Genomic and Functional Approaches to Genetic Adaptation

Elena Carnero Montoro

TESI DOCTORAL UPF / 2013

Thesis Director

Dra. ELENA BOSCH

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

UNIVERSITAT POMPEU FABRA

Fitxer PDF de la tesi dividit en 7 parts

Part 1 de 7	pàg 0 - 17	Introduction Cap. 1 – Cap 2
Part 2 de 7	pàg 18 - 20	Introduction Cap. 3 : 3.1, 3.2
Part 3 de 7	pàg. 20 – 25	Introduction Cap. 3 : 3.2.1
Part 4 de 7	pàg. 25 – 28	Introduction Cap. 3: 3.2.2 – 3.2.3
Part 5 de 7	pàg. 29 - 64	Introduction Cap. 4 – Cap. 6
Part 6 de 7	pàg. 65 - 183	Objectives, Results, Discussion, Concluding remarks
Part 7 de 7	pàg. 184 – 233	References, Annexes

Part 1 de 7

- Front Cover by Marc Torres Ciuró -

al uno,

a la dos,

a la tres,

(al cuatro).

Acknowledgments

*(En una Mezquita descalza, acurrucada en un rincón, la luz entra directa desde una gran ventana, Danny enfrente es cómplice del momento en silencio. Con un Rosario entre los dedos de una mano. La otra mano pasa las cuentas. No hay rezo, ni plegarias, ni perdón. Solo el mundo se paró y es momento para dar las **GRACIAS**).*

Siempre, cuando uno logra algún objetivo o llega a alguna meta mira a su alrededor y busca quien está a su lado. La victoria para mí es poder compartir este momento con ese bello círculo de familia, amigos, compañeros, que sienten conmigo la alegría de este momento. A todos ellos gracias por estar cerca.

Agradezco a **Elena Bosch**, directora de esta tesis, todas las enseñanzas, la confianza, la libertad y la paciencia que me ha brindado en estos cuatro años de doctorado. **A mi compañero Johannes** por ser tan original, por ser tan creativo, por ir más allá, por velar por mi en muchos momentos y sobre todo por creer en lo que hace. A todo **BioEvo** porque han sido unos años muy bonitos con todos vosotros que no voy a olvidar. Gracias a **Ruben V** y todo su lab por acoger nuestro proyecto con los brazos abiertos.

Mi **gran amiga Vero** ha tenido un papel fundamental en la creación de este documento final. Significa mucho para mi que hayas querido ser parte de ésto y te hayas paseado por cada una de estas páginas con tanto entusiasmo, hasta el fondo y sin vértigo. Sólo una amiga como tú…Gracias!

Siempre-gracias-siempre a los ominipresentes **Iarita** y **Pierin**, por darle color a estas páginas.

A la **meva familia catalana Torres-Ciuró**. Per tot el carinyo que m'heu donat desde el primer dia, per la motivació, per cuidar-me, per fer-me sentir que aquesta és també la meva llar i que soc una més. Moltissimes gràcies a tots!

A **mis abuelos** que están tan dentro de mi.

A **mis padres** que me lo han dado todo. Me siento llena de orgullo e infinitamente afortunada de pertenecer a donde pertenezco. Gracias por el el amor que no tiene límites. Por transmitirnos el espíritu libre, la visión crítica, el valor del esfuerzo y el no tener miedo a superarnos a nosotras mismas. Gracias por todas las celebraciones. Gracias por estar siempre detrás de todas las circunstancias. Por compartir. Por recorrer kilómetros sin medida. Por ser amigos. Por ser tan generosos. Por animarnos a hacer todo aquello que queremos. Por impulsarnos. Y gracias sobre todo por transmitirnos lo más bonito que tenemos, que es la alegría y las ganas de vivir.

A mi **gran hermanoide, Luciita,** por ser la mejor de las amigas, por su admiración, por el apoyo incondicional, por creer en mí más que nadie, por querer estar a mi lado. Porque me encanta que seas yo y me encanta ser tú, y sobre todo, ser contigo. No sabes cuánto te admiro.

A **Marc**, en especial a Marc, al **marcmeu**, por ser parte de todo. Por el gran equipo que formamos. Por llenar cada dia de mi vida de una felicidad que considero absoluta. Por compartir el entusiasmo. Por regalarte entero. También por todas las contribuciones a esta tesis, que son muchas. Porque me encanta llegar a casa y decirte que vengo contenta … y oirte contento … millones de gracias.

La participación de muchas personas a este trabajo hace que esta tesis tenga un valor añadido muy especial. ¡Gracias a la colectividad! (Véase más en "more and more and more and more" en los Anexos)

Summary

The genetic basis of phenotypes that have contributed to the adaptation of species and organisms to new environments is a central question in evolutionary genetics. The recent accumulation of genetic variability data has allowed a genome-wide search for different signatures of positive selection which has led to the discovery of hundreds of putative candidate genes that may have played a role in adaptation. However, such hypothesis-free approaches do not reveal either causal variants or the actual biological mechanisms that have made each adaptation possible. Furthermore, the detection of molecular signatures is limited both by the complex architecture of the genome and by the possible polygenic nature of the selected trait.

In this thesis, through different evolutionary and functional approaches, we have disentangled the adaptive role of two non-synonymous variants in two different candidate genes encoding for a lymphocyte receptor and a zinc transporter, respectively. In past human adaptation, they were most likely selected as more effective means to fight pathogens. We have also revealed differences in the action of natural selection between different pathways and different coding and non-coding genomic elements in the chimpanzee lineage by analyzing polymorphisms and divergence data together.

Resumen

La base genética de los carácteres que han contribuido a la adaptación de los organismos y las especies ha sido siempre una pregunta central en biología evolutiva. Gracias a la acumulación masiva de datos de variabilidad genética, se ha permitido detectar en el genoma diferentes señales de selección positiva y también localizar cientos de genes candidatos que han podido tener un papel en la adaptación de las poblaciones a diferentes ambientes. Sin embargo en estos estudios, donde no hay una hipótesis a priori, se desconoce qué variantes dentro de estos genes fueron realmente las que proporcionaron una ventaja selectiva y por qué. Además, la compleja arquitectura del genoma y la naturaleza poligénica de muchos carácteres hace que sea difícil detectar casos más complejos de adaptación.

En esta tesis se intenta resolver alguno de estos problemas. En primer lugar, mediante un enfoque evolutivo y funcional, hemos descifrado el rol adaptativo de dos variantes genéticas, una en un receptor linfocitario y la otra en un transportador de zinc, que probablemente fueron seleccionadas por conferir resistencia a patógenos. En segundo lugar, mediante el análisis de datos de polimorfismo y divergencia conjuntamente, también se han detectado distintas actuaciones de la selección natural en distintos pathways y entre elementos codificantes y elementos no codificantes reguladores en chimpancé.

Preface

In the not-so-long-ago past, an obvious phenotypic difference between different human populations served as starting point in the search for the genes that were underlying adaptation. This work led to the identification of clear-cut genetic variants related to light skin color, the ability to digest milk as an adult, high altitude adaptation or an increased resistance against pathogens such as malaria. Hence, a largely known phenotype found its gene, which in turn helped to refine the phentype and explain it on the molecular level.

To the contrary in the early days of this thesis, the two chosen non-synonymous variants did not have any attached phenotype. Instead, they were found to be interesting by themselves based on evolutionary grounds. More precisely they were interesting as they appeared to have evolved under natural selection, which was then reinforced through statistical analysis. And since natural selection acts on the phenotype, a key step in this work was to reach out to collaborators in molecular and clinical biology in order to find a link between genotype and phenotype. Fittingly, this thesis draws on an impressive variety of current methods from the “dry lab”, especially population genetics and comparative genomics as well as the “wet lab”, especially next-generation sequencing and cell culture.

Currently there is only a limited set of phenotypes that are known to have evolved under positive selection in the recent history of human populations. Therefore, it is remarkable that this thesis made accessible a novel phenotype under selection which seems related to the thriving field of zinc biology. Particularly, it describes a common human variant with reduced intracellular zinc uptake in SubSaharan African

populations. This evolutionary analysis and molecular phenotype open the door to exciting questions such as:

May there exist a general difference in zinc biology between SubSaharan Africans and other modern humans? Given the observed differences in disease risk in African Americans in diseases such diabetes, cancer and neurological diseases and given the role of zinc in the same diseases, may this variant explain part of these differences? Given the widespread phenomenon of mild zinc deficiency in humans, could such findings be relevant for supplementation programs with zinc in developing countries?
What effect may this variant have on the hundreds of zinc-binding proteins in a cell? What effect may it have on different cell types and organs?
What has been the selective force in the past? Has it been an African pathogen as this thesis proposes?

Every good piece of work opens more questions than it closes, and it will be exciting to see how this beautiful snapshot of the current knowledge plays out in the future. It is encouraging to see how this thesis succeeded to learn about our evolutionary past by combining collaborations and know-how from different fields of research including evolutionary biology, membrane physiology and the clinical field.

-Johannes Engelken-

Index

Introduction

1. Natural selection and adaptation

> "This preservation of favorable variations and the rejection of injurious variations, I call Natural Selection" (Darwin 1859)

A heritable trait that increases the chances for an organism to survive, and/or that benefits its reproduction in a given environmental context is considered an adaptive trait, and it is, thus, the target of natural selection.

By the process of Darwinian natural selection, beneficial alleles (heritable units) responsible for adaptive traits increase their frequency in the population they are segregating. Ultimately, this process increases the fitness of the population and allows a better fitting to its habitat.

This generally accepted and apparently simple concept has been the object of extensive debate concerning the prevalence of adaptive traits, the mechanisms by which they are originated, and the levels at which selection operates.

Only by observing diversity in nature, proto evolutionary naturalists in the 18th century, such as Erasmus Darwin, Buffon, or Montboddo, already stated that living beings have the faculty of improving, transforming and/or acquiring new parts over long time intervals in response to stimuli as well as to endlessly deliver those evolutionary adaptive changes to the following generations. However, none of them really discovered the causes or means of such transformations. It was not until Darwin and Wallace in 1858, that they communicated their independent theory of natural selection, when an explanation on how organisms change or evolve through the generations appeared. In summary, three basic requirements are necessary for natural selection to happen:

First, there should be, in Darwin's own words, a "struggle for existence" since more individuals are born than those that can

actually survive. Second, there should be variation in the ability of individuals to survive and reproduce, so that the fittest are those who survive, "the survival of the fittest". Third, some of this variation should be heritable, allowing each generation to be better fitted to its ecological niche.

While in his extensive work Darwin carefully explained and demonstrated the first two requirements of natural selection, he could not provide, at that time, suitable inheritance models nor a mechanism suitable to generate the observed variability. He first suggested a theory of blending inheritance in which the offspring present a fusion of the parental characters, but he soon realized that such explanation is not valid because it removes variability by producing a uniform generations where selection can no longer act. The absence of convincing models of inheritance brought some skepticism in the field about the natural selection theory until the early 20th century, when Mendel's work was brought to light. This rediscovery coupled with the formulation of the Hardy-Weinberg Law provided mechanisms by which allele frequencies are maintained over generations and, thus, variability is not removed as long as random mating exists in a sufficiently large population. It is only when mating is not random, but depends on the fitness of an individual, that selection can operate.

After the union of ideas from different biological fields, the 20th century brought the Modern Synthesis of Evolution: a consensus theory on how evolution proceeds based on a gene-center view that provided the theoretical mathematical framework that would give birth to the field of population genetics, in charge of predicting how allele frequencies change over generations due to evolutionary forces. Over that time, evolutionary thinking was basically selectionist, meaning that natural selection, in its different modes of action, was the main force considered in determining the observed phenotypic variability and preserving the adapted acquisitions.

1.1 Types of natural selection

Although the genetic determination of a trait is important to explain its heritability, and thus, its ability to be subjected to natural selection, it is crucial to understand that natural selection does not act directly on genotypes but rather on phenotypic traits, which can also be influenced by environmental factors.

Depending on which direction natural selection affects phenotypic variability and, depending on how genetic variability is shaped as a consequence, different modes of selection can be distinguished:

- *Purifying selection*, also referred to as *negative selection*, or *stabilizing selection*. It eliminates deleterious mutations from the population's genetic pool. It preserves function in well-adapted organisms and avoids the spread of damaging mutations over generations. It is thought to be the most common type of selection since mutations impairing basic biological functions seem to be much more frequent than those that increase fitness.
- *Positive selection,* also referred to as *Darwinian selection* or *directional selection.* It is the increase in frequency of a beneficial mutation that somehow enhances the fitness of individuals in a particular environment. This type of selection allows the adaptation of individuals to new environments.
- *Balancing selection.* It sustains the segregation of different alleles in a population. Unlike in positive or negative selection, alleles never reach fixation; consequently, this favors genetic diversity. Alleles under balancing selection cannot be strictly classified as deleterious or beneficial. Different processes can lead to an excess of polymorphisms, such as overdominance (when the heterozygote is the fittest), frequency-dependent selection (when an allele becomes deleterious or beneficial depending on its frequency), fluctuating selection (when

selection coefficients vary over time and/or space) or pleiotropy (when the selective variant affects multiple traits with different effects).

1.2 Neutralism as a theoretical background

According to the selectionist hypotheses, since balancing selection is considered an extremely unlikely event and no other process would maintain genetic polymorphisms in the population, then polymorphism should be expected to be rare. However, the accumulation of molecular and sequencing data in the late 20th century led to the observation that genetic variability was more frequent than expected and that the model in which selectionism was based, suffered from some notorious limitations.

From this observation, Kimura, in the 1960s, provided an explanation to the excess of observed polymorphisms with the neutral theory of molecular evolution (M. Kimura 1968). The neutral theory claims that most of the genetic variation observed is neutral, and thus, does not have a phenotypic effect on its carrier, and that, as a consequence, changes in frequency are governed by random changes also referred to as genetic drift. Although the model considers that negative selection plays a role in the elimination of deleterious mutations, it postulates that beneficial mutations rarely appear, and that adaptive events are rare. According to the neutral theory, the expected amount of diversity is thus governed by the interplay of mutation and genetic drift as evolutionary forces (see figure 1).

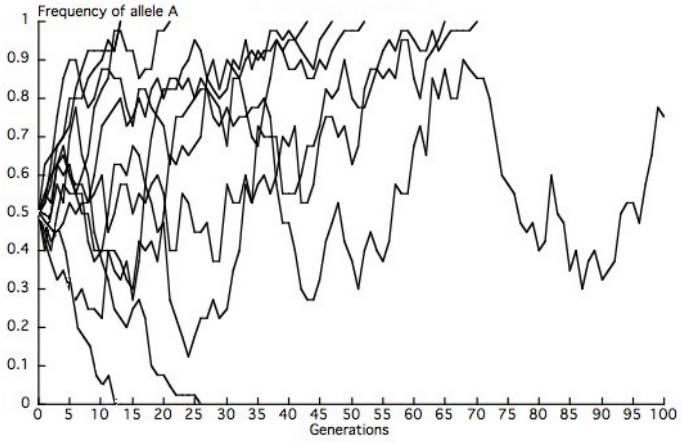

Figure 1. Simulated changes in allele frequency under genetic drift. The fate of the change is random and only depends on the initial frequency of the allele. From Nature Education publishing.

An important outcome of the neutral theory is that it provides the theoretical background to detect the action of the evolutionary forces that disrupt expectations from neutral evolution and affect genomic diversity patterns, as, for example, positive selection.

Today it is generally accepted that the neutral explanation cannot solely explain the genomic diversity detected in organisms, because of two main arguments: first, because part of the bulk of mutations cannot be considered strictly neutral and do somehow affect fitness, and second, because selection, indeed, plays a major role in the adaptation of species and populations as it will be further discussed.

2. Methods to detect positive natural selection

Positive selection shapes the variability of our genomes leaving distinct characteristic footprints, which are identifiable and can be used to explore the genome and reveal the action of natural selection. Moreover, since not all the molecular signatures of selection persist equally in the genome, a timeframe for natural selection can also be inferred from it (see figure 2).

There are two ways to proceed: one can explore the diversity between populations of the same species and detect regions with a special variability pattern of segregating alleles shaped by positive selection, or one can identify fixed substitutions in the genome that have changed between different species due to past positive selection events that have contributed to their divergence.

Statistical tests detecting such molecular signatures are mainly based on the detection of particular regions that do not follow the variability pattern expected under a neutral evolution model. Those are known as neutrality tests.

Selection acts on the genetic basis of a given phenotype, which is usually coded for by one or a few genes. This property is important when disentangling natural selection from other evolutionary forces such as genetic drift or gene flow, which affect the whole genome.

Neutrality tests can be broadly divided in three categories based on the type of data they use: tests based on polymorphism, tests based on divergence between species, and tests that use both polymorphism and divergence data.

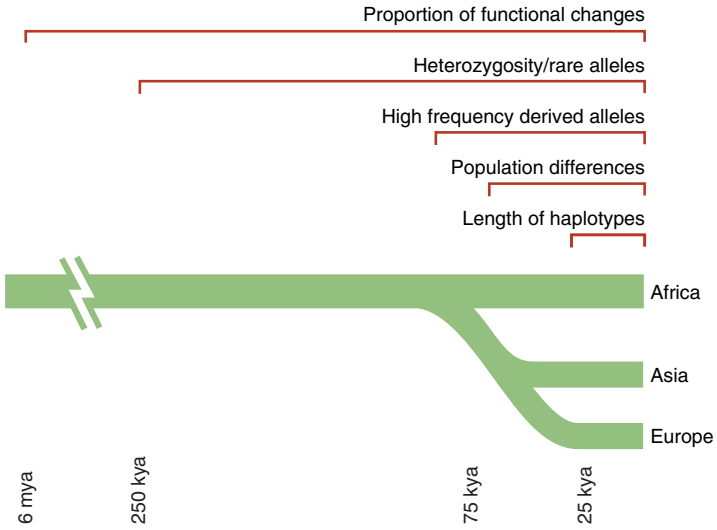

Figure 2. Time scales at which 5 different molecular signatures of positive selection can be detected. Adapted from Sabeti et al. (2006)

This section is dedicated to a brief description of the different features of each molecular signature, the statistical tests that are available to detect them, the evolutionary time at which they can detect positive selection, and the explanation of some of their limitations.

2.1 Tests based on intra-specific variability. Detecting *selective sweeps.*

If a novel advantageous mutation appears in a population, it will cause a molecular signature known as "selective sweep" or "hard sweep". This molecular pattern differs from the background distribution of genetic variation, which generally evolves neutrally (M. Kimura 1968). The signal is characterized by a drastic decrease in variability and an increase in linkage disequilibrium. This can be explained as follows: as the beneficial allele increases

in frequency, it will also drag up the neutral linked variants to high prevalence by a phenomenon called hitchhiking effect (Smith and Haigh 1974) (see figure 3). Eventually, new mutations will slowly restore the variability of the region, and recombination will break the linkage disequilibrium among selected and nearby loci.

The strength of selection will determine the velocity at which the adaptive allele reaches fixation and also the distance at which a neutral site can be affected by a sweep. However, how genomic diversity is affected by a selective sweep is also determined by the local recombination rate. If a recombination hotspot is located near the adaptive site, the genetic diversity may not be noticeably depleted. A rich collection of powerful neutrality tests was devised to detect different molecular patterns of the selective sweeps.

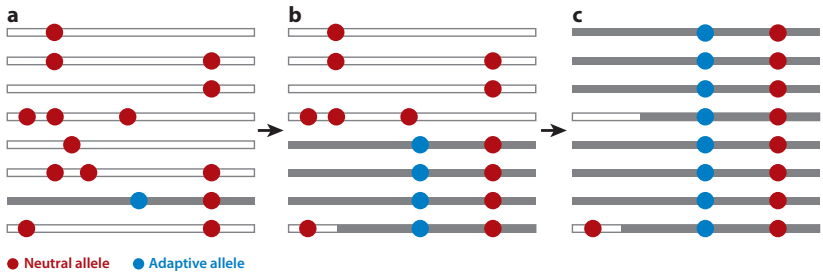

Figure 3. Molecular signatures of a selective sweep over 8 chromosomes. A) Before the selective sweep in a neutrally evolving region, an adaptive mutation (blue circle) appears in one chromosome. B) During the selective sweep, the frequency of the adaptive allele and its linked variants rapidly increase in frequency. C) After the sweep, adaptive and linked alleles are fixed, variability is lost and new mutations begin to appear there or in different chromosomic backgrounds by recombination and mutation. Figure adapted from Kelley and Swanson (2008).

Tests based on long haplotypes. Haplotypes are allelic combinations that are inherited together. In each meiosis recombination breaks such combinations by shuffling the allelic variation of the two homolog chromosomes. Two polymorphic loci are in linkage equilibrium if their allele frequencies in the population are statistically independent from each other. On the

contrary, if one haplotipic combination is more frequent than expected at random, these sites are in linkage disequilibrium (LD). As recombination hotspots are heterogeneously distributed along the genome, allelic combinations are shaped in a haplotype-block manner. Positive selection creates high levels of LD in the region where the adaptive variant is located. This occurs because neutral nearby variants would also become more frequent, as they are dragged together with the selected allele. Thus, recent selective sweeps are characterized by the presence of long haplotype blocks that can be detected by applying LD methods.

These haplotypes can span regions of more than 1Mb. The long-range haplotypes are footprints of very recent selective events (<30,000 years in humans), as eventually recombination will shuffle the variation breaking down the haplotype blocks in fragments that are not long enough to be detectable. The LRH (Long Range Haplotype) test, haplotype similarity and other haplotype-sharing methods are commonly used to detect such signals (Sabeti et al. 2002). One limitation of some of the haplotype-based early tests arose from the fact that recombination rates vary widely over the genome and this variation was not always taken into account in the analysis. However, some tests such as the iHS (Voight et al. 2006)(Tang, Thornton, and Stoneking 2007) and others are meant to overcome this problem by contrasting the LD pattern between the ancestral and derived alleles of polymorphic sites.

Tests based on population differentiation. In humans, genetic differentiation exists mainly after the human migrations out of Africa some 50,000 years ago as a result of genetic drift operating in reproductively isolated populations. The most commonly used statistic to identify differentiation is $F_{ST,}$ the fixation index (Lewontin and Krakauer 1973). As a result of geographically restricted environmental pressures, we can observe extreme allele frequencies between geographically and reproductively isolated populations at those loci conferring local geographical adaptation.

However, an extreme F_{ST} value alone is not a valid evidence of a selective sweep, as random changes caused by demography and genetic drift may cause population differentiation as well. But, unlike natural selection, it would affect all loci equally.

Tests based on the Site Frequency Spectrum. The site frequency spectrum (SFS) is the distribution of the number of alleles at different frequency classes for any given set of polymorphic sites in a sample. Due to hitchhiking, regions under positive selection contain many high frequency derived alleles, a pattern that is not common in a neutrally evolving region (see figure 5). The most widely used neutrality test to detect such excess of derived alleles is the Fay and Wu's *H* statistic, which is able to detect adaptive events up to 80,000 years ago (J C Fay and Wu 2000). It might be difficult sometimes to perform this test, as the unfolded site frequency spectra -which includes information about ancestral states for all sites- is required. Another limitation of this test is that it loses power as high frequency alleles affected by the sweep reach fixation.

The site frequency spectrum created by a selective sweep is also characterized by a reduction of diversity and a skew towards an excess of rare alleles. New mutations will appear slowly in the region swapped and will be present at low frequencies (see figure 4). These signatures persist longer in the genome (< 250,000 years) and help recognizing older selective events close to the origins of modern humans. Statistical tests used to detect such excesses of rare alleles and reductions of diversity are Tajima's D, the Hudson-Kreitman-Aguadé (HKA) test, and Fu and Li's D* (Tajima 1989) (Hudson, Kreitman, and Aguadé 1987) (Fu and Li 1993).

These methods do not need prior knowledge of ancestrality since they are based on the folded site frequency spectra. Unfortunately, changes in population size can also generate a similar pattern of reduced diversity and excess of rare alleles and thus be mistaken

for selective sweeps. Although these phenomena will affect all the genome, the effects of selection can be seen only in the loci conferring adaptation.

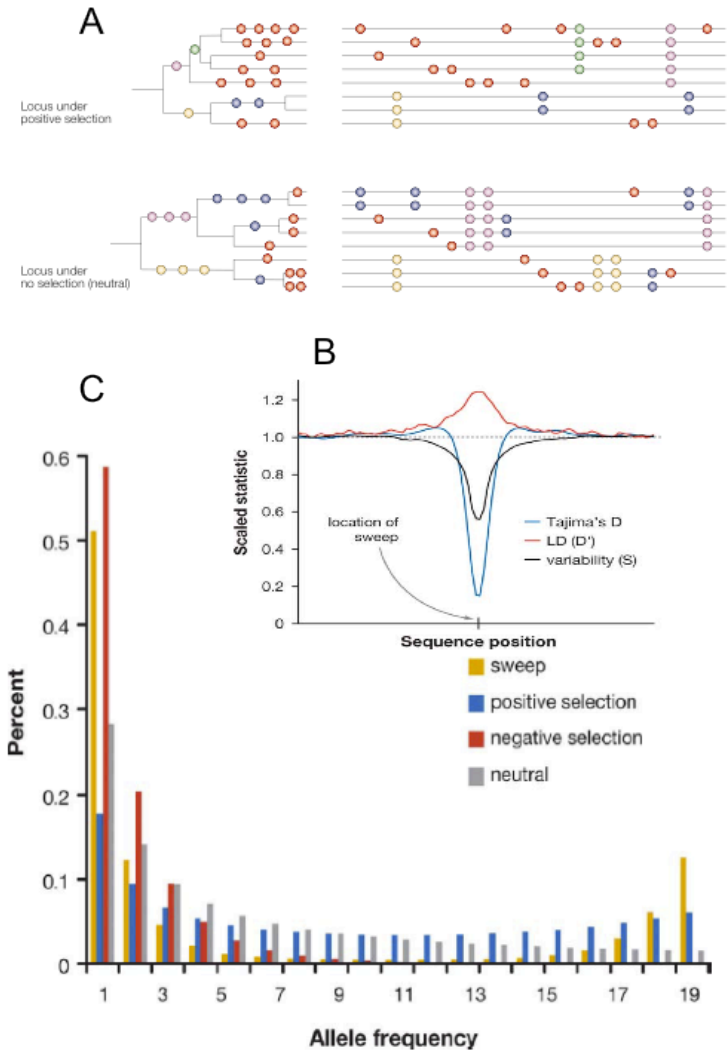

Figure 4. Effect of positive selection on the distribution of genetic variation. Positive selection results in lower levels of diversity, in an excess of low-frequent high-derived variants and in longer haplotype blocks. A) Genealogy of loci under positive selection and neutral evolution plus its corresponding haplotype representation. The hitchhiking effect can be observed in the gene genealogy under positive selection (Bamshad and Wooding 2003). B) Tajima´s D, diversity and LD values (Nielsen 2005). C) Effect of selective sweep on the site frequency spectra (Nielsen 2005).

Significance of tests based on population data and distinction from demographic effects is usually estimated through simulations that model neutrality under a number of demographic models in a genomic region with similar molecular characteristics (size, recombination rate, mutation rate, CpG content, etc) to those we are studying, and by comparing where the values of the studied data are located within the values of the simulated distribution. Alternatively, in a genome-wide study, significance of a test can be evaluated based on the genome-wide empirical distribution of the same statistic. Since positive selection is expected to be a non-frequent process, it will lead to outlier values in the tail of the genomic distribution.

2.2 Adaptation not only by classical selective sweeps

The "hard sweep" mode of selection refers to the classical sweep model as described by Maynard Smith and Haight (2007), where a new mutation increases rapidly its frequency towards fixation due to selection and affects the variability pattern of linked loci. Most tests intended to search for signatures of selection are based on the identification of the molecular signatures left in the genome by this hard sweep.

This topic has been largely discussed, and it was argued that this vision is too limited. It is important to recognize that adaptation can leave other types of genetic footprints beyond the classical hard sweep models due to several aspects:

- First, it was shown that hard sweeps are not equally distributed in the genome, for example, genomic regions with low recombination rates are enriched in hard sweeps (Cai et al. 2009). Such observations do not imply that selection does not occur in regions with higher recombination rates, but more likely that,

because recombination can obscure the signal of selection by shuffling the variation, hard sweeps are not detectable anymore in regions with higher recombination rates. The lack of haplotypic structure around regions with high-F_{ST} SNPs (Single Nucleotide Polymorphisms) supports this scenario (Coop et al. 2009).

- Second, the demography and structure of the population can influence the genetic patterns of regions under selection. For example, hard sweeps are difficultly recognizable in African populations, and are, on the contrary, more frequently detected in Asians (Pritchard, Pickrell, and Coop 2010). This, again, does not imply that selective events have not happened with the same frequency in all populations, but rather that population demographic history hides the signal in Africans and makes it more difficult to recognize (Alves et al. 2012)(Granka et al. 2012).

- Third, other modes of selection are possible beyond the "hard sweep" model, and they are generally referred to as "soft sweeps". For example, selection could have occurred not due to a new variant, but due to some already existing standing variation that was neutral or mildly deleterious prior to the presence of the selective pressure. In this possible scenario, we would not find the typical footprint of a hard sweep, which consequently could not be detected by classical tests (Pritchard, Pickrell, and Coop 2010).

- Last, due to the non-monogenic character of many traits that could affect population fitness, polygenic adaptation is also expected to be a frequent mode of selection that has not been taken into account in classic population genetic methods. Classic quantitative genetics and recent genome wide analysis (GWASs) support the idea that many phenotypically variable traits are genetically determined by a large number of loci producing small phenotypic changes. This goes against the theory of single mutations being responsible for advantageous phenotypes. Polygenic adaptation would allow rapid local adaptation by

modestly changing frequencies of many loci. Signatures of hitchhiking coming from such modest changes in frequency are not likely to be detectable, so other methods that collect information about differences in frequency of loci mapped to the phenotypic variation are needed in order to recognize which genetic variants underlying the phenotypic variation could be a target of selection (Pritchard and Di Rienzo 2010) (see figure 5).

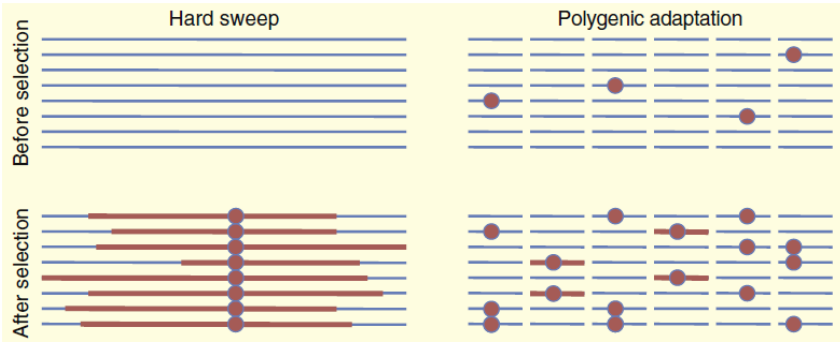

Figure 5. Hard sweep and polygenic adaptation models. Hard sweeps are produced by a new mutation in a particular locus that rapidly reaches high frequency in the haplotype pool. Polygenic adaptation is due to small frequency changes of standing variation at many loci that were already present in the population prior to selection. From (Pritchard, Pickrell, and Coop 2010)

In summary, although the recognition of hard sweeps has given important insights of many loci contributing to the adaptation of human population, they are not sufficient to explain all the genetic adaptation that our species have undergone. Detecting hidden hard sweeps or soft sweeps by population genetic methods is difficult because they do not leave a clear genomic signature that can be easily distinguished from neutral evolution. Simulations of what is expected in these selective scenarios might help recognizing new adaptive events. Also, the incorporation of phenotypic information in cases where polygenic adaptation is suspected and performing quantitative analysis can help detecting those genetic variants that have undergone positive selection and have influenced multigenic adaptive phenotypes.

2.3 Tests based on inter-specific variability. Detection of increased proportion of functional mutations.

The regions that have participated in adaptation at older time scales (millions of years ago) can be identified trough inter-specific comparisons by measuring the proportion of potentially functional to putative neutral substitutions. A mutation that impacts the function of a particular gene, in the protein-coding region or in a regulatory element, is functional, and thus potentially subject to natural selection. Neutrality tests using divergence data differ in whether they are based solely on divergence data or they also include polymorphic data.

Divergence-based tests. In a neutral evolving region, the ratio of the substitution rate in functional sites compared to the rate in neutral sites is close to one, as measured by the d_N/d_S (non-synonymous to synonymous) ratio in coding regions. On the contrary, functional elements are usually constrained by purifying selection, and since mutations altering their function are commonly deleterious and thus eliminated, they typically present d_N/d_S values lower than 1. Furthermore, a high proportion of functional mutations can increase the d_N/d_S ratio, a signature that can be taken as evidence for relaxed purifying selection or can even be interpreted as a signature of accelerated or adaptive evolution.

Divergence and polymorphism-based tests. Methods such as the McDonald-Kreitman test (MKT) and related tests allow contrasting the levels of polymorphism and divergence at putative neutral and selected sites (figure 6); and also to estimate both the fraction of substitutions at functional sites that were driven to fixation by positive selection (denoted as α), as well as the adaptive substitution rate at which they appeared (denoted as ω_α). The application of these tests has been fundamental to answer a central question in evolutionary biology: *what proportion of the differences we see among species is adaptive*?

Because the MK test uses alternated sites as neutral reference, it takes into account possible variations in coalescence histories and potential mutational rate differences among the genomic regions included in the analysis. Therefore, it is a robust test to perform and rule out this potential source of error. However, it can be mistaken for the presence of slightly deleterious mutations segregating in the population as a result of particular demography events or linkage effects as these mutations will contribute to polymorphism.

Over the last years, much effort has been undertaken to overcome such methodological limitations with extensions of the MK tests, as for example, those using the SFS of neutral polymorphisms to estimate the demography history, to assess the distribution of fitness effects at functional sites by comparisons with the SFS of selected polymorphisms, and to estimate the rate of adaptation (Schneider et al. 2011)(Boyko et al. 2008). Recently, a new work has shown that most of the polymorphisms segregating and affecting α and ω_α values are produced by linkage effects and genetic draft, and proposed a new method based on an asymptotic extension of MK-based tests (Messer and Petrov 2013).

	Nonsynonymous (N)	Synonymous (S)	N/S
Polymorphism	6	25	0.24
Divergence	16	40	0.40

$$\alpha = 1 - (pN/pS)/(dN/dS) = 1 - (0.24)/(0.40) = 0.40$$

Figure 6. The MacDonald-Kreitman test (MKT) formulation. The MKT estimates whether the ratio of functional (i.e. non-synonymous) to neutral (i.e. synonymous) polymorphisms differs statistically from the ratio of functional to neutral divergence. Excess of functional divergence compared to polymorphism is attributable to positive selection. The parameter α estimates the proportion of functional substitutions driven by positive selection. From Fay (2011).