-4}$, $N = 62$). Furthermore, as expected if CNS are indicators of regulatory elements, the correlation between expression pattern index and CNS length is positive and significant ($r_{\mathrm{Spearman}} = 0.579$, $p < 10^{-4}$, $N = 47$). Nelson et al. (2004) showed that the amount of noncoding DNA between a gene and its nearest neighbor correlates positively with the regulatory complexity of that gene. We found a similar correlation between conserved noncoding sequence length within introns and regulatory complexity of a gene, extending the results of Nelson et al. (2004) from length of the intergenic noncoding sequences to length of introns. Our results show that genes with a higher expression pattern index are also genes with longer introns and lower nonsynonymous polymorphism.

Duret and Mouchiroud (2000) analyzed the protein evolution of a wide set of human/rodent and mouse/rat orthologous genes among 19 tissues from three developmental stages. They showed that substitution rates at nonsynonymous sites are negatively correlated with tissue distribution breath. Moreover, they reported that these broadly expressed genes have more regulatory elements in their 3′ UTR than tissue specific genes. The interpretation of these observations was that the efficiency of selection increases with tissue distribution for coding sequences as for regulatory elements. Accordingly, our results indicated that genes with more mutant phenotypes detected in different tissues and developmental stages have more putative regulatory sequences within introns and are more constrained.

Why Do Proteins with High Regulatory Complexity Evolve Slowly?

We have shown that genes with longer introns have lower levels of nonsynonymous polymorphism and that this association can be explained by the regulatory content of introns. Genes with high regulatory complexity could have a wide range of functions in different tissues and developmental stages and therefore be more constrained. We tested for possible differences in the biological function of the genes with versus without CNS in their introns. We used the FatiGo server (Al-Shahrour et al. 2005; see Methods) to test unequal distribution of Gene Ontology terms between the two groups of genes (with or without CNS). The results indicated that the group of genes with CNS in their introns is significantly enriched in gene ontology terms: ‘‘organ development,’’ ‘‘tube development,’’ ‘‘mesoderm development,’’ ‘‘organ morphogenesis,’’ ‘‘cell fate determination,’’ ‘‘migration,’’ ‘‘motility,’’ ‘‘locomotion,’’ ‘‘localization of cell,’’ ‘‘regulation of cellular physiological process,’’ ‘‘regulation of cellular process,’’ and ‘‘regulation of physiological process’’ (Fisher exact test adjusted to multiple test, $p = 0.03$; genes with CNS, $N = 19$; genes without CNS, $N = 52$). Therefore, genes belonging to the group with CNS seem be functionally complex and involved in the regulatory and developmental process, which would constrain their evolution.

Functionally complex genes have longer coding sequences and longer introns than housekeeping genes in humans (Vinogradov 2004). Therefore, coding length is not a negligible variable in our analysis and is expected to be related to nonsynonymous polymorphism and, also, regulatory complexity. In fact we found a significant positive correlation between total coding length and total intron length and between coding length and FBxbin index ($r_{\mathrm{Spearman}} = 0.669$, $p < 10^{-5}$, $N = 85$, and $r_{\mathrm{Spearman}} = 0.468$, $p < 10^{-3}$, $N = 62$, respectively) and a significant negative correlation between coding length and π_{n} ($r_{\mathrm{Spearman}} = -0.453$, $p < 10^{-4}$, $N = 107$). Thus, our results indicate that longer proteins are related to longer introns and high regulatory complexity. However, highly expressed genes, such as housekeeping genes, have shorter introns (Castillo-Davis 2002; Vinogradov 2004) and evolve slowly in several species (Akashi 2001; Pál et al. 2001; Zhang and Li 2004; Rocha and Danchin 2004; Marais et al. 2004; Drummond et al. 2006). The latter associations seem to be contradictory with our results. However, the dot distribution between total intron length or FBxbin and π_{n} in Figs. 1 and 3 is L-shaped, indicating that low π_{n} values can be found in genes both with and without CNS. The point is that genes with higher regulatory complexity and CNS within introns will almost indefectibly have low nonsynonymous variation.

Conclusions

Our results extend the negative correlation between coding evolution and intron length found by Marais et al. (2005) from nonsynonymous divergence to nonsynonymous polymorphism and give support to the hypothesis that the correlation could be accounted for the regulatory content of introns. We show that intronic CNS content could explain the association between intron length and nonsynonymous variation. All the evidence together indicates that longer introns seem to contain regulatory elements that modulate the expression of genes. Supporting this view is the example of Pappu et al. (2005), who reported an intronic enhancer within an intron of *Drosophila melanogaster*, which directs the eye-specific expression of the *dac* locus. Higher amounts of conserved noncoding sequences within introns could, therefore, be indicating higher levels of regulatory complexity. Overall, proteins with a higher regulatory complexity are longer and seem to be functionally complex and more constrained by purifying selection. Our results emphasize the importance of intron content in the evolution of coding sequences, suggest that purifying selection is the principal force acting in the evolution of genes with high regulatory complexity, and support the emerging view that genetic variation within and among species results from the coupled evolution of the proteome and the transcriptome.

Acknowledgments. The authors would like to thank Raquel Egea and Jordi Pijoan for their help in the manual revision of the alignments and testing of the dataset; Brian Charlesworth for helpful discussions and comments on this work during his stay at our lab; and Marta Puig, Dmitri Petrov, Hiroshi Akashi, and an anonymous reviewer for their useful comments. This work was funded by the Ministerio de Ciencia y Tecnología (Grant BES–2003–0416) and Ministerio de Educación y Ciencia (Grant BFU-2006-08640). N.P. was supported by a grant from the Departament de Genètica i Microbiologia of the Universitat Autònoma de Barcelona, and S.C. by the Ministerio de Ciencia y Tecnología (Grant BES-2003-0416).

References

Akashi H (2001) Gene expression and molecular evolution. Curr Opin Genet Dev 1:660–666

Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Nucleic Acids Res 33:W460–W464

Andolfatto P (2005) Adaptative evolution of non-coding DNA in Drosophila. Nature 437:1149–1152

Ashburner MC, Ball A, Blake JA, Botstein D, The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29

Begun DJ, Aquadro CF (1992) Levels of naturally ocurring DNA polymorphism correlate with recombination rates in Drosophila melanogaster. Nature 5:19–52

Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, Stapleton M, Wan K, George RA, de Jong PJ, Botas J, Rubin GM, Celniker SE (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. Genome Biol 3:research0086.20

Carvalho AB, Clark AG (1999) Intron size and natural selection. Nature 401:344–345

Casillas S, Barbadilla A (2004) PDA: a pipeline to explore and estimate polymorphism in large DNA datasets. Nucleic Acids Res 32:W166–W169

Casillas S, Petit N, Barbadilla A (2005) DPDB: a database for the storage, representation and analysis of polymorphism in the Drosophila genus. Bioinformatics 21:ii26–ii30

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA (2002) Selection for short introns in highly expressed genes. Nat Genet 31:415–418

Castillo-Davis CI, Hartl DL, Achaz G (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res 14:1530–1536

Charlesworth B, Coyne JA, Barton NH (1987) The relative rates of evolution of sex-chromosomes and autosomes. Am Nat 130:113–146

Clark AG (2001) The search for meaning in noncoding DNA. Genome Res 11:1319–1320

Comeron JM, Kreitman M (2000) The correlation between Length of intron and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. Genetics 156:1175–1190

Comeron JM, Kreitman M, Aguadé M (1999) Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics 151:239–249

Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin E, Pachter L, Dubchak I (2003) Strategies and tools for whole-genome alignments. Genome Res 13:73–80

Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23:327–337

Drysdale RA, Crosby MA, The Flybase Consortium (2005) FlyBase: genes and gene models. Nucleic Acids Res 33:D390–D395

Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression patterns affect selection intensity but not mutation rates. Mol Biol Evol 17:68–74

Fay JC, Wu C (2003) Sequence divergence, functional constraint, and selection in protein evolution. Annu Rev Genom Hum Genet 4:213–235

Haddrill P, Charlesworth B, Halligan DL, Andolfatto P (2005) Patterns of intronic sequence evolution in Drosophila are dependent upon length and GC content. Genome Biol 6:R67

Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16:369–372

Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ (2005) Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res 15:1373–1378

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309:1850–1854

Kliman RM, Hey J (1993) Reduced natural selection associated with low recombination in Drosophila melanogaster. Mol Biol Evol 10:1239–1258

Marais G, Domazet-Loso T, Tautz D, Charlesworth B (2004) Correlated evolution of synonymous and nonsynonymous sites in Drosophila. J Mol Evol 59:771–779

Marais G, Nouvellet P, Keightley PD, Charlesworth B (2005) Intron size and exon evolution in Drosophila. Genetics 170:481–485

Moriyama EN, Powell JR (1996) Intraspecific nuclear DNA variation in Drosophila. Mol Biol Evol 13:261–277

Negre B, Casillas S, Suzanne M, Sanchez-Herrero E, Akam M, Nefedov M, Barbadilla A, de Jong P , Ruiz A (2005) Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex. Genome Res 15:692–700

Nei M, Gojobori T (1986) Simple methods for estimating the number of synonymous and no synonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Nelson CE, Hersh BM, Carroll SB (2004) The regulatory content of intergenic DNA shapes genome architecture. Genome Biol 5:R25

Nuzhdin SV, Wayne ML, Harmon KL, Mcintyre LM (2004) Common pattern of evolution of gene expression level and protein sequence in Drosophila. Mol Biol Evol 21:1308–1317

Pal C, Papp B, Hurst LD (2001) Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. Mol Biol Evol 18:2323–2326

Pappu KS, Ostrin EJ, Middlebrooks BW, Sili BT, Chen R, Atkins MR, Gibbs R, Mardon G (2005) Dual regulation and redundant function of two eye-specific enhancers of the Drosophila retinal determination gene dachshund. Development 132:2895–2905

Presgraves DC (2005) Recombination enhances protein adaptation in *Drosophila melanogaster*. Curr Biol 15:1651–1656

Rocha EP, Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21:108–116

Seoighe C, Gehring C, Hurst LD (2005) Gametophytic selection in Arabidopsis thaliana supports the selective model of intron length reduction. PLoS Genet 1:e13

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionary conserved elements in vertebrate, insect, worm and yeast genomes. Genome Res 15:1034–1050

Smith NG, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415:1022–1024

Sokal RR, Rholf FJ (1981) Biometry: the principles and practice of statistics in biological research. 2nd ed. W. H. Freeman, New York

Vinogradov AE (2004) Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet 20:248–253

Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol 21:236–239

Appendix III

Conservation of regulatory sequences and gene expression patterns in the disintegrating *Drosophila Hox* gene complex

Bárbara Negre,[1] Sònia Casillas,[1] Magali Suzanne,[2] Ernesto Sánchez-Herrero,[2] Michael Akam,[3] Michael Nefedov,[4] Antonio Barbadilla,[1] Pieter de Jong,[4] Alfredo Ruiz[1,5]

[1]Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain; [2]Centro de Biología Molecular "Severo Ochoa", Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain; [3]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, United Kingdom; [4]Children's Hospital Oakland Research Institute, Oakland 94609, California, USA

Homeotic (*Hox*) genes are usually clustered and arranged in the same order as they are expressed along the anteroposterior body axis of metazoans. The mechanistic explanation for this colinearity has been elusive, and it may well be that a single and universal cause does not exist. The *Hox*-gene complex (HOM-C) has been rearranged differently in several *Drosophila* species, producing a striking diversity of *Hox* gene organizations. We investigated the genomic and functional consequences of the two HOM-C splits present in *Drosophila buzzatii*. Firstly, we sequenced two regions of the *D. buzzatii* genome, one containing the genes *labial* and *abdominal A*, and another one including *proboscipedia*, and compared their organization with that of *D. melanogaster* and *D. pseudoobscura* in order to map precisely the two splits. Then, a plethora of conserved noncoding sequences, which are putative enhancers, were identified around the three *Hox* genes closer to the splits. The position and order of these enhancers are conserved, with minor exceptions, between the three *Drosophila* species. Finally, we analyzed the expression patterns of the same three genes in embryos and imaginal discs of four *Drosophila* species with different *Hox*-gene organizations. The results show that their expression patterns are conserved despite the HOM-C splits. We conclude that, in *Drosophila*, *Hox*-gene clustering is not an absolute requirement for proper function. Rather, the organization of *Hox* genes is modular, and their clustering seems the result of phylogenetic inertia more than functional necessity.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. AY900631–AY900632 and AY897430–AY897434.]

Homeotic (*Hox*) genes were discovered in *Drosophila melanogaster* as mutations that transform one body part into another. Lewis (1978) and Kaufman et al. (1980) found that these genes are clustered and arranged in the chromosome in the same order as their domains of action in the body of flies. Homologous *Hox* genes were subsequently found in many other animals and their arrangement in complexes (HOM-C) shown to be the general rule (McGinnis and Krumlauf 1992; Ruddle et al. 1994). *Hox* genes encode transcription factors involved in the determination of segment identity along the anteroposterior body axis, and thus, play a fundamental role in animal development. The conserved colinearity between *Hox* gene chromosomal arrangement and expression domain is a basic notion of developmental biology, yet this is an enigmatic phenomenon for which no single satisfactory explanation exists (Kmita and Duboule 2003). Furthermore, HOM-C splits have been observed in *Drosophila* (Von Allmen et al. 1996; Lewis et al. 2003; Negre et al. 2003), *Bombyx* (Yasukochi et al. 2004), nematodes (Aboobaker and Blaxter 2003), and tunicates (Ikuta et al. 2004; Seo et al. 2004).

Ten genes arranged in a single complex comprised the ancestral HOM-C of arthropods (Cook et al. 2001; Hughes and Kaufman 2002; Hughes et al. 2004). In winged insects, including *Drosophila*, the genes *Hox3* and *fushi tarazu* (*ftz*) lost their homeotic function, and thus, only eight truly homeotic genes remain. Three different splits of the ancestral HOM-C have been found so far in the *Drosophila* genus (Fig. 1A). In *D. melanogaster*, the complex is split between the genes *Antennapedia* (*Antp*) and *Ultrabithorax* (*Ubx*), leaving two separate gene clusters as follows: the Antennapedia complex, ANT-C (Kaufman et al. 1990) that specifies the identity of the mouth parts and anterior thorax, and the Bithorax complex, BX-C (Duncan 1987; Martin et al. 1995) involved in the development of the posterior thorax and abdomen. In *D. pseudoobscura*, the HOM-C is also similarly divided in the ANT-C and BX-C complexes (Lewis et al. 2003). A different split between *Ubx* and *abdominal A* (*abdA*) occurs in *D. virilis* (Von Allmen et al. 1996), *D. repleta* (Ranz et al. 2001), *D. buzzatii*, and other species of the *Drosophila* subgenus (Negre et al. 2003; Fig. 1B). Finally, an additional split, between *labial* (*lab*) and *proboscipedia* (*pb*), is present in *D. buzzatii* and other species of the *repleta* group (Negre et al. 2003). This third split separated the gene *lab* far from *pb* and the anterior genes of the *Hox* complex and relocated it near the posterior genes *abdA* and *Abdominal B* (*AbdB*) in a flagrant violation of the colinearity rule. The functional consequences of these splits are unknown.

[5]**Corresponding author.**
E-mail Alfredo.Ruiz@uab.es; fax 0034-93-581-23-87.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3468605.

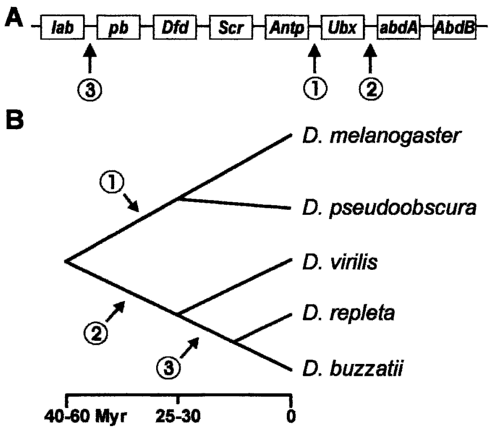


Figure 1. Genomic (*A*) and phylogenetic (*B*) localization of the three *Hox* gene complex splits observed in the *Drosophila* genus. (*A*) Ancestral arrangement of the eight *Hox* genes within the insects is as follows: *labial* (*lab*), *proboscipedia* (*pb*), *Deformed* (*Dfd*), *Sex combs reduced* (*Scr*), *Antennapedia* (*Antp*), *Ultrabithorax* (*Ubx*), *abdominal A* (*abdA*), and *Abdominal B* (*AbdB*). (*B*) Phylogenetic relationships and divergence times for the five *Drosophila* species included in this study. *D. melanogaster* and *D. pseudoobscura* belong to the *Sophophora* subgenus. *D. repleta* and *D. buzzatii* (both in the *repleta* species group) and *D. virilis* (*virilis* species group) belong to the *Drosophila* subgenus (see Negre et al. 2003 for details).

In order to ascertain the consequences of *Drosophila* HOM-C splits, we have carried out a genomic and functional characterization of the two splits present in *D. buzzatii*. We isolated and sequenced two BAC clones containing the *lab-abdA* and *pb* chromosomal regions of *D. buzzatii*. The gene organization in these regions is compared with that of the homologous regions in *D. melanogaster* and *D. pseudoobscura* to map the precise site of the two splits. None of the two splits has altered the coding regions of *Hox* genes. We then searched for Conserved Noncoding Sequences (CNS), which are putative regulatory sequences, around the genes *lab*, *pb*, and *abdA*, to find out whether the splits removed or altered any *Hox*-gene enhancer. The position of CNS around *Hox* genes is compared with experimentally identified *Hox*-gene enhancers, and the arrangement of CNS is compared between *Hox* and non-*Hox* genes. Finally, we analyzed the expression patterns of three *Hox* genes, *lab*, *pb*, and *abdA*, in four *Drosophila* species with different *Hox*-gene organizations (with and without the splits) in whole-mount embryos and imaginal discs. The results show that, in *Drosophila* species, *Hox* genes, as well as their regulatory regions and expression patterns, are conserved, despite the *Hox* complex breaks. Thus, the functional significance of the *Hox*-gene clustering in *Drosophila* is questionable.

Results

Molecular characterization of *Hox*-gene complex breakpoints

To characterize the two HOM-C splits present in *D. buzzatii*, we isolated and sequenced two BAC clones, one (5H14, 124,024 bp) containing the *lab-abdA* region, and another (40C11, 132,938 bp) including the *pb* region (see Methods). The organization of the two regions of *D. buzzatii* chromosome 2 is shown in Figure 2 along with the homologous regions of *D. melanogaster* and *D. pseudoobscura* for comparison. *D. melanogaster* and *D. pseudoobscura* are homosequential in the analyzed regions, except where indicated. The sequenced *pb* region (Fig. 2A) contains 16 ORFs including *Dbuz\pb*, *Dbuz\zerknüllt* (*Dbuz\zen*), *Dbuz\zerknüllt-related* (*Dbuz\zen2*), and *Dbuz\bicoid* (*Dbuz\bcd*). These four genes are present in the ANT-C of *D. melanogaster* and also in the homologous region of *D. pseudoobscura* (Fig. 2B,C). The orientation of *Dbuz\zen2* is the same as that of *Dpse\zen2*, but inverted with regard to *Dmel\zen2*. The remaining 12 genes in this region are orthologous to *D. melanogaster* genes from four different regions (84D1–2, 89D2, 84E5, and 91D4–5) of chromosomal arm 3R. One of the genes, *CG14609*, is represented by six copies, in contrast to the single copy present in *D. melanogaster* or *D. pseudoobscura*. A total of four breakpoints are fixed in this region between *D. buzzatii* and *D. melanogaster* beside the *zen2* microrearrangement. That corresponding to the *lab-pb* split is located in the ~3-kb intergenic segment between *Dbuz\pb* and *Dbuz\CG17836* (Fig. 2A).

The sequenced *lab-abdA* region contains 11 ORFs, including *Dbuz\lab*, the cuticular cluster genes (*Dbuz\Ccp*), and *Dbuz\abdA*

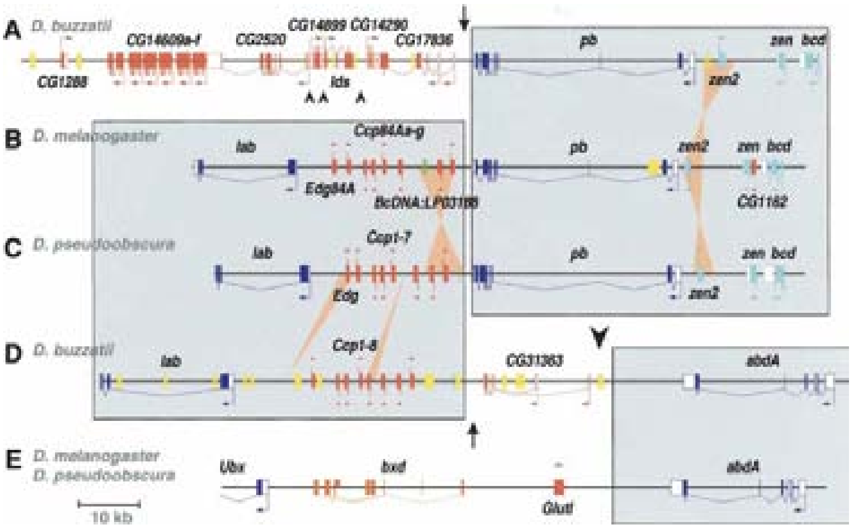


Figure 2. Gene organization of the *lab-abdA* and *pb* genomic regions of *D. buzzatii* compared with the homologous regions of *D. melanogaster* and *D. pseudoobscura*. The localization of the *lab-pb* split (arrow) and the *Ubx-abdA* split (large arrowhead) are indicated. (*A*) Sequence of *D. buzzatii* BAC 40C11 containing the *pb* region. (*B*) Organization of the *lab-pb* region in *D. melanogaster*. (*C*) *Idem in D. pseudoobscura*. (*D*) Sequence of *D. buzzatii* BAC 5H14 containing the *lab-abdA* region. (*E*) Organization of the *abdA* region in *D. melanogaster* and *D. pseudoobscura*. Genes are represented as open (UTRs) and filled boxes (coding sequences) with arrows indicating the sense of transcription. *Hox* genes are colored in dark blue, *Hox*-derived genes in light blue, non-*Hox* genes in red, noncoding RNA genes in orange, and the BcDNA:LP03188 and orthologous sequences in green. Transposable element insertions (usually *ISBu* elements, see Negre et al. 2003) are shown as yellow boxes. Large shaded rectangles include homologous *Hox*-gene regions in different species. Ochre triangles denote small inversions and insertions or deletions. Small arrowheads show breakpoints between *D. buzzatii* and *D. melanogaster* in non-*Hox* regions.

(Fig. 2D). The number of *Ccp* copies (including the gene *Edg*) is eight in the three species, but there is a small inversion encompassing two copies (plus the cDNA *BcDNA:LP03188*) in *D. melanogaster* in comparison to *D. buzzatii* or *D. pseudoobscura*, as well as one gain and one loss (Fig. 2B–D). These 11 genes come from three different regions (84A2–5, 86E11–13, and 89E2) of *D. melanogaster* chromosomal arm 3R, which means two fixed breakpoints between *D. buzzatii* and *D. melanogaster*, beside the small inversion of *Ccp* genes. One breakpoint corresponds to the *lab-pb* split and is found ~40 kb upstream of *Dbuz\lab*, in the 5-kb between the sequence similar to *BcDNA:LP03188* and the gene *Dbuz\CG31363*. The second breakpoint is that of the *Ubx-abdA* split and is located between 11 and 15 kb downstream of *Dbuz\abdA*. The two breakpoints are separated by a DNA segment of only ~22 kb encoding a single gene, *Dbuz\CG31363* (Fig. 2D).

Conserved noncoding sequences in *Hox* gene regions

We analyzed the conservation of noncoding sequences around the three *Hox* genes *lab*, *pb*, and *abdA* by comparing the sequences of the three species *D. buzzatii*, *D. melanogaster*, and *D. pseudoobscura* as done previously by other authors (Bergman and Kreitman 2001; Bergman et al. 2002) (see Methods). Figure 3 shows the VISTA graph, where the conservation between the aligned sequences is plotted (when higher than 50%) and the regions that meet the selected criteria (75% identity in a 25-bp window) are highlighted for both coding and noncoding sequences. A preliminary analysis showed no differences between intergenic and intronic regions, in agreement with previous studies (Bergman and Kreitman 2001). Thus, CNS are defined as intergenic (excluding UTRs) or intronic sequences that meet the above criteria. The characteristics of observed CNS are given in

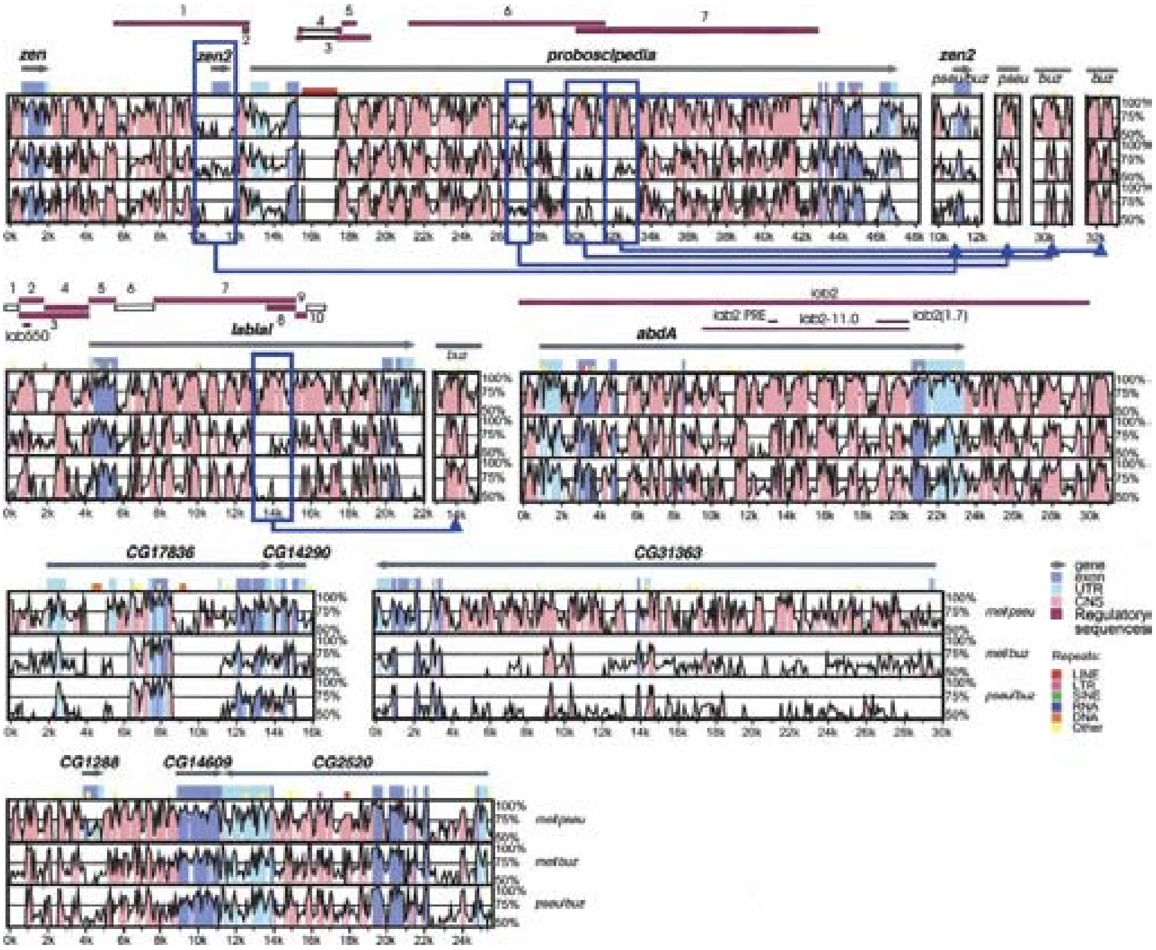


Figure 3. Nucleotide sequence conservation in the *lab-abdA* and *pb* regions between *Drosophila* species. The three panels in each VISTA plot represent pairwise comparisons between *D. melanogaster* and *D. pseudoobscura* (*mel/pse*), *D. melanogaster* and *D. buzzatii* (*mel/buz*) and *D. pseudoobscura* and *D. buzzatii* (*pse/buz*). The *x*-axis represents *D. melanogaster* coordinates, and *y*-axis sequence identity (50%–100%). Gray arrows show the location and orientation of genes. Conservation in exons and UTRs is shown in dark and light blue, respectively. Pink regions represent CNS. Experimentally identified regulatory sequences (solid purple bars) or segments with negative results (empty bars) are indicated on *top* of each plot. Five microinversions detected in the *lab* or *pb* regions are enclosed in blue frames, and the VISTA graphs generated with the inverted sequences shown to the *right* of the main plots. VISTA plots for the *CG17836-CG14290*, *CG31363*, and *CG1288-CG14609-CG2520* regions (adjacent to *Hox* genes) are shown at the *bottom* of the figure for comparison.

Tables 1 and 2, and the results of statistical analysis are shown in Supplemental Tables S1 and S2.

When *D. buzzatii* is compared with *D. melanogaster* or *D. pseudoobscura*, 395 and 440 CNS are found, respectively, around the three *Hox* genes (Table 1). This gives a density of 4.5 and 5 CNS per kilobase, respectively. These conserved blocks show a mean size of 44 bp with 86.5% nucleotide identity and represent 20%–22% of the analyzed noncoding sequence. When *D. melanogaster* and *D. pseudoobscura* are compared, 563 CNS are detected (6.5/kb) with a mean size of 55 bp and an average identity of 87.4%. In this comparison, the sequence in CNS represents 36% of noncoding sequence. In all three comparisons, the three regions around the *Hox* genes *lab*, *pb*, and *abdA* are homogeneous with little variation either in CNS density, size, or nucleotide identity (Supplemental Table S1). It is worth noting that CNS are coincident in all three comparisons (Fig. 3), which means that all CNS detected when comparing *D. buzzatii* with either *D. melanogaster* or *D. pseudoobscura* are also found in the comparison between the latter two species. Although most CNS keep colinearity (relative position and orientation), we could identify four microinversions, around 1–2 kb in size. One is located within the large intron of *lab* and the other three in introns 2 and 3 of *pb* (Fig. 3).

D. buzzatii is equally distant phylogenetically from either *D. melanogaster* or *D. pseudoobscura* (Fig. 1). The latter two species belong to the same subgenus and are phylogenetically closer. We compared the characteristics of the CNS found in the three pairwise comparisons. As expected, there are no statistical differences between the CNS found when comparing *D. buzzatii* with either *D. melanogaster* or *D. pseudoobscura* (Supplemental Table S2). The CNS density and the proportion of sequence in CNS are significantly higher when comparing the phylogenetically closer species *D. melanogaster* and *D. pseudoobscura*. Increasing divergence time does not seem to affect the nucleotide identity of the CNS, although the size of the CNS detected in the *Hox*-gene regions shows a significant decrease (Supplemental Table S2).

Conserved noncoding sequences in non-*Hox* gene regions

To find out whether the observed pattern of CNS is a particular feature of *Hox* genes, we also analyzed the presence of CNS in regions of the sequenced BACs adjacent, but unrelated, to *Hox* genes. We used the three microsyntenic regions between *D. buzzatii*, *D. melanogaster*, and *D. pseudoobscura* longer than 10 kb, i.e., the *CG31363* gene region, between *lab* and *abdA,* and the *CG17836-CG14290* and *CG1288-CG2520* regions, near *pb* (Fig. 2). These regions include one, two, and three genes, respectively. The pattern of CNS detected is shown in Figure 3 and summarized in Table 2. In the comparisons with *D. buzzatii*, we found around 100 CNS (~2/kb), which represents <8% of noncoding sequence. Thus, in these non-*Hox* regions, a much smaller number of CNS is observed and the proportion of sequence in CNS is also significantly lower than in *Hox*-gene regions (Supplemental Table S1). In the *D. melanogaster–D. pseudoobscura* comparison, there are 326 CNS (5.7/kb) which represents a 23% of noncoding sequence. Thus, in this case, the density is similar between *Hox* and non-*Hox*-gene regions, but the size of CNS is significantly smaller in the latter regions (Supplemental Table S1). Consequently, the proportion of sequence in CNS is also significantly lower in the non-*Hox*-gene regions. It should be noted that non-*Hox* regions show a significant variation for CNS density and also for the proportion of sequence in CNS that is not observed in *Hox*-gene regions (Supplemental Table S1). The higher variation observed between non-*Hox* regions is probably due to the heterogeneity of the sample from a functional point of view. There is little information available on the function and expression pattern of the six non-*Hox* genes analyzed, which probably represent a mixture of genes with different regulatory needs and number of enhancers.

Conservation of known regulatory sequences

Regulatory sequences of the genes *lab*, *pb*, and *abdA* have been experimentally identified in *D. melanogaster* (Karch et al. 1985; Chouinard and Kaufman 1991; Kapoun and Kaufman 1995; Martin et al. 1995). We compared their position with the pattern of CNS found around *Hox* genes. As shown in Figure 3, the regulatory sequences identified in *D. melanogaster* generally contain or correspond to CNS in *D. buzzatii*. For instance, CNS are found in the sites corresponding to the *iab2 PRE* and *iab2(1.7)* enhancers of *abdA* (Shimell et al. 1994, 2000). Similarly, a prominent conservation peak is observed at the site of the *lab550* enhancer, which directs the expression of *lab* in the embryo midgut (Marty et al. 2001). Also, the inverted segment found in the large intron of the *lab* gene roughly corresponds to the segment responsible for *lab* expression in the posterior midgut. Sequence details of the *lab550* and *iab2(1.7)* enhancer and binding site conservation are shown in Supplemental Figure S1. The Homeotic Response Ele-

Table 1. Characteristics of conserved noncoding sequences (CNS) detected with mVISTA in comparisons of Hox gene regions between *D. melanogaster* (mel), *D. pseudoobscura* (pse), and *D. buzzatii* (buz)

Region	Noncoding nucleotides	Species pair	Number of CNS	Density[a] (SD)	Mean size (nt) (SD)	Mean nucleotide identity (%)	Sequence in CNS (%)
lab	19,227	mel/pse	129	6.71 (0.59)	53.20 (34.03)	87.69	35.69
		mel/buz	73	3.80 (0.44)	46.70 (24.35)	87.33	17.73
		pse/buz	84	4.37 (0.48)	44.54 (27.05)	88.08	19.46
pb	42,056	mel/pse	265	6.30 (0.39)	55.04 (35.79)	87.29	34.68
		mel/buz	196	4.66 (0.33)	41.88 (22.38)	87.09	19.52
		pse/buz	215	5.11 (0.35)	42.92 (24.65)	86.38	21.94
abdA	26,043	mel/pse	169	6.49 (0.50)	59.29 (38.43)	87.44	38.45
		mel/buz	126	4.84 (0.43)	45.98 (26.15)	86.18	22.25
		pse/buz	141	5.41 (0.46)	46.11 (24.97)	85.44	24.97
Total	87,326	mel/pse	563	6.45 (0.27)	55.89 (36.23)	87.42	36.03
Hox gene regions		mel/buz	395	4.52 (0.23)	44.08 (24.04)	86.83	19.94
		pse/buz	440	5.04 (0.24)	44.25 (25.53)	86.39	22.30

[a]Density = number of CNS per kilobase.

Table 2. Characteristics of conserved noncoding sequences (CNS) detected with mVISTA in comparisons of non-*Hox* gene regions between *D. melanogaster* (mel), *D. pseudoobscura* (pse), and *D. buzzatii* (buz)

Region	Noncoding nucleotides	Species pair	Number of CNS	Density[a] (SD)	Mean size (nt) (SD)	Mean nucleotide identity (%)	Sequence in CNS (%)
CG1288-CG2520	18,333	mel/pse	127	6.93 (0.61)	43.48 (28.16)	86.31	30.12
		mel/buz	65	3.55 (0.44)	45.26 (31.51)	86.30	16.05
		pse/buz	67	3.65 (0.45)	42.44 (29.40)	86.81	15.59
CG17836-CG14290	10,921	mel/pse	46	4.21 (0.62)	45.02 (33.27)	82.67	18.96
		mel/buz	18	1.65 (0.39)	39.61 (22.67)	82.88	6.53
		pse/buz	22	2.01 (0.43)	42.09 (24.34)	84.34	8.48
CG31363	27,510	mel/pse	153	5.56 (0.45)	35.51 (14.95)	87.17	19.75
		mel/buz	22	0.80 (0.17)	26.09 (3.94)	82.93	2.09
		pse/buz	23	0.84 (0.17)	28.78 (7.70)	83.23	2.41
Total	56,764	mel/pse	326	5.74 (0.32)	39.96 (24.15)	86.09	22.95
non-Hox gene regions		mel/buz	105	1.84 (0.18)	40.28 (27.50)	85.27	7.45
		pse/buz	112	1.97 (0.19)	39.59 (25.88)	85.76	7.83

[a]Density = number of CNS per kilobase.

ment (HOMRE) of the *lab550* enhancer contains four binding sites; all of them are conserved in the three species. In the *iab2(1.7)* enhancer, there are five Hunchback (HB)-binding sites, three of which are conserved in the three species, whereas the other two vary in position between species. This enhancer also contains a unique Krüppel (KR)-binding site, where point mutations in *D. melanogaster* cause gain-of-expression mutants (Hab1 and Hab2) (Shimell et al. 1994). This binding site is conserved in all three species (Supplemental Fig. S1). The conservation between *D. melanogaster* and *D. buzzatii* around *abdA* ends 9 kb (in *D. melanogaster*) and 11 kb (in *D. buzzatii*) downstream of this gene (Fig. 3). This boundary lies between the *iab2* and *pbx* regulatory sequences, which control the expression of *abdA* and *Ubx*, respectively (Karch et al. 1985). We have shown that the *iab2* region downstream of *abdA* is conserved in *D. buzzatii*. We have not sequenced the *Ubx* region in *D. buzzatii*, but we assume that the *pbx* regulatory sequence will conserve its position upstream of *Ubx*, i.e., there are no rearrangements between *Ubx* and its regulatory sequences (see below).

It is worth noting though, that CNS were also found in fragments not experimentally tested or described as with no effect on expression (Fig. 3). This observation suggests that the regulation of these genes may be even more complex than currently envisaged, and that more regulatory modules may be operative in nature than those experimentally identified in the laboratory.

Hox gene expression patterns

The conservation of regulatory sequences suggests that splits of the HOM-C had no consequences on *Hox*-gene expression. To test this prediction, we compared the expression patterns of the *Hox* genes *lab*, *pb*, and *abdA* between *D. melanogaster*, *D. virilis*, *D. buzzatii*, and *D. repleta*. These four *Drosophila* species represent three different *Hox*-gene organizations (Figs. 1,2). *D. melanogaster* possess the *Antp-Ubx* split only, whereas *D. virilis* has the *Ubx-abdA* split instead. Both *D. buzzatii* and *D. repleta* present the *Ubx-abdA* and *lab-pb* splits. We used in situ hybridization and antibody staining to whole-mount embryos and to imaginal discs from third instar larvae and prepupae (see Methods). Detailed results are given in Figure 4 and Supplemental Figures S2–S6. The expression patterns of the four species closely follow those described for *D. melanogaster* (for review, see Hughes and Kaufman 2002). Interspecific variation was detected only in the *pb* gene, which in *D. virilis* presents an extra domain in the embryo mesoderm (Fig. 4). As this expression domain is not shared by *D. melanogaster*, it is seemingly not related with the *lab-pb* split. Although our analysis is qualitative, and slightly quantitative changes or domain changes of a few cells may remain undetected, it shows that the reorganization of the HOM-C caused no major alterations of the expression patterns of the three *Hox* genes adjacent to the splits, in good agreement with the conservation of regulatory sequences (see above). Likewise, Bomze and

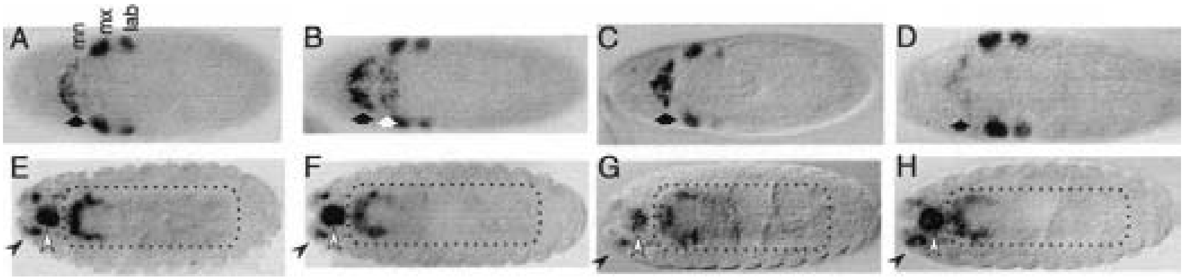


Figure 4. Expression pattern of *pb* in embryos. (*A–D*) stage 11 embryos, (*E–H*) stage 17 embryos. (*A,E*) *D. melanogaster*, (*B,F*) *D. virilis*, (*C,G*) *D. buzzatii*, and (*D,H*) *D. repleta*. (*A–D*) Expression on the ectoderm of the maxillary and labial lobes. Later in development (*E–H*) *pb* is detected in the derivatives of the maxillary (white arrowhead) and labial (black arrowhead) lobes, and in the ventral nervous system (boxed area). (*A–D*) *pb* expression is detected in the mesodermal layer of the mandibular segment (black arrow) in all four species. In *D. virilis* only (*B*), *pb* is also expressed in the mesodermal layer of the maxillary segment (white arrow). The mandibular (mn), maxillary (mx), and labial (lab) segments are shown in *A*.

López (1994) found that the expression pattern of *Ubx* in embryos is conserved between *D. melanogaster*, *D. pseudoobscura*, *D. virilis*, and *D. hydei* (a species of the *repleta* group), despite their different *Hox*-gene organization (Figure 1).

Discussion

zen2 predates the *Drosophila* radiation

The *zen* and *bcd* genes come from a duplication of *Hox3* in the ancestor of Cyclorraphan flies (Stauber et al. 2002). A second duplication of *zen* gave birth to *zen2*, which was thought to be a recent event in *D. melanogaster* (Randazzo et al. 1993), where it has no discernible function. However, the existence of *Dpse\zen2* and *Dbuz\zen2* shows that the *zen–zen2* duplication must predate the divergence of the *Sophophora* and *Drosophila* subgenus, and that this gene has been kept during at least 40–60 Myr of evolution. Whether this gene is also present in other flies outside of the *Drosophila* genus is still unknown.

Patterns of conserved noncoding sequence evolution

Cis-Regulatory Modules (CRM) are transcription regulatory DNA segments (from a few hundred base pair to 1 kb in size) that control gene expression in higher eukaryotes (Wray et al. 2003). CRM have a complex structure still not fully understood. They contain one or several binding sites for different transcription factors, which act cooperatively to activate or repress transcription of the target gene. As CRM are functionally constrained to maintain the expression of the target gene, they evolve slower than nonfunctional sequences. Therefore, the conservation of noncoding sequences between phylogenetically distant species may be used as a guide for identification of regulatory sequences. Several recent studies (Bergman and Kreitman 2001; Bergman et al. 2002; Cooper and Sidow 2003; Nobrega et al. 2003; Santini et al. 2003) support the use of comparative sequence analysis and characterization of CNS as a useful approach to detect putative CRM in *Drosophila* and other organisms. The clustering of previously characterized transcription-factor binding sites may be also used for detection of CRM (Berman et al. 2004). However, the absence of high-quality binding data for most *Drosophila* transcription factors represent a great current limitation in the widespread application of this method.

We exhaustively searched for CNS around *lab*, *pb*, and *abdA* and around adjacent non-*Hox* genes by comparing three species pairs. A plethora of highly conserved blocks was found surrounding the three *Hox* genes in the comparison between the phylogenetically distant species *D. buzzatii* and *D. melanogaster* or *D. pseudoobscura* (Fig. 1). The proportion of noncoding sequence included in CNS was 20%–22%. In most cases, these CNS keep their relative position and colinearity, although a few microrearrangements were found. The interpretation of these CNS as regulatory sequences is supported by the high neutral substitution rate (Moriyama and Gojobori 1992) and intrinsic rate of DNA loss (Petrov et al. 1996; Singh and Petrov 2004) in *Drosophila*. Noncoding sequences are not expected to be conserved between such distantly related species unless they are functionally constrained. The coincidence between CNS and known enhancers such as *iab2 PRE* or *lab550* (Supplemental Fig. S1) further supports this interpretation.

A lower CNS density was observed around non-*Hox* genes. This result fits well with previous observations showing that genes with complex developmentally regulated expression show a higher degree of conservation in noncoding regions than more simple genes with metabolic or housekeeping functions (Bergman and Kreitman 2001; Bergman et al. 2002; Halligan et al. 2004). Moreover, *Hox* genes are associated with larger noncoding regions. *Hox* genes harbor some of the longest introns of any *Drosophila* gene (Moriyama et al. 1998) and mean intron size is significantly greater in the *Hox* than in the non-*Hox* genes analyzed here ($F = 4.69$, $df = 1$, $P < 0.05$). This observation also fits with the notion that the amount of noncoding DNA must be larger in those genes with complex developmental functions in order to harbor the required CRM (Nelson et al. 2004).

HOM-C evolution in *Drosophila*

In *Drosophila*, *Hox* genes are arranged in the same 5′→3′ orientation (with only one exception, the *Deformed* gene in *D. melanogaster*). Their regulatory sequences are usually located upstream of each gene and in the introns. If we look at the three HOM-C splits known in *Drosophila*, a common pattern arises. As can be seen in Figure 2, the *lab–pb* split took place close to the 3′ end of *pb* and far from the *lab* 5′ end. Likewise, the split between the genes *Ubx* and *abdA* took place near the *abdA* 3′ end and far from the *Ubx* 5′ end, in the short space between their respective regulatory sequences *pbx* and *iab2*. This is approximately the same position where an experimental break that does not affect development has been observed (Struhl 1984), although the deficiencies used in the complementation tests both carry a fraction of the *pbx* and *iab2* regions. Finally, sequence comparison between *D. melanogaster* and *D. virilis* (Lewis et al. 2003) show that both the insertion of the *CG31217* gene and the *Antp–Ubx* split took place close to the *Ubx* 3′ end, and far from the *Antp* 5′ end (results not shown). Thus, all three splits seem to have occurred far from the 5′ end of one gene and much closer to the 3′ end of the next one, in such a way as to keep in place the regulatory sequences of both genes. In this way, rearrangements did not alter any of the known regulatory sequences of these *Hox* genes; this would explain the absence of gene expression changes.

In the repleta group species, the anterior gene *lab* is located near the posterior genes *abdA* and *AbdB*. The sequence analysis shows that *lab* and *abdA* are only 75 kb apart and show the same orientation. The breakpoint of the *lab–pb* split occurred at ~22 kb from that of the *Ubx–abdA* split. None of those splits seem to have affected the regulatory regions of the *Hox* genes, because the expression patterns of *lab* and *abdA* are unaffected. Although it is intriguing, the proximity between these genes in the *D. buzzatii* genome seems purely accidental and lacking any functional significance.

The most likely mechanisms for the generation of the HOM-C splits are paracentric inversions (Ranz et al. 2001; Gonzalez et al. 2002). A plausible reconstruction of HOM-C evolution in the *Drosophila* subgenus that accounts for the current organization of *Hox* genes in *D. buzzatii* is shown in Figure 5. In lower Dipterans, such as *Anopheles gambiae*, the eight *Hox* genes, plus *Hox3* and *ftz*, are arranged as a single cluster (Powers et al. 2000). Before the radiation of the *Drosophila* genus, two transpositions occurred as follows: the *Ccp* gene cluster between *lab* and *pb*, and the gene *CG31217* between *Antp* and *Ubx* (Lewis et al. 2003). Also, *zen*, *zen2*, and *bcd* evolved from the *Hox3* gene (see above). In the lineage of the *Drosophila* subgenus, an inversion took place with one breakpoint between *Ubx* and *abdA* (split 2 in Fig. 1) and the other one between *CG31363* and an unknown ORF (*X*). This HOM-C structure is now present in species of the *Drosophila* sub-

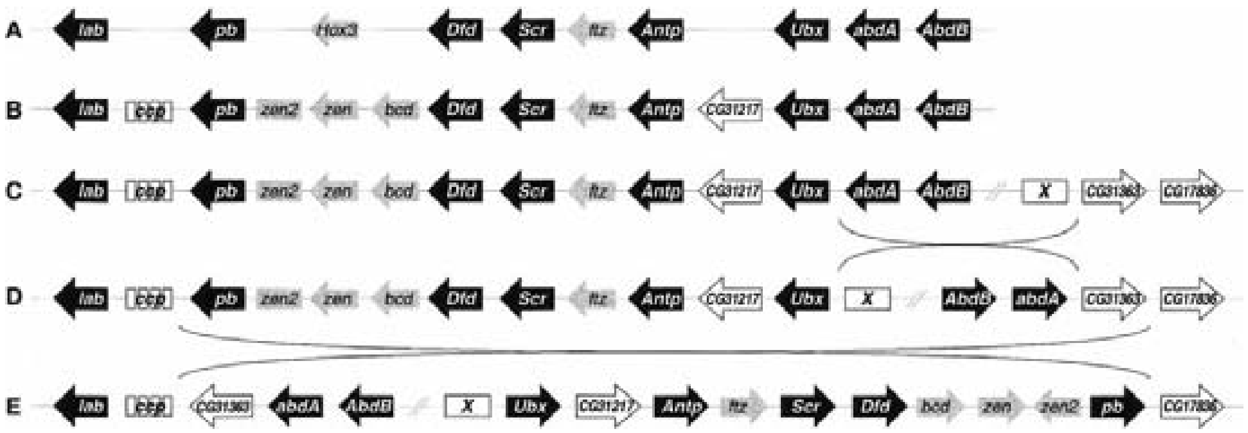


Figure 5. Reconstruction of the *Hox* gene complex evolution in the *Drosophila* subgenus. Genes are shown as arrows when the orientation 5′→3′ is known, and as rectangles otherwise. *Hox* genes are in black, *Hox*-related genes in gray, and non-*Hox* genes in white. (*A*) Lower Dipterans. (*B*) Before the radiation of the *Drosophila* genus. (*C*) *Drosophila* subgenus after its separation from that of the *Sophophora* subgenus. (*D*) Ancestor of the *repleta* group. (*E*) Present arrangement of *Hox* genes in *Drosophila buzzatii* (cf. Fig. 2A,D).

genus outside the repleta group, such as *D. virilis* (see Fig. 1). A second inversion, in the ancestor of the *repleta* group, split the HOM-C between *lab* and *pb* (split 3 in Fig. 1). This inversion, which relocated *lab* close to *abdA*, had one breakpoint between *pb* and the *Ccp* cluster genes and the second breakpoint between *CG31363* and *CG17836*. These two genes are not adjacent in the *D. melanogaster* genome, but we infer that they were so in the ancestor of the *Drosophila* subgenus.

Do flies have a *Hox* gene complex?

Despite the striking conservation of *Hox*-gene clustering in metazoans, if we compare two of the most deeply studied organisms, *Drosophila* and vertebrates, important differences arise (Ferrier and Minguillon 2003; Santini et al. 2003; Wagner et al. 2003). *Drosophila Hox*-gene regions (1) are much larger than those of vertebrates, e.g., the human *Hox*A cluster is only 110 kb long, whereas the *D. melanogaster* HOM-C spans 665 kb; (2) contain transposable element insertions, which are remarkably absent in those of vertebrates; (3) contain also non-*Hox* genes that are inserted between the *Hox* genes, and tandem duplications within the complex, such as those of the *zen*-related genes; (4) allow for small inversions of *Hox* genes, such as *Dfd* (Randazzo et al. 1993), and non-*Hox* genes, such as *zen2* (Fig. 2); and (5) are split in three ways in different lineages, apparently without consequences on gene expression. These observations suggest a highly dynamic evolution in *Drosophila* that contrasts with the compact structure seen in vertebrates. Thus, the splits of HOM-C in *Drosophila* indicate a release of functional requirements present in other metazoan.

Moreover, *Drosophila* is not the only organism known to have a split HOM-C. Split *Hox*-gene complexes were also known in nematodes, and recently have been described in *Bombyx* and tunicates. What do those organisms have in common in addition to the split HOM-C? Vertebrate development follows a rostral-to-caudal temporal progression, and the colinearity of *Hox* genes is not only spatial, but also temporal (the *Hox* clock) (Kmita and Duboule 2003). In the tunicate *Oikopleura*, *Hox* gene expression still evokes spatial colinearity but not temporal (Seo et al. 2004), which favors the argument that the constraining force of HOM-C structure conservation is temporal colinearity (Ferrier and Minguillon 2003). In nematodes, the pattern of *Hox*-gene evolution seems indicative of the move to a deterministic developmental mode (Aboobaker and Blaxter 2003). *Bombyx* embryogenesis, which is difficult to assign to a short or a long germ insect, is characterized by a quick development (Davis and Patel 2002). *Drosophila* is a long germ insect, where all *Hox* genes are activated almost simultaneously during the cellular blastoderm stage. Thus, none of these organisms seems to show temporal colinearity. A common feature between all organisms shown so far to have a split *Hox* complex seems to be a derived mode of embryogenesis characterized by a fast early development.

The loss of temporal progression in the activation of *Hox* genes in a very rapid mode of embryogenesis could be the ultimate cause for the modular organization of those *Hox* "clusters," where modules can be taken apart without loss of function. Given the high rate of chromosomal rearrangement in the genus *Drosophila* (Ranz et al. 2001; Gonzalez et al. 2002), we anticipate that an even greater variety of *Hox*-gene organizations will be discovered when more species are investigated. It is ironical that *Hox*-gene colinearity was discovered in *Drosophila*, an organism with a partially disassembled complex, which may be the by-product of phylogenetic inertia more than that of functional necessity.

Methods

Flies

D. buzzatii stock st-1 was used for construction of a genomic BAC library (González et al. 2005). The following species and stocks were used for gene expression experiments: *D. buzzatii* (j19), *D. repleta* (1611.2), *D. virilis* (Tokyo-Japan), and *D. melanogaster* (Canton S and Oregon R).

BAC sequencing

The genomic BAC library was screened with probes from the *lab*, *pb*, and *abdA* genes (González et al. 2005). Positive clones were used to build physical maps for the *lab-abdA* and *pb* chromosomal regions, and one BAC clone from each region was chosen for sequencing. Shotgun sublibraries were constructed for each BAC using the vector TOPO, and enough plasmid clones were sequenced by both ends to reach an ~6× redundancy. Reads were assembled with the PHRED-PHRAD-CONSED software (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998) and sequences finished with one round of AUTOFINISH (Gordon et al. 2001), followed by PCR to bridge the remaining gaps. A continuous high-quality sequence (PHRED score >40) was obtained for BAC clones 5H14 (124,024 bp), and 40C11 (132,938 bp). Statistic details of the sequencing process are given in Supplemental Table S3.

Sequence annotation

Nucleotide sequences were annotated with the aid of GENESCRIPT (Hudek et al. 2003) and ARTEMIS (Berriman and Rutherford 2003). Predicted ORFs were corroborated with GOFIGURE (Khan et al. 2003) for automatic Gene Ontology (Harris et al. 2004) annotation, and BLAST (McGinnis and Madden 2004) for similarity searches. *D. buzzatii* sequences were compared with those of homologous regions in *D. melanogaster* (Celniker et al. 2002) and *D. pseudoobscura* (Richards et al. 2005) genomes. *D. melanogaster* sequences used were as follows: AE001572 (ANT-C), DMU31961 (BX-C), and AE003692, AE003672, AE003713, AE003676, and AE003724 (other regions). *D. pseudoobscura* contigs AADE01000437 (*lab*), AADE01000149 (*pb*), AADE01000036 (*abdA*), and AADE01000014, AADE000175, AADE01002495, AADE01000322 (non-*Hox* genes) were identified with Genome VISTA (Dubchak et al. 2000) and the regions of interest annotated.

Analysis of regulatory sequences

Pairwise alignments of six homologous genomic regions between *D. buzzatii*, *D. melanogaster*, and *D. pseudoobscura* were performed with the AVID global-alignment tool using default parameters (Bray et al. 2003). CNS were identified in the alignments with mVISTA (Mayor et al. 2000) using a window size of 25 bp and a minimum identity of 75%. Statistical tests were carried out to compare the characteristics of the CNS found in the different regions. Comparisons of CNS size distributions, which depart significantly from normality, were conducted using the G-test (Sokal and Rohlf 1995). The number of CNS and the proportion of sequence within CNS was scored for 1-kb windows along the analyzed regions (masking out exons). The resulting variables (density and percent sequence in CNS) as well as the nucleotide identity (per CNS) were tested using ANOVA (Sokal and Rohlf 1995). A complete list of CNS detected is provided in Supplemental Table S5.

Gene-expression experiments

In situ hybridizations and antibody staining were performed to whole-mount embryos and to imaginal discs from third-instar larvae and prepupae as described (Alonso and Akam 2003; Suzanne et al. 2003). cDNA clones were obtained for *lab* from the four species, *pb* from *D. buzzatii* and *D. melanogaster* and *abdA* from *D. buzzatii*, *D. repleta*, and *D. virilis* as described (Negre et al. 2003) (for primers see Supplemental Table S4). Sense and antisense RNA probes were produced as described (Suzanne et al. 2003). When no species-specific probe was available, at least two different ones were used in independent experiments, and the results were always consistent. Specific antibodies against the protein were used for *abdA* (Macias et al. 1990).

Acknowledgments

We thank Claudio Alonso, Joaquin Ariño, Anna Barceló, James Castelli-Gair, Jose Felix de Celis, Susan Celniker, Ariel Chipman, Josefa González, Roger Hoskins, Barret Pfeiffer, Theresa Ren, Armand Sánchez, Chung Li Shu, and three anonymous referees for help or comments on the manuscript. This work was supported by grants BMC2002-01708 awarded to A.R. and BMC2002-00300 to E.S.-H. from the Dirección General de Enseñanza Superior e Investigación Científica (MEC, Spain), an Institutional Grant from the Fundación Ramón Areces, a doctoral FI/DGR fellowship from the Generalitat de Catalunya awarded to B.N., and a doctoral FPI fellowship from the Ministerio de Ciencia y Tecnología (BES-2003–0416) awarded to S.C.

References

Aboobaker, A. and Blaxter, M. 2003. *Hox* gene evolution in nematodes: Novelty conserved. *Curr. Opin. Genet. Dev.* **13:** 593–598.

Alonso, C.R. and Akam, M. 2003. A *Hox* gene mutation that triggers nonsense-mediated RNA decay and affects alternative splicing during *Drosophila* development. *Nucleic Acids Res.* **31:** 3873–3880.

Bergman, C.M. and Kreitman, M. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11:** 1335–1345.

Bergman, C., Pfeiffer, B., Rincon-Limas, D., Hoskins, R., Gnirke, A., Mungall, C., Wang, A., Kronmiller, B., Pacleb, J., Park, S., et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* **3:** research0086.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. 2004. Computational identification of developmental enhancers: Conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.* **5:** R61.

Berriman, M. and Rutherford, K. 2003. Viewing and annotating sequence data with Artemis. *Brief Bioinform.* **4:** 124–132.

Bomze, H.M. and López, A.J. 1994. Evolutionary conservation of the structure and expression of alternatively spliced *Ultrabithorax* isoforms from *Drosophila*. *Genetics* **136:** 965–977.

Bray, N., Dubchak, I., and Pachter, L. 2003. AVID: A global alignment program. *Genome Res.* **13:** 97–102.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S.P., Frise, E., et al. 2002. Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3:** research0079.

Chouinard, S. and Kaufman, T.C. 1991. Control of expression of the homeotic labial (*lab*) locus of *Drosophila melanogaster*: Evidence for both positive and negative autogenous regulation. *Development* **113:** 1267–1280.

Cook, C.E., Smith, M.L., Telford, M.J., Bastianello, A., and Akam, M. 2001. *Hox* genes and the phylogeny of the arthropods. *Curr. Biol.* **11:** 759–763.

Cooper, G.M. and Sidow, A. 2003. Genomic regulatory regions: Insights from comparative sequence analysis. *Curr. Opin. Genet. Dev.* **13:** 604–610.

Davis, G.K. and Patel, N.H. 2002. Short, long, and beyond: Molecular and embryological approaches to insect segmentation. *Annu. Rev. Entomol.* **47:** 669–699.

Dubchak, I., Brudno, M., Loots, G.G., Pachter, L., Mayor, C., Rubin, E.M., and Frazer, K.A. 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10:** 1304–1306.

Duncan, I. 1987. The bithorax complex. *Annu. Rev. Genet.* **21:** 285–319.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Ferrier, D.E. and Minguillon, C. 2003. Evolution of the *Hox*/Para*Hox* gene clusters. *Int. J. Dev. Biol.* **47:** 605–611.

González, J., Ranz, J.M., and Ruiz, A. 2002. Chromosomal elements evolve at different rates in the *Drosophila* genome. *Genetics* **161:** 1137–1154.

González, J., Nefedov, M., Bosdet, I., Casals, F., Calvete, O., Delprat, A., Shin, H., Chiu, R., Mathewson, C., Wye, N., et al. 2005. A BAC-based physical map of the *Drosophila buzzatii* genome. *Genome Res.* (in press).

Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

Gordon, D., Desmarais, C., and Green, P. 2001. Automated finishing with autofinish. *Genome Res.* **11:** 614–625.

Halligan, D.L., Eyre-Walker, A., Andolfatto, P., and Keightley, P.D. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res.* **14:** 273–279.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32:** D258–D261.

Hudek, A.K., Cheung, J., Boright, A.P., and Scherer, S.W. 2003. Genescript: DNA sequence annotation pipeline. *Bioinformatics* **19:** 1177–1178.

Hughes, C.L. and Kaufman, T.C. 2002. *Hox* genes and the evolution of the arthropod body plan. *Evol. Dev.* **4:** 459–499.

Hughes, C.L., Liu, P.Z., and Kaufman, T.C. 2004. Expression patterns of the rogue *Hox* genes *Hox3/zen* and *fushi tarazu* in the apterygote insect *Thermobia domestica*. *Evol. Dev.* **6:** 393–401.

Ikuta, T., Yoshida, N., Satoh, N., and Saiga, H. 2004. *Ciona intestinalis Hox* gene cluster: Its dispersed structure and residual colinear expression in development. *Proc. Natl. Acad. Sci.* **101:** 15118–15123.

Kapoun, A.M. and Kaufman, T.C. 1995. A functional analysis of 5′, intronic and promoter regions of the homeotic gene *proboscipedia* in *Drosophila melanogaster*. *Development* **121:** 2127–2141.

Karch, F., Weiffenbach, B., Peifer, M., Bender, W., Duncan, I., Celniker, S., Crosby, M., and Lewis, E.B. 1985. The abdominal region of the bithorax complex. *Cell* **43:** 81–96.

Kaufman, T.C., Lewis, R., and Wakimoto, B. 1980. Cytogenetic analysis of chromosome 3 in *Drosophila melanogaster*: The homoeotic gene complex in polytene chromosome interval 84A-B. *Genetics* **94:** 115–133.

Kaufman, T.C., Seeger, M.A., and Olsen, G. 1990. Molecular and genetic organization of the *antennapedia* gene complex of *Drosophila melanogaster*. *Adv. Genet.* **27:** 309–362.

Khan, S., Situ, G., Decker, K., and Schmidt, C.J. 2003. GoFigure: Automated Gene Ontology annotation. *Bioinformatics* **19:** 2484–2485.

Kmita, M. and Duboule, D. 2003. Organizing axes in time and space; 25 years of colinear tinkering. *Science* **301:** 331–333.

Lewis, E.B. 1978. A gene complex controlling segmentation in *Drosophila*. *Nature* **276:** 565–570.

Lewis, E.B., Pfeiffer, B.D., Mathog, D.R., and Celniker, S.E. 2003. Evolution of the homeobox complex in the Diptera. *Curr. Biol.* **13:** R587–R588.

Macias, A., Casanova, J., and Morata, G. 1990. Expression and regulation of the *abd-A* gene of *Drosophila*. *Development* **110:** 1197–1207.

Martin, C.H., Mayeda, C.A., Davis, C.A., Ericsson, C.L., Knafels, J.D., Mathog, D.R., Celniker, S.E., Lewis, E.B., and Palazzolo, M.J. 1995. Complete sequence of the bithorax complex of *Drosophila*. *Proc. Natl. Acad. Sci.* **92:** 8398–8402.

Marty, T., Vigano, M.A., Ribeiro, C., Nussbaumer, U., Grieder, N.C., and Affolter, M. 2001. A *Hox* complex, a repressor element and a 50 bp sequence confer regional specificity to a DPP-responsive enhancer. *Development* **128:** 2833–2845.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. 2000. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046–1047.

McGinnis, W. and Krumlauf, R. 1992. Homeobox genes and axial patterning. *Cell* **68:** 283–302.

McGinnis, S. and Madden, T.L. 2004. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32:** W20–W25.

Moriyama, E.N. and Gojobori, T. 1992. Rates of synonymous substitution and base composition of nuclear genes in *Drosophila*. *Genetics* **130:** 855–864.

Moriyama, E.N., Petrov, D.A., and Hartl, D.L. 1998. Genome size and intron size in *Drosophila*. *Mol. Biol. Evol.* **15:** 770–773.

Negre, B., Ranz, J.M., Casals, F., Cáceres, M., and Ruiz, A. 2003. A new split of the *Hox* gene complex in *Drosophila*: Relocation and evolution of the gene *labial*. *Mol. Biol. Evol.* **20:** 2042–2054.

Nelson, C.E., Hersh, B.M., and Carroll, S.B. 2004. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* **5:** R25.

Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.

Petrov, D.A., Lozovskaya, E.R., and Hartl, D.L. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384:** 346–349.

Powers, T.P., Hogan, J., Ke, Z., Dymbrowski, K., Wang, X., Collins, F.H., and Kaufman, T.C. 2000. Characterization of the *Hox* cluster from the mosquito *Anopheles gambiae* (Diptera: Culicidae). *Evol. Dev.* **2:** 311–325.

Randazzo, F.M., Seeger, M.A., Huss, C.A., Sweeney, M.A., Cecil, J.K., and Kaufman, T.C. 1993. Structural changes in the antennapedia complex of *Drosophila pseudoobscura*. *Genetics* **134:** 319–330.

Ranz, J.M., Casals, F., and Ruiz, A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* **11:** 230–239.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15:** 1–18.

Ruddle, F.H., Bartels, J.L., Bentley, K.L., Kappen, C., Murtha, M.T., and Pendleton, J.W. 1994. Evolution of *Hox* genes. *Annu. Rev. Genet.* **28:** 423–442.

Santini, S., Boore, J.L., and Meyer, A. 2003. Evolutionary conservation of regulatory elements in vertebrate *Hox* gene clusters. *Genome Res.* **13:** 1111–1122.

Seo, H.C., Edvardsen, R.B., Maeland, A.D., Bjordal, M., Jensen, M.F., Hansen, A., Flaat, M., Weissenbach, J., Lehrach, H., Wincker, P., et al. 2004. *Hox* cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. *Nature* **431:** 67–71.

Shimell, M.J., Simon, J., Bender, W., and O'Connor, M.B. 1994. Enhancer point mutation results in a homeotic transformation in *Drosophila*. *Science* **264:** 968–971.

Shimell, M.J., Peterson, A.J., Burr, J., Simon, J.A., and O'Connor, M.B. 2000. Functional analysis of repressor binding sites in the iab-2 regulatory region of the *abdominal-A* homeotic gene. *Dev. Biol.* **218:** 38–52.

Singh, N.D. and Petrov, D.A. 2004. Rapid sequence turnover at an intergenic locus in *Drosophila*. *Mol. Biol. Evol.* **21:** 670–680.

Sokal, R.R. and Rohlf, F.J. 1995. *Biometry: The principles and practice of statistics in biological research*. W.H. Freeman and Co., New York.

Stauber, M., Prell, A., and Schmidt-Ott, U. 2002. A single *Hox3* gene with composite *bicoid* and *zerknullt* expression characteristics in non-Cyclorrhaphan flies. *Proc. Natl. Acad. Sci.* **99:** 274–279.

Struhl, G. 1984. Splitting the bithorax complex of *Drosophila*. *Nature* **308:** 454–457.

Suzanne, M., Estella, C., Calleja, M., and Sánchez-Herrero, E. 2003. The *hernandez* and *fernandez* genes of *Drosophila* specify eye and antenna. *Dev. Biol.* **260:** 465–483.

Von Allmen, G., Hogga, I., Spierer, A., Karch, F., Bender, W., Gyurkovics, H., and Lewis, E. 1996. Splits in fruitfly *Hox* gene complexes. *Nature* **380:** 116.

Wagner, G.P., Amemiya, C., and Ruddle, F. 2003. *Hox* cluster duplications and the opportunity for evolutionary novelties. *Proc. Natl. Acad. Sci.* **100:** 14603–14606.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano, L.A. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20:** 1377–1419.

Yasukochi, Y., Ashakumary, L., Wu, C., Yoshido, A., Nohata, J., Mita, K., and Sahara, K. 2004. Organization of the *Hox* gene cluster of the silkworm, *Bombyx mori*: A split of the *Hox* cluster in a non-*Drosophila* insect. *Dev. Genes Evol.* **214:** 606–614.

Received November 15, 2004; accepted in revised form January 26, 2005.

Bibliography

Acampora, D., M. D'Esposito, A. Faiella, M. Pannese, E. Migliaccio *et al.* (1989) The human HOX gene family. *Nucleic Acids Res* **17**(24)**:** 10385-10402.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.* (2000) The genome sequence of Drosophila melanogaster. *Science* **287**(5461)**:** 2185-2195.

Aguade, M., N. Miyashita and C. H. Langley (1989) Reduced Variation in the Yellow-Achaete-Scute Region in Natural Populations of Drosophila Melanogaster. *Genetics* **122**(3)**:** 607-615.

Ahituv, N., Y. Zhu, A. Visel, A. Holt, V. Afzal *et al.* (2007) Deletion of Ultraconserved Elements Yields Viable Mice. *PLoS Biol* **5**(9)**:** e234.

Akashi, H. (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics* **139**(2)**:** 1067-1076.

Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver *et al.* (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* **2**(10)**:** e286.

Albrecht, A. and S. Mundlos (2005) The other trinucleotide repeat: polyalanine expansion disorders. *Curr Opin Genet Dev* **15**(3)**:** 285-293.

Andolfatto, P. (2005) Adaptive evolution of non-coding DNA in Drosophila. *Nature* **437**(7062)**:** 1149-1152.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J. M. Chia *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**(5585)**:** 1301-1310.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**(1)**:** 25-29.

Ashburner, M. and C. M. Bergman (2005) Drosophila melanogaster: a case study of a model genomic sequence and its consequences. *Genome Res* **15**(12)**:** 1661-1667.

Averof, M. (2002) Arthropod Hox genes: insights on the evolutionary forces that shape gene functions. *Curr Opin Genet Dev* **12**(4)**:** 386-392.

Avery, O. T., C. M. MacLeod and M. McCarty (1944) Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *J.Exp.Med* **79:** 137-158.

Ayala, F. J., M. L. Tracey, L. G. Barr, J. F. McDonald and S. Perez-Salas (1974) Genetic variation in natural populations of five Drosophila species and the hypothesis of the selective neutrality of protein polymorphisms. *Genetics* **77**(2)**:** 343-384.

Bamshad, M. and S. P. Wooding (2003) Signatures of natural selection in the human genome. *Nat Rev Genet* **4**(2)**:** 99-111.

Barbadilla, A., L. M. King and R. C. Lewontin (1996) What does electrophoretic variation tell us about protein variation? *Mol Biol Evol* **13**(2)**:** 427-432.

Barrier, M., C. D. Bustamante, J. Yu and M. D. Purugganan (2003) Selection on rapidly evolving proteins in the Arabidopsis genome. *Genetics* **163**(2)**:** 723-733.

Bartolome, C., X. Maside and B. Charlesworth (2002) On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. *Mol Biol Evol* **19**(6)**:** 926-937.

Bateson, W. (1894) *Materials for the Study of Variation: Treated with Especial Regard to Discontinuity in the Origin of Species.* Macmillan & Company, London.

Bazin, E., L. Duret, S. Penel and N. Galtier (2005) Polymorphix: a sequence polymorphism database. *Nucleic Acids Res* **33**(Database issue)**:** D481-484.

BAZIN, E., S. GLEMIN and N. GALTIER (2006) Population size does not influence mitochondrial genetic diversity in animals. *Science* **312**(5773)**:** 570-572.

BEGUN, D. J. and C. F. AQUADRO (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356**(6369)**:** 519-520.

BEJERANO, G., M. PHEASANT, I. MAKUNIN, S. STEPHEN, W. J. KENT *et al.* (2004) Ultraconserved elements in the human genome. *Science* **304**(5675)**:** 1321-1325.

BELTING, H. G., C. S. SHASHIKANT and F. H. RUDDLE (1998) Modification of expression and cis-regulation of Hoxc8 in the evolution of diverged axial morphology. *Proc Natl Acad Sci U S A* **95**(5)**:** 2355-2360.

BENSON, D. A., I. KARSCH-MIZRACHI, D. J. LIPMAN, J. OSTELL and D. L. WHEELER (2007) GenBank. *Nucleic Acids Res* **35**(Database issue)**:** D21-25.

BERGER, J., T. SUZUKI, K. A. SENTI, J. STUBBS, G. SCHAFFNER *et al.* (2001) Genetic mapping with SNP markers in Drosophila. *Nat Genet* **29**(4)**:** 475-481.

BERGMAN, C. M. (2001) Evolutionary analyses of transcriptional control sequences in Drosophila. Doctoral Thesis, pp. 132 in *Department of Ecology and Evolution*. The University of Chicago, Chicago, Illinois.

BERGMAN, C. M., J. W. CARLSON and S. E. CELNIKER (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics* **21**(8)**:** 1747-1749.

BERGMAN, C. M. and M. KREITMAN (2001) Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res* **11**(8)**:** 1335-1345.

BERGMAN, C. M., B. D. PFEIFFER, D. E. RINCON-LIMAS, R. A. HOSKINS, A. GNIRKE *et al.* (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol* **3**(12)**:** RESEARCH0086.

BERGMAN, C. M., H. QUESNEVILLE, D. ANXOLABEHERE and M. ASHBURNER (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biol* **7**(11)**:** R112.

BERLETH, T., M. BURRI, G. THOMA, D. BOPP, S. RICHSTEIN *et al.* (1988) The role of localization of bicoid RNA in organizing the anterior pattern of the Drosophila embryo. *Embo J* **7**(6)**:** 1749-1756.

BERMAN, B. P., Y. NIBU, B. D. PFEIFFER, P. TOMANCAK, S. E. CELNIKER *et al.* (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A* **99**(2)**:** 757-762.

BERRY, A. J., J. W. AJIOKA and M. KREITMAN (1991) Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. *Genetics* **129**(4)**:** 1111-1117.

BIERNE, N. and A. EYRE-WALKER (2004) The genomic rate of adaptive amino acid substitution in Drosophila. *Mol Biol Evol* **21**(7)**:** 1350-1360.

BIRD, A. P. (1995) Gene number, noise reduction and biological complexity. *Trends Genet* **11**(3)**:** 94-100.

BOFFELLI, D., J. MCAULIFFE, D. OVCHARENKO, K. D. LEWIS, I. OVCHARENKO *et al.* (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**(5611)**:** 1391-1394.

BONNETON, F. (2003) Divergence évolutive extrême d'un gène homéotique: le cas bicoid. *Medicine/Sciences* **19:** 1265-1270.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**(2)**:** 783-796.

BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**(5568)**:** 752-755.

BRIDGES, C. B. and T. H. MORGAN (1923) The third-chromosome group of mutant characters of *Drosophila melanogaster*. *Carnegie Inst. Washington Publ.* **327:** 93.

BRITTEN, R. J. and E. H. DAVIDSON (1969) Gene regulation for higher cells: a theory. *Science* **165**(891)**:** 349-357.

BROWN, L. Y. and S. A. BROWN (2004) Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends Genet* **20**(1)**:** 51-58.

BUSTAMANTE, C. D., A. FLEDEL-ALON, S. WILLIAMSON, R. NIELSEN, M. T. HUBISZ *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* **437**(7062)**:** 1153-1157.

BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN *et al.* (2002) The cost of inbreeding in Arabidopsis. *Nature* **416**(6880)**:** 531-534.

CAMERON, R. A., S. H. CHOW, K. BERNEY, T. Y. CHIU, Q. A. YUAN *et al.* (2005) An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. *Proc Natl Acad Sci U S A* **102**(33)**:** 11769-11774.

CARROLL, S. B. (2005) Evolution at two levels: on genes and form. *PLoS Biol* **3**(7)**:** e245.

CARROLL, S. B., J. K. GRENIER and S. D. WEATHERBEE (2001) *From DNA to diversity: Molecular Genetics and the Evolution of Animal Design*. Blackwell Science, Malden, Massachusetts.

CARVALHO, A. B., M. D. VIBRANOVSKI, J. W. CARLSON, S. E. CELNIKER, R. A. HOSKINS *et al.* (2003) Y chromosome and other heterochromatic sequences of the Drosophila melanogaster genome: how far can we go? *Genetica* **117**(2-3)**:** 227-237.

CASILLAS, S. and A. BARBADILLA (2004) PDA: a pipeline to explore and estimate polymorphism in large DNA databases. *Nucleic Acids Res* **32**(Web Server issue)**:** W166-169.

CASILLAS, S. and A. BARBADILLA (2006) PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. *Nucleic Acids Res* **34**(Web Server issue)**:** W632-634.

CASILLAS, S., A. BARBADILLA and C. M. BERGMAN (2007a) Purifying selection maintains highly conserved noncoding sequences in Drosophila. *Mol Biol Evol* **24**(10)**:** 2222-2234.

CASILLAS, S., R. EGEA, N. PETIT, C. M. BERGMAN and A. BARBADILLA (2007b) Drosophila Polymorphism Database (DPDB): a portal for nucleotide polymorphism in Drosophila. *Fly* **1**(4)**:** 205-211.

CASILLAS, S., B. NEGRE, A. BARBADILLA and A. RUIZ (2006) Fast sequence evolution of Hox and Hox-derived genes in the genus Drosophila. *BMC Evol Biol* **6:** 106.

CASILLAS, S., N. PETIT and A. BARBADILLA (2005) DPDB: a database for the storage, representation and analysis of polymorphism in the Drosophila genus. *Bioinformatics* **21**(Suppl.2)**:** ii26-ii30.

CASTELLI-GAIR, J. and M. AKAM (1995) How the Hox gene Ultrabithorax specifies two different segments: the significance of spatial and temporal regulation within metameres. *Development* **121**(9)**:** 2973-2982.

CASTELLI-GAIR, J., S. GREIG, G. MICKLEM and M. AKAM (1994) Dissecting the temporal requirements for homeotic gene function. *Development* **120**(7)**:** 1983-1995.

CASTILLO-DAVIS, C. I., D. L. HARTL and G. ACHAZ (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res* **14**(8)**:** 1530-1536.

CELNIKER, S. E. and G. M. RUBIN (2003) The Drosophila melanogaster genome. *Annu Rev Genomics Hum Genet* **4:** 89-117.

CELNIKER, S. E., D. A. WHEELER, B. KRONMILLER, J. W. CARLSON, A. HALPERN *et al.* (2002) Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol* **3**(12)**:** RESEARCH0079.

CLARK, A. G. (2001) The search for meaning in noncoding DNA. *Genome Res* **11**(8)**:** 1319-1320.

CLARK, A. G., S. GLANOWSKI, R. NIELSEN, P. D. THOMAS, A. KEJARIWAL *et al.* (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**(5652)**:** 1960-1963.

CODD, E. F. (1970) A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM* **13**(6)**:** 377-387.

COLLINS, F. S., L. D. BROOKS and A. CHAKRAVARTI (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8**(12)**:** 1229-1231.

COLLINS, F. S., E. D. GREEN, A. E. GUTTMACHER and M. S. GUYER (2003) A vision for the future of genomics research. *Nature* **422**(6934)**:** 835-847.

CONSORTIUM, D. C. G. S. A. A. (2007) Evolution of genes and genomes in the context of the Drosophila phylogeny. *Nature***:** In press.

CONSORTIUM, T. C. E. S. (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**(5396)**:** 2012-2018.

CONSORTIUM, T. H. G. S. (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **443**(7114)**:** 931-949.

CONSORTIUM, T. I. H. (2003) The International HapMap Project. *Nature* **426**(6968)**:** 789-796.

CONSORTIUM, T. I. H. (2004) Integrating ethics and science in the International HapMap Project. *Nat Rev Genet* **5**(6)**:** 467-475.

CONSORTIUM, T. I. H. (2005) A haplotype map of the human genome. *Nature* **437**(7063)**:** 1299-1320.

COOPER, G. M. and A. SIDOW (2003) Genomic regulatory regions: insights from comparative sequence analysis. *Curr Opin Genet Dev* **13**(6)**:** 604-610.

COSTAS, J., P. S. PEREIRA, C. P. VIEIRA, S. PINHO, J. VIEIRA *et al.* (2004) Dynamics and function of intron sequences of the wingless gene during the evolution of the Drosophila genus. *Evol Dev* **6**(5)**:** 325-335.

CROSBY, M. A., J. L. GOODMAN, V. B. STRELETS, P. ZHANG and W. M. GELBART (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res* **35**(Database issue)**:** D486-491.

CHAN, S. K., M. HSING, F. HORMOZDIARI and A. CHERKASOV (2007) Relationship between insertion/deletion (indel) frequency of proteins and essentiality. *BMC Bioinformatics* **8:** 227.

CHARLESWORTH, B. (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res* **63**(3)**:** 213-227.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**(4)**:** 1289-1303.

CHARLESWORTH, D., B. CHARLESWORTH and M. T. MORGAN (1995) The pattern of neutral molecular variation under the background selection model. *Genetics* **141**(4)**:** 1619-1632.

CHARLESWORTH, J. and A. EYRE-WALKER (2006) The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol* **23**(7)**:** 1348-1356.

CHEN, C. T., J. C. WANG and B. A. COHEN (2007a) The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet* **80**(4)**:** 692-704.

CHEN, F. C., C. J. CHEN and T. J. CHUANG (2007b) INDELSCAN: a web server for comparative identification of species-specific and non-species-specific insertion/deletion events. *Nucleic Acids Res* **35**(Web Server issue)**:** W633-638.

CHEN, K., M. D. MCLELLAN, L. DING, M. C. WENDL, Y. KASAI *et al.* (2007c) PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Res* **17**(5)**:** 659-666.

CHEN, K. and N. RAJEWSKY (2006) Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet* **38**(12)**:** 1452-1456.

CHENNA, R., H. SUGAWARA, T. KOIKE, R. LOPEZ, T. J. GIBSON *et al.* (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**(13)**:** 3497-3500.

DARWIN, C. (1859) *On the origin of species by means of natural selection. Or the preservation of favoured races in the struggle for life.* John Murray, London.

DAVIDSON, E. H. (2006) *The Regulatory Genome. Gene Regulatory Networks In Development and Evolution.* Elsevier, Burlington, Massachusetts.

DAVIES, S. R., L. W. CHANG, D. PATRA, X. XING, K. POSEY *et al.* (2007) Computational identification and functional validation of regulatory motifs in cartilage-expressed genes. *Genome Res* **17**(10)**:** 1438-1447.

DAVIS, J. C., O. BRANDMAN and D. A. PETROV (2005) Protein evolution in the context of Drosophila development. *J Mol Evol* **60**(6)**:** 774-785.

DE MEAUX, J., U. GOEBEL, A. POP and T. MITCHELL-OLDS (2005) Allele-specific assay reveals functional variation in the chalcone synthase promoter of Arabidopsis thaliana that is compatible with neutral evolution. *Plant Cell* **17**(3)**:** 676-690.

DE ROSA, R., J. K. GRENIER, T. ANDREEVA, C. E. COOK, A. ADOUTTE *et al.* (1999) Hox genes in brachiopods and priapulids and protostome evolution. *Nature* **399**(6738)**:** 772-776.

DE VELASCO, B., J. SHEN, S. GO and V. HARTENSTEIN (2004) Embryonic development of the Drosophila corpus cardiacum, a neuroendocrine gland with similarity to the vertebrate pituitary, is controlled by sine oculis and glass. *Dev Biol* **274**(2)**:** 280-294.

DENNIS, C. (2002) Mouse genome: a forage in the junkyard. *Nature* **420**(6915)**:** 458-459.

DERMITZAKIS, E. T., C. M. BERGMAN and A. G. CLARK (2003) Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. *Mol Biol Evol* **20**(5)**:** 703-714.

DI VENTURA, B., C. LEMERLE, K. MICHALODIMITRAKIS and L. SERRANO (2006) From in vivo to in silico biology and back. *Nature* **443**(7111)**:** 527-533.

DOBZHANSKY, T. (1937) *Genetics and the Origin of Species*. Columbia University Press, New York.

DOBZHANSKY, T. (1970) *Genetics of the Evolutionary Process*. Columbia University Press.

DOBZHANSKY, T. and A. H. STURTEVANT (1938) Inversions in the Chromosomes of Drosophila Pseudoobscura. *Genetics* **23**(1)**:** 28-64.

DOEBLEY, J. and L. LUKENS (1998) Transcriptional regulators and the evolution of plant form. *Plant Cell* **10**(7)**:** 1075-1082.

DONIGER, S. W. and J. C. FAY (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* **3**(5)**:** e99.

DRAKE, J. A., C. BIRD, J. NEMESH, D. J. THOMAS, C. NEWTON-CHEH *et al.* (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet* **38**(2)**:** 223-227.

DUBOULE, D. (1994) Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl***:** 135-142.

EDGAR, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5)**:** 1792-1797.

EFSTRATIADIS, A., J. W. POSAKONY, T. MANIATIS, R. M. LAWN, C. O'CONNELL *et al.* (1980) The structure and evolution of the human beta-globin gene family. *Cell* **21**(3)**:** 653-668.

EGEA, R., S. CASILLAS, E. FERNANDEZ, M. A. SENAR and A. BARBADILLA (2007) MamPol: a database of nucleotide polymorphism in the Mammalia class. *Nucleic Acids Res* **35**(Database issue)**:** D624-629.

EICHLER, E. E., D. A. NICKERSON, D. ALTSHULER, A. M. BOWCOCK, L. D. BROOKS *et al.* (2007) Completing the map of human genetic variation. *Nature* **447**(7141)**:** 161-165.

ELLEGREN, H. (2004) Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**(6)**:** 435-445.

EMBERLY, E., N. RAJEWSKY and E. D. SIGGIA (2003) Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics* **4:** 57.

ENRIGHT, A. J., B. JOHN, U. GAUL, T. TUSCHL, C. SANDER *et al.* (2003) MicroRNA targets in Drosophila. *Genome Biol* **5**(1)**:** R1.

ERWIN, D., J. VALENTINE and D. JABLONSKI (1997) The origin of animal body plans. *American Scientist* **85:** 126-137.

EYRE-WALKER, A. (2002) Changing effective population size and the McDonald-Kreitman test. *Genetics* **162**(4)**:** 2017-2024.

EYRE-WALKER, A. (2006) The genomic rate of adaptive evolution. *Trends Ecol Evol* **21**(10)**:** 569-575.

FAY, J. C. and C. I. WU (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155**(3)**:** 1405-1413.

FAY, J. C., G. J. WYCKOFF and C. I. WU (2001) Positive and negative selection on the human genome. *Genetics* **158**(3)**:** 1227-1234.

FAY, J. C., G. J. WYCKOFF and C. I. WU (2002) Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature* **415**(6875)**:** 1024-1026.

FERRIER, D. E. and C. MINGUILLON (2003) Evolution of the Hox/ParaHox gene clusters. *Int J Dev Biol* **47**(7-8)**:** 605-611.

FEUK, L., A. R. CARSON and S. W. SCHERER (2006) Structural variation in the human genome. *Nat Rev Genet* **7**(2)**:** 85-97.

FONDON, J. W., 3RD and H. R. GARNER (2004) Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* **101**(52)**:** 18058-18063.

FORD, E. B. (1971) *Ecological Genetics*. 3rd ed. Chapman and Hall, London.

FOX, J. A., S. MCMILLAN and B. F. OUELLETTE (2006) A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res* **34**(Web Server issue)**:** W3-5.

FRAZER, K. A., D. G. BALLINGER, D. R. COX, D. A. HINDS, L. L. STUVE *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164)**:** 851-861.

FREEMAN, J. L., G. H. PERRY, L. FEUK, R. REDON, S. A. MCCARROLL *et al.* (2006) Copy number variation: new insights in genome diversity. *Genome Res* **16**(8)**:** 949-961.

FROMENTAL, C., M. KANNO, H. NOMIYAMA and P. CHAMBON (1988) Cooperativity and hierarchical levels of functional organization in the SV40 enhancer. *Cell* **54**(7)**:** 943-953.

FU, Y. X. and W. H. LI (1993) Statistical tests of neutrality of mutations. *Genetics* **133**(3)**:** 693-709.

GALANT, R. and S. B. CARROLL (2002) Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* **415**(6874)**:** 910-913.

GALPERIN, M. Y. (2007) The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Res* **35**(Database issue)**:** D3-4.

GARCIA-BELLIDO, A. (1975) Genetic control of wing disc development in Drosophila. *Ciba Found Symp* **0**(29)**:** 161-182.

GARCIA-FERNANDEZ, J. (2005) The genesis and evolution of homeobox gene clusters. *Nat Rev Genet* **6**(12)**:** 881-892.

GAUCHAT, D., F. MAZET, C. BERNEY, M. SCHUMMER, S. KREGER *et al.* (2000) Evolution of Antp-class genes and differential expression of Hydra Hox/paraHox genes in anterior patterning. *Proc Natl Acad Sci U S A* **97**(9)**:** 4493-4498.

GELLON, G. and W. MCGINNIS (1998) Shaping animal body plans in development and evolution by modulation of Hox expression patterns. *Bioessays* **20**(2)**:** 116-125.

GILBERT, S. F. (2003) The morphogenesis of evolutionary developmental biology. *Int J Dev Biol* **47**(7-8)**:** 467-477.

GILLESPIE, J. H. (2000a) Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* **155**(2)**:** 909-919.

GILLESPIE, J. H. (2000b) The neutral theory in an infinite population. *Gene* **261**(1)**:** 11-18.

GILLESPIE, J. H. (2001) Is the population size of a species relevant to its evolution? *Evolution Int J Org Evolution* **55**(11)**:** 2161-2169.

GILLESPIE, J. H. (2004) *Population Genetics. A concise guide.* The Johns Hopkins University Press, Baltimore and London.

GLINKA, S., L. OMETTO, S. MOUSSET, W. STEPHAN and D. DE LORENZO (2003) Demography and natural selection have shaped genetic variation in Drosophila melanogaster: a multi-locus approach. *Genetics* **165**(3)**:** 1269-1278.

GOFFEAU, A., B. G. BARRELL, H. BUSSEY, R. W. DAVIS, B. DUJON *et al.* (1996) Life with 6000 genes. *Science* **274**(5287)**:** 546, 563-547.

GUTTMACHER, A. E. (2001) Human genetics on the web. *Annu Rev Genomics Hum Genet* **2:** 213-233.

HADDRILL, P. R., B. CHARLESWORTH, D. L. HALLIGAN and P. ANDOLFATTO (2005) Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. *Genome Biol* **6**(8)**:** R67.

HALLIGAN, D. L. and P. D. KEIGHTLEY (2006) Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res* **16**(7)**:** 875-884.

HANCOCK, J. M., E. A. WORTHEY and M. F. SANTIBANEZ-KOREF (2001) A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Mol Biol Evol* **18**(6)**:** 1014-1023.

HARRIS, H. (1966) Enzyme polymorphisms in man. *Proc R Soc Lond B Biol Sci* **164**(995)**:** 298-310.

HARRISON, P. M., D. MILBURN, Z. ZHANG, P. BERTONE and M. GERSTEIN (2003) Identification of pseudogenes in the Drosophila melanogaster genome. *Nucleic Acids Res* **31**(3)**:** 1033-1037.

HARTL, D. L. and A. G. CLARK (1997) *Principles of Population Genetics*. Sinauer Associates, Inc., Sunderland, Massachusetts.

HILL, W. G. and A. ROBERTSON (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38:** 226-231.

HINDS, D. A., L. L. STUVE, G. B. NILSEN, E. HALPERIN, E. ESKIN *et al.* (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* **307**(5712)**:** 1072-1079.

HOLT, R. A., G. M. SUBRAMANIAN, A. HALPERN, G. G. SUTTON, R. CHARLAB *et al.* (2002) The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**(5591)**:** 129-149.

HOLLAND, P. W., L. Z. HOLLAND, N. A. WILLIAMS and N. D. HOLLAND (1992) An amphioxus homeobox gene: sequence conservation, spatial expression during development and insights into vertebrate evolution. *Development* **116**(3)**:** 653-661.

HOSKINS, R. A., C. D. SMITH, J. W. CARLSON, A. B. CARVALHO, A. HALPERN *et al.* (2002) Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome Biol* **3**(12)**:** RESEARCH0085.

HUDSON, R. R. (1987) Estimating the recombination parameter of a finite population model without selection. *Genet Res* **50**(3)**:** 245-250.

HUDSON, R. R. and N. L. KAPLAN (1995) Deleterious background selection with recombination. *Genetics* **141**(4)**:** 1605-1617.

HUDSON, R. R., M. KREITMAN and M. AGUADE (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**(1)**:** 153-159.

HUGHES, C. L., P. Z. LIU and T. C. KAUFMAN (2004) Expression patterns of the rogue Hox genes Hox3/zen and fushi tarazu in the apterygote insect Thermobia domestica. *Evol Dev* **6**(6)**:** 393-401.

IHLE, S., I. RAVAOARIMANANA, M. THOMAS and D. TAUTZ (2006) An analysis of signatures of selective sweeps in natural populations of the house mouse. *Mol Biol Evol* **23**(4)**:** 790-797.

IVANOV, A. I., A. C. ROVESCALLI, P. POZZI, S. YOO, B. MOZER *et al.* (2004) Genes required for Drosophila nervous system development identified by RNA interference. *Proc Natl Acad Sci U S A* **101**(46)**:** 16216-16221.

JACKSON, D. G., M. D. HEALY and D. B. DAVISON (2003) Bioinformatics: not just for sequences anymore. *Biosilico* **1**(3)**:** 103-111.

JOHNSON, F. M., C. G. KANAPI, R. H. RICHARDSON, M. R. WHEELER and W. S. STONE (1966) An analysis of polymorphisms among isozyme loci in dark and light Drosophila ananassae strains from American and Western Samoa. *Proc Natl Acad Sci U S A* **56**(1)**:** 119-125.

JUKES, T. H. and C. R. CANTOR (1969) Mammalian protein metabolism, pp. 21-132 in *Evolution of Protein Molecules*, edited by H. N. MUNRO. Academic Press, New York.

KAMINKER, J. S., C. M. BERGMAN, B. KRONMILLER, J. CARLSON, R. SVIRSKAS *et al.* (2002) The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol* **3**(12)**:** RESEARCH0084.

KANEHISA, M. and P. BORK (2003) Bioinformatics in the post-sequence era. *Nat Genet* **33**(3s)**:** 305-310.

KARLIN, S., L. BROCCHIERI, A. BERGMAN, J. MRAZEK and A. J. GENTLES (2002a) Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci U S A* **99**(1)**:** 333-338.

KARLIN, S. and C. BURGE (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc Natl Acad Sci U S A* **93**(4)**:** 1560-1565.

KARLIN, S., C. CHEN, A. J. GENTLES and M. CLEARY (2002b) Associations between human disease genes and overlapping gene groups and multiple amino acid runs. *Proc Natl Acad Sci U S A* **99**(26)**:** 17008-17013.

KATOH, K., K. KUMA, H. TOH and T. MIYATA (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**(2)**:** 511-518.

KATZMAN, S., A. D. KERN, G. BEJERANO, G. FEWELL, L. FULTON *et al.* (2007) Human genome ultraconserved elements are ultraselected. *Science* **317**(5840)**:** 915.

KAUER, M. O., D. DIERINGER and C. SCHLOTTERER (2003) A microsatellite variability screen for positive selection associated with the "out of Africa" habitat expansion of Drosophila melanogaster. *Genetics* **165**(3)**:** 1137-1148.

KAUFMAN, T. C., R. LEWIS and B. WAKIMOTO (1980) Cytogenetic Analysis of Chromosome 3 in *Drosophila melanogaster*: the Homeotic Gene Complex in Polytene Chromosome Interval 84A-B. *Genetics* **94**(1)**:** 115-133.

KAUFMAN, T. C., D. W. SEVERSON and G. E. ROBINSON (2002) The Anopheles genome and comparative insect genomics. *Science* **298**(5591)**:** 97-98.

KEIGHTLEY, P. D., G. V. KRYUKOV, S. SUNYAEV, D. L. HALLIGAN and D. J. GAFFNEY (2005a) Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res* **15**(10)**:** 1373-1378.

KEIGHTLEY, P. D., M. J. LERCHER and A. EYRE-WALKER (2005b) Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol* **3**(2)**:** e42.

KELLY, J. K. (1997) A test of neutrality based on interlocus associations. *Genetics* **146**(3)**:** 1197-1206.

KENT, W. J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res* **12**(4)**:** 656-664.

KIDA, Y., Y. MAEDA, T. SHIRAISHI, T. SUZUKI and T. OGURA (2004) Chick Dach1 interacts with the Smad complex and Sin3a to control AER formation and limb development along the proximodistal axis. *Development* **131**(17)**:** 4179-4187.

KIM, Y. and W. STEPHAN (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**(2)**:** 765-777.

KIMURA, M. (1968) Evolutionary rate at the molecular level. *Nature* **217**(129)**:** 624-626.

KIMURA, M. (1980) Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods. *Proc Natl Acad Sci U S A* **77**(1)**:** 522-526.

KIMURA, M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, Massachusetts.

KMITA, M. and D. DUBOULE (2003) Organizing axes in time and space; 25 years of colinear tinkering. *Science* **301**(5631)**:** 331-333.

KONDRASHOV, A. S. (2005) Evolutionary biology: fruitfly genome is not junk. *Nature* **437**(7062)**:** 1106.

KOPCZYNSKI, C. C., J. N. NOORDERMEER, T. L. SERANO, W. Y. CHEN, J. D. PENDLETON *et al.* (1998) A high throughput screen to identify secreted and transmembrane proteins involved in Drosophila embryogenesis. *Proc Natl Acad Sci U S A* **95**(17)**:** 9973-9978.

KREITMAN, M. (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. *Nature* **304**(5925)**:** 412-417.

KRUGLYAK, L. and D. A. NICKERSON (2001) Variation is the spice of life. *Nat Genet* **27**(3)**:** 234-236.

KRUGLYAK, S., R. T. DURRETT, M. D. SCHUG and C. F. AQUADRO (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A* **95**(18)**:** 10774-10778.

KRYUKOV, G. V., S. SCHMIDT and S. SUNYAEV (2005) Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet* **14**(15)**:** 2221-2229.

KUHN, R. M., D. KAROLCHIK, A. S. ZWEIG, H. TRUMBOWER, D. J. THOMAS *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res* **35**(Database issue)**:** D668-673.

LAI, E. C., P. TOMANCAK, R. W. WILLIAMS and G. M. RUBIN (2003) Computational identification of Drosophila microRNA genes. *Genome Biol* **4**(7)**:** R42.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822)**:** 860-921.

LANGLEY, C. H., J. MACDONALD, N. MIYASHITA and M. AGUADE (1993) Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in Drosophila melanogaster and Drosophila simulans. *Proc Natl Acad Sci U S A* **90**(5)**:** 1800-1803.

LAUGHON, A. and M. P. SCOTT (1984) Sequence of a Drosophila segmentation gene: protein structure homology with DNA-binding proteins. *Nature* **310**(5972)**:** 25-31.

LEWIS, E. B. (1978) A gene complex controlling segmentation in Drosophila. *Nature* **276**(5688)**:** 565-570.

Lewis, E. B., B. D. Pfeiffer, D. R. Mathog and S. E. Celniker (2003) Evolution of the homeobox complex in the Diptera. *Curr Biol* **13**(15)**:** R587-588.

Lewis, R. A., T. C. Kaufman, R. E. Denell and P. Tallerico (1980a) Genetic Analysis of the Antennapedia Gene Complex (ANT-C) and Adjacent Chromosomal Regions of *Drosophila melanogaster*. I. Polytene Chromosome Segments 84B-D. *Genetics* **95**(2)**:** 367-381.

Lewis, R. A., B. T. Wakimoto, R. E. Denell and T. C. Kaufman (1980b) Genetic Analysis of the Antennapedia Gene Complex (ANT-C) and Adjacent Chromosomal Regions of *Drosophila melanogaster*. II. Polytene Chromosome Segments 84A-84B1,2. *Genetics* **95**(2)**:** 383-397.

Lewontin, R. C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49:** 49-67.

Lewontin, R. C. (1974) *The genetic basis of evolutionary change*. Columbia University Press, New York.

Lewontin, R. C. (2002) Directions in evolutionary biology. *Annu Rev Genet* **36:** 1-18.

Lewontin, R. C. and J. L. Hubby (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics* **54**(2)**:** 595-609.

Lewontin, R. C. and K. Kojima (1960) The evolutionary dynamics of complex polymorphisms. *Evolution* **14:** 458-472.

Li, W.-H. (1997) *Molecular Evolution*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Li, W. H., Z. Gu, H. Wang and A. Nekrutenko (2001) Evolutionary analyses of the human genome. *Nature* **409**(6822)**:** 847-849.

Li, W. H., C. C. Luo and C. I. Wu (1985a) Evolution of DNA sequences, pp. 1-94 in *Molecular Evolutionary Genetics*, edited by R. J. MacIntyre. Plenum Press, New York.

Li, W. H., C. I. Wu and C. C. Luo (1985b) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* **2**(2)**:** 150-174.

Long, M., E. Betran, K. Thornton and W. Wang (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**(11)**:** 865-875.

Ludwig, M. Z., C. Bergman, N. H. Patel and M. Kreitman (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**(6769)**:** 564-567.

Lunter, G. (2007) Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics* **23**(13)**:** i289-296.

Lunter, G., C. P. Ponting and J. Hein (2006) Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2**(1)**:** e5.

Lynch, M. (2006) The origins of eukaryotic gene structure. *Mol Biol Evol* **23**(2)**:** 450-468.

Lynch, M. (2007) *The origins of genome architecture*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Lynch, M. and J. S. Conery (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**(5494)**:** 1151-1155.

Lynch, M. and J. S. Conery (2003) The origins of genome complexity. *Science* **302**(5649)**:** 1401-1404.

Lynch, M. and A. Force (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**(1)**:** 459-473.

Lynch, V. J., J. J. Roth and G. P. Wagner (2006) Adaptive evolution of Hox-gene homeodomains after cluster duplications. *BMC Evol Biol* **6:** 86.

Mann, R. S. (1994) Engrailed-mediated repression of Ultrabithorax is necessary for the parasegment 6 identity in Drosophila. *Development* **120**(11)**:** 3205-3212.

Marais, G., P. Nouvellet, P. D. Keightley and B. Charlesworth (2005) Intron size and exon evolution in Drosophila. *Genetics* **170**(1)**:** 481-485.

Markstein, M., P. Markstein, V. Markstein and M. S. Levine (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc Natl Acad Sci U S A* **99**(2)**:** 763-768.

Marotta, C. A., J. T. Wilson, B. G. Forget and S. M. Weissman (1977) Human beta-globin messenger

RNA. III. Nucleotide sequences derived from complementary DNA. *J Biol Chem* **252**(14)**:** 5040-5053.

MARTIN-CAMPOS, J. M., J. M. COMERON, N. MIYASHITA and M. AGUADE (1992) Intraspecific and interspecific variation at the y-ac-sc region of Drosophila simulans and Drosophila melanogaster. *Genetics* **130**(4)**:** 805-816.

MATTHEWS, K. A., T. C. KAUFMAN and W. M. GELBART (2005) Research resources for Drosophila: the expanding universe. *Nat Rev Genet* **6**(3)**:** 179-193.

MATTICK, J. S. (2004) RNA regulation: a new genetics? *Nat Rev Genet* **5**(4)**:** 316-323.

MAYR, E. (1942) *Systematics and the Origin of Species*. Columbia University Press, New York.

MAYR, E. (1963) *Animal species and evolution*. Harvard University Press, Cambridge, Massachusetts.

MAYR, E. (1976) Typological versus population thinking, pp. 26-29 in *Evolution and the diversity of life: selected essays*. Harvard University Press, Cambridge, Massachusetts.

MCDONALD, J. H. and M. KREITMAN (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**(6328)**:** 652-654.

MCGINNIS, W. (1994) A century of homeosis, a decade of homeoboxes. *Genetics* **137**(3)**:** 607-611.

MCGINNIS, W. and R. KRUMLAUF (1992) Homeobox genes and axial patterning. *Cell* **68**(2)**:** 283-302.

MCGINNIS, W., M. S. LEVINE, E. HAFEN, A. KUROIWA and W. J. GEHRING (1984) A conserved DNA sequence in homeotic genes of the Drosophila Antennapedia and bithorax complexes. *Nature* **308**(5958)**:** 428-433.

MCVEAN, G., C. C. SPENCER and R. CHAIX (2005) Perspectives on human genetic variation from the HapMap Project. *PLoS Genet* **1**(4)**:** e54.

MCVEAN, G. A., S. R. MYERS, S. HUNT, P. DELOUKAS, D. R. BENTLEY *et al.* (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* **304**(5670)**:** 581-584.

MIKKELSEN, T. S., L. W. HILLIER, E. E. EICHLER, M. C. ZODY, D. B. JAFFE *et al.* (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055)**:** 69-87.

MIKLOS, G. L. and G. M. RUBIN (1996) The role of the genome project in determining gene function: insights from model organisms. *Cell* **86**(4)**:** 521-529.

MISRA, S., M. A. CROSBY, C. J. MUNGALL, B. B. MATTHEWS, K. S. CAMPBELL *et al.* (2002) Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol* **3**(12)**:** RESEARCH0083.

MITA, K., M. KASAHARA, S. SASAKI, Y. NAGAYASU, T. YAMADA *et al.* (2004) The genome sequence of silkworm, Bombyx mori. *DNA Res* **11**(1)**:** 27-35.

MIYASHITA, N. and C. H. LANGLEY (1988) Molecular and phenotypic variation of the white locus region in Drosophila melanogaster. *Genetics* **120**(1)**:** 199-212.

MONGIN, E., C. LOUIS, R. A. HOLT, E. BIRNEY and F. H. COLLINS (2004) The Anopheles gambiae genome: an update. *Trends Parasitol* **20**(2)**:** 49-52.

MORGAN, T. H., A. H. STURTEVANT, H. J. MULLER and C. B. BRIDGES (1915) *The Mechanism of Mendelian Heredity*. Henry Holt & Company, New York.

MORGENSTERN, B. (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acids Res* **32**(Web Server issue)**:** W33-36.

MULLER, H. J. (1927) Artificial Transmutation of the Gene. *Science* **66:** 84-87.

MULLER, H. J. and W. D. KAPLAN (1966) The dosage compensation of Drosophila and mammals as showing the accuracy of the normal type. *Genet Res* **8**(1)**:** 41-59.

NACHMAN, M. W. (2001) Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet* **17**(9)**:** 481-485.

NAM, J. and M. NEI (2005) Evolutionary change of the numbers of homeobox genes in bilateral animals. *Mol Biol Evol* **22**(12)**:** 2386-2394.

NARDELLI-HAEFLIGER, D. and M. SHANKLAND (1992) Lox2, a putative leech segment identity gene, is expressed in the same segmental domain in different stem cell lineages. *Development* **116**(3)**:** 697-710.

NEGRE, B. (2005) Caracterización genómica y funcional de las reorganizaciones del complejo de genes Hox en Drosophila. Doctoral Thesis, pp. 245 in *Departament* *de Genètica i de Microbiologia*. Universitat Autònoma de Barcelona, Bellaterra, Barcelona.

NEGRE, B., S. CASILLAS, M. SUZANNE, E. SANCHEZ-HERRERO, M. AKAM *et al.* (2005) Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex. *Genome Res* **15**(5)**:** 692-700.

NEGRE, B. and A. RUIZ (2007) HOM-C evolution in Drosophila: is there a need for Hox gene clustering? *Trends Genet* **23**(2)**:** 55-59.

NEI, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

NEI, M. and T. GOJOBORI (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**(5)**:** 418-426.

NELSON, C. E., B. M. HERSH and S. B. CARROLL (2004) The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol* **5**(4)**:** R25.

NEVO, E., A. BEILES and R. BEN-SHLOMO (1984) The evolutionary significance of genetic diversity: ecological, demographic and life history correlates, pp. 13-213 in *Lecture notes in biomethemathics, S. Levin, Ed., vol. 53, Evolutionary dynamics of genetic diversity.* G. S. Mani, Ed. (Springer-Verlag), Berlin.

NIELSEN, R., S. WILLIAMSON, Y. KIM, M. J. HUBISZ, A. G. CLARK *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Res* **15**(11)**:** 1566-1575.

NIELSEN, R. and Z. YANG (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol* **20**(8)**:** 1231-1239.

NJOLSTAD, P. R. and A. FJOSE (1988) In situ hybridization patterns of zebrafish homeobox genes homologous to Hox-2.1 and En-2 of mouse. *Biochem Biophys Res Commun* **157**(2)**:** 426-432.

NJOLSTAD, P. R., A. MOLVEN, H. G. EIKEN and A. FJOSE (1988) Structure and neural expression of a zebrafish homeobox sequence. *Gene* **73**(1)**:** 33-46.

NORDBORG, M. and S. TAVARE (2002) Linkage disequilibrium: what history has to tell us. *Trends Genet* **18**(2)**:** 83-90.

NOTREDAME, C., D. G. HIGGINS and J. HERINGA (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**(1)**:** 205-217.

NUSSLEIN-VOLHARD, C. and E. WIESCHAUS (1980) Mutations affecting segment number and polarity in Drosophila. *Nature* **287**(5785)**:** 795-801.

O'HARA, R. J. (1998) Population thinking and tree thinking in systematics. *Zool. Scr.* **26**(4)**:** 323-329.

OHLER, U., G. C. LIAO, H. NIEMANN and G. M. RUBIN (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol* **3**(12)**:** RESEARCH0087.

OHTA, T. (1995) Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J Mol Evol* **40**(1)**:** 56-63.

OMETTO, L., S. GLINKA, D. DE LORENZO and W. STEPHAN (2005a) Inferring the effects of demography and selection on Drosophila melanogaster populations from a chromosome-wide scan of DNA variation. *Mol Biol Evol* **22**(10)**:** 2119-2130.

OMETTO, L., W. STEPHAN and D. DE LORENZO (2005b) Insertion/deletion and nucleotide polymorphism data reveal constraints in Drosophila melanogaster introns and intergenic regions. *Genetics* **169**(3)**:** 1521-1527.

ONDEK, B., L. GLOSS and W. HERR (1988) The SV40 enhancer contains two distinct levels of organization. *Nature* **333**(6168)**:** 40-45.

ORENGO, D. J. and M. AGUADE (2004) Detecting the footprint of positive selection in a european population of Drosophila melanogaster: multilocus pattern of variation and distance to coding regions. *Genetics* **167**(4)**:** 1759-1766.

PANDEY, A. and F. LEWITTER (1999) Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem Sci* **24**(7)**:** 276-280.

PANFILIO, K. A. and M. AKAM (2007) A comparison of Hox3 and Zen protein coding sequences in taxa that span the Hox3/zen divergence. *Dev Genes Evol* **217**(4)**:** 323-329.

PETIT, N., S. CASILLAS, A. RUIZ and A. BARBADILLA (2007) Protein Polymorphism Is Negatively Correlated with Conservation of Intronic Sequences and Complexity of Expression Patterns in Drosophila melanogaster. *J Mol Evol* **64**(5)**:** 511-518.

PETROV, D. A. and D. L. HARTL (1998) High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. *Mol Biol Evol* **15**(3)**:** 293-302.

PETROV, D. A., E. R. LOZOVSKAYA and D. L. HARTL (1996) High intrinsic rate of DNA loss in Drosophila. *Nature* **384**(6607)**:** 346-349.

PIGANEAU, G. and A. EYRE-WALKER (2003) Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. *Proc Natl Acad Sci U S A* **100**(18)**:** 10335-10340.

POWELL, J. R. (1997) *Progress and prospects in Evolutionary Biology: The Drosophila model.* Oxford University Press, New York.

POWELL, J. R., A. CACCONE, J. M. GLEASON and L. NIGRO (1993) Rates of DNA evolution in Drosophila depend on function and developmental stage of expression. *Genetics* **133**(2)**:** 291-298.

PRIBNOW, D., D. C. SIGURDSON, L. GOLD, B. S. SINGER, C. NAPOLI *et al.* (1981) rII cistrons of bacteriophage T4. DNA sequence around the intercistronic divide and positions of genetic landmarks. *J Mol Biol* **149**(3)**:** 337-376.

QUESNEVILLE, H., C. M. BERGMAN, O. ANDRIEU, D. AUTARD, D. NOUAUD *et al.* (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* **1**(2)**:** 166-175.

RAND, D. M. and L. M. KANN (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Mol Biol Evol* **13**(6)**:** 735-748.

RANZ, J. M., D. MAURIN, Y. S. CHAN, M. VON GROTTHUSS, L. W. HILLIER *et al.* (2007) Principles of genome evolution in the Drosophila melanogaster species group. *PLoS Biol* **5**(6)**:** e152.

REDON, R., S. ISHIKAWA, K. R. FITCH, L. FEUK, G. H. PERRY *et al.* (2006) Global variation in copy number in the human genome. *Nature* **444**(7118)**:** 444-454.

RICHARDS, S., Y. LIU, B. R. BETTENCOURT, P. HRADECKY, S. LETOVSKY *et al.* (2005) Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res* **15**(1)**:** 1-18.

RIEDL, R. (1978) *Order in living organisms: A systems analysis of evolution.* Wiley, New York.

ROBERTS, D. B. (2006) *Drosophila melanogaster*: the model organism. *Entomologia Experimentalis et Applicata* **121:** 93-103.

ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**(18)**:** 2496-2497.

RUBIN, G. M. (1996) Around the genomes: the Drosophila genome project. *Genome Res* **6**(2)**:** 71-79.

RUBIN, G. M. and E. B. LEWIS (2000) A brief history of Drosophila's contributions to genome research. *Science* **287**(5461)**:** 2216-2218.

RUBIN, G. M., M. D. YANDELL, J. R. WORTMAN, G. L. GABOR MIKLOS, C. R. NELSON *et al.* (2000) Comparative genomics of the eukaryotes. *Science* **287**(5461)**:** 2204-2215.

RUSSO, C. A., N. TAKEZAKI and M. NEI (1995) Molecular phylogeny and divergence times of drosophilid species. *Mol Biol Evol* **12**(3)**:** 391-404.

SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. LEVINE, D. J. RICHTER *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**(6909)**:** 832-837.

SAWYER, S. A. and D. L. HARTL (1992) Population genetics of polymorphism and divergence. *Genetics* **132**(4)**:** 1161-1176.

SAWYER, S. A., R. J. KULATHINAL, C. D. BUSTAMANTE and D. L. HARTL (2003) Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. *J Mol Evol* **57**(Suppl 1)**:** S154-164.

SCOTT, M. P. and A. J. WEINER (1984) Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. *Proc Natl Acad Sci U S A* **81**(13)**:** 4115-4119.

SCHLOTTERER, C. (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* **160**(2)**:** 753-763.

SCHLOTTERER, C. (2003) Hitchhiking mapping--functional genomics from the population genetics perspective. *Trends Genet* **19**(1)**:** 32-38.

SCHUMMER, M., I. SCHEURLEN, C. SCHALLER and B. GALLIOT (1992) HOM/HOX homeobox genes are present in hydra (Chlorohydra viridissima) and are differentially expressed during regeneration. *Embo J* **11**(5)**:** 1815-1823.

SEARLS, D. B. (2000) Bioinformatics tools for whole genomes. *Annu Rev Genomics Hum Genet* **1:** 251-279.

SHABALINA, S. A., A. Y. OGURTSOV, V. A. KONDRASHOV and A. S. KONDRASHOV (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet* **17**(7)**:** 373-376.

SHARAKHOVA, M. V., M. P. HAMMOND, N. F. LOBO, J. KRZYWINSKI, M. F. UNGER *et al.* (2007) Update of the Anopheles gambiae PEST genome assembly. *Genome Biol* **8**(1)**:** R5.

SIEPEL, A., G. BEJERANO, J. S. PEDERSEN, A. S. HINRICHS, M. HOU *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**(8)**:** 1034-1050.

SIMPSON, G. G. (1944) *Tempo and Mode in Evolution*. Columbia University Press, New York.

SINGH, R. S. and L. R. RHOMBERG (1987) A Comprehensive Study of Genic Variation in Natural Populations of Drosophila melanogaster. II. Estimates of Heterozygosity and Patterns of Geographic Differentiation. *Genetics* **117**(2)**:** 255-271.

SLACK, J. M., P. W. HOLLAND and C. F. GRAHAM (1993) The zootype and the phylotypic stage. *Nature* **361**(6412)**:** 490-492.

SMITH, J. M. and J. HAIGH (1974) The hitch-hiking effect of a favourable gene. *Genet Res* **23**(1)**:** 23-35.

SMITH, N. G. and A. EYRE-WALKER (2002) Adaptive protein evolution in Drosophila. *Nature* **415**(6875)**:** 1022-1024.

STAUBER, M., H. JACKLE and U. SCHMIDT-OTT (1999) The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene. *Proc Natl Acad Sci U S A* **96**(7)**:** 3786-3789.

STAUBER, M., A. PRELL and U. SCHMIDT-OTT (2002) A single Hox3 gene with composite bicoid and zerknullt expression characteristics in non-Cyclorrhaphan flies. *Proc Natl Acad Sci U S A* **99**(1)**:** 274-279.

STEBBINS, G. L. (1950) *Variation and Evolution in Plants*. Columbia University Press, New York.

STEIN, L. (2002) Creating a bioinformatics nation. *Nature* **417**(6885)**:** 119-120.

STEIN, L. (2003) Foreword in *Mastering Perl for Bioinformatics*, edited by J. D. TISDALL. O'Reilly & Associates, Inc., Sebastopol, California.

STEIN, L. D. (2001) Using Perl to Facilitate Biological Analysis, pp. 413-449 in *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, edited by A. D. BAXEVANIS and B. F. F. OUELLETTE. John Wiley & Sons, Inc., New York.

STEPHAN, W. (1995) An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Mol Biol Evol* **12**(5)**:** 959-962.

STEPHAN, W. and C. H. LANGLEY (1989) Molecular genetic variation in the centromeric region of the X chromosome in three Drosophila ananassae populations. I. Contrasts between the vermilion and forked loci. *Genetics* **121**(1)**:** 89-99.

STEPHAN, W. and S. J. MITCHELL (1992) Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of Drosophila ananassae. *Genetics* **132**(4)**:** 1039-1045.

STORZ, J. F., B. A. PAYSEUR and M. W. NACHMAN (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Mol Biol Evol* **21**(9)**:** 1800-1811.

STUDIER, F. W., A. H. ROSENBERG, M. N. SIMON and J. J. DUNN (1979) Genetic and physical mapping in the early region of bacteriophage T7 DNA. *J Mol Biol* **135**(4)**:** 917-937.

TAFT, R., J, and J. S. MATTICK (2003) Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biol* **5:** P1.

TAJIMA, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**(2)**:** 437-460.

TAJIMA, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3)**:** 585-595.

TAJIMA, F. (1993) Mesurement of DNA polymorphism in *Mechanisms of molecular evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinauer Associates Inc., Suderland, Massachusetts.

TAJIMA, F. (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics* **143**(3)**:** 1457-1465.

TAMURA, K., S. SUBRAMANIAN and S. KUMAR (2004) Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol* **21**(1)**:** 36-44.

TELFORD, M. J. (2000) Evidence for the derivation of the Drosophila fushi tarazu gene from a Hox gene orthologous to lophotrochozoan Lox5. *Curr Biol* **10**(6)**:** 349-352.

TESHIMA, K. M., G. COOP and M. PRZEWORSKI (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Res* **16**(6)**:** 702-712.

THORISSON, G. A., A. V. SMITH, L. KRISHNAN and L. D. STEIN (2005) The International HapMap Project Web site. *Genome Res* **15**(11)**:** 1592-1593.

TUPY, J. L., A. M. BAILEY, G. DAILEY, M. EVANS-HOLM, C. W. SIEBEL *et al.* (2005) Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **102**(15)**:** 5495-5500.

VALENCIA, A. (2002) Bioinformatics: biology by other means. *Bioinformatics* **18**(12)**:** 1551-1552.

VENTER, J. C., M. D. ADAMS, E. W. MYERS, P. W. LI, R. J. MURAL *et al.* (2001) The sequence of the human genome. *Science* **291**(5507)**:** 1304-1351.

VERAKSA, A., M. DEL CAMPO and W. MCGINNIS (2000) Developmental patterning genes and their conserved functions: from model organisms to humans. *Mol Genet Metab* **69**(2)**:** 85-100.

VILELLA, A. J., A. BLANCO-GARCIA, S. HUTTER and J. ROZAS (2005) VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics* **21**(11)**:** 2791-2793.

VOIGHT, B. F., S. KUDARAVALLI, X. WEN and J. K. PRITCHARD (2006) A map of recent positive selection in the human genome. *PLoS Biol* **4**(3)**:** e72.

WANG, B. B., M. M. MULLER-IMMERGLUCK, J. AUSTIN, N. T. ROBINSON, A. CHISHOLM *et al.* (1993) A homeotic gene cluster patterns the anteroposterior body axis of C. elegans. *Cell* **74**(1)**:** 29-42.

WASSERMAN, W. W. and A. SANDELIN (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**(4)**:** 276-287.

WATERSTON, R. H., K. LINDBLAD-TOH, E. BIRNEY, J. ROGERS, J. F. ABRIL *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915)**:** 520-562.

WATTERSON, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**(2)**:** 256-276.

WEATHERBEE, S. D. and S. B. CARROLL (1999) Selector genes and limb identity in arthropods and vertebrates. *Cell* **97**(3)**:** 283-286.

WEBB, C. T., S. A. SHABALINA, A. Y. OGURTSOV and A. S KONDRASHOV (2002) Analysis of similarity within 142 pairs of orthologous intergenic regions of Caenorhabditis elegans and Caenorhabditis briggsae. *Nucleic Acids Res* **30**(5)**:** 1233-1239.

WELCH, J. J. (2006) Estimating the genome-wide rate of adaptive protein evolution in Drosophila. *Genetics* **173**(2)**:** 821-837.

WHEELER, D. L., T. BARRETT, D. A. BENSON, S. H. BRYANT, K. CANESE *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **35**(Database issue)**:** D5-12.

WIEHE, T., V. NOLTE, D. ZIVKOVIC and C. SCHLOTTERER (2007) Identification of selective sweeps using a dynamically adjusted number of linked microsatellites. *Genetics* **175**(1)**:** 207-218.

WIEHE, T. H. and W. STEPHAN (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster. *Mol Biol Evol* **10**(4)**:** 842-854.

WILLIAMSON, S. (2003) Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Mol Biol Evol* **20**(8)**:** 1318-1325.

WITTKOPP, P. J., B. K. HAERUM and A. G. CLARK (2004) Evolutionary changes in cis and trans gene regulation. *Nature* **430**(6995)**:** 85-88.

WRAY, G. A., M. W. HAHN, E. ABOUHEIF, J. P. BALHOFF, M. PIZER *et al.* (2003) The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **20**(9)**:** 1377-1419.

WRIGHT, S. (1931) Evolution in Mendelian populations. *Genetics* **16:** 97-159.

WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.* (2005) The effects of artificial selection on the maize genome. *Science* **308**(5726)**:** 1310-1314.

XIA, Q., Z. ZHOU, C. LU, D. CHENG, F. DAI *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (Bombyx mori). *Science* **306**(5703)**:** 1937-1940.

YANG, Z. and J. P. BIELAWSKI (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* **15**(12)**:** 496-503.

ZHANG, J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**(6)**:** 292-298.

ZHANG, L. and W. H. LI (2005) Human SNPs reveal no evidence of frequent positive selection. *Mol Biol Evol* **22**(12)**:** 2504-2507.

ZHANG, Z., S. SCHWARTZ, L. WAGNER and W. MILLER (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**(1-2)**:** 203-214.

ZHU, W. and C. R. BUELL (2007) Improvement of whole-genome annotation of cereals through comparative analyses. *Genome Res* **17**(3)**:** 299-310.

ZUCKERKANDL, E. and L. PAULING (1962) Molecular disease, evolution, and genetic heterogeneity, pp. 189-225 in *Horizons in Biochemistry*, edited by M. KASHA and B. PULLMAN. Academic Press, New York.

Web References

1. Web Portals

- **AAA *Drosophila***, Assembly, Alignment and Annotation of 12 related *Drosophila* species: http://rana.lbl.gov/drosophila/
- **Affymetrix/NCI Human Transcriptome Project:** http://transcriptome.affymetrix.com/
- **DPGP**, the *Drosophila* Population Genomics Project: http://www.dpgp.org/
- **NCBI**, National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/
- **The FlySNP Project**: http://flysnp.imp.ac.at/

2. Data banks

- ***Entrez* dbSNP**, database of single nucleotide polymorphisms in the NCBI: http://www.ncbi.nlm.nih.gov/SNP/
- ***Entrez* Nucleotide - GenBank**, sequence archive from the NCBI: http://www.ncbi.nlm.nih.gov/Genbank/index.html
- ***Entrez* PopSet**, polymorphic sets archive from the NCBI: http://www.ncbi.nlm.nih.gov/sites/entrez?db=popset
- **DPDB**, Drosophila Polymorphism Database: http://dpdb.uab.cat/
- **FlyBase**, sequence archive database and other resources for Drosophila: http://flybase.bio.indiana.edu/
- **GO**, Gene Ontology: http://www.geneontology.org/
- **HapMap,** haplotype map of the human genome: http://www.hapmap.org/
- **Yeastract,** Yeast Search for Transcriptional Regulators and Consensus Tracking: http://www.yeastract.com/

3. INTERNET RESOURCES FOR WHOLE-GENOME COMPARATIVE ANALYSIS

- **UCSC Genome Browser**: http://genome.ucsc.edu/

4. SOFTWARE AND SERVICES

- **BLAT:** http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
- **CLUSTALW:** http://www.ebi.ac.uk/clustalw/
- **DIALIGN:** http://dialign.gobics.de/
- **DNASP:** http://www.ub.es/dnasp/
- **MAFFT:** http://align.bmr.kyushu-u.ac.jp/mafft/software/
- **MUSCLE:** http://www.drive5.com/muscle/
- **MySQL:** http://www.mysql.com/
- **PDA:** http://pda.uab.cat/
- **PERL:** http://www.perl.com/
- **PHASTCONS:** http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=99458819&c=chr2L&g=phastConsElements15way
- **T-COFFEE:** http://www.tcoffee.org/Projects_home_page/t_coffee_home_page.html

Abbreviations

α	The proportion of substitutions driven by positive selection
μ	Mutation rate
A/P	Anteroposterior
abd-A	*abdominal-A* gene
Abd-B	*Abdominal-B* gene
Adh	*Alcohol dehydrogenase* gene
AMNIS	Algorithm for the Maximization of the Number of Informative Sites in the alignments
ANT-C	Antennapedia Complex
bcd	*bicoid* gene
BX-C	Bithorax Complex
Ccp	Cuticle Cluster Proteins
CDS	Coding Sequence
CNS, C	Conserved Noncoding Sequence
CNV	Copy Number Variation
CON	Constructed
CRM	*Cis*-Regulatory Module
DAF	Derived Allele Frequency
DIP	Deletion-Insertion Polymorphisms
D_n	Number of nonsynonymous substitutions
d_n, K_a	Number of nonsynonymous substitutions per nonsynonymous site
DPDB	Drosophila Polymorphism Database
D_s	Number of synonymous substitutions
d_s, K_s	Number of synonymous substitutions per synonymous site
EST	Expressed Sequence Tag
ftz	*fushi tarazu* gene
GSS	Genome Sequence Scan
HKA test	Hudson-Kreitman-Aguade test

HOM-C	Homeotic gene complex
Hox	Homeotic gene
HTC	High Throughput cDNA Sequencing
HTG	High Throughput Genome Sequencing
IF	Indel Fixed difference
Indels	Insertions and Deletions
IP	Indel Polymorphism
kb	kilo bases
lab	*labial* gene
LD	Linkage Disequilibrium
LRH	Long-range linkage disequilibrium test
Mb	Mega bases
MK test	McDonald-Kreitman test
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
MYA	Million years ago
N, N_e	(Effective) population size
ncDNA	Noncoding DNA
ncRNA	Noncoding RNA
NI	Neutrality Index
Non-CNS, NC	Non-Conserved Noncoding Sequence
PAT	Patents
pb	Base pairs
pb	*proboscipedia* gene
PDA	Pipeline Diversity Analysis
P_n	Number of nonsynonymous polymorphisms
P_s	Number of synonymous polymorphisms
SNF	Single Nucleotide Fixed difference
SNP	Single Nucleotide Polymorphism
SSR	Short Sequence Repeats, also called microsatellites or short tandem repeats (STR)
STS	Sequence Tagged Site
SYN	Synthetic

TE	Transposable element
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
tgDNA	Total genomic DNA
TPA	Third Party Annotation
UTR	Untranslated Region
VNTR	Variable Number of Tandem Repeats
WGS	Whole-Genome Shotgun
zen	*zerknüllt* gene
zen2	*zerknüllt-related* gene
π	Nucleotide diversity
π_n	Nonsynonymous polymorphism
π_s	Synonymous polymorphism
ω	d_n/d_s ratio (measure of functional constraint)

Index of Tables

Part 1 | Introduction

Part 2 | Results

Index of Figures

Part 1 | Introduction

Part 2 | Results

PART 3 | DISCUSSION

Index of Boxes

Part 1 | Introduction

Part 3 | Discussion