INCORPORATING RECOMBINATION INTO THE STUDY OF RECENT HUMAN EVOLUTIONARY HISTORY

Marta Melé Messeguer

TESI DOCTORAL UPF / 2011

Directors de la Tesi:

Dr. Jaume Bertranpetit Busquets. Departament de Ciències Experimentals i de la Salut.

Dr. Francesc Calafell Majó. Departament de Ciències Experimentals i de la Salut.

UNIVERSITAT POMPEU FABRA

A l'Uri

Acknowledgements

La realització d'aquesta tesi no hagués estat possible sense l'ajuda de nombroses persones que m'han donat suport, m'han ajudat i han confiat en mi:

Als pares, que sempre m'han donat tot el suport i la llibertat per fer el que volgués. Puc haver pres decisions més o menys encertades, però al final, he acabat fent el que volia, el que m'agradava, i sempre m'he sentit recolzada per vosaltres. Aquesta tesi és en gran part mèrit vostre.

A la Marina, que en aquest temps hem passat de viure juntes a un poble, a estar separades a la ciutat i seguim tan unides com sempre. Per les hores que m'has escoltat, sentit riure i plorar, per totes les passejades i les excursions, per venir a veure'm a NY, i sobretot, per creure tant en mi.

A la Maria, per estar sempre disposada a ajudar-me a estudiar, un trosset dels meus triomfs són gracies a tu. A l'Oriol i la Magda, que m'han acollit amb els braços oberts i confiat en mi durant tot aquest temps.

A tots els amics de la Uni, com me n'alegro d'haver escollit biologia!! Per tot el que hem passat aquests anys: a Sidreries, al Diamant, a Sant Iscle, a Arrós... Perquè sou el millor remei contra l'estrés, la millor combinació per sortir de festa, els millors companys de carrera i post-carrera i això durarà ja per sempre. Al Zariguey, al Contoni, al Rober, al Setum, al Torrent, al Jordi i al Pau; a la Judit, la Clàudia, la Glòria, l'Aurora i la Montse.

A tot Biologia Evolutiva, amb qui he compartit moltíssims bons moments dins i fora de la feina, jugant a futbol, a voley, anant a córrer, als retreats, als congressos i als crazy Fridays, que m'heu acceptat encara que "canti per dins" i tingui fred si em menjo un yogurt! A la Ixa, a la Mònix, a l'Angel, a l'Urko, al Johannes, al

Martin, a l'Òscar, als Ferran-David-Roger, a la Judit, al Hafid, a la Graciela, a la Valeria, a la Ludovica, i a tot Bioevo en general, m'he trobat com a casa amb vosaltres.

A la Belén, pels bons moments que hem passat juntes al PRBB, pels teus consells durant la tesi, que sempre l'has clavat!! M'has ajudat moltíssim en moments en que ningú més no podia. I tesi a part, ens ho hem passat molt bé a Roma, a Nova York, a Seattle i a Chicago, de festa per Barcelona i dinant fora perquè si.

També vull donar gràcies a l'Arcadi i al Tomàs, pel seu interès en la trajectòria del meu projecte durant aquests quatre anys i pels seus consells. També, als meus companys de grup i de despatx (que han sigut molts!) i a tots els que heu seguit la meva feina, vingut als meus seminaris, i donat feedback i ànims sobre el que hem estat fent.

A l'Asif, per estar sempre disposat a donar un cop de mà i escoltar-me. A la Laxmi per donar-me la oportunitat d'anar a IBM.

Al Marc, per l'empenta que vas donar al projecte quan més ho necessitava; amb tu he après a treballar en equip. De cada 5 dies, 4 de dolents i un de molt bo, han acabat donant resultats boníssims! IRiS és tant meu com teu, i els dos hem fet que funcioni.

A les dues persones que han fet possible aquesta tesi: al Francesc i al Jaume. Al Francesc perquè m'has estat ajudant des del primer dia, resolent qualsevol dubte, ensenyant, corregint, discutint, he après moltíssim al teu costat. Al Jaume, en primer lloc, per haver-me donat l'oportunitat de fer la tesi. Perquè sempre has trobat un forat per mi a la teva agenda impossible, perquè has sabut disfrutar de les nostres discussions (i la meva tossuderia), i, sobretot, perquè m'has valorat i donat suport quan més calia.

I finalment, a l'Uri, amb qui més he compartit aquesta tesi, amb qui més n'he parlat i discutit, sempre disposat a escoltar les meves idees

i a fer-me mirar les coses des de nous punt de vista. Perquè has sabut fer-me parar i respirar quan calia, m'has deixat treballar fos quan fos i m'has ajudat a fer més quan ja no podia. Tu més que ningú m'has donat suport i a tu dedico aquesta tesi, per totes les coses que hem viscut junts aquests quatre anys.

Abstract

The aim of this work is to use the information left by recombination in our genomes to make inferences on the recent evolutionary history of human populations. For that, a novel method called IRiS has been developed that allows detecting specific past recombination events in a set of extant sequences. IRiS is extensively validated and studied in whole set of different scenarios in order to assess its performance. Once recombination events are detected, they can be used as genetic markers to study the recombinational diversity patterns of human populations. We apply this innovative approach to a whole set of different human populations within the Old World that were specifically genotyped for this end and we provide new insights in the recent human evolutionary history of our species.

Resum

En aquest treball es pretén utilitzar la informació que deixa la recombinació al nostres genomes per fer inferències sobre la història evolutiva recent de les poblacions humanes. Per fer-ho, s'ha desenvolupat un mètode novedós, anomenat IRiS, que permet la detecció de recombinacions antigues específiques en un conjunt de seqüències. Hem validat extensivament IRiS i l'hem sotmès a diferents escenaris per tal d'avaluar-ne l' eficàcia. Un cop els events de recombinació són detectats, es poden utilitzar com a marcadors genètics per estudiar els patrons de diversitat de les poblacions humanes. Finalment, hem aplicat aquesta innovadora aproximació a un conjunt de poblacions humanes del Vell Món, que varen ser genotipades específicament amb aquesta finalitat, aportant nous coneixements en la història evolutiva recent dels humans.

Preface

The paradigm on how variation is inherited was solved in the early twentieth century, when Mendel's rules of inheritance were reconciled with a Darwinian theory of evolution. We now have an understanding on which are the forces that shape the variation present in our genomes; which are their dynamics, and how they can be modeled.

Although recombination is one of the main sources of this genetic diversity, the footprint it leaves is far more difficult to detect than the one left by mutation. Therefore, the study of human genetic variation and the inferences about our evolutionary history have been based on the study of markers generated by mutation such as SNPs and STRs whereas recombination has been put aside for its complexity and lack of informative data.

Nowadays, data is produced at an astonishing rate thanks to fast and cost-efficient high throughput technologies, and new computational methods are being developed to analyze such data. The era in which complete genomes are sequenced in a matter of days has just arrived.

It is in such revolutionary context that our project was born: trying to incorporate recombination in the study of human genetic variation was a very challenging project but was now feasible. I believe that incorporating recombination into the study of human diversity may have a whole set of different applications that will not be restricted only to population genetics but also to study the mechanisms underlying recombination at the genomic level.

.

Index

Abstract ix

Resum ix

Preface xi

1. INTRODUCTION 1

1.1. Sources of genetic diversity 3

1.1.1. Mutation 3

1.1.2. Recombination 4

1.1.3. Random Genetic drift 9

1.1.4. Migration 13

1.1.5. Selection 15

1.2. Making inferences from diversity 19

1.2.1. Phylogeography: mtDNA and Y chromosome 19

1.2.2. Summary statistics 23

1.2.3. Coalescent-based inference 25

1.2.4. Bayesian clustering analysis and Principal Component Analysis 27

1.3. The recent human evolutionary history 31

1.3.1. Origin of Anatomically Modern Humans 31

1.3.2. Routes of the Out of Africa migration 32

1.3.3. Tempo and mode of the Out of Africa 34

1.4. Recombination in the study of human population history 37

1.5. How to detect recombination 39

1.5.1. Computational methods to detect the presence of recombination 39

1.5.2. Methods to infer recombination rates 41

1.6. Recent findings on recombination 45

2. OBJECTIVES 49

3. RESULTS 53

3.1. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns 55
3.2. A New Method to Reconstruct Recombination Events at a Genomic Scale 79

Supplementary information 95

3.3. The footprint of recombination gives a new insight in the effective population size and the history of the Old World human populations 105

Supplementary information 124

3.4. SNPs, haplotypes and recombination: a human variation study of the Old World 133

SNPs and haplotypes: a human variation study of the Old World 135

4. DISCUSSION 159

4.1. Challenges of the study of recombination 161
4.2. Inferring the Ancestral Recombination Graph? 162
4.3. Sensitivity, false discovery rate, accuracy in placing the breakpoint 163
4.4. Which recombinations are detected? 164
4.5. Gene conversion, recurrent mutation, genotyping errors, and phasing errors 167
4.6. Recombination in human population genetics 168
4.7. IRiS applied to other organisms 170
4.8. IRiS from SNPs to sequences and the genomic scale 173
4.9. Recombination and selection 174
4.10. Recombinations detected by IRiS, recombination rates and the evolution of hotspots 175
4.11. Concluding remarks 176

MAIN REFERENCES FOR THE INTRODUCTION 179

REFERENCES 181

APPENDIX A. Contributions to other articles..... 201

A1. Minimizing recombinations in consensus networks for phylogeographic studies..... 203
A2. On recombination rates and with genetic differentiation in humans..... 217

APPENDIX B. Contributions to other articles as a Genographic Consortium member..... 241

1. INTRODUCTION

1.1. Sources of genetic diversity

Two main processes generate diversity in genomes: mutation and recombination. Mutation is the sole source of new alleles whereas recombination creates new combinations of alleles at different loci. Other processes shape diversity, such as genetic drift, selection, gene conversion and migration.

1.1.1. Mutation

By definition, any change in a DNA molecule producing a new allele is called a mutation. Only those mutations occurring on the germ-line will be heritable and may carry evolutionary consequences. Changes produced by mutation can be substitutions, insertions or deletions (indels) of single bases or small segments, expansions or contractions of the number of tandemly repeated DNA motifs, insertions of transposable elements, duplications, deletions and inversions of megabase segments of DNA, translocations of chromosomal segments and even changes in chromosome number. These mutation events are caused by different mechanisms and they may have very different dynamics and rates.

Base substitutions are generally caused either by the misincorporation of nucleotides during replication, or due to mutagenesis caused by chemical agents or physical damage such as ultraviolet radiation. By definition, they will give rise to single nucleotide polymorphism (SNPs) once the new variant reaches a frequency higher than 1% in the population. In humans, the estimated substitution rate per generation and nucleotide has been estimated to be around 10^{-8} per nucleotide per generation (Roach et al. 2010; The 1000 Genomes Project Consortium 2010). On the other hand, it has been shown that the substitution rate is not uniform; for example, CpG dinucleotides mutate at a rate ten times faster than that of any other dinucleotide. Moreover, the average mutation rate of mitochondrial DNA is 3.33×10^{-7} substitutions per generation (Soares et al. 2009), with a rate an order of magnitude faster in some segments of the mtDNA molecule. Another type of

genetic variation are the Variable Number of Tandem Repeats (VNTRs), which are sequences arranged in tandem arrays and depending on the length of their arrays are classified into microsatellites, minisatellites and satellites. Mutation at the VNTR involves both changes in the number of repeats and in repeat composition.

Microsatellites or STRs are tandem arrays of repeats of 1- 6 bp. They have been largely used to study human genetic variation, basically, because they are highly informative: they are multiallelic and they have a high mutation rate (around 10^{-3}-10^{-4} per locus per generation). The mechanism underlying the mutation process is thought to involve replication slippage.

Mutation rates are different between males and females; and in general terms males have higher substitution rates within the germline cells, which are thought to be related with the higher number of cell divisions that sperm undergoes, compared to oocites. Further, since in the sperm line the number of cell divisions increases with age, the number of substitutions also increases with the father's age. On average, the male:female substitution ratio has been estimated to be 5:1 (Makova and Li 2002). In the case of microsatellites, microsatellite mutations occur three to five times more often in fathers than in mothers although a smaller paternal age effect is observed (Gusmão et al. 2005; Jobling et al. 2004).

1.1.2. Recombination

Recombination is defined as the process by which a molecule of DNA is broken and then joined to a different one. Recombination can occur between similar molecules of DNA, as in homologous recombination, or dissimilar molecules, as in non-homologous end joining.

The process of homologous recombination starts with a double strand break (DSB) that will then be repaired using the sister chromatid as a template. This process can end in crossing over

when flanking markers have exchanged or in a non crossing-over event named gene conversion in which the initializing chromatid acquires a short sequence from its homologous partner with no exchange of flanking markers. See Figure 1 for a more detailed explanation of the process.

The combination of different alleles within a chromosome is called a haplotype. Recombination, by creating new allele combinations, increases haplotype diversity in a population. Similarly, recombination can break up combinations of alleles unless they are very close to each other. In a population sample, the non-random association of alleles at different loci more or less often than would be expected by chance is called linkage disequilibrium (LD). When a new allele appears in a population, it does so in a particular haplotype, meaning certain alleles will be associated to it. As generations go by, this new combination of alleles may be broken depending on the recombination rate between these two markers.

There are several measures of LD. One of them is the D statistic, which is calculated as the observed frequency of the haplotype formed by alleles A and B minus the expected frequency if those two alleles were statistically independent. Also, D' can be calculated as the value of D divided by the maximum possible value of D given A and B allele frequencies. Then, if D' is equal to one, it means that evidence of recombination is absent in the population sample analyzed. Another measure of LD is r^2, which is the square of the correlation coefficient between the two loci. r^2 can be equal to one if only the two complementary haplotypes (i.e. AB and ab) are present; then one locus carries complete information about the contents of the other locus, as the allelic state of locus A can be perfectly predicted from locus B, and vice versa. This measure of LD is less inflated by small sample sizes than D'.

The average recombination rate over the genome is around 1 cM/Mb which corresponds to 10^{-8} recombinations per locus per generation. However, recombination rates vary greatly at the megabase scale, at the sequence level, and between sexes.

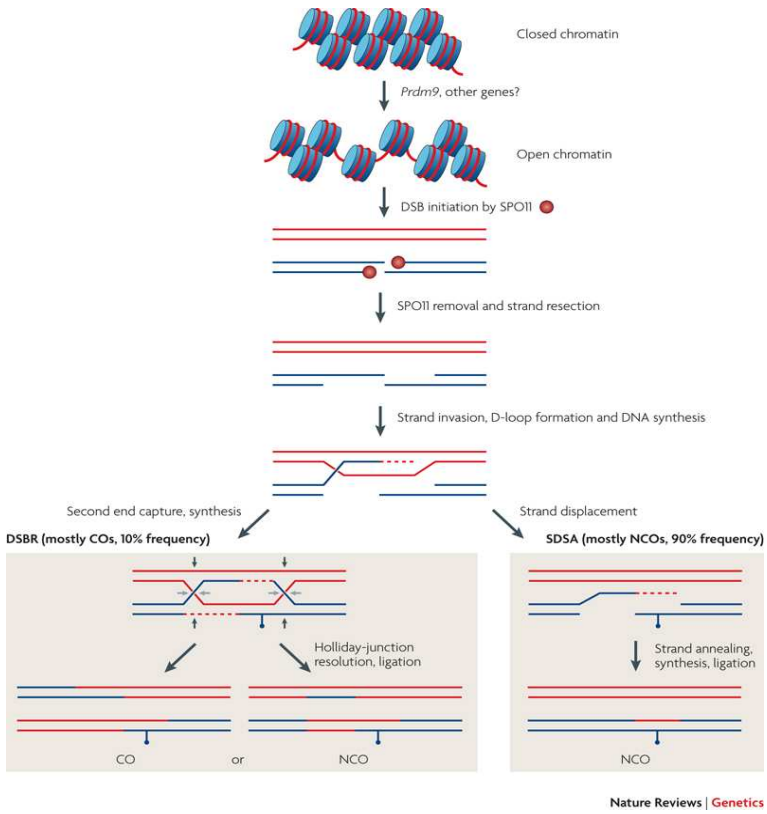


Figure 1. Recombination begins when the products of trans-acting genes, such as PR domain-containing 9 (Prdm9), locally activate chromatin, permitting the topoisomerase sporulation-specific 11 (SPO11) to catalyse a DNA double-strand break (DSB) on one of the four chromatids. This is followed by resection of the 5′ strand to leave a 3′ overhang, which in turn invades a non-sister chromatid. The resulting strand overlaps to form so-called Holliday junctions, which then migrate outwards away from the original site. This interaction promotes pairing of the non-sister chromatids along their length. The DSBs are subsequently repaired by the process of homologous recombination, yielding either crossovers (COs), with an exchange of flanking markers, or non-crossover (NCO) gene conversions in which the initiating chromatid acquires a short sequence from its homologous partner without the exchange of flanking markers. In either case the site of the original DSB is repaired using the opposite chromatid as a template; when SNPs are available in the middle of the hot spot this fact can be used to determine which chromatid initiated recombination. Current evidence suggests that the alternative CO and NCO products arise by distinct recombination pathways: DSB repair (DSBR), which yields predominantly COs, and synthesis-dependent strand annealing (SDSA), which yields predominantly NCOs. The SDSA pathway predominates, and only about 10% of original DSBs result in COs. Figure taken from the review Paigen et. al 2010.

Recombination rates have been shown to be much higher than average on the telomeres (3 cM/Mb), lower in the centromeres (0.1 cM/Mb) and, on average, about twice as high on the smallest chromosomes compared with the largest ones. In fact, it is thought that one recombination per chromosomal arm per generation is necessary for the correct separation of the chromosomes during meiosis (Figure 2). Moreover, rates are strongly positively correlated with GC content and with other genomic properties, notably gene density.

At the sequence level differences are much stronger. In fact, it has been shown that only 20% of the sequence undergoes 80% of all recombinations (Myers et al. 2005). This is due to the fact that there are 1-2 kb regions in the genome called hotspots that have recombination rates that are four orders of magnitude higher than neighboring regions (called coldspots). See below for a discussion of recent findings on recombination mechanisms.

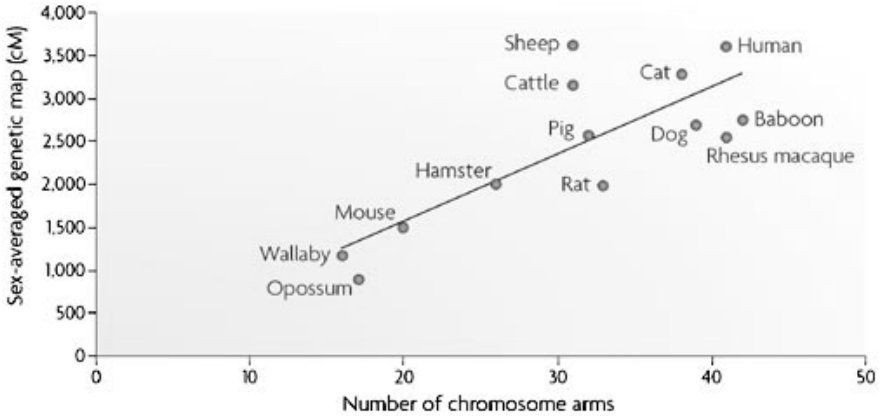


Figure 2. Relationship between total sex-averaged genetic-map length and total number of chromosomal arms in different species, excluding the small arms of acrocentric chromosomes. This correlation suggests that recombination is necessary to occur in each chromosome for the correct separation of homolog chromatides in the meiotic process. Figure taken from Coop et. al 2007.

Recombination rate is higher in females than males being about two fold higher in females (Kong et al. 2010). The distribution of crossover locations also differs between sexes, tending to be lower at the telomeres and higher near the centromere in females compared to males. At the fine scale, females tend to recombine in location between genes and males between exons (Figure 3) and it has been recently estimated that 15% of the hotspots are sex-specific (Kong et al. 2010). Finally, recombination rates are also different among individuals of the same sex, and these differences are inheritable (Broman et al. 1998; Kong et al. 2002; Kong et al. 2008). In females, the number of crossovers varies enormously among the oocytes for the same women (Lenzi et al. 2005).

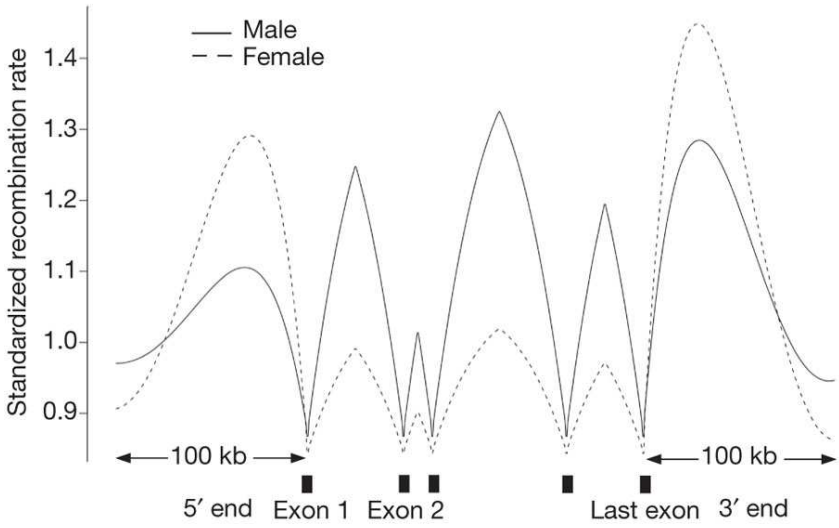


Figure 3. Schematic picture summarizing general trends on the recombination location in males and females from an extensive pedigree study of 15,257 parent-offspring pairs. It is not meant to reflect the recombination rate pattern around a specific gene. Male recombination rate, although low at exons, tends to be high at intronic regions that are distant from exons. Male and female recombination rates both tend to be high at intergenic regions around 40 kb from the first or last exon of a gene, but it is higher for females. Also, for both sexes, intergenic regions close to 3′ ends tend to have higher recombination rates than those close to 5′ ends. Figure taken from Kong et. al 2010.

1.1.3. Random Genetic drift

Genetic drift is defined as the source of variation on allele frequencies given the sampling process of gametes from one generation to the next. Although changes in allele frequencies due to random genetic drift in any individual population cannot be predicted, the average behavior of allele frequencies in a large number of populations can be. The population model to study these effects is called the Wright-Fisher Model and was described by Wright and Fisher independently (Fisher 1930; Wright 1931). This model has several assumptions:

- Non overlapping generations
- Constant population size
- Random mating (panmixia)
- A random Poisson-distributed number of offspring per individual

Under the Wright-Fisher model, it can be shown that the lower the population size the stronger the genetic drift. In fact, a new allele arisen in a small population will not only have higher probability of becoming fixed but it will also be fixed more rapidly than it would in a larger population (Figure 4). Specifically, it can be assumed that $T= 4N$ where T is time to fixation and N is the size of the population.

Moreover, the model implicitly assumes that the population has persisted over a long period of time such that it has reached an equilibrium state. Under this equilibrium state, the diversity present in the population will be constant over time since the same amount of new variants that appear through mutation are removed from the population due to genetic drift. This equilibrium value of diversity is known as the population mutation parameter or theta (θ) and it combines information on the mutation rate (μ) and the effective population size (N_e) of the population:

$$\theta = 4N_e\mu$$

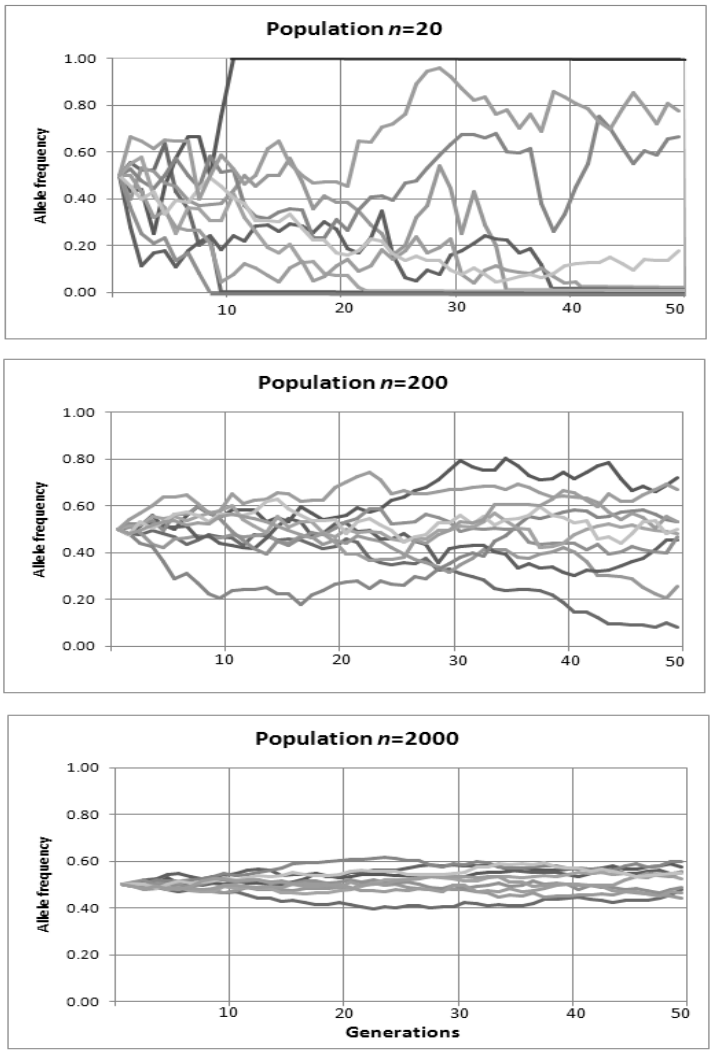


Figure 4. Effect of population size on genetic drift with varying population sizes. Each figure shows ten simulations of random change in the frequency distribution of a single hypothetical allele over 50 generations. (A) Population size = 20. (B) Population size = 200. (C) Population size = 2000. In the population of constant size of 20, alleles can either become fixed or lost very rapidly whereas more subtle variations are seen in the populations with larger sizes. Figure by *professor Marginalia* in http://upload.wikimedia.org/wikipedia/commons/a/a0/Random_genetic_drift_chart.png

The same equilibrium will be reached regarding the diversity generated by recombination. Recombination generates new haplotypes by breaking up linkage disequilibrium between alleles whereas random genetic drift will remove those haplotypes from the population. Therefore, the amount of recombination that we will find in the population will be determined by the population recombination parameter or rho (ρ) which combines information on the effective population size (N_e) and recombination rate (r).

$$\rho = 4N_e r$$

Moreover, rho can be directly related with the amount of linkage disequilibrium measured by r^2 within a population by the approximation (Hill 1975):

$$E(r^2) \sim 1/(2+\rho) \sim 1/(2+4N_e r)$$

The assumptions made for the Wright- Fisher Model, however, are unrealistic and, in most of the cases, the census size and the effective population size are different. Then, for any population, its effective population size represents the size of an idealized Wright-Fisher population that experiences the same amount of genetic drift. Therefore, the effective population size gives a hint on the magnitude of genetic drift that a population may have undergone. In general terms, the effective population size is less than the census size. For example, when generations overlap, which is the case for humans; it has been shown that N_e is roughly N/3.

There are several scenarios in which some of the assumptions of the Wright-Fisher model are not fulfilled. For example, when the population size is not usually constant over time, the effective population size has been shown to be equal to the harmonic mean of the population sizes over time. This means that the effective population size is extremely influenced by the low values of previous population sizes.

Another case in which the Wright-Fisher assumptions are not fulfilled is when the variance in the reproductive success is high, which means that the number of offspring each individual has is highly variable. The higher the variance, the lower the effective population size is, compared to the census size. Moreover, because differences in the variance on the number of offspring can be different between males and females, the effective population sizes of the two sexes may be different as well.

Individuals of a population do not mate randomly due to several reasons. Individuals may choose their mates to be more similar to themselves than randomly expected; this phenomenon is called assortative mating and increases genetic drift as long as the features a mate chooses on are inheritable. For instance, human couples show a positive significant correlation for height. The reverse phenomenon is called dissortattive mating and decreases genetic drift. Choosing mates having an HLA type different from one's own would be an example.

Census size and effective size can also be different when the population is substructured, meaning that what we call a population is made of partially isolated subpopulations and individuals belonging to one subpopulation will tend to mate among them. This will increase random genetic drift and decrease the effective population size. Note, however that whereas random genetic drift decreases diversity in each of the subpopulations, it will increase the genetic differentiation between them.

One of the most common statistics to measure the degree of substructure between subpopulations is the F_{ST} statistic, which specifically measures the difference in allele frequencies between the subpopulations. For one locus, $F_{ST} = var\ (p_i) / p\ (1-p)$ where $var\ (p_i)$ is the variance of allele frequencies in the subpopulations i and p is the average allele frequency across all subpopulations.

Several demographic events may have an effect on the effective population size. Two of the most relevant are bottlenecks and founder effects. A bottleneck however, refers to a reduction on size

of a previously large population whereas the founder effect is related to the process of colonization and the genetic separation of a group of individuals from a source population. Both imply a reduction of the effective population size and a loss of diversity.

Finally, different regions of the genome may have different effective sizes. If we consider a single mating couple, together they have four copies of the autosomes, three copies of the X chromosome, one copy of a Y chromosome, two copies of mtDNA from which only one will pass to the next generation. Therefore, the effective size of the X is ¾ of that of the autosomes and for the Y and the mtDNA it is ¼.

1.1.4. Migration

Migration can be defined as the movement of individuals from one population to another and their contribution to the gene pool of the receptor population. The consequence of migration therefore is gene flow and the higher the gene flow between two populations, the less differentiated they will be.

There are several models of migration (Figure 5). The simplest is called the island model in which a meta-population splits into islands of equal size *N* which exchange genes at the same rate *m* per generation (Wright 1940). Under this model, the amount of differentiation between populations (measured with the F_{ST} statistic) depends only on the size of each subpopulation and the migration rate (which are all the same for all subpopulations):

$$F_{ST} = 1/(1+4Nm)$$

The stepping stone model (Kimura and Weiss 1964) specifically includes the space element, and therefore, migration can only happen between neighboring populations. This model also assumes equal rates of migration between subpopulations.

The isolation by distance model (Malecot 1969; Wright 1943) models migration as occurring within a continuous population by considering that mating choices are limited by distance. Then, on average, individuals will be related as a function of their geographical distance. Some recent studies have shown that worldwide patterns of genetic variation in humans can be explained under an isolation-by-distance model (Conrad et al. 2006; Relethford 2004). For example, Relethford et al. (2004), looked at genetic variation between human populations across the world using data on red blood cell polymorphisms, microsatellite DNA markers, and craniometric traits and showed how the isolation-by-distance model provided a good fit to the patterns observed.

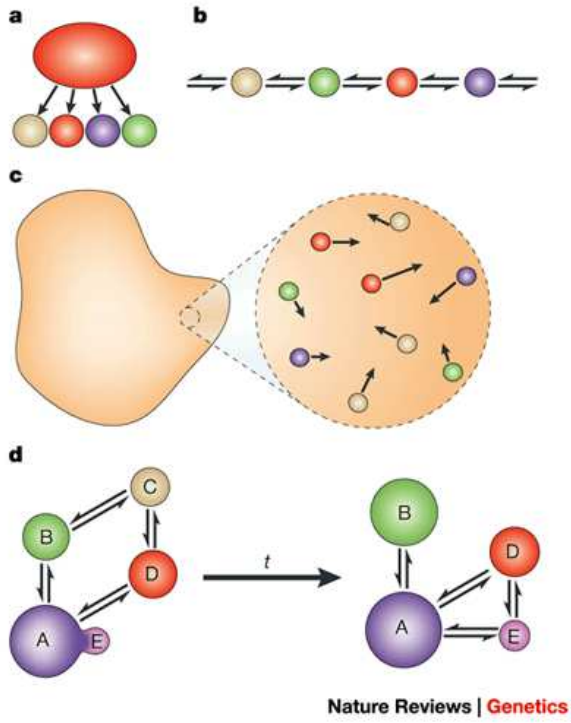


Figure 5. Different models of migration. (a) Island model of migration. (b) Stepping-stone model. (c) Isolation by distance model. (d) Metapopulation model in which populations come and go over time (t) with the founding and extinction of entire populations being an important component of population structure. Figure from Hey and Machado 2003.

These three models of migration, however, assume that migration rates have been constant for a long period of time and that the system has reached equilibrium. These assumptions may not be true in most of the cases. Other models incorporate parameters that can change as a function of time and they do not assume populations to have reached equilibrium (Slatkin 1977; Wade and McCauley 1988).

On the other hand, all these models consider that the migrants are a random sample of the source population although it has been shown that migrants tend to be sex-biased, age-structured and related to one another (Jobling et al. 2004). For example, it has been estimated to be 70% of modern societies are patrilocal (Jobling et al. 2004). This implies that in marriages between different villages, it will be the females that will migrate to the men's village to live with them.

1.1.5. Selection

Natural selection is defined as the differential reproduction of genotypes in succeeding generations. The ability to detect the footprint of natural selection in the genetic record has arisen a considerable excitement. First, this will allow studying the evolutionary processes that lead to adaptation and, second, because information regarding selection may provide important functional information that could potentially be related to basic biology and/or disease.

Several challenges are faced when trying to detect selection, one of the most important ones being that some demographic processes may lead to very similar patterns of diversity. Several types of selection can be defined depending on the fate of the alleles when acting on them.

Positive selection on a particular allele will lead to an increase of the probability of a particular new variant to be fixed. Therefore, there may be evidence of a rapid increase in frequency of a derived

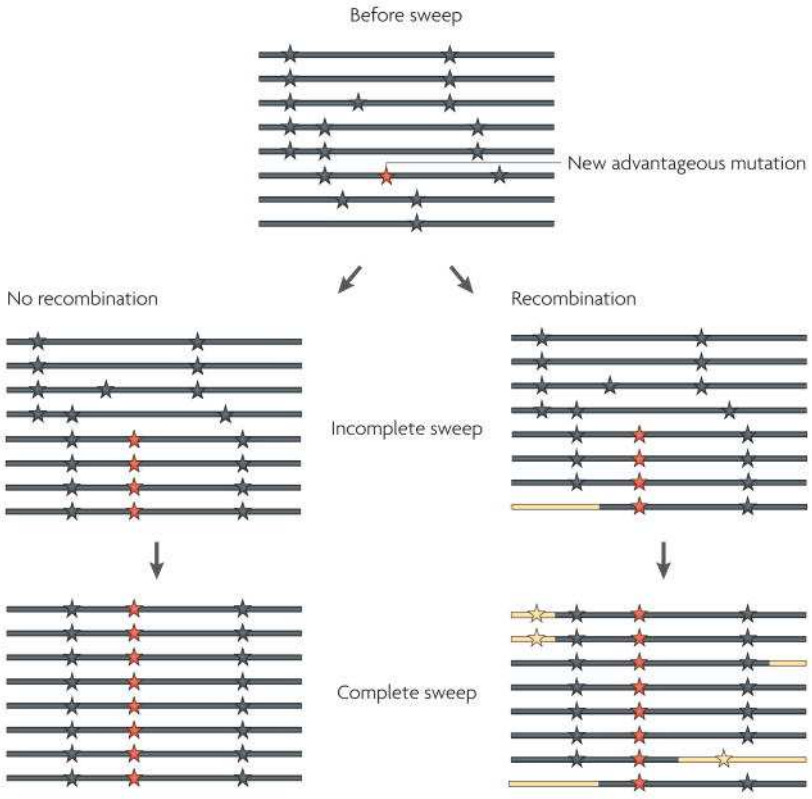


Figure 6. The process of a selective sweep. The lines indicate individual DNA sequences or haplotypes, and derived SNP alleles are depicted as stars. A new advantageous mutation (indicated by a red star) appears initially on one haplotype. In the absence of recombination, all neutral SNP alleles on the chromosome in which the advantageous mutation first occurs will also reach a frequency of 100% as the advantageous mutation become fixed in the population. Likewise, SNP alleles that do not occur on this chromosome will be lost, so that all variability has been eliminated in the region in which the selective sweep occurred. However, new haplotypes can emerge through recombination, allowing some of the neutral mutations that are linked to the advantageous mutation to segregate after a completed selective sweep. As the rate of recombination depends on the physical distance among sites, the effect of a selective sweep on variation in the genomic regions around it diminishes with distance from the site that is under selection. Chromosomal segments that are linked to advantageous mutations through recombination during the selective sweep are coloured yellow. Data that are sampled during the selective sweep at a time point when the new mutation has not yet reached a frequency of 100% represent an incomplete selective sweep. Figure taken from Nielsen et. al 2007.

variant together with the linked loci (selective sweep) and consequent accumulation of rare variants in the selected haplotype after a while (Figure 6). A population expansion may mimic the effect of selection by creating an increased number of rare variants but, whereas the selective sweep will take place within a specific locus, the footprint of demographic events should be spread throughout the genome. Negative or purifying selection acts by removing deleterious variants. The main consequence is a decrease in diversity since new deleterious alleles are systematically removed from the population.

Balancing selection is a particular case in which the selective advantage is conferred to the heterozygous individuals. The consequence is an increase of diversity since it promotes that both copies of the alleles are at intermediate frequencies. Population substructure will lead to similar diversity values since one of the alleles may be at higher frequency in one population and the other in the other one producing an average allele frequency of intermediate values when taking the two populations together. Again, evidence of substructure should be seen over all the genome whereas balancing selection should be located in specific regions of the genome.

Several tests can detect the footprint of selection acting in some populations compared to others. Tests could be divided in those that consider at population differentiation such as the Lewontin-Krakauer (Lewontin and Krakauer 1973) or F_{ST} - based, tests based on the allele frequency spectra such as Tajima's D (Tajima 1989) or Fu and Li (Fu 1997), and finally test based on linkage disequilibrium and haplotype structure, which take advantage of the pattern left by a selective sweep. Most of the different tests to detect selection have been reviewed by Nielsen (2005; 2007).

1.2. Making inferences from diversity

The study of population genetics has traditionally been divided between gene-tree based phylogeographic methods (Avise (1987) coined the word phylogeography) and a more traditional mathematical approach that relies on explicit models and summary statistics. The development of the coalescent theory, however, presented a coherent statistical framework for analysis of genetic polymorphism that allowed modeling processes such as recombination, demography and selection, something that could not easily be done before. Finally, with the increase of molecular data, the number of computational methods and more explicit model-based approaches, mostly based on likelihood, has increased significantly.

1.2.1. Phylogeography: mtDNA and Y chromosome

Phylogeography refers to the study of the geographical distribution of the clades within a phylogeny. Therefore, this approach necessarily implies the construction of a tree representing the phylogenetic relationships among the individuals or, in fact, of the DNA sequences.

Several methods can be used to construct phylogenetic trees based on genetic data (reviewed in Holder and Lewis (2003). The simplest method is the UPGMA, in which a tree is constructed based on a distance matrix by putting together the taxa with the lowest distance in an iterative process. The problem with UPGMA is that it considers that the evolution rate is the same in all branches. Conversely, the Neighbor-Joining method attempts to construct the tree with the shortest sum of branch lengths but it allows different branch lengths. On the other hand, the maximum parsimony method finds the tree with the smallest number of evolutionary changes whereas the maximum likelihood method chooses, given an evolutionary model, the tree which has the maximum likelihood of producing the data. Maximum parsimony is very accurate with phylogenetically close taxa but can not be suitable for divergent

taxa. Maximum likelihood is computationally much more intensive but can be more reliable provided that the evolutionary model used is suitable. Finally, Bayesian approaches have recently been introduced to phylogenetics. With these methods, the optimal hypothesis is the one that maximizes the posterior probability. They allow complex models of evolution to be implemented and they provide measures of support faster than maximum likelihood bootstrapping.

In some cases, however, the phylogenetic relationships are best represented with a network rather than a tree because networks can contain information of several trees. For example, it is very common to build networks when trying to represent phylogenetic relationships using mtDNA in human populations since by building up networks, all possible recurrent mutations can be represented as reticulations and no assertion is made on which is the true tree. Choosing which method to use may depend on the kind of data that we have and the computational power we may have available.

In any tree, time is intrinsically represented in the sense that events occurring in the tips of the tree may represent later events than those closer to the root. The molecular clock hypothesis states that for any given DNA sequence, the rate of evolution is approximately constant over all lineages. Taking this into consideration, we potentially could date all events in the tree once the molecular clock has been calibrated using some external information such as the fossil record. However, to date, calibration on mutation rates is not very accurate.

All those methods, however, intrinsically imply that recombination is absent, since recombination will put together two different lineages. Therefore, most of the phylogeographic approaches to study the recent human history and migrations have been restricted to mtDNA or the non recombining portion of the Y-chromosome (Torroni et al. 2006; Underhill and Kivisild 2007).

However, some issues should be taken into account when making inferences from these two compartments (see Balloux (2009)). First,

each of them represents a single locus and therefore they represent a single realization of the many possible outcomes within a given demographic history (Ballard and Whitlock 2004). Second, they both contain genes and therefore, they could be subject to selection, something that would affect the whole tree structure since all genes are linked and any selective effect in one gene would affect all the others. Finally, in the case of mtDNA, the mutation rate is very high at some loci (Soares et al. 2009) and homoplasy can highly affect the inferential process.

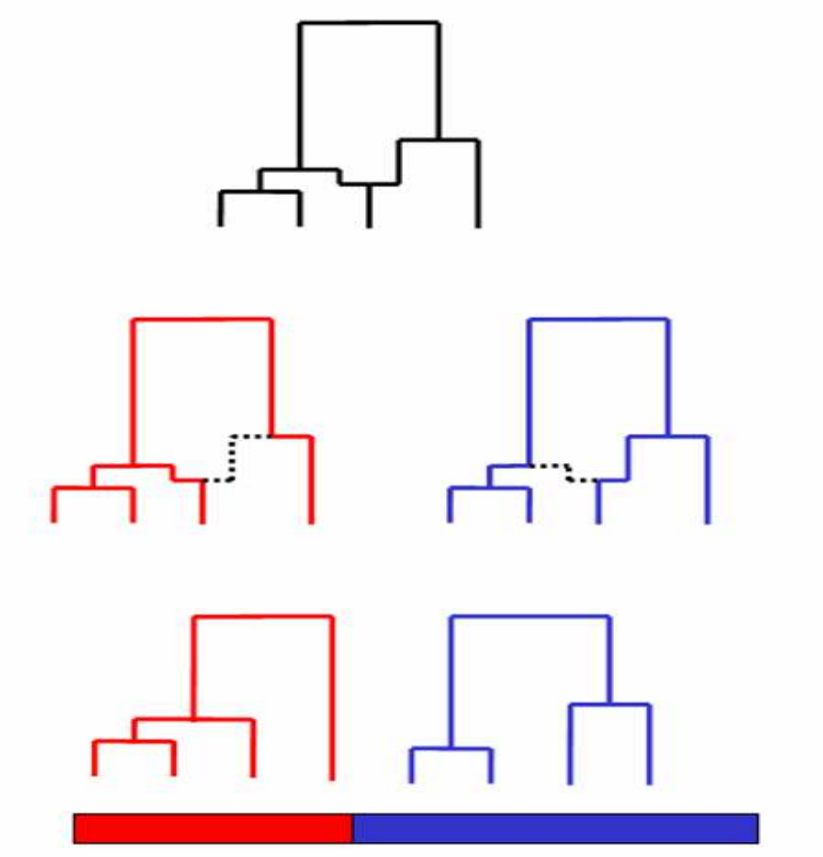


Figure 7. If recombination is present each locus may have a different phylogenetic history. The way of representing information of more than one tree is by means of a network in which the recombinant sequences will have two parental nodes representing their two different phylogenetic histories.

The rest of the genome (except mtDNA and Y chromosome), however, undergoes recombination. Each locus could potentially have a different history and the way to represent the phylogenetic relationships is then a complex network named Ancestral Recombinational Graph (Figure 7 and Figure 8). Moreover, it has also been demonstrated that attempts to construct trees ignoring recombination would lead to different types of biases (Schierup and Hein 2000).

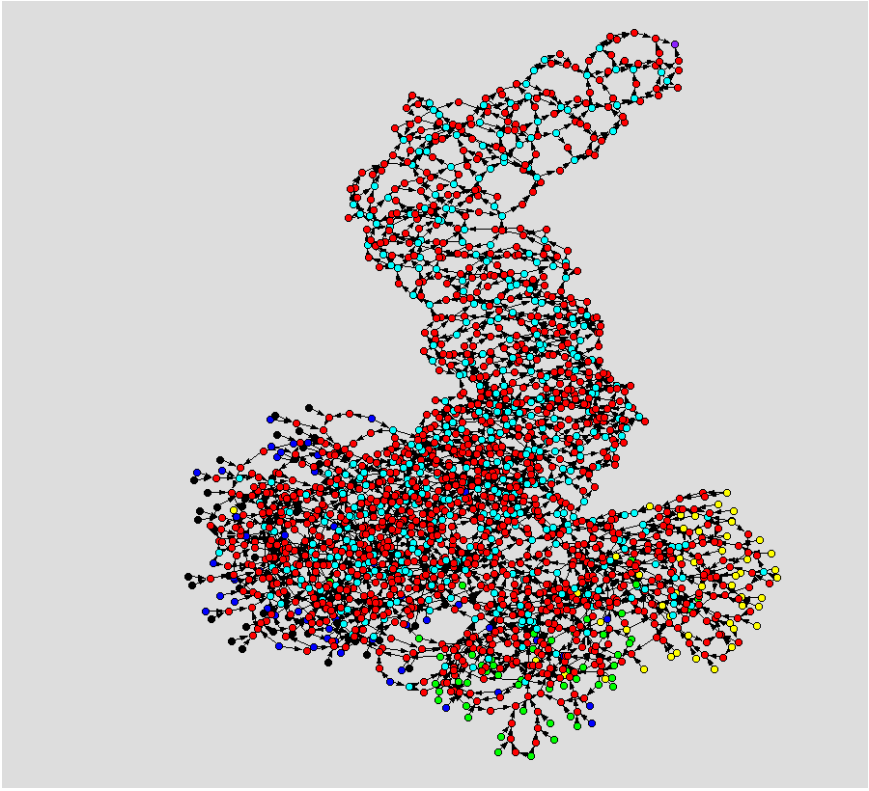


Figure 8. Ancestral recombination graph (ARG) generated with coalescent simulations with a human like demography and varying recombination rates along the sequences. This ARG represents the genealogical relationship of 210 human sequences belonging to four different populations: Africans (black), African Americans (dark blue), Europeans (green), Asians (yellow) for a region spanning 200 kb. The red nodes are the recombinant nodes and the light blue nodes the coalescent ones. Software used Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

Recently, some models have been developed that try to reconstruct the history of a set of sequences allowing for recombination events to take place by means of inferring a network structure consistent with the data (Gusfield et al. 2007; Parida et al. 2008; Song and Hein 2005; Wiuf 2002). Some of these models are reviewed in Huson and Bryant (2006).

However, the problem with inferring the Ancestral Recombinational Graph is basically that there are huge numbers of possible ARGs that could have created the data, especially when rates of mutation and recombination are comparable as is the case for humans (McVean and Cardin 2005). As a result, and despite some attempts, trying to infer the complete sequence of recombination events in a genealogy has been computationally intractable for realistic datasets (Hellenthal and Stephens 2006).

1.2.2. Summary statistics

Summary statistics capture in one figure the variation present in the data in order to compare the observed value with that expected under a particular population genetics model such as the Wright – Fisher Model.

The simplest summary statistic to measure the amount of variation present in a sample is the number of segregating sites (S) but is highly dependent on the sample sizes. Nucleotide diversity or pi (π) is a measure of the degree of polymorphism within a population. Specifically, it describes the probability that two copies of the same nucleotide drawn at random from a set of sequences will be different from one another.

$$\pi = \sum_{i=1}^{n} \sum_{j=i+1}^{n} x_i x_j p_{ij}$$

Where x_i and x_j are the frequencies of haplotypes *i* and *j* respectively, and x_{ij} is the proportion of differences between them. Interestingly, under the Wright-Fisher model in which population size is constant, populations are panmictic and do not overlap, a population reaches an equilibrium in which the number of novel variants created by mutation is balanced by the number of variants lost by drift. As stated before, this equilibrium value of diversity is known as the population mutation parameter or theta (θ) and these two measures of diversity *S* and π are good statistics to estimate it.

Further, Tajima's D statistic is based on the expectation that under a non-equilibrium situation, the number of segregating sites and the nucleotide diversity will differ significantly. Under neutrality, Tajima's D is expected to be zero because S and π are equivalent. However, positive values of this statistic indicate that the number of alleles at intermediate frequencies is higher than expected, something that is generally caused by population subdivision or balancing selection. Conversely, negative values of the statistic indicate an excess of rare variants which can be caused by positive selection or population growth.

Heterozigosity is another measure of diversity and it is calculated locus by locus and then averaged over the whole sequence. For a single locus with *n* alleles, heterozigosity is: $H = 1 - \sum_{i=1}^{n} p_i^2$. Then for the whole sequence with *m* loci: $H = \frac{1}{m}\sum_{i=1}^{m} H_i$

Another common way of representing diversity is the mismatch distribution, which depicts the number of pairwise comparisons between haplotypes that have a certain number of differences. This distribution can provide some information on past demographic events of the samples. Note that the mean of the pairwise distribution divided by the sequence length is the same as the nucleotide diversity. Further, the variance of the pairwise differences between haplotypes can be interpreted as a measure of LD and it can be used to estimate the population recombination

parameter (Hudson 1987) since recombination decreases this variance.

Statistics that summarize the amount of variation, however, do not contain all information present on the data and different evolutionary processes could give rise to similar values of the chosen statistic, as we have seen for the population growth and positive selection effects.

1.2.3. Coalescent-based inference

The mathematical theory of the coalescent was developed in the early 1980s by John Kingman (1982) and Richard Hudson (1983) independently and it is nowadays one of the basic tools of population genetics studies. The coalescent theory is based on the Wright – Fisher neutral model and it simulates backwards in time the genealogy of only those chromosomes that appear in the sample, up to the ancestor of all lineages (MRCA). Because it allows simulating only the genealogy of the sampled sequences and not all the populations and because it models the genealogical process independently of the mutational process, it is extremely efficient computationally (Figure 9).

The coalescent is a natural extension of the classical population genetics models and it is very different from phylogenetic methods. Phylogenetic methods estimate trees whereas the coalescent is used to estimate parameters of the random genealogical process that has given rise to each tree. The tree itself has no inherent interest. Thus, the coalescent provides a coherent statistical framework to study the effect of the process that shape the diversity found in the genomes such as recombination, migration, selection and so on.

The coalescent has several applications. For example, it can provide useful guidance about how many individuals, populations and loci are needed to be sampled to answer questions of interest. Secondly, it is a simulation tool for hypothesis testing. In this direction, Shaffner and colleagues (2005) presented a model based on the

coalescent that mimicked human data for three populations in allele frequency, linkage disequilibrium and population differentiation. This model can then be used to compare empirical measures of sequence variation, linkage disequilibrium and selection expectations under a null distribution that already takes into account a simplified version of the complex demographic history of human populations.

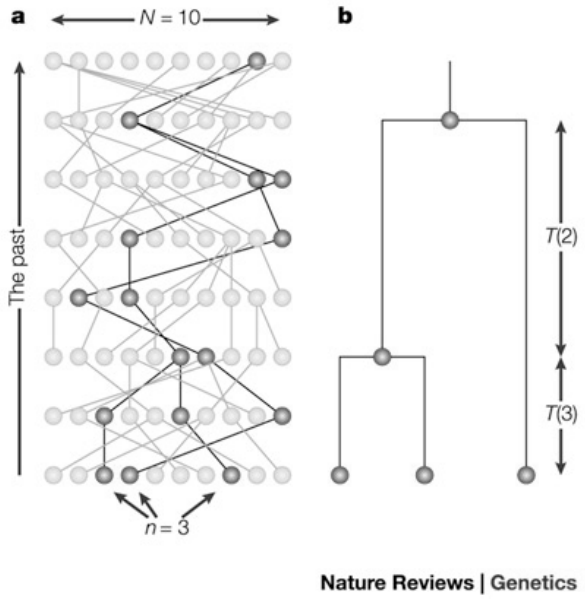


Figure 9 The basic coalescent principle: only the gene genealogy of the sampled chromosomes will be inferred (a)The complete genealogy for a population of ten haploid individuals is shown (diploid populations of N individuals are typically studied using a haploid model with 2N individuals). The black lines trace the ancestries of three sampled lineages back to a single common ancestor. (b) The subgenealogy for the three sampled lineages. In the basic version of the coalescent, it is only necessary to keep track of the times between coalescence events (T(3) and T(2)) and the topology — that is, which lineages coalesce with which. N is the number of allelic copies in the population and n the sample size. Figure taken from Rosenberg and Nordborg 2002.

Finally, the coalescent approach can be used as the basis for full-likelihood inference. Basically, after collecting the data, all possible genealogies and their probabilities under models of interest should be considered. For each genealogy, the likelihood of the data is

calculated and the parameters estimated by finding values that maximize the likelihood of the data. Finally, the models should be tested by comparing the likelihoods under different hypothesis. Unfortunately, this process is computationally very intense and advanced computational techniques such as importance sampling or Markov-chain Montecarlo should be applied (Stephens and Donnelly 2000).

Several recent studies have used coalescent simulations to find the evolutionary model that is most likely to produce the observed data or the observed summary of the data to make inferences on recent human demographic history. Some examples can be found (DeGiorgio et al. 2009; Liu et al. 2006; 2005). For example, DeGiorgio and colleagues (2009) use a coalescent-based serial founder model to explain patterns of human genetic variation and the process of migration out of Africa.

Other studies do not use coalescent approaches, for example Gutenkunst and colleagues (2009) use a diffusion approach. In fact, with the explosive growth of both the amount of molecular data being generated and the computational power available to analyze it, an increasing variety of computational methods are available to analyze and interpret such data. Most of these new methods are model-based approaches based on likelihood, which are becoming more and more used in the field (see Beaumont (2004) and Marjoram and Tavaré (2006) for reviews on this subject).

1.2.4. Bayesian clustering analysis and Principal Component Analysis

One goal of population genetics analysis is to identify the genetic structure that exists within a set of genotyped individuals, which may give some insights into population relationships and help to minimize false-positive results in association mapping studies.

One of the most popular methods is the Bayesian clustering algorithm implemented in the software STRUCTURE (Pritchard et

al. 2000). The method assumes a model in which K populations exist, each of which is characterized by a set of allele frequencies at each locus. Individuals in the sample are assigned (probabilistically) to populations, or jointly to two or more populations if their genotypes indicate that they are admixed. It can be applied to microsatellites, SNPs and RFLPs and the model assumes that markers are either not in linkage disequilibrium or weakly linked (Falush et al. 2003). STRUCTURE is quite computationally intensive and other methods have been developed that allow for a much higher number of genetic markers to be taken into account such as *frappe* used in Li et al. (2008) with 650,000 markers. Many of the most relevant studies of patterns of human genetic variation using a genome-wide set of genetic markers have applied Bayesian clustering approaches (Conrad et al. 2006; Jakobsson et al. 2008a; Li et al. 2008; Rosenberg et al. 2002; Tishkoff et al. 2009). For an example see Figure 10A.

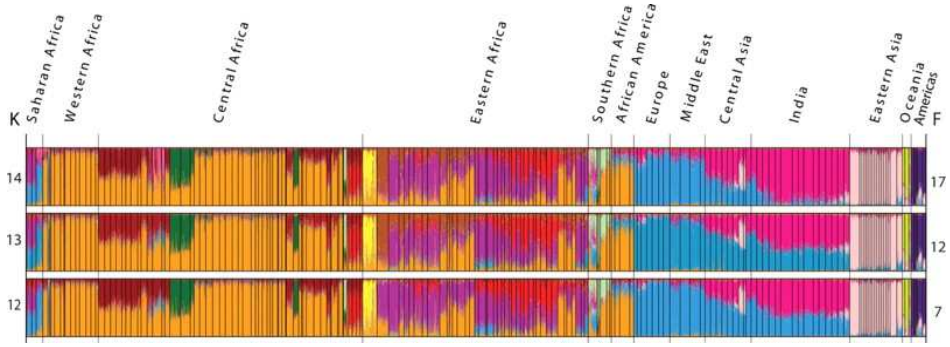
Another interesting method to study the underlying structure of population is Principal Component Analysis (PCA). This method involves a mathematical procedure that transforms a number of possibly correlated variables into a number of uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has as high a variance as possible (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible.

Although it was introduced to population genetics by Cavalli-Sforza in the late 1980s, renewed interest in this approach was taken with the implementation the Eigensoft package by Patterson et al. (2006). One of the main reasons was that they allowed statistical validation of the inferred structure and that it can deal with a larger amount of markers than STUCTURE. Several relevant studies of human genetic variation have used PCA to study the underlying structure of human populations (Li et al. 2008; Tishkoff et al. 2009). See Figure 10B for an example. One of the most striking results was found by Novembre et al. (2008) and Lao et al. (2008) in which their Principal Component Analysis of Europeans based on genome-wide SNP data, reconstructed the geographic map of

Europe. The same result could be seen in a recent study with sub-Saharan African populations (Sikora et al. 2010).

However, interpretation of PCA results is still not clear (Novembre and Stephens 2008) and is generally a first analysis aimed at defining the genetic relationships among groups or even better, the relative overall similarity among them.

A



B

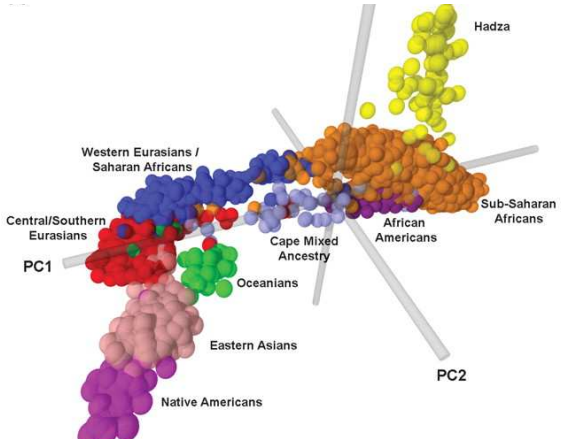


Figure 10. STRUCTURE and PCA analysis of the global data set with 1327 microsatellites genotyped in 3945 worldwide individuals. Individuals are clustered by major geographic region (Tishkoff et. al 2009). (A) STRUCTURE analysis. Each vertical line represents an individual. Colors represent the inferred

ancestry from *K* ancestral populations. STRUCTURE results for *K* = 12 to 14 (left) are shown with the number of similar runs (F) for the primary mode of 25 STRUCTURE runs at each *K* value (right**). (**B**)** Principal components analysis created on the basis of individual genotypes.

1.3. The recent human evolutionary history

1.3.1. Origin of Anatomically Modern Humans

The oldest fossil remains that show clearly anatomically modern human traits were identified in East Africa and dated around 195-150 Kya (McDougall et al. 2005; Stringer 2003; White et al. 2003). Therefore, the basic morphology of anatomically modern humans was established in Africa about 200 kya.

Genetic studies have confirmed the origins of anatomically modern humans in Africa based on Y-chromosome, mtDNA and genomewide (Underhill and Kivisild 2007). First, DNA markers typically have shown higher diversity (heterozygosity and nucleotide diversity) in sub-Saharan Africa populations. This has been seen for mtDNA (Cann et al. 1987) nuclear microsatellites (Relethford and Jorde 1999), Alu insertion markers (Watkins et al. 2001), and SNPs (Tishkoff et al. (2009) among others).

Moreover, there is also a clear geographic pattern in regional diversity. Specifically, genetic diversity outside Africa tends to be a subset of the diversity within Africa (Behar et al. 2008; Tishkoff et al. 1996; Watkins et al. 2001). For example, mtDNA sequences outside Africa fall into two clades, M and N, which both are rare in sub-Saharan Africa where the mtDNA sequences belong to the ancestral clade L. Moreover, distinct M variants are present in a frequency of 20% in Ethiopia, which lead to propose East Africa as the source of a migration out of Africa (Quintana-Murci et al. 1999).

Finally, global analysis of microsatellite data has shown that diversity decreases with distance from East Africa (Prugnolle et al. 2005a; Ramachandran et al. 2005a; Tishkoff et al. 2009) (Figure 11). This observation is best explained by a model of serial founder effect starting at a single origin in which the migration of populations across much of the globe occurred in many small steps with each migration event involving a sampling of variation from

the previous population (Conrad et al. 2006; DeGiorgio et al. 2009; Ramachandran et al. 2005a).

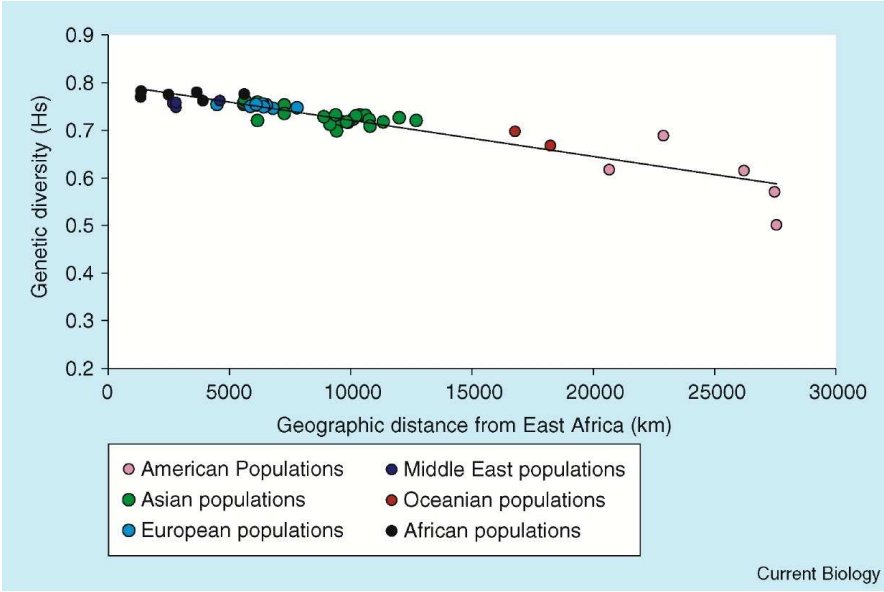


Figure 11. Relationship between mean genetic diversity of 51 human populations computed over 377 autosomal microsatellite markers and their geographic distances from East Africa. The percentage of variance explained by geographic distance is R^2=85% ($p<10^{-4}$). The different colours correspond to the different ethnic groups. Figure taken from Prugnolle et. al 2005.

1.3.2. Routes of the Out of Africa migration

The routes followed and the number of the first anatomically modern humans that left Africa however is still a subject of debate.

Traditionally, two main migratory routes Out of Africa (OoA) have been hypothesized for anatomically modern humans, initially based with archaeological record (Lahr and Foley 1994; Lahr and Foley 1998) and later supported by phylogenetic trees constructed from data on a limited number of protein markers (Cavalli-Sforza and Feldman 2003). This model involves a northern migration via North Africa and the Nile Valley into the Levant with subsequent

dispersal into both Europe and Asia. Moreover, there would have been an earlier southern coastal route that took place earlier in time in which anatomically modern humans left Africa by crossing the Bab el Mandeb strait in the mouth of the Red Sea and then rapidly migrated along the South Asia coastline to Australia and Melanesia (Figure 12).

Recent studies, have shown evidence from both mtDNA and Y chromosome that an early rapid migration OoA took place following the southern coastal route, through India and into Southeast Asia and Australasia taking place around 65,000 years ago (Forster and Matsumura 2005; Macaulay et al. 2005; Mellars 2006; Thangaraj et al. 2005).

Figure 12. Map of possible dispersal routes of anatomically and genetically modern human populations from Africa to Asia and Australia according to Forster and Matsumura (Forster and Matsumura et al. 2005).The models assume an origin in eastern Africa, and dispersal either via the Nile Valley and Sinai Peninsula (the "northern" route) or via the mouth of the Red Sea to Arabia and Australia (the "southern" route). The oldest human traces outside of Africa and the Levant are at Lake Mungo in Australia (>46,000 years old) and in the Niah Cave of Borneo (>45,000 years ago). New mtDNA data, from Malaysians and aboriginal Andaman islanders, suggest that human settlements appeared along the Indian Ocean coastline 60,000 years ago (Macaulay et al. 2005; Thangaraj et al. 2005),. Figure taken from Forster and Matsumura 2005.

It is not still clear, however, whether there were two migratory routes or a single southern route. Paul Mellars, in a recent review (2006) provides some plausible explanations to reconcile the archaeological record with a fast and single migration via the southern route by anatomically modern humans. Moreover, a study based on studying six hundred thousand loci in East and South East samples (The Hugo Pan-Asian SNP Consortium 2009) seems to point out that South East Asian populations were the major geographic source of East Asian populations and that there was a single primary wave of entry to the Asian continent. This means that later expansions of East Asian populations were based on offshoots of this initial main migration.

1.3.3. Tempo and mode of the Out of Africa

Several studies suggest that a strong bottleneck occurred in the populations that left Africa around 40,000 and 80,000 years ago (Marth et al. 2003; Reich et al. 2001; Voight et al. 2005; Wall and Przeworski 2000).

All the recent studies that have used genetic variation to infer human effective population size attribute a higher long-term effective population size to African populations. This again is explained by African populations having an older origin and a higher number of effective individuals, compared to non-African populations which underwent a strong bottleneck when leaving Africa. Different studies, however, have found slightly different estimates. Zhao et al. (2006) used nucleotide diversity estimates to estimate effective population size in three continental human populations and found it to be around 15,000 for Africans and around 7,500 for non-Africans. Conversely, Tenesa et al. (2007) used LD patterns seen in the four HapMap II populations and they found lower estimates being 7,500 for the Africans and 3,100 for non-Africans. Finally, Cox et al. (2008) used genetic diversity at twenty X chromosome loci to determine the most likely effective population size under an isolation-with-migration model. They found that African population sizes tend to have larger (2,300–

9,000) effective population sizes than non-African populations (300–3,300).

Moreover, the size of the ancestral population(s) that left Africa is estimated to be around 1000 effective founding males and females based on autosomal microsatellite loci (Liu et al. 2006) or 1500 effective founding males and females based on mtDNA, Y chromosome and X chromosome re-sequencing data (Garrigan et al. 2007).

Two studies tried to assess differences in the effective population size of founding females and males by looking at diversity in the X chromosome compared to the autosomes. Hammer et al. (2008) studied population substructure biases and found higher than expected levels of diversity in the X chromosome, suggesting lower male versus female effective population size. Conversely, Keinan et al. (2009) who used nucleotide diversity estimates, detected lower diversity in the X chromosome compared to the autosomes suggesting the opposite results (Bustamante and Ramachandran 2009). Finally, Emery et al. (2010) showed that the two estimators detected biases that have occurred in different time-scales and that these results can be explained by a recent male higher effective population size compared to that of females and an earlier and persistent female higher effective population size. A different study showed that lower diversity patterns found in the X chrosomosme can be explained by a model of primarily male migration during the out of Africa (Keinan and Reich 2010).

1.4. Recombination in the study of human population history

The use of recombination in the study of human population history has been very limited, although recombination is, together with mutation, the main force shaping our genome. Precisely, most of what is known about human population history has been inferred by looking at non-recombining portions of the genome such as the mtDNA and Y chromosome because the lack of recombination makes the inference of phylogenetic history easier.

One indirect way of using the information provided by recombination in the study of human population genetics would be to take the haplotypes as genetic markers, since haplotypes, unlike SNPs, are the consequence of the action of both mutation and recombinational processes. Although there has been a decrease in the cost of genotyping arrays and a huge number of studies have looked at the patterns of human genetic variation based on thousands of markers, those markers have been mostly SNPs or microsatellites but not haplotypes.

Recently, however, a relevant study compared haplotypes and SNPs as genetic markers using 500,000 markers in 52 worldwide populations (Jakobsson et al. 2008b). Results showed that haplotypes contained more information regarding population structure than SNPs. Specifically, the analysis of haplotypes allowed to detect additional genetic structure in Africa. Moreover, a study still in preparation further confirms this observation by studying 2Mb at high SNP density in 33 populations of the Old world. Further, in this study a method to extract most of the information by defining optimal haplotype lengths is provided (Javed et al. in preparation, see section 3.4.).

Other studies have made use of haplotypes to make some inferences on human demographic history. Haplotype sharing patterns between populations have been used to reconstruct the colonization of the major landmasses by anatomically modern humans (Hellenthal et

al. 2008) and to estimate parameters of a population split model (Davison et al. 2009). Moreover, Lohmueller et al. (2009) use the join distribution of haplotype number and major haplotype frequency in empirical and simulated data to estimate population size changes.

LD patterns have also been used to study differences between populations. For example, Tenesa et al. (2007) used patterns of LD between three human populations to infer human effective population size and Plangol et al. (2006) uses a measure of LD to study possible archaic structure in human populations.

Finally, the footprint of recombination has sometime been used to study human adaptation trying to detect regions under positive selection. Some tests of selection are based on the increase of LD when a selective sweep takes place. For example, two genes, glucose -6-phosphate dehydrognease and CD40L, which are associated to malaria resistance, showed this pattern of extended haplotype (Sabeti et al. 2002).

All these studies make an indirect use, however, of the footprint of recombination, either by looking at haplotypes or at LD patterns. Over fifty years ago, Sir Ronald A. Fisher (1954) pointed out that recombination, when shuffling together sequences from different lineages, leaves a signal or junction that will be passed to the subsequent generations. This observation opened the door to the use of recombination as a genetic marker although the necessary tools to carry out such an attempt would not be available until decades later. With the advent of high density SNP data, the higher number of individuals being genotyped and an increase in computational power, now it has been possible to develop a method aimed at detecting and using recombination events to study human population history (Melé et al. 2010 and Melé et al. submitted).

1.5. How to detect recombination

1.5.1. Computational methods to detect the presence of recombination

Several methods are aimed at detecting the presence or absence of recombination. The most widely used is the four-gamete test which is based in the observation that when considering two loci, if the four possible combinations of alleles are observed, this is evidence of recombination or recurrent mutation (Weir 1979). (Figure 13)

Carrying out the four gamete test on all pairs of sites, it is possible to identify intervals in which recombination must have occurred. Then, the minimum number of recombination events that have occurred in the history of a sample of chromosomes (R_m) can be inferred assuming that only one recombination event occurred in the overlapping intervals (Hudson and Kaplan 1985).This assumption is very conservative and it may well be the case that there is more than one recombination event occurring in those intervals.

By comparing the number of haplotypes with the number of polymorphic sites it is possible to estimate the number of recombination events. If *M* haplotypes are observed in a region with *N* segregating sites, then at least *M*-*N* recombination events must have occurred (Myers and Griffiths 2003). *M*-*N* is therefore a local lower bound and combining these local bounds allows the construction of the global minimum number of recombination events that have occurred in a region.

However, the minimum number of recombination events and the real number can be very different. In fact, only a small proportion of recombination events in simulated genealogies can be detected in population-genetic data since different conditions need to be met in order to detect recombination: divergence between ancestral sequences, age of the event, sample size, etc (Figure 13c).

In the last decades, many different methods have been developed to detect presence of recombination, which can be found in:

http://www.bioinf.manchester.ac.uk/recombination/

Some methods not only detect presence or absence of recombination but either detect the breakpoint location, the

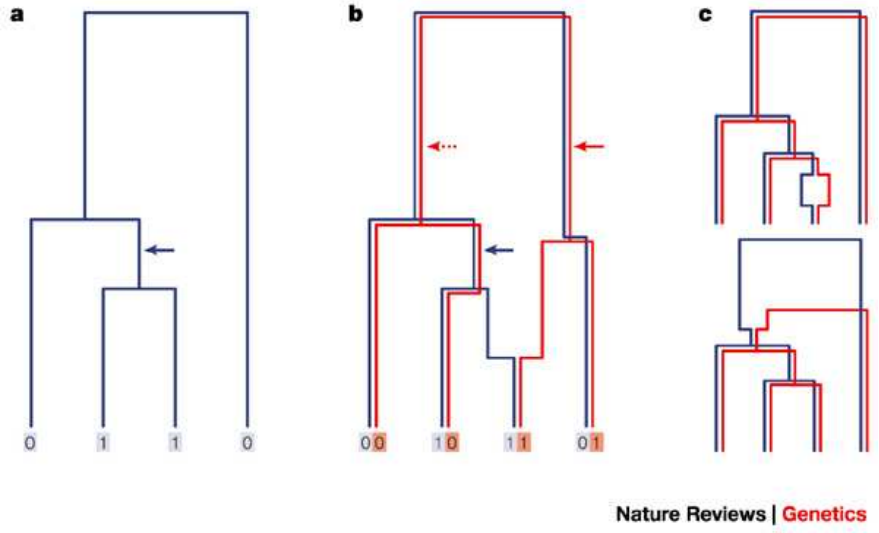


Fig13. The effect of recombination on neigboring loci. In the legend, 0 and 1 denote ancestral and derived alleles, respectively. (a) The genealogy of a single hypothetical locus is represented by a single bifurcating tree. A mutation event of 0 to 1 gives rise to a derived allele. (b) The genealogy of a second locus (red) that is physically close to the locus depicted in part a is shown; its genealogy is partially correlated with the original (blue) genealogy. If mutations occur along the two lineages (indicated by the solid arrows) then the recombination event will be detected in the resulting two-locus gametes, because, as shown here, all four possible gametes (0,0; 0,1; 1,0; 1,1) are observed in the sample. It should be noted that there are two lineages along the red genealogy for which a mutation event can cause the recombination event to be detected (red solid and dashed arrows). (c) In these two genealogies the recombination event cannot be detected from the resulting data, no matter on which lineages mutations occur. This is because there is no combination of lineages among the two marginal genealogies along which mutations will give rise to all four possible two-locus gametes. For this reason smaller samples are less informative about recombination than larger samples. Figure from Mc vean et. al 2003.

recombinant sequences and the ancestral sequences, or try to infer the underlying ARG present in the data.

Only two simulation studies have tried to generally evaluate some of these methods (Posada and Crandall 2001; Wiuf et al. 2001). Both studies consistently showed that the power of the evaluated methods was generally quite low. For example, Wiuf et al.(2001) stated as one of the main conclusions of the study that "all of the investigated methods detected far less recombination than is theoretically possible".

Different methods cover different needs, some approaches may be more suitable to treat some problems than others and so far there is no consensus on which method should be used when trying to detect recombination.

1.5.2. Methods to infer recombination rates

A huge interest has arisen in developing methods to estimate recombination rates in an effort to try to understand the nature and causes of the recombinational process in humans and other organisms. Traditionally, the estimates of recombination rate of the human genome have come by means of pedigree studies but, initially, they only had resolution at the megabase scale. Sperm typing techniques allowed studying specific hotspots at the individual level but they are too costly to be used at the genome wide scale. Finally, several computational methods have been developed with the aim of inferring fine–scale recombination maps of the human genome.

1) Recombination Pedigrees

Using large pedigree families has enabled to create genome-wide maps of recombination in humans. Some of the advantages of using pedigree-based maps are that differences between males and females can be assessed, that it is possible to ascertain

interindividual differences, and that it is possible to evaluate the heritability of certain recombinational patterns.

The first map at the megabase scale was published by deCODE Genetics (Kong et al. 2002) who used 5,136 microsatellite markers genotyped in 146 Icelandic families. Recently, a much finer pedigree–based recombination map has been extended by deCODE (Kong et al. 2010) which has a resolution that goes down to 10 kb.

2) Recombination detection using sperm typing

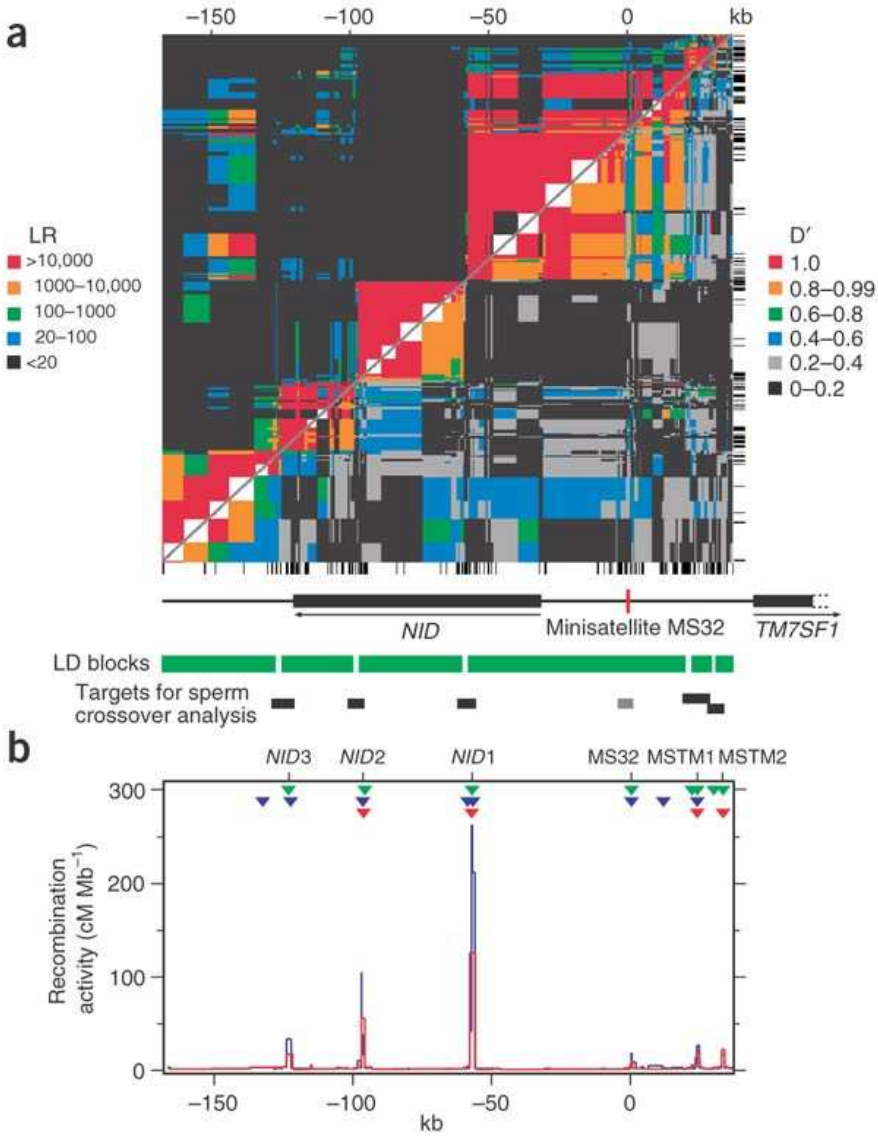
Sperm typing consists on amplifying and genotyping the sperm of one male in order to assess where crossover have taken place. It can be done either from single sperm or pooled sperm. Singled sperm is subject to a round of whole genome amplification to permit multiple loci to be typed from the same molecule and it is mostly used to construct genetic maps (typing distant genetic markers). Pooled sperm allows examining large number of individual sperm within a specific region (generally a hotspot) and counting the number of crossing over events versus the non-crossing over. Generally, samples are diluted so that aliquots mostly contain one sperm molecule and a quantitative value is estimated based on the number of positive samples.

3) LD based methods

It is expected from basic population genetics theory that the expected amount of LD between two markers depends on the recombination rate between them. It follows that by looking at LD patterns in natural populations, the underlying recombination rate can be inferred (Figure 14). Moreover, with the development of the coalescent theory (Hudson 1983; Kingman 1982), modeling the underlying process in a sample of sequences incorporating recombination was feasible (Griffiths and Marjoram 1996; Hudson and Kaplan 1988).

Several methods have been devised to estimate recombination rates from population data (see Table 1 in Stumpf and Mc.Vean (2003) for a list of some of them). Basically, they could be divided in two types: full and approximate likelihood approaches. Approximate likelihood methods try to avoid the computational burden of the full likelihood methods by either ignoring low frequency markers or else by considering a small number of markers at a time. Then, separate likelihoods are calculated for these subsets of the data and are combined to obtain the approximate likelihood estimator. If these subsets consist of only two pair of sites, a "composite likelihood" is then obtained by multiplying all pairwise likelihoods. This is the approach implemented in the LDhat software (McVean et al. 2004) that was used to estimate the recombination rates genome-wide using the HapMap phase II dataset (Myers et al. 2005). Conversely, full likelihood methods estimate the probability of observing a given dataset under an assumed population-genetics model using all the information present in the dataset. However, full likelihood approaches are computationally very intensive and still cannot be used genome-wide. One of the most recent examples of this approach is the full likelihood MCMC method developed by Wang and Ranalla (2008; 2009).

Figure 14. Patterns of LD and the corresponding historical recombination rates in a 206-kb interval around minisatellite MS32. (a) LD profile across MS32 and the neighboring gene *NID* established from 200 SNPs genotyped in a panel of 80 UK semen donors of north European origin. Maximum likelihood haplotype frequencies for each pair of SNPs were used to estimate |D'| levels of LD (lower right), as well as the associated LR (likelihood ratio) versus free association (upper left), and are color-coded as indicated. The locations of the remaining 175 SNPs are shown below and to the right of the plot, with positions centered on the middle of MS32 at coordinate 0. LD blocks were identified visually as regions where most marker pairs are in strong (|D'| > 0.8) and highly significant (LR > 10^4) association. Regions of LD breakdown targeted for sperm crossover analysis are shown. (b) Historical recombination rates and positions of putative recombination hot spots (marked above plot) estimated from coalescent analyses of genotype data. Population recombination rates ρ, defined as $\rho = 4N_e r$ where N_e is the effective population size and r is per-generation recombination rate, were estimated across the region using LDhat (red) and PHASE (blue). These were converted to r assuming that N_e = 10,000 and used to estimate the local sex-averaged recombination activity in cM per Mb. Colored triangles show putative recombination hot spots significant at $P < 0.01$ for three different hot-spot detection methods: LDhot (red), Hotspotter (blue) and Fearnhead's method (green). All coalescent analyses were undertaken after sperm typing and without knowledge of the sperm typing results. Figure taken from Jeffreys et al 2005.


a
−150
−100
−50
0
kb
LR
>10,000
1000–10,000
100–1000
20–100
<20
D′
1.0
0.8–0.99
0.6–0.8
0.4–0.6
0.2–0.4
0–0.2
NID
Minisatellite MS32
TM7SF1
LD blocks
Targets for sperm crossover analysis
b
NID3
NID2
NID1
MS32
MSTM1
MSTM2
300
200
100
0
Recombination activity (cM Mb−1)
−150
−100
−50
0
kb

1.6. Recent findings on recombination

Recombination is essential for the correct separation of chromosomes during meiosis and it has been shown that too little recombination can result in aneuplody, which is mainly lethal in humans (with very few exceptions), or chromosomal rearrangements, which have been associated to disease (Coop and Przeworski 2007). Moreover, it is the main mechanism that creates new allele combinations, something vital to generate the necessary diversity that will allow adaptation of individuals to their environment. These constraints suggest that the frequency and location of recombination events should be a highly regulated process and the genes that regulate such processes should be under strong purifying selection. There is evidence however, that despite the high sequence similarity between humans and chimpanzees (which differ only in 1% of the sequence), hotspot location does not overlap (Ptak et al. 2005; Winckler et al. 2005). Moreover, recombination location varies greatly between different individuals (Coop et al. 2008) and sexes (Kong et al. 2010), suggesting hotspot location has a faster mechanism of evolution than sequences.

Another question that surrounded recombination evolution was defined as the hotspot paradox (Boulton et al. 1997; Jeffreys and Neumann 2002; Jeffreys and Neumann 2005). During recombination, it is the initiating chromatid of the crossing over the one that acquires the DNA sequence of its opposite partner. Then, if the initiating chromatid contains an allele that promotes the initialization of recombination in that location, this allele is doomed to extinction. From this observation it follows that there should be a mechanism that gives rise to new hotspots since recombination is essential for sexual reproduction. Although it had been shown that some alleles were more prone to initialize recombination process than others (Jeffreys and Neumann 2002) it was not until the study by Myers and colleagues (Myers et al. 2005) that a small motif enriched in some of the human hotspots (10%) was found, suggesting signals promoting recombination existed at particular locations. Later, a 13 bp motif was identified to be present in 41% of the human hotspots (Myers et al. 2008).

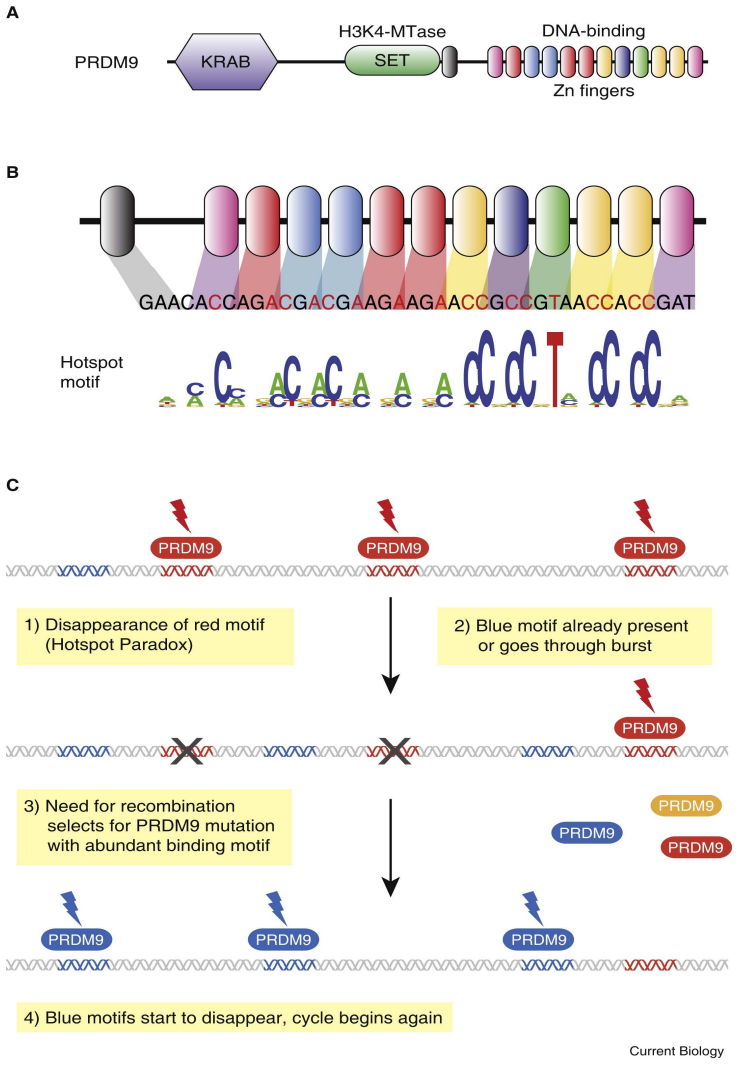


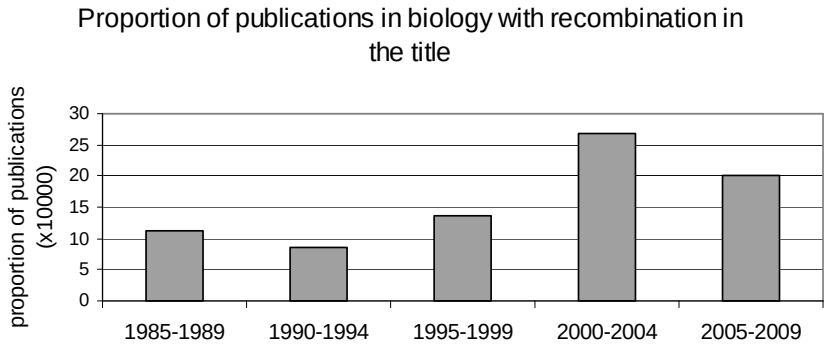
Figure 15 Function and evolution of PRDM9. (A) Functional domains of PRDM9. The SET domain has H3K4 trimethyltransferase activity. KRAB is a domain of unknown function found in many zinc-finger DNA-binding proteins. (B) Alignment of the 13-mer hotspot motif in humans and the predicted PRDM9-binding motif. Bases in red are those aligning with the motif. Degeneracy in the hotspot motif is shown. (C) Co-evolution of motifs and PRDM9. Recurrent changes in the PRDM9/motif pair imply fast evolution of hotspot distribution as well as interspecies differences and possibly incompatibilities. Figure taken from Hochwagen et. al 2010.

During 2009, three independent studies identified the zinc finger protein PRDM9 (Figure 15a) as a major determinant of hotspot activation in mammals. First, computational analysis predicted that PRDM9 could bind this 13-mer motif (Baudat et al. 2009; Myers et al. 2009) (Figure 15b). Second, differences in crossover distribution between laboratory mouse strains identified that differences in hotspot usage could be explained by sequence differences in Prdm9 (Baudat et al. 2009; Parvanov et al. 2009) and, the same correlation between PRDM9 alleles and hotspot usage held true in humans (Baudat et al. 2009). Finally, Berg et al. (2010), provided evidence that PRDM9 may define hotspot location even without binding to the known 13 bp motif. Overall, this provided convincing evidence that this gene is a central regulator of mammalian crossover distribution.

PRDM9 seems to evolve very fast since the zinc-finger domain numbers and sequences vary considerably among species (Oliver et al. 2009; Thomas et al. 2009). This could explain why hotspot location in chimpanzees and humans does not overlap: their PRDM9 proteins bind to different motifs. Moreover, this fast evolvavility of PRDM9 could provide some explanation for the hotspot paradox. Once the hotspot-promoting motif starts to be very rare in the genome, selection will favor the appearance of a new motif by producing a selective advantage to any of the new PRDM9 variants that bind to different motifs. A small change in the PRDM9 gene may trigger the change in hotspot location without need of other changes at the sequence level. (Figure 15c)

This places recombination as one of the fastest evolving systems, much faster than sequence evolution, and raises new questions on which role has recombination played in recent human evolution. The number of studies devoted to the study of recombination has increased in the last decades, and more importantly, the number of publications devoted to recombination in journals of high scientific impact has also increased (Figure 16). Studying how recombination has shaped our genomes is nowadays one of the most interesting questions in evolutionary biology and thus the number of studies related to recombination will likely continue to grow in the future.

A



B

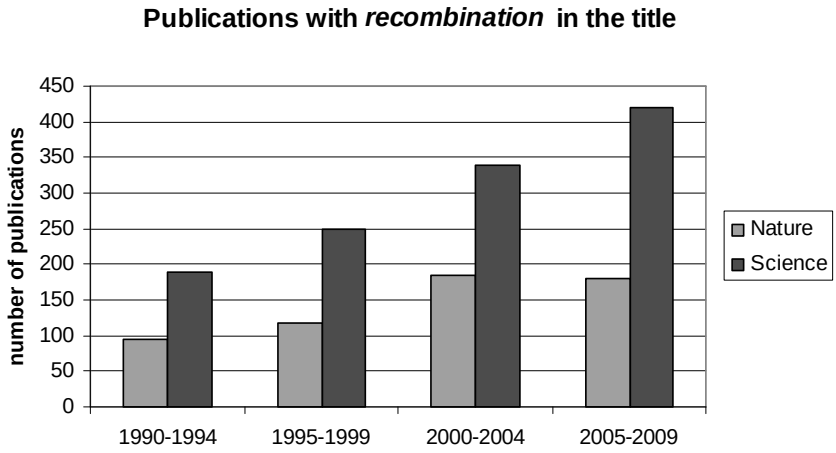


Figure 16. Publications with recombination in the title (A) All publications with recombination in the title in the field of "genetics" normalized by all publications in the field in the last 100 years. (B) Publications in the Science and Nature journals in the last 20 years. Source is the ISI web of knowledge for A, Google scholar for B.

2. OBJECTIVES

OBJECTIVES

The main aim of the work is the incorporation of recombination into the study of human population history by using recombination as a genetic marker. In order to do so, we had the following objectives:

1) Develop an algorithm capable of detecting recombination given a set of extant sequences. Specifically, this algorithm had to be both fast and able to detect the breakpoint location and the recombinant sequences.

2) Fine-tune the method as to make it suitable to analyze recombination in a set of human sequences.

3) Assess sensitivity and false discovery rate of the method relative to parameters such as age of the recombinations, recombination rate of the region, informativity of the two ancestral haplotypes… by means of using extensive simulations.

4) Select optimal regions and SNPs on the X chromosome for the application of the method.

5) Undertake a novel project of SNP typing in a new set of populations from the old world that was optimal for the study of recombination.

6) Characterize the recombinational landscape of these regions in a wide set of the method.

7) Interpret the results in terms of human population history

8) Revisit the question on what information can be extracted from haplotypes rather than SNPs when studying the patterns of human genetic variation.

3. RESULTS

3.1. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns

Laxmi Parida, **Marta Melé**, Francesc Calafell and Jaume Bertranpetit.

Journal of Computational Biology 15: 1133-1154 (2008)

http://www.liebertonline.com/doi/abs/10.1089/cmb.2008.0065

Parida L, Melé M, Calafell F, Bertranpetit J, Genographic Consortium. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. J Comput Biol. 2008; 15(9): 1133-54.

3.2. A New Method to Reconstruct Recombination Events at a Genomic Scale

Marta Melé, Asif Javed, Marc Pybus, Francesc Calafell, Laxmi Parida, Jaume Bertranpetit and The Genographic Consortium

PLoS Computational Biology 6**:** e1001010.

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1001010

Melé M, Javed A, Pybus M, Calafell F, Parida L, Bertranpetit J, et al. A new method to reconstruct recombination events at a genomic scale. PloS Comput Biol. 2010; 6(11): e1001010.

Supplementary information

Tables

Table S1. Evaluation of IRiS with the optimal parameters for different SNP ascertainments. SNP selection process is explained in the methods ssction. Mean SNP density values are calculated over all simulations. SNP sel = SNP selection method; MP = mergepats parameter; MM = minimum MAF, NR = number of runs, MSD = mean SNP density, FDR = false discovery rate, S= sensitivity, 90%CI = 90% Confidence Interval.

SNP sel	MP	MM	NR	LS (Kb)	MSD (SNP/bp)	FDR (%)	S (%)	90% CI
TAG aggressive	inactive	0.1	69	400	1/ 2758	5.64	18.61	4.75
TAG pairwise	inactive	0.1	69	400	1 / 2079	5.52	19.18	4.8
1SNP/5Kb	inactive	0.1	69	400	1 / 5014	7.24	19.94	5.14
1SNP/2Kb	inactive	0.1	100	200	1 / 2106	8.57	22.77	5.53
1SNP/Kb	inactive	0.1	100	200	1/ 1233	7.58	23.92	5.83
all SNPs	inactive	0.1	100	200	1 / 512	12.72	24	7.28
TAG pairwise	inactive	0.1	1000	200	1 / 1545	7.2	17.83	5.54
TAG pairwise	active	0.1	100	200	1/1980	5.65	18.67	5.52
1SNP/2Kb	inactive	0.01	100	200	1/2000	16.01	21.04	6.69
1SNP/Kb	inactive	0.01	100	200	1/1013	20.56	21.76	7.51

Table S2. Percentage values on the number of times each of the simulated event is either not detected, detected as 1 recombination or as 2 recombinations. The percentage values are calculated over 1000 in silico simulations.

	% not detected as recombination		% detected as 1 recombination		% detected as 2 recombination	
mergepats parameter	active	inactive	active	inactive	active	inactive
gene conversion (1 SNP)	99.2	97.7	0.8	2.3	0	0
gene conversion (3 SNPs)	96.2	92.6	3.7	7.3	0.1	0.1
gene conversion (5 SNPs)	90	88	9.8	12	0.2	0
gene conversion (10 SNPs)	76.2	74	23.7	25.6	0.1	0.4
recurrent mutation	89	78.5	10.8	21.4	0.2	0.1
phasing errors	57.3	50.9	12.1	14.6	30.4	34.5

Table S3. Number of recombinations detected in each of the 18 regions in the male dataset, female dataset and female dataset when removing putative phasing errors. Females were phased using both PHASE and fastPHASE without using male phase information.

REGIONS	MALE	FEMALE	FEMALE	MALE cleaned	FEMALE cleaned	FEMALE cleaned
phasing method		PHASE	fastPHASE		PHASE	fastPHASE
reg 1	442	432	473	364	376	359
reg 2	237	246	290	221	234	248
reg 3	58	77	77	58	75	73
reg 4	57	59	64	55	59	62
reg 5	269	269	319	257	255	293
reg 6	24	31	28	24	31	26
reg 7	149	166	178	139	166	162
reg 8	224	204	232	216	198	228
reg 9	298	312	353	284	300	315
reg 10	99	110	117	97	110	103
reg 11	126	111	133	114	107	123
reg 12	293	285	321	283	279	293
reg 13	75	77	78	73	73	76
reg 14	44	38	44	44	38	42
reg 15	370	324	388	326	308	326
reg 16	262	256	287	236	242	243
reg 17	252	264	293	228	240	257
reg 18	319	322	399	305	306	351
ALL	3598	3583	4074	3324	3397	3580

Table S4. The main characteristics of 18 X-chromosome regions. From left to right: start position and end position in base pairs (based on NCBI Build 36 assembly), length of each in base pairs, number of SNPs (N SNPs), number of haplotypes (N haplo), recombination rate calculated by means of Ldhat, Number of recombinations detected, number of recotypes, average number of recombinations detected by IRiS per Kb.

region	start (bp)	end (bp)	Length (bp)	N snps	N haplo	Rec Rate (4Ne/bp)	N rec	N reco	n_rec /Kb
reg1	22505979	22728622	222643	95	485	1.34	442	367	1.99
reg2	23071760	23213016	141256	97	375	1.06	237	208	1.68
reg3	25715611	26016381	300770	83	208	0.27	58	59	0.19
reg4	35038017	35504132	466115	84	170	0.23	57	57	0.12
reg5	38875482	39480082	604607	179	473	0.44	269	211	0.44
reg6	84704863	84952842	247979	80	81	0.11	24	24	0.1
reg7	86338463	86609425	270962	91	372	0.65	149	146	0.55
reg8	87288915	87838907	549992	205	453	0.54	224	187	0.41
reg9	93522874	94555707	1032833	183	478	0.37	298	223	0.29
reg10	112181012	112602418	421406	92	241	0.24	99	98	0.23
reg11	116631417	116865805	234388	82	324	0.53	126	123	0.54
reg12	120875730	121450338	574608	157	401	0.46	293	237	0.51
reg13	125833172	126301999	468827	91	169	0.19	75	74	0.16
reg14	126499106	126892013	392907	84	84	0.09	44	44	0.11
reg15	140883556	141050268	166712	99	494	1.68	370	327	2.22
reg16	141376625	141647366	270741	89	462	0.95	262	226	0.97
reg17	143563468	143896320	332852	97	414	0.61	252	235	0.76
reg18	144769060	145266667	497607	164	480	0.64	319	248	0.64
ALL			7197205	2052			3598		

Figure legends

Figure S1. Mean values taken from the analysis of 100 simulations with different IRiS settings: grain sizes (5, 10, 15, 20 and 30), different thresholds, defined as number of detections to be considered as true divided by the grain size or the double of the grain size in the cases in which the algorithm is run in two directions. For each setting the algorithm could be run only on the forward direction (F) or in both directions (FR). Figure S1A False discovery rate (%). Figure S1B Sensitivity (%). Figure S1C 90% confidence interval of the distance (measured in number of SNPs) between the inferred breakpoint position and the real location. Figure S1D, median age of the detected recombinations.

Figure S2. Mean values taken from the analysis of 100 simulations with different IRiS settings that combine different grain sizes (indicated with different colors), different thresholds (defined as number of detections to be considered as true divided by the sum of the different grain size and multiplied by two since the algorithm is run in the two directions). All settings included running the algorithm in the two possible senses. Figure S2A False discovery rate (%). Figure S2B Sensitivity (%).Figure S2C. 90th percentile distance from the breakpoint location measured in number of SNPs.

Figure S3. Plot showing the relationship between the false discovery rate and the number of COSI simulations under a scenario in which IRiS is given a different dataset than the one used to compare it with the COSI results.

Figure S4. Each dot represents mean values of false discovery rate and median age of the detected recombinations taken from the analysis of 100 simulations with different IRiS settings that combine different grain sizes (indicated with different colors) and different thresholds. All settings included running the algorithm in the two possible senses.

Figure S5. Plot showing values of the number of times *in silico* recombination events were detected by IRiS run with no threshold depending on the breakpoint location along the sequence. Different colors indicate different ways to produce the recombinant sequence, from light gray to black: “random” indicates that parental haplotypes were taken at random, “1dif near bkp” indicates that parental sequences had to be different near the breakpoint region (plus minus 10 SNPs), “ 2 dif near bkp” indicates that parental sequences had to be different near the breakpoint regions at both sides of the breakpoint, and “ unique” indicates that the parental sequences had to be different near the breakpoint region and the recombinant sequence had to be unique within the breakpoint region. Below, the recombination rate estimated by LDhat is shown, following the right axis.

Figure S6. MDS 2D plot based on a recombinational distance matrix. The stress is 0.081 which is below the 0.16 stress obtained with 1% probability with random data sets (Sturrock and Rocha 2000).

1. Sturrock K, Rocha J (2000) A Multidimensional Scaling Stress Evaluation Table. Field Methods 12: 49-60.

Figures

Figure S1A

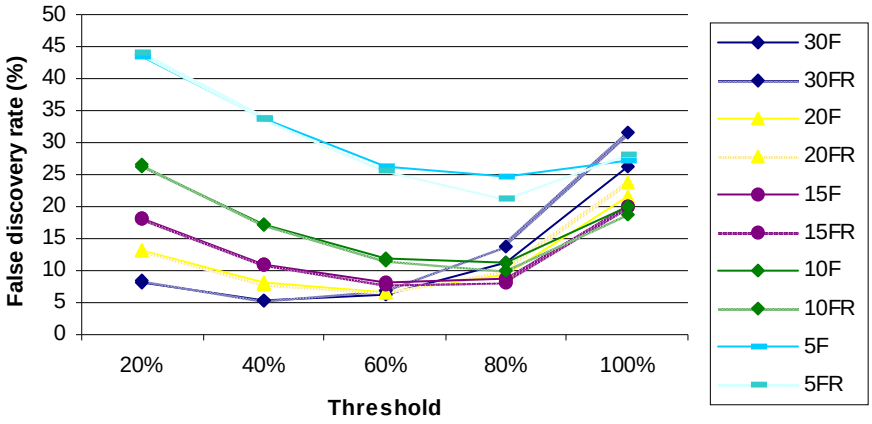


Figure S1B

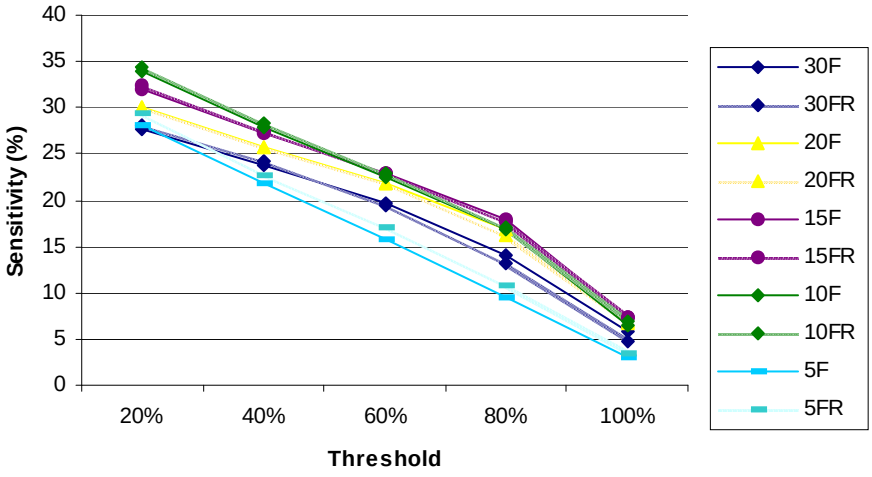


Figure S1C

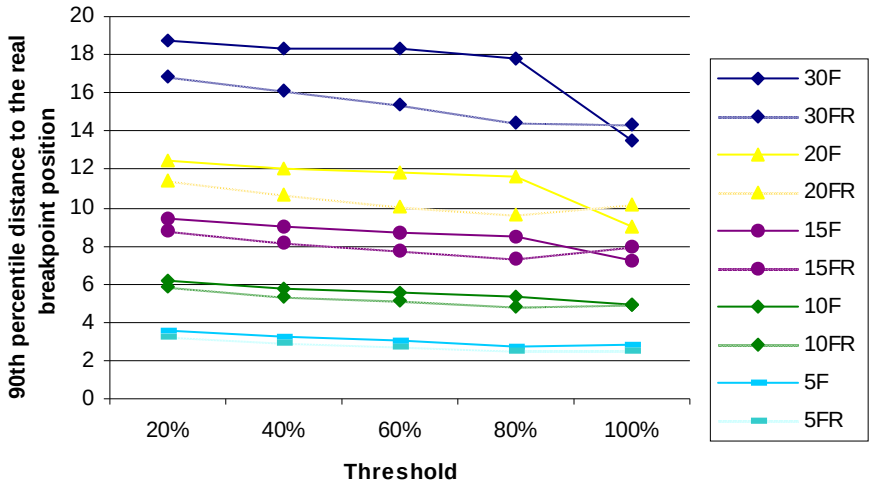


Figure S1D

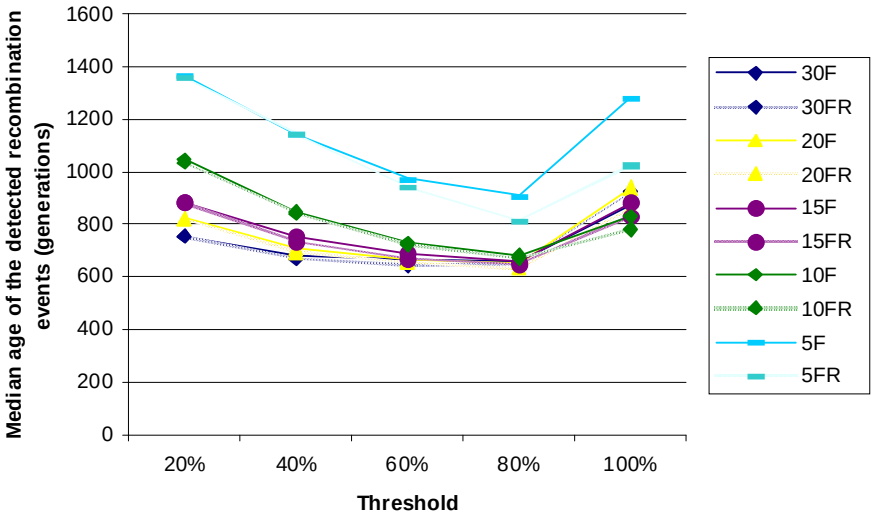


Figure S2A

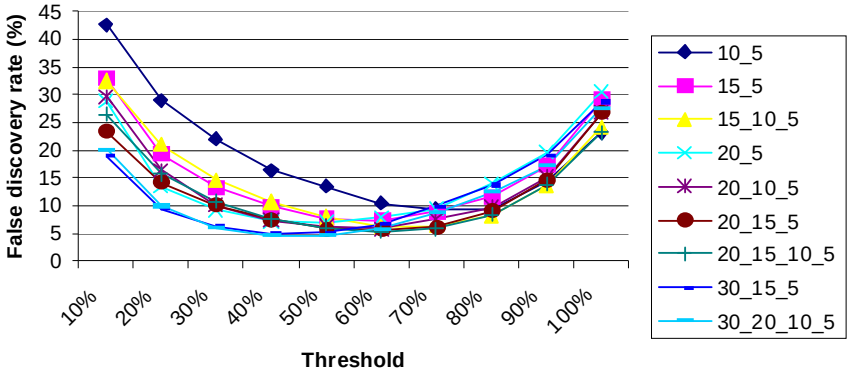


Figure S2B

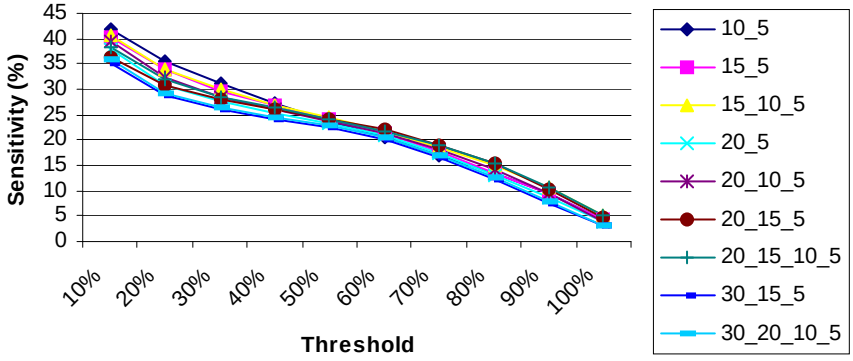


Figure S2C

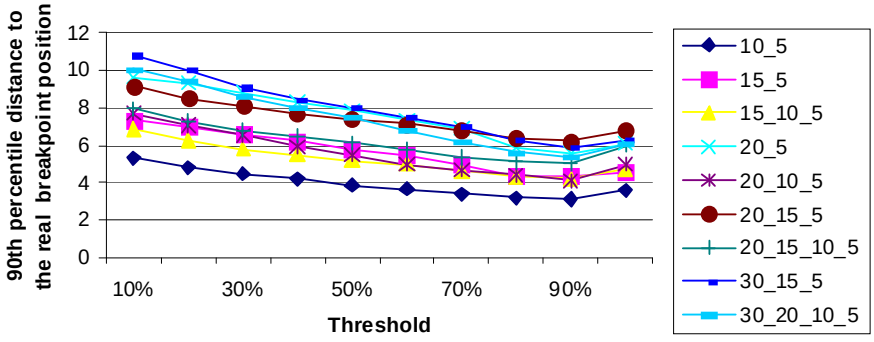


Figure S3.

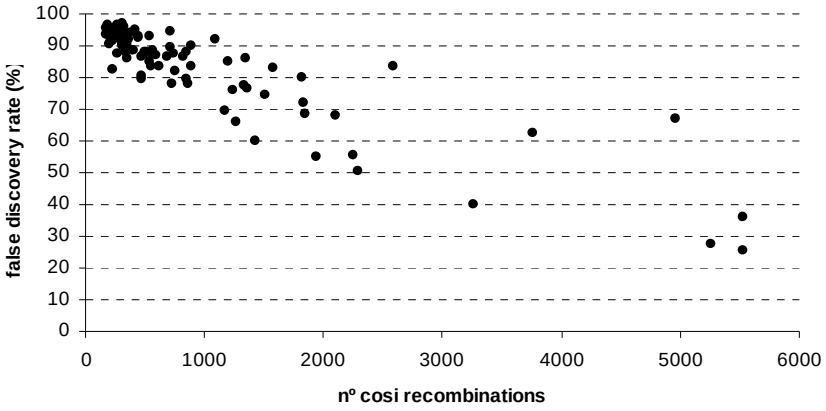
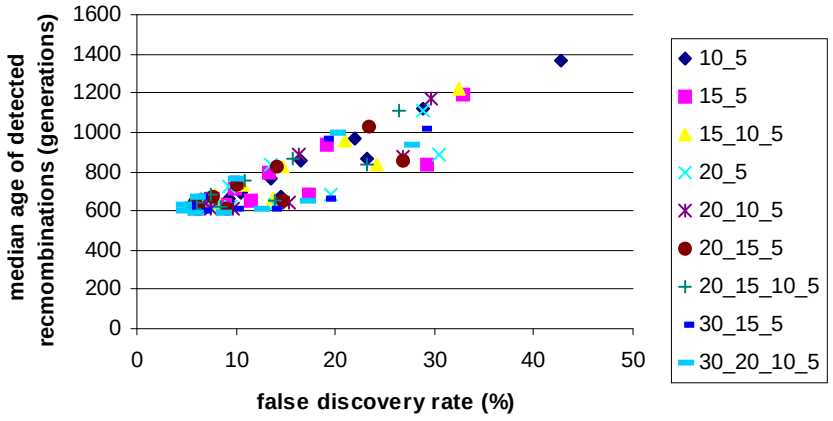


Figure S4.



FigureS5.

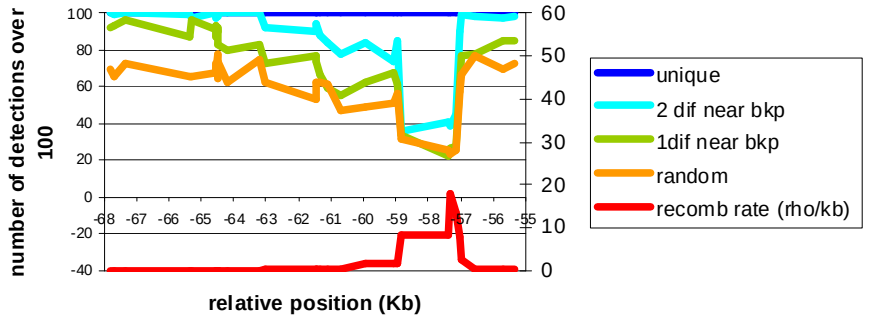
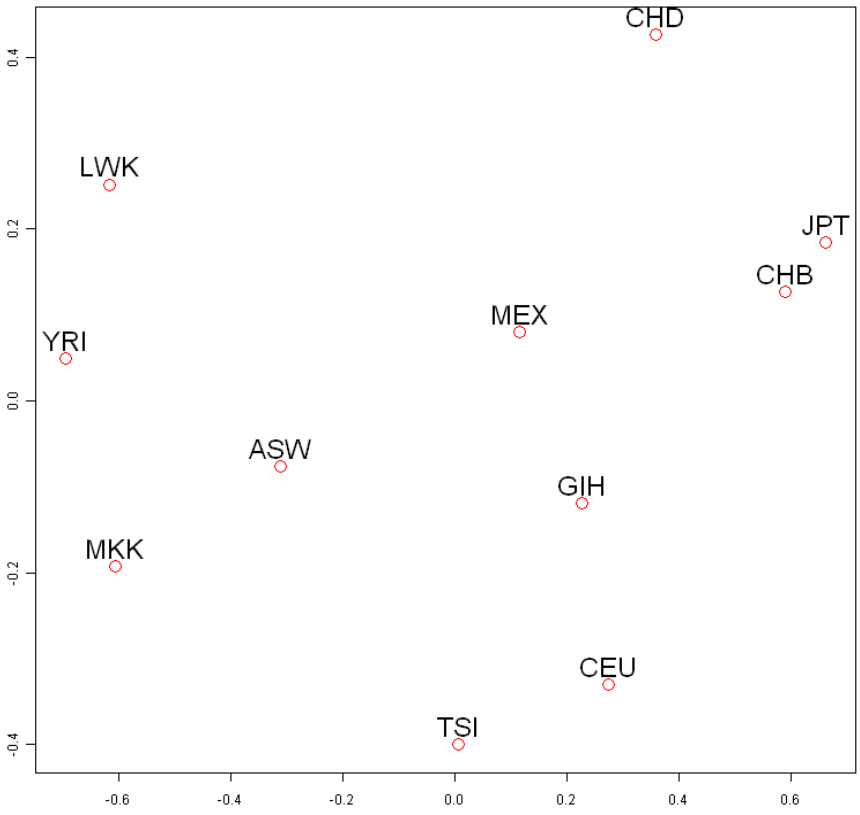


Figure S6


CHD
LWK
JPT
CHB
MEX
YRI
ASW
GIH
MKK
CEU
TSI
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6

3.3. The footprint of recombination gives a new insight in the effective population size and the history of the Old World human populations

Marta Melé', Asif Javed', Marc Pybus, Pierre Zalloua, Marc Haber, David Comas, Mihai G. Netea, Oleg Balanovsky, Elena Balanovska, Li Jin, Yajun Yang, RM. Pitchappan, G. Arunkumar, Laxmi Parida, Francesc Calafell, Jaume Bertranpetit and The Genographic Consortium.

Submitted.

The footprint of recombination gives a new insight in the effective population size and the history of the Old World human populations

Marta Melé[1,], Asif Javed[2,], Marc Pybus[1], Pierre Zalloua[3], Marc Haber[3], David Comas[1], Mihai G. Netea[4], Oleg Balanovsky[5, 6], Elena Balanovska[5], Li Jin[7], Yajun Yang [7], RM. Pitchappan[8,9], G. Arunkumar[9], Laxmi Parida[2], Francesc Calafell[1], Jaume Bertranpetit[1] and The Genographic Consortium[¶].

[1] IBE, Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Catalonia, Spain. Computational Biology Center, IBM T J Watson Research, Yorktown, USA. [3] Lebanese American University, School of Medicine, Beirut, Lebanon. [4]Department of Medicine and Nijmegen Institute for Infection, Inflammation, and Immunity, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands. [5] Research Centre for Medical Genetics, Moscow, Russia. [6] Vavilov Institute for General Genetics, Moscow, Russia. [7] MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China. [8] Chettinad Academy of Research & Education, Chettinad Health City, Rajiv Gandhi Salai, Kelampakkam, Chennai, India. [9] School of Biological Sciences, Madurai Kamaraj University, Madurai, 625021 India.

Correspondence: FC, JB

¶ Membership of the Genographic Consortium is provided in the Acknowledgments

Abstract

Effective population size captures in a single parameter the cumulative effects of drift in a population. While estimates are available for the human species or for broad continental groups, a detailed survey for single human populations has not been produced. Here we provide such figures, and interpret them in terms of the demographic history of anatomically modern humans by means of a recombination-based analysis. We have genotyped a set of 1250 SNPs in five regions of the X chromosome in 1240 males from 30 Old World populations. We have counted the number and location of recombination events and have detected the sequences that carry them by means of a combinatorial algorithm implemented in the IRiS program. The number of recombinations can be used to estimate effective population size through the $\rho=4N_e r$ parameter. We have found, in line with other studies, that African populations have effective population sizes that are ~3 times greater than those of non-African populations. Outside of Africa, South Asian populations had the largest effective sizes. Additionally, recombinational diversity correlated with distance out of Africa through a southern, but not a northern, route, and, in Eurasian populations, recombinational distance correlated with distance from Southern India. These findings suggest a larger role than previously envisaged for South Asia in the demographic history and population expansions of anatomically modern humans out of Africa.

The estimation of effective population size in human evolution has been a subject of intense research in the recent past. The seminal papers by Takahata (reviewed in Kim et al (2010)) established the highly cited figure of 10,000 individuals for the past human evolutionary history, which has been lately revised to 15,000 with a much larger genetic dataset (Kim et al. 2010). These figures have been derived through gene diversity estimates of 31 autosomal loci. This led to an estimated average time to the most recent common ancestor (TMRCA) of 1.24 Myr. The effective population size estimate, thus, captures an extremely long period of time, much beyond the existence of our own species. This presumably denotes greatly fluctuating biological phenomena both in space and time, which could not be captured by any of the present methodologies.

Laval et al. (2010) have formulated a detailed "Historical and Demographic Model" of recent human evolution. This is based on diversity (or heterozygosity) measures on resequencing data for noncoding autosomal regions, with interesting co-estimation of parameters. Their results on effective population sizes, estimated for three main continental populations, offer figures of ~31,200 for Europeans and ~14,500 for Asians after the split of these populations ~22,500 years ago, with a more complex picture for Africa. It is difficult to disentangle the importance of the shared ancestral polymorphisms and the accuracy and robustness of the many estimated parameters in the estimates based on population-specific sequence diversity.

Earlier, Hayes et al. (2003) and Tenesa et al. (2007), considering the importance of the temporal and spatial framework for effective population size estimates, have proposed an independent method based on Linkage Disequilibrium (LD) data. Their analysis on four HapMap populations results in much lower estimates, on the order of ~7,500 for African Yoruba (YRI) and ~3,100 for each of the Eurasian populations. Their lower values are justified by the time frame of the genetic events analyzed. While gene diversity would reflect average population size for long periods of time, LD depends to a greater extent on the population size in more recent times. The method relies on computing r^2 on inferred haplotype frequencies, which may not be accurate. Moreover, although LD is indeed primarily determined by recombination rate, demography and natural selection can also modify LD and alter recombination rate estimates based on LD.

Finally, Cox et. al (2008) have used an isolation-with-migration model to independently estimate effective population sizes and migration rates using resequencing data of noncoding regions on the X chromosome. They found low (albeit somewhat imprecise) values of effective population sizes as well, which were higher for Africans (2,300–9,000) than for non-Africans (300–3,300).

We have developed a method called IRiS (Identifying Recombination in Sequences) to detect specific past recombination

events from extant sequences (Melé et al., 2010) based on a combinatorial algorithm (Parida et al. 2009; Parida et al. 2008). The algorithm yields which sequences are descendants of ancient recombination events, which sequences carry the ancestral patterns that were involved in the recombination event, and where is the breakpoint located in the genome. Here we propose to use the data produced in a large SNP survey (Javed et al., in preparation) to estimate the historical recombinations produced, and from them estimate the effective population sizes of the diverse populations. Several points in this approach are novel: the detection of recombinations is not based on LD, the time distribution of the reconstructed recombinations is known (Melé et al. 2010), and the recombination rate is available at a very narrow scale in the human genome (Kong et al. 2010).

It is well known that recombination is not evenly distributed across the genome; 80% of the recombination events take place in 20% of the sequence, in recombination hotspots (Myers et al. 2005). Therefore, newer recombinations may overwrite traces of past events and the main consequence of this process is that allocating specific recombinations to specific sequences becomes harder for older events. In our previous study (Melé et al. 2010), we showed that recent recombinations are detected by IRiS with greater sensitivity. Specifically, we inferred that 90% of the events detected by IRiS occurred after the Out of Africa migration. Therefore, recombinations can be used as recent genetic markers and they can potentially help to make inferences on the most recent events of human evolutionary history, such as the estimation of population-specific effective population size. In fact, most of the reconstructed recombinations are population-specific (93.13%). The historical recombinations that can be detected are a fraction of the total recombinations that occurred. This fraction, that is, the sensitivity achieved by IRiS, can be estimated at 7.3% with simulations (see Supplementary Text for details) and used to obtain an indirect estimate of the total number of recombinations that have taken place along the genealogy. Sensitivity estimations may be affected by a stochastic variation that may compound into the absolute values of effective population sizes. However, since we can indeed detect actual recombinations in each of the populations, estimation of the

relative effective population sizes among the 30 analyzed populations will be more robust than more indirect methods.

The dataset used consisted of 1250 SNPs in five regions (Table S1) spanning 2 Mb of presumed gene-free regions of the X chromosome, genotyped in 1240 males from 30 Old World populations (Figure 1, Table S2). High, uniform SNP coverage was necessary to detect as many recombination events as possible; by choosing only male samples, we could overcome the uncertainty associated with phasing haplotypes; finally, regions known to contain genes were avoided in order to prevent the possible confounding effects of natural selection. Further details about region selection and genotyping can be found in the Supplementary Text.

We used the expression $\rho = 3N_e r$ (Hartl and Clark 1997) to infer the effective population size, where $\rho = \frac{R}{\sum_{i=1}^{n-1} \frac{1}{i}}$ where R is the number of recombinations inferred for each population, and n is the number of sequences analyzed. R values were calculated by dividing the number of recombinations detected by IRiS (Table 1) by the corresponding sensitivity. r stands for the recombination rate, which was calculated as the weighted average of the rates of each region based on the deCODE map (Kong et. al 2010). Finally, equal male and female effective population sizes were considered, which, for the X chromosome, implies that $\rho = 3N_e r$ ($\rho = 4N_e r$ for autosomes).

Estimates of the effective population sizes for each of the populations are given in Table 1 (in relative and absolute values) and plotted in Figure 2. As expected, results consistently show that Sub-Saharan Africans have much higher effective population sizes than all other populations; values are roughly four-fold larger, or, in absolute terms, of ~4000 for African populations and of ~1000 for the rest. This result is in line with the low values obtained with LD-based estimates (Laval et al. 2010; Tenesa et al. 2007), but, as mentioned above, the relative population sizes, and, in particular,

the ratio of the sizes between African and non-African populations are more reliable figures, and that was also found to be >2.5 both from genetic diversity (Laval et al. 2010) and from LD (Hayes et al. 2003; Tenesa et al. 2007).

For the first time, we provide specific effective sizes for a wide range of Old World populations in relative and absolute values (Table 1). Besides the Sub-Saharan African / non sub-Saharan chasm in population sizes, a number of interesting patterns are revealed. The populations with the largest sizes other than Sub-Saharan Africans are North Africans (Moroccans and Egyptians), as could be expected due to their known Sub-Saharan admixture (Bosch et al. 2001; Brakez et al. 2001; Krings et al. 1999). Outside of Africa, the largest effective population sizes are found in South Asia; only recently, the high internal diversity of Indian populations is being appreciated (Xing et al. 2010). Europeans and East Asians have similar effective population sizes. Tibetans and Basques showed the lowest values, a direct measure of small population size and isolation.

We further investigated the geographic variation of both SNPs and recombinations to understand the general pattern of genetic variation and population history (Table 1). In order to compare patterns of diversity across populations, we used Nei's nucleotide diversity statistic to calculate the standard gene diversity using either SNP allele frequencies or population frequencies of each recombination event using the whole dataset. With this approach, we can apply the same, widely used measure of diversity both to the SNP alleles in a classical fashion, and to our new data on detected recombination events.

We provide a geographic framework to these values, by plotting them against the geographic distance of each population to Eastern Africa, the presumed place of origin of modern humans (Quintana-Murci et al. 1999; Tishkoff et al. 2009). As expected, gene diversity was found to be highly correlated with geographical distance with East Africa (Spearman's r = -0.596; p = 0.00050) (Figure 3) (Prugnolle et al. 2005b; Ramachandran et al. 2005b); even if

African samples were removed (Spearman's r = -0.445, p = 0.023). With recombinational diversity, nonetheless, a marginal correlation is found (Spearman's r = -0.363; p = 0.048) which completely disappears if African samples are removed from the analysis (Spearman's r = -0.0352; p = 0.86) (Figure 4). The main differences between the two plots are that African populations show significantly higher recombinational diversity than any other population (Mann-Whitney test; p = 0.0015), in a proportion that goes to a four- or five-fold higher diversity than the mean for non Africans; European populations show similar diversity values as East Asian populations, whereas Indian populations showed significantly more diversity than Europeans (Mann Whitney test p = 0.0055) and East Asians (Mann Whitney test p = 0.011).

The present results stress the wide differences between Sub-Saharan Africans and the rest of the Old World populations and point to a special role for South Asia in the Out of Africa expansion of modern humans; this role could have been more significant than those of places located on the possible corridor out of Africa, be it the posited Northern route through the Middle East, or the Southern route through Arabia. It is debatable whether this is an argument for India having had a role in a maturation phase prior to the expansion of modern humans to the whole of Eurasia.

Given the fact that recombinational diversity was not related with the distance from East Africa and that effective population size was notably higher in India compared to other Eurasian populations, we tested whether recombinational diversity was correlated with the geographical distance of Eurasians from South Asia, particularly south India (Figure 5). This correlation turned out to be significant (Spearman's r = -0.495; p = 0.010). One of the clear outliers of the regression were the Moroccan, which is somehow expected if they have a high proportion of sub- Saharan ancestry, as discussed above. If this population is removed the correlation coefficient increases (r = -0.682; p = 0.0002).

Finally, in order to assess whether a southern route out of Africa could better explain the relationship between recombinational

diversity and distance from East Africa, we calculated the distance from East Africa considering that non-African populations left Africa through the Bab-el-Mandeb Strait and going through the Red Sea following the coastline. The correlation became stronger and highly significant (Spearman's r = -0.592, p = 0.00057). If African samples (including the Moroccans) are removed from the analysis, the result is maintained (Spearman's r = - 0.523, p = 0.0073). Interestingly, the southern route explained better the patterns of recombinational diversity, but not the patterns based on nucleotide diversity values (Spearman's r = - 0.463, p =0.010), which correlate less with geographical distance through the southern than through the northern route. The difference between nucleotide and recombinational distance may just be a reflection of the time frame of both approximations, with the recombination analysis detecting events that happened more recently.

We have thus presented a new method of analyzing SNP-based genetic information that uncovers one of the main (albeit often neglected) processes generating genetic diversity, namely recombination. By directly counting recombinations, we have provided effective relative and absolute population size estimates for a number of interesting populations studied here. We have also described geographic patterns of genetic diversity based on recombinations that are less clear if nucleotide diversity is considered; by focusing on recombination, we seem to have overcome the effect of SNP ascertainment bias and have focused the analysis on the timeframe of recent human history, since the Out of Africa expansion. We have thus managed to recover the known higher effective population size of Sub-Saharan Africans, but we also have found high population sizes in North Africa and South Asia. While the former may be the trivial consequence of Sub-Saharan African admixture, the latter may open the avenue for the exploration of a larger role than previously envisaged for South Asia in the path that led modern humans from Africa to the rest of the world.

Acknowledgements

We are grateful to Mònica Vallés, UPF, for excellent technical support and to Sònia Sagristà for her help. Funding for this project was provided by the Spanish Ministry of Science and Innovation project BFU BFU2007-63657; and by National Geographic and IBM within the Genographic Project initiative. MM was supported by grant AP2006-03268. Genotyping and bioinformatic services were provided respectively by CEGEN (Centro Nacional de Genotipado) and INB (National Bioinformatics Insitute), Spain. HapMap phase III population samples were obtained from the Coriell Cell Repository.

The Genographic Consortium includes: Syama Adhikarla (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Christina J. Adler (University of Adelaide, South Australia, Australia), Danielle A. Badro (Lebanese American University, Chouran, Beirut, Lebanon), Andrew C. Clarke (University of Otago, Dunedin, New Zealand), Alan Cooper (University of Adelaide, South Australia, Australia), Clio S. I. Der Sarkissian (University of Adelaide, South Australia, Australia), Matthew C. Dulik (University of Pennsylvania, Philadelphia, Pennsylvania, United States), Christoff J. Erasmus (National Health Laboratory Service, Johannesburg, South Africa), Jill B. Gaieski (University of Pennsylvania, Philadelphia, Pennsylvania, United States), Wolfgang Haak (University of Adelaide, South Australia, Australia), Angela Hobbs (National Health Laboratory Service, Johannesburg, South Africa), Matthew E. Kaplan (University of Arizona, Tucson, Arizona, United States), Shilin Li (Fudan University, Shanghai, China), Begoña Martínez-Cruz (Universitat Pompeu Fabra, Barcelona, Spain), Elizabeth A. Matisoo-Smith (University of Otago, Dunedin, New Zealand), Nirav C. Merchant (University of Arizona, Tucson, Arizona, United States), R. John Mitchell (La Trobe University, Melbourne, Victoria, Australia), Amanda C. Owings (University of Pennsylvania, Philadelphia, Pennsylvania, United States), Daniel E. Platt (IBM, Yorktown Heights, New York, United States), Lluis Quintana-Murci (Institut Pasteur, Paris, France), Colin Renfrew (University of Cambridge, Cambridge, United Kingdom), Daniela R. Lacerda (Universidade Federal de Minas Gerais, Belo Horizonte,

Minas Gerais, Brazil), Ajay K. Royyuru (IBM, Yorktown Heights, New York, United States), Fabrício R. Santos (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil), Theodore G. Schurr (University of Pennsylvania, Philadelphia, Pennsylvania, United States), Himla Soodyall (National Health Laboratory Service, Johannesburg, South Africa), David F. Soria Hernanz (National Geographic Society, Washington, District of Columbia, United States), Pandikumar Swamikrishnan (IBM, Somers, New York, United States), Chris Tyler-Smith (The Wellcome Trust Sanger Institute, Hinxton, United Kingdom), Kavitha Valampuri John (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Arun Varatharajan Santhakumari (Madurai Kamaraj University, Madurai, Tamil Nadu, India), Pedro Paulo Vieira (Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil), R. Spencer Wells (National Geographic Society, Washington, District of Columbia, United States), Janet S. Ziegle (Applied Biosystems, Foster City, California, United States).

References

Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, Bertranpetit J (2001) High-Resolution Analysis of Human Y-Chromosome Variation Shows a Sharp Discontinuity and Limited Gene Flow between Northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68:1019-1029

Brakez Z, Bosch E, Izaabel H, Akhayat O, Comas D, Bertranpetit J, Calafell F (2001) Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. Annals of Human Biology 28:295-307

Cox M, Woerner A, Wall J, Hammer M (2008) Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. BMC Genetics 9:76

Hartl DL, Clark AG (1997) Principles in Population Genetics. SinauerAssociates Inc.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME (2003) Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. Genome Research 13:635-643

Kim HL, Igawa T, Kawashima A, Satta Y, Takahata N (2010) Divergence, demography and gene loss along the human lineage. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2451-2457

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, Gudjonsson SA, Frigge ML, Helgason A, Thorsteinsdottir U, Stefansson K (2010) Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467:1099-1103

Krings M, Salem A-eH, Bauer K, Geisert H, Malek AK, Chaix L, Simon C, Welsby D, Di Rienzo A, Utermann G, Sajantila A, Pääbo S, Stoneking M (1999) mtDNA Analysis of Nile River Valley Populations: A Genetic Corridor or a Barrier to Migration? Am J Hum Genet 64:1166-1176

Laval G, Patin E, Barreiro LB, Quintana-Murci L (2010) Formulating a Historical and Demographic Model of Recent Human Evolution Based on Resequencing Data from Noncoding Regions. PLoS ONE 5:e10284

Melé M, Javed A, Pybus M, Calafell F, Parida L, Bertranpetit J, The Genographic C (2010) A New Method to Reconstruct Recombination Events at a Genomic Scale. PLoS Comput Biol 6:e1001010

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321-324

Parida L, Javed A, Mele M, Calafell F, Bertranpetit J (2009) Minimizing recombinations in consensus networks for phylogeographic studies. BMC Bioinformatics 10 Suppl 1:S72

Parida L, Mele M, Calafell F, Bertranpetit J (2008) Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. J Comput Biol 15:1133-1154

Prugnolle F, Manica A, Balloux F (2005) Geography predicts neutral genetic diversity of human populations. Curr Biol 15:R159 - 160

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet 23:437-441

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 102:15942 - 15947

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM (2007) Recent human effective population size estimated from linkage disequilibrium. Genome Research 17:520-526

Tishkoff SA, Reed FA, Friedlaender FoR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM (2009) The Genetic Structure and History of Africans and African Americans. Science 324:1035-1044

Xing J, Watkins WS, Hu Y, Huff C, Sabo A, Muzny D, Bamshad M, Gibbs R, Jorde L, Yu F (2010) Genetic diversity in India and the inference of Eurasian population expansion. Genome Biology 11:R113

Tables

Table 1. Diversity calculated as in Nei's gene diversity formula for alleles (Nuc Div) and recombination junctions (Rec Div); mean number of recombinations detected over 100 runs on datasets created by randomly selecting 18 chromosomes per population per region and their standard deviations (Mean nº rec (stand dev)). Effective population size (Ne) and relative population size (Relative Ne) which is calculated based on the lowest value (that of Tibetans).

Population	Nuc Div	Rec Div (x 1000)	Mean nº rec (stand dev)	Ne	Relative Ne
Yoruba(YRI)	0.29	4.79	112.0 (8.0)	4287	7.5
Maasai (MKK)	0.29	4.67	84.1 (7.9)	3217	5.6
Luuya (LWK)	0.29	5.72	113.5 (9.4)	4344	7.6
Chad (CHA)	0.31	5.14	106.4 (7.9)	4072	7.1
Lebanese (LEB)	0.26	1.49	27.5 (3.9)	1438	2.5
Kuwaiti (KUW)	0.28	1.64	32.9 (4.5)	1461	2.5
Iranian (IRA)	0.26	1.16	25.9 (4.0)	1054	1.8
Egyptian (EGY)	0.27	2.03	38.2 (5.1)	1260	2.2
Moroccan (MOR)	0.3	2.38	37.6 (3.7)	993	1.7
N.and W. European (CEU)	0.27	1.01	19.9 (3.7)	761	1.3
British (BRI)	0.27	1.05	21.8 (3.5)	832	1.5
Dutch (DUT)	0.27	1.01	20.2 (3.2)	772	1.3
Basque (BAS)	0.26	0.51	15.5 (3.0)	594	1
Tuscan (TSI)	0.26	0.93	19.7 (3.8)	752	1.3
Romanian (ROM)	0.28	0.65	18.0 (3.3)	689	1.2
Chechen (CHE)	0.27	1.44	23.0 (3.6)	881	1.5
Russian (RUS)	0.27	1.11	21.7 (4.4)	831	1.5
Tatar (TAT)	0.26	0.82	18.0 (3.0)	688	1.2
Altaian (ALT)	0.27	1.32	24.6 (3.8)	940	1.6
Uighur (UIG)	0.27	1.09	22.8 (4.3)	873	1.5
Gujarati (GIH)	0.27	1.75	31.9 (4.2)	1222	2.1
Nadar (CAN)	0.26	1.95	31.9 (4.4)	1219	2.1
Parayar (NTN)	0.28	2.35	40.2 (5.0)	1539	2.7
Kalita (KAL)	0.27	1.46	32.5 (4.5)	1242	2.2
Adi (ADI)	0.24	1.1	26.2 (4.2)	1001	1.7
Tibetan (TIB)	0.24	0.52	15.0 (3.5)	573	1
Laotian (LAO)	0.24	1.21	23.8 (4.1)	911	1.6
Ati (ATI)	0.25	1.22	25.6 (2.6)	980	1.7
Chinese (CHB)	0.24	1.26	28.8 (4.6)	1103	1.9
Japanese (JPT)	0.23	0.98	23.0 (3.4)	879	1.5

Figures

Figure 1. Populations of the study and their geographic region of ancestry. Abbreviations as in Table 1.

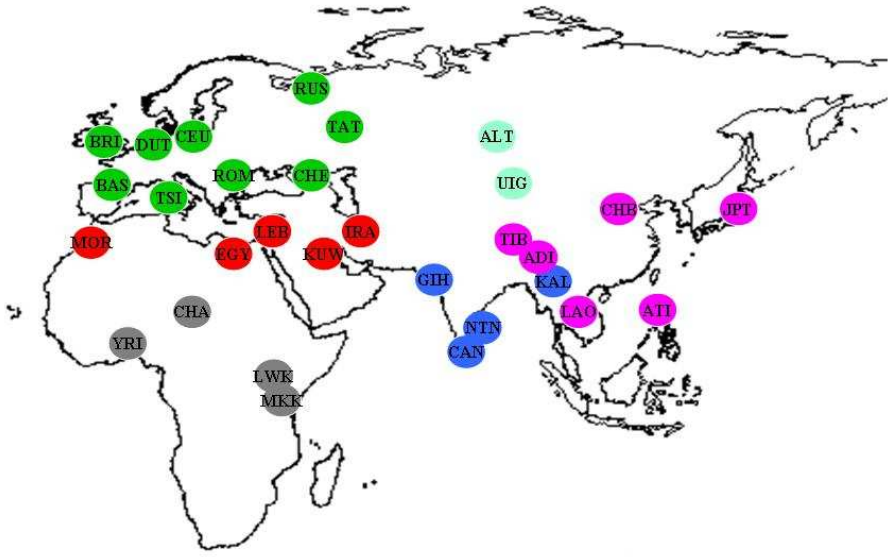


Figure 2. Inferred effective population sizes from the number of recombinations detected. Standard deviations and population abbreviations as in Table 1.

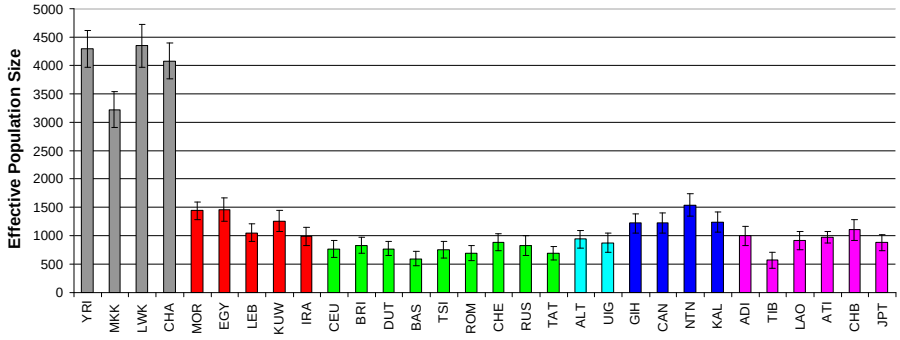


Figure 3. Nucleotide diversity and geographic distance from East Africa (in Km), through the northern route. Populations are color-coded by continent as in Figure 2.

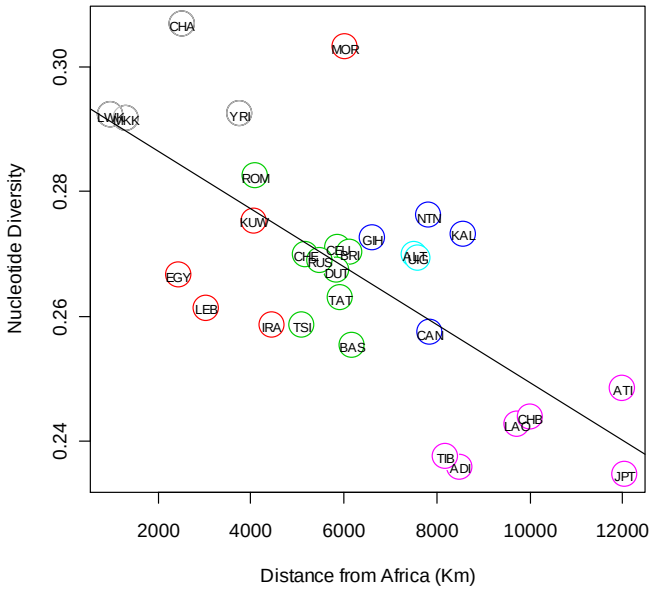


Figure 4. Recombiantional diversity and geographic distance from East Africa (in Km), through the northern route. Populations are color-coded by continent as in Figure 2.

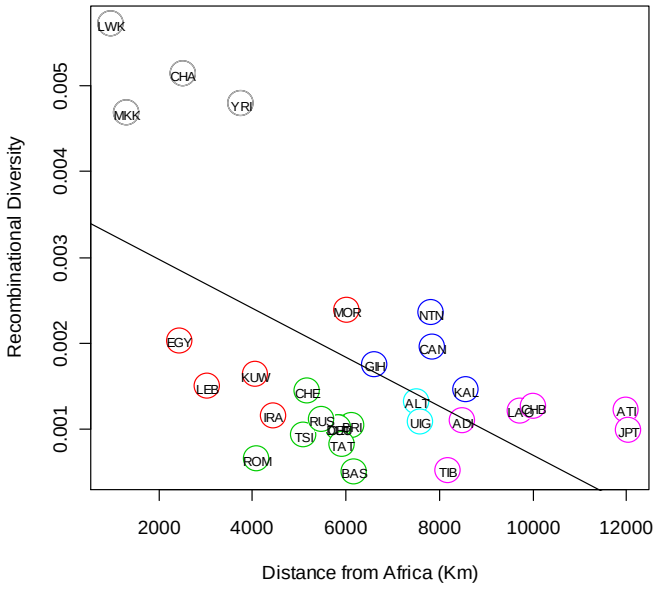
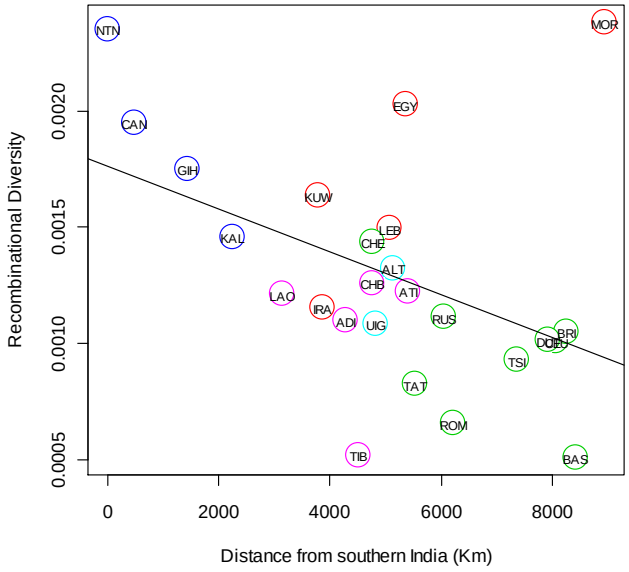


Figure 5. Recombinational diversity and distance from South India. Populations are color-coded by continent as in Figure 2.



Supplementary information

Materials and Methods

Genotype selection

Data was obtained from a previous study (Javed. et al. in preparation) in which five gene-free regions of the X chromosome spanning more than 2Mb were genotyped. Genotyping was performed using the Illumina GoldenGate custom Oligos and SNPs were selected based on the HapMap phase II release 24 as to obtain the highest possible density provided they meet some technical genotyping conditions.

In the mentioned study, quality control processes included removing SNPs with more than 15 % of missing data and those having a cluster of heterozygous positions in male samples. Samples with missing data higher than 10% or male samples with more than 3 heterozygous positions were excluded.

For our study, only male genotypes were taken in order to avoid phasing errors Informed consent was obtained for all the subjects. Recently admixed populations (namely, African Americans, Mexicans, and Gypsies) were not included in the present study. Details on the origin of samples, and number of individuals per population can be found in Table S2.

Recombinational analysis

The IRiS method was run with the optimal parameters defined in Melé et. al (2010) with the mergepats parameter on to ensure the robustness to any present genotyping errors, recurrent mutations or gene conversion events.

In order to overcome the uncertainty in the number of recombinations detected due to sampling, 100 datasets were created in which 18 different chromosomes per population per region were selected randomly. IRiS was run on each of the datasets and the average number of recombinations detected and standard deviations per population was extracted.

One single run of IRiS was performed with all the samples together in order to calculate recombinational diversity in the equivalent way as gene diversity was calculated over the whole dataset. Basically, using all the recombinations detected on our set of sequences, a matrix in which each detected recombination has a column and the names of the sequences are in rows. Presence or absence of particular events is marked with ones or zeroes respectively. Then, each sequence is defined by a string of zeroes and ones that is called a recotype in which one indicates presence of a specific recombination event and zero absence. Then we can perform the equivalent calculation of gene diversity using the recotype matrix

Sensitivity estimation

Sensitivity of the method was estimated using the coalescent simulator COSI (Schaffner et al. 2005) in the same way as described in Melé et al (2010). COSI has been calibrated in order to simulate data that resembles the extant human data by means of simulating a human genealogy modeling variable recombination rate and hotspots.

Since COSI simulates three human populations based on a human demography, we simulated the number of chromosomes that matched our data: 90 for the African, 180 for the European and 108 for East Asians considering we took 18 chromosomes for each population. The lower effective population size given the X chromosome (3/4 of the N_e of the model for autosomes) was also taken into account. In order to obtain a similar SNP density, all datasets were ascertained in order to have an approximate density of 1 SNP every 1600 bp. We run 1000 simulations and estimated sensitivity values for IRiS over 1000 different genealogies.

The model implemented in COSI named “the best fit model” generates, in each simulation, a different recombination rate distribution including varying hotspots location. Therefore, the actual number of recombinations in each simulation has a large stochastic variation. Since the sensitivity of IRiS is highly affected by the recombination rate (Melé et. al 2010), we calculated a regression line between sensitivity and the recombination rate based on the 1000 simulations. The function, which had a correlation coefficient of 0.72, was:

$$Sensitivity = 0.00143 \times (RecombinationRate^{-0.479})$$

Then, based on the deCODE map of recombination rate (Table S1) (Kong et. al 2010), we could calculate a weighted average of the recombination rate over the five regions and estimate the sensitivity of IRiS in our dataset to be 7.3% according to the equation above.

Geographic distance calculations

Geographic distances from Eastern Africa were calculated as in Jackobson et. al and Ramachandran et. al 2005 which start at Addis Ababa (9N, 38E) and for all non African populations travel through Egypt (30N, 31E). Paths to Europe also passed through Turkey (41N, 28E). The alternative southern route for non-African populations was calculated as if crossing the Bab-el-Mandeb straits going directly to Iran (26N, 58E). Distances from south India were calculated starting from Villupuram (12N, 80E) where the Parayars had been collected. The Himalaya Mountains were taken into account by forcing the path to the Uygur and the Altaians to pass through Dushanbe in Tajikistan (39N, 69E) and the Ati and Lao through Chengdu (31N, 104E). Other paths to East Asian populations out of South Asia passed through Dacca in Bangladesh (24N, 90E) in order to get out of the Indian peninsula. Distances were calculated using the great-circle distance implemented in the http://williams.best.vwh.net/gccalc.htm site with the default Earth model (WGS84/NAD83/GRS80) (Table S1).

All correlation analyses were performed using SPSS software (SPSS Inc., Chicago IL). Since recombinational diversity values did not follow a normal distribution, all tests performed were non-parametric.

Supplementary tables

Table S1. Regions of the X chromosome selected for the study. Start and end positions are based on NCBI's build 36 of the Human Genome.

Chromosomal region	Start (bp)	End (bp)	Length (bp)	Number of SNPs	Rec rate (cM/Mb)
region 1	22509816	22728031	218215	205	4.22
region 2	39100654	39237964	137310	129	1
region 3	93525304	94555531	1030227	382	0.82
region 4	140885581	141035312	149731	158	5.77
region 5	144772688	145266246	493558	376	2.24
SUM			2029041	1250	

Table S2. Information on the origin and number of samples per population and their assignment to a specific continental region.

Population Name	Acronym	Continental Group	Sampling Region	N males
Yoruba	YRI	Africa	Ibadan, Nigeria	53
Maasai	MKK	Africa	Kinyawa, Kenya	46
Luhya	LWK	Africa	Webuye, Kenya	46
Chadian (Laal and Sara)	CHA	Africa	Southern Chad	43
African American	ASW	Africa	Southwest USA	45
Moroccan	MOR	Middle East and North Africa	Assa-zag, Morocco	20
Egyptian	EGY	Middle East and North Africa	Egypt	46
Lebanese	LEB	Middle East and North Africa	Lebanon	42
Kuwaitis	KUW	Middle East and North Africa	Kuwait	43
Iranian	IRA	Middle East and North Africa	Kordestan, Iran	32
N. and W. European (CEPH)	CEU	Europe	Utah, USA	44
British	BRI	Europe	Great Britain, UK	32
Dutch	DUT	Europe	Netherlands	29
Basque	BAS	Europe	Guipuzcoa, Spain	45
Gypsies	GYP	Europe	La Mina, Sant Adrià del Besòs, Spain	24
Toscans	TSI	Europe	Toscana, Italy	46
Romanian	ROM	Europe	Romania	33
Chechen	CHE	Europe	Chechnya, Ingushetia and Dagestan, Russia	37
North Russian	RUS	Europe	Arkhangel, Kostroma and Pskov regions, Russia	42
Tatar (Kazan and Mishar)	TAT	Europe	Tatarstan, Russia	46
Altaian (Tubalar, Altai-Kizhi, Telengit,and Chelkans)	ALT	Central Eurasia	Gorniy Altay, Russia	30
Uigur	UIG	Central Eurasia	Xinjiam, China	45
Gujarati	GIH	Southern Asia	Houston, Texas, USA	46
Nadar	CAN	Southern Asia	Cape Comorin, Tamil Nadu, South India	47
Parayar	NTN	Southern Asia	Villupuram, Northern Tamil Nadu, South India	32
Kalita	KAL	Southern Asia	Guwahati, Assam, NE India	41
Adi	ADI	East Asia	Siang region, Arunachal Pradesh, NE India	31
Tibetan	TIB	East Asia	Tibetan from Tibet, China	47
Laotian	LAO	East Asia	Laos	43

Ati	ATI	East Asia	Phillipines	18
Han Chinese	CHB	East Asia	Beijing, China	33
Japanese	JPT	East Asia	Tokyo, Japan	33

Table S3. Information on the geographic coordinates (latitude and longitude) used to calculate the distance from East Africa either through Egypt (Northern route=EAN) or Iran (Southern route=EAS and distance from south India for the non-African populations (SI).

Population Name	**Location for coordinates**	**Lat**	**Lon**	**EAN (Km)**	**EAS (Km)**	**SI (Km)**
Yoruba	Ibadan, Nigeria	7.40	3.92	3758	3758	
Maasai	Loitokitok, Kenia	-2.91	37.52	1318	1318	
Luhya	Webuye, Kenia	0.62	34.77	994	994	
Chadian	N'Djamena, Chad	12.11	15.04	2536	2536	
Moroccan	Rabat, Morocco	34.02	-6.83	6018	9061	8945
Egyptian	Cairo, Egypt	30.06	31.25	2461	5515	5356
Lebanese	Beirut, Lebanon	33.89	35.50	3041	5212	5073
Kuwaitis	Kuwait City, Kuwait	29.37	47.98	4079	3871	3789
Iranian	Teheran, Iran	35.70	51.42	4445	4123	3860
N. and W. European (CEPH)	Paris, France	48.86	2.35	5877	8247	8079
British	London, UK	51.50	-0.13	6127	8431	8261
Dutch	Amsterdam, Netherlands	52.37	4.89	5844	8092	7921
Basque	Tolosa, Spain	43.14	-2.07	6173	8524	8432
Toscans	Florence, Italy	43.77	11.26	5096	7484	7356
Romanian	Cluj-Napoca, Romania	44.43	23.66	4098	6332	6211
Chechen	Makhachkala, Russia	47.15	45.67	5166	5008	4759
North Russian	Arkhangelsk, Russia	55.76	37.62	5471	6531	6054
Tatar	Kazan, Russia	55.70	49.11	5931	6239	5522
Altaian	Gorno-Altaysk, Russia	51.96	85.97	7511	6553	5134
Uigur	Ürümqi, Russia	43.83	87.62	7602	6187	4814
Gujarati	Ahmedabad, India	23.04	72.57	6627	4384	1431
Nadar	Nagercoil, India	8.18	77.43	7837	5660	474
Parayar	Villupuram, India	11.94	79.50	7814	5608	0
Kalita	Jorhat, India	26.76	94.21	8563	6408	2250
Adi	Siang Region,	28.24	94.07	8497	6396	4276

	Arunachal Pradesh, India					
Tibetan	Lhasa, China	29.65	91.14	8178	6094	4520
Laotian	Vientiane, Laos	17.96	102.61	9735	7537	3132
Ati	Panay, Phillipines	11.56	122.79	11985	9780	5400
Han Chinese	Beijing, China	39.90	116.41	10018	8366	4768
Japanese	Tokyo, Japan	35.69	139.69	12040	10513	6637

3.4. SNPs, haplotypes and recombination: a human variation study of the Old World

Asif Javed*[2,], **Marta Melé***[1,], Marc Pybus[1], Pierre Zalloua[3], Marc Haber[3], David Comas[1], Mihai G. Netea[4], Oleg Balanovsky[5, 6], Elena Balanovska[5], Li Jin[7], Yajun Yang [7], RM. Pitchappan[8,9], G. Arunkumar[9], Jaume Bertranpetit[1], Francesc Calafell[1], Laxmi Parida[2], and The Genographic Consortium[¶].

***** These two authors contributed equally to this work

[1] IBE, Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Catalonia, Spain. [2]Computational Biology Center, IBM T J Watson Research, Yorktown, USA. [3] Lebanese American University, School of Medicine, Beirut, Lebanon. [4]Department of Medicine and Nijmegen Institute for Infection, Inflammation, and Immunity, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands. [5] Research Centre for Medical Genetics, Moscow, Russia. [6] Vavilov Institute for General Genetics, Moscow, Russia. [7] MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China. [8] Chettinad Academy of Research & Education, Chettinad Health City, Rajiv Gandhi Salai, Kelampakkam, Chennai, India. [9] School of Biological Sciences, Madurai Kamaraj University, Madurai, 625021 India.

In preparation

The following section contains only the part of this work that deals with SNPs and haplotypes which was mainly done by **Marta Melé**. The part dealing with recombination is not yet finished since part of the work is currently being developed at IBM by Asif Javed and Laxmi Parida.

SNPs and haplotypes: a human variation study of the Old World

Introduction

In the genome, genetic variation is organized as haplotypes, which are, by definition, the combination of allelic states at neighboring polymorphisms. With the exception of mtDNA and the non-recombining portion of the Y chromosome, haplotypes are the expression of the action of both mutation and recombination. If polymorphisms are previously ascertained and subsequently genotyped, as is often the case with SNPs, the threshold imposed to allele frequencies implies that the age of these polymorphisms will be biased toward older values. On the contrary, the action of recombination will be detectable only after those polymorphisms have appeared, and more recent recombination events will have a greater impact on genetic variation. Thus, the analysis of haplotypes widens the time window that can be explored from the extant genetic variation.

The vast majority of studies of worldwide human genetic variation have been based on the study of uncorrelated SNPs (Auton et al. 2009; Li et al. 2008; Rosenberg et al. 2002; Xing et al. 2010). An exception is found in Conrad et al. (2006), where several uncorrelated SNPs are used to study the haplotype structure of several world wide populations. Lohmueller et al. (2009) and Jakobsson et al. (2008) used haplotypes to make inferences on demography. Specifically, Jakobsson et al. (2008) addressed the question of whether haplotypes can be used as a genetic marker to study human genetic variation at a global scale.

In the sequencing era, more and more data will be available, with a density of SNPs high enough to perform such kind of complementary analysis. One of the difficulties associated with using haplotypes as genetic markers, however, is to determine how many SNPs define a haplotype. The longer the haplotypes the larger the diversity but, beyond a certain point, all individuals carry

different haplotypes, and no genetic information can be extracted at the population level. Different studies have reached different solutions: either one or more fixed arbitrary lengths are used (Auton et al. 2009, Xing et al. 2010, among others) or the variation at each SNP is locally summarized by computing a probability of belonging to any of a fixed number of haplotype clusters (Jakobsson et al. 2008).

In this work, we revisit the question of whether haplotypes are more informative than SNPs and develop a method to define the best haplotype length for our study. For that, we analyze the human genetic variation in 33 populations of the Old World collected within the Genographic project. A total of five gene-free regions of the X chromosome spanning 2Mb were genotyped. SNPs were genotyped at high density, and independently of the underlying linkage disequilibrium (LD) structure. Samples were mostly from males in order to minimize the errors introduced in haplotype reconstruction (phasing). The X chromosome conferred an additional advantage: its effective population size is three-quarters of that of the autosomes and therefore demographic processes will leave a slightly deeper record on it.

Our results show first, that haplotypes are indeed more informative than SNPs for the study human genetic variation. Second, we provide a robust methodology to define haplotype length in order to obtain the maximum informativity out of them. Finally, we anticipate that this method could be extrapolated to other studies providing a way to extract more information than previously envisaged.

Results

Populations, SNPs and haplotypes

A total of 1255 SNP of 5 regions of the X chromosome spanning 2 Mb (Table 1) were analyzed in 1318 individuals from 33 different human populations (Figure 1, Table 2). In order to study the genetic

structure of the populations we not only used SNPs but we incorporated a haplotype-based approach. Haplotype length (*L*, in number of SNPs) was determined as that that maximized a length-specific index of informativeness (Figure 2). This index balances the fact that as *L* increases, the number of different haplotypes that appear increases as well but the number of sequences that each haplotype harbours will decrease. Ideally, for population structure analysis, we would like different haplotypes to be present in the dataset but not reaching the extreme in which all individuals are different from each other. For some analysis, the haplotype estimation procedure used by Jakobsson et al. (2008) was also calculated.

Inferring the genetic structure based on SNPs and haplotypes

In order to assess the fraction of the variation that could be explained within and between populations and continental regions, we performed an AMOVA analysis. For SNP data, differences among groups were 9.40%, among populations within groups, 1.78%, and, finally, within groups, 88.81%. For haplotypes, differences among groups explain 4.52% of the variation, differences among populations within groups 1.58 % and, finally, differences within groups explain 93.89%.

We performed two types of analyses: Principal Component Analysis (PCA) and a Bayesian Clustering Analysis, and assessed whether SNPs or haplotypes performed better at recovering the population structure of our dataset. In the PCA results, population areas overlapped less in the haplotype-based analysis than in the SNP based-analysis (Figures 3a and 3b). In the clustering analysis (Figure 4), we evaluated which method better classified individuals into specific clusters, by calculating the average Shannon's Diversity index (H) for each individual at each K. H grows with the information needed to classify individuals into specific populations, and a method with lower H provides sharper classifications. For all K from 2 to 4, the cluster memberships obtained with haplotypes gave lower mean H values (Table 3). The haplotype estimation procedure used by Jakobsson et al. (2008) classified individuals

better than SNPs but worse than our optimal fixed-length method, as assessed with H (Table 3).

Genetic structure of human populations

We next describe the results obtained with the haplotypes analysis. As expected, the first principal component separates African from non-African populations, leaving some of the North African and Middle Eastern populations leaning towards Africa. The second component separates East Asian from European and Middle Eastern populations leaving Indian, Mexicans and two of the Central Eurasian populations between those two groups.

The populations with the highest areas (and thus, with the highest internal diversities) are all the African populations and the Mexicans (Figure 3a and 3b). In particular, African Americans show the largest area, something that could be explained by their admixed origin. Chadian, Luhyan, and Yoruban individuals overlap only with other Sub-Saharan African populations, while the Maasai are closer to Middle Eastern, North African, and European populations. Conversely, Egyptians and Moroccans in North Africa (as well as Kuwaitis) have lower values of the first PC, which places them closer to Sub-Saharan Africans. In contrast, Lebanese and Iranians are indistinguishable in this plot from Europeans. Small, widely overlapping areas mark the European populations, which can be interpreted as a hugely homogeneous set.

East Asians including the Adi, Indians, and Europeans are clearly separated among them especially when looking at their centroids. Interestingly, the Gypsies are the Europeans that are closest to the Indian samples and the Gujarati from NW India appear between the rest of the Indian populations and European samples. Moreover, Tatars are almost indistinguishable from Europeans whereas the other Central Eurasian populations (Uygurs and Altaians) appear between the Indians and East Asians. Finally, the Mexicans have a very large area that overlaps with Europeans, Indians, and East Asians possibly explained by their admixed origin.

Regarding the clustering analysis (Figure 4), K=2 clearly separates African and non-African populations. At K=3 three groups separate: Sub-Saharan Africans, Europeans and populations from the Middle East and North Africa, and East Asians. Indians, Mexicans and two of the Central Eurasian populations appear as admixed between East Asians and Europeans. Egypt and Morocco show a greater Sub-Saharan African contribution than any other of the Middle Eastern populations. Finally, At K=4, the Indian populations have a specific cluster and are separated from the rest, something that did not happen in the SNP-based analysis. At K=5, the new component is restricted to European, Middle Eastern, and North African populations with an apparently random distribution (data not shown).

Table 4 shows the fraction of ancestry for each population at K=4. The four clusters than can be assigned to four continental regions in which they are predominant: an African Associated Ancestral Cluster (AAC) present in 93.5% on average in Sub-Saharan African populations, a West Eurasian AAC with 91.3% in European populations on average, an Indian AAC (68.6%) and an East Asian AAC (89.7%).

The African AAC is present outside Sub-Saharan Africa populations in Morocco, Egypt, and Kuwait, and rare elsewhere. The West Eurasian AAC reaches its highest value in the Basques (95.5%), and makes an unexpected appearance in the East African Maasai (11.4%). The three Central Eurasian populations had complementary levels of West Eurasian and East Asian AAC; the former was predominant in the Tatar (79.6%), and the latter in the Uighur (62.1%) and Altaian (62.1%). The Indian AAC was more frequent in the two Southern Indian populations, namely the Parayar (78%) and Cape Nadar (80.3%), while in the Gujarati the West Eurasian AAC reached 33.7%, and in the Kalita it was the East Asian component that was somewhat elevated (23.4%). The Tibeto-Burman speaking Adi, although sampled in NE India, clustered closely with East Asia (91.6% East Asian AAC). The Indian AAC is present in non-negligible frequencies in the SE Asian Lao, and, more notably, in the Ati, an isolated Philippino population, part of a group of peoples of low stature and dark pigmentation called

Negritos. Some known cases of mixed ancestry that are confirmed in the cluster analysis. The West Eurasian AAC is present at low frequencies in African Americans; Spanish Gypsies appear mostly as West Eurasian (63.2%) but have a significant amount of Indian AAC (33.9 %). Finally, Mexicans contain contributions from the East Asian (55.5%) and West Eurasian (28.2%) AAC components.

Discussion

Haplotypes versus SNPs

The present work provides evidence that haplotypes are significantly more informative than SNPs in the analysis of the genetic structure of populations. Both in the PC analysis and in Bayesian clustering population substructure appears defined in higher resolution when using haplotypes as genetic markers than with SNPs. Interestingly, however, the fraction of variance explained is lower, which may be just a reflection that haplotype frequencies are more constrained in their range than allele frequencies, and rare haplotypes may drive F_{ST} down.

The main difference between haplotypes and SNPs is given by the fact that haplotypes incorporate information on the recombinational history of the sample. Recombination is one of the main forces shaping the genome but its information is not used when performing analysis based solely on uncorrelated SNPs. By taking the haplotype information we are able some how to extract information on both mutation and recombinational history of the samples analyzed.

The field of population genetics is now moving from genotyping to complete sequencing projects such as the 1000 Genome Project (The 1000 Genomes Project Consortium 2010) and a method that can deal with such high SNP density data will be necessary in order to study the genetic variation of human populations at a finer level. Our method to find the optimal haplotype length turned out to allow a more informative analysis than the one used in Jakobsson et al.

(2008). Besides, only when using our SNP length definition, a new cluster appeared at K=4 in the STRUCTURE analysis. We have provided a method that is able to define haplotypes to obtain the maximum information in terms of population structure and we show that it performs better than previously defined methods. How much more information will our method provide compared to using SNPs as independent markers in different studies remains to be further studied. However, we believe that this new definition should be taken into account for future approaches.

Population structure

This study has been performed with a large number of populations within the Old World and it contains one of the largest surveys of human genetic variation. Although our data set overlaps in geographic coverage with other sets such as HGDP, it contains particular features such as the representation of India as well as singular populations such as the Gypsies and the Ati.

Our results are, first of all, consistent with the Out of Africa hypothesis since African populations are the most differentiated and most internally diverse from other populations in both analyses. Within Africa, both the Maasai and the African American seem to have some West Eurasian or Middle Eastern component. In African American this could be explained by their known recent admixture, which may be slightly underestimated by the X chromosome: a male-mediated European admixture would result in a 1:2 ratio of European to African X chromosomes being transmitted. The West Eurasian component in the Maasai can either be explained by them being descendants of populations ancestral to non-Africans and / or gene flow from non-Africans into Africa.

In Europe, the most outstanding result is the clear demonstration of the Indian origin and West Eurasian admixture of Gypsies, which had been shown before using unilinearly transmitted markers (see Mendizabal et al. (2011) and references therein). In the Central Asian continuum of genetic variation, Tatars showed the smallest

East Asian contribution, which was higher in the more easterly located Uighur and Altai.

In the Bayesian clustering analysis, a component that was predominant in Indian populations was revealed only when our optimal fixed-length haplotypes were used. Our data set contained two populations from Southern India, where this component reached its higher frequencies. Thus, it is possible that it captures a predominantly S. Indian dimension of genetic variation. This would explain why this component appears somewhat diluted in the NW Indian, Indo-European speaking Gujarati, as well as in the NE Indian Kalita, with Western and Eastern genetic contributions, respectively. The fact the the Indian AAC appears in the Lao of SE Asia and in the Ati Negritos of the Philippines may imply that this AAC captured some of the contribution of the southern route out of Africa (Melé et al. submitted). Still, the Ati were clearly linked to East Asian populations, as was shown also with unilinear markers (Delfin et al. 2011; Gunnarsdóttir et al. 2011).

The admixed nature of the Mexican general population was also revealed; the lack of Native American reference samples could explain why the predominant component was East Asian; sex-biased gene flow may have led to an overestimate of this component. The diversity in individual histories, with various degrees of Native American vs. European ancestry is apparent both in the Bayesian cluster results and in the wide area occupied by Mexicans in the PC graph.

In this manuscript, we have shown that a particular haplotype approximation can allow extracting a large amount of genetic information from genomic data; while we have analyzed only a small part of the genome, we have been able to recover many of the patterns seen with much larger datasets.

Materials and methods

DNA samples

DNA samples were selected based on geographic distribution trying to sample equally from the different regions of the Old World (Figure 1 and Table 1). In order to avoid phasing errors, the genotyping was performed on the X chromosome and, the whenever possible, we selected male over female samples.

Samples from 23 worldwide populations were collected within the Genographic Project. Informed consent was obtained from all study subjects. Individuals from another six populations were obtained from the Coriell Cell Repository. Details on the origin of samples can be found in Table 1. Overall we sampled 1455 individuals, 1283 of which were males.

SNP selection

Five regions on the X chromosome that were at least 50Kb distant from known genes, copy number variants and segmental duplication were selected. These conditions were meant to avoid selection, genotyping errors, and to ensure sufficient precision to detect recombination. These 5 regions correspond to some of the regions studied in Melé et al (2010) in which the X chromosome was screened to find the optimal regions for a recombination based analysis (Table 2). SNPs were selected based on the HapMap phase II release 24 as to obtain the highest possible density. For the 5 selected regions, all SNPs appearing in the HapMap database were selected provided they meet some technical genotyping conditions (Illumina designability rank higher than 0.5, and a minimum distance of 60 bp between SNPs). We also downloaded genotypes of four HapMap phase II samples (www.hapmap.org).

SNP genotyping, quality control, phasing and imputing

Genotyping was performed using the Illumina GoldenGate custom Oligos array of 1536 SNPs. After the genotyping process, SNPs with more than 15 % of missing data were removed as well as those having a cluster of heterozygous positions in male samples (80 SNPs). Those samples with missing data higher than 10% (123

samples), or male samples with more than 3 heterozygous positions were removed (14 samples) (heterozygous positions in male samples with 1 or 2 heterozygous positions were recoded as missing and imputed, see below). Monomorphic SNPs were removed from the analysis (201 SNPs). The final dataset consisted of 1255 SNPs genotyped in 1318 samples (1269 were males) belonging to 33 worldwide populations (Table 1). None of our 22 internal replicate samples showed inconsistencies. Missing values were imputed using fastPHASE (Scheet and Stephens 2006) and the female samples were phased using PHASE (Stephens and Scheet 2005; Stephens et al. 2001), using the very completed haplotypes given by males. Thus the amount of inferred information is extremely low.

Haplotype definition

Haplotype length was defined as the length L in number of SNPs that gave the highest Informativeness. Informativeness was calculated for lengths 5, 10, 20, 30, 40, 60, 80, 100, and 120 SNPs (Figure 2) with the equation:

$$I_L = \frac{1}{ncol} \sum_{i=1}^{ncol} h_i \sum_{i=1}^{ncol} \frac{N}{h_i}$$

where *ncol* is the number of columns obtained when dividing each sequence into windows of *L* SNPs, h_i is the number of different haplotypes found in each column *i* and *N* is the number of sequences.

For each length L, 5 different informativeness values were calculated by changing the starting position of the first window from position 1 and increasing it in L/5 steps. Then all 5 values were averaged. The highest average informativeness value laid between lengths 30 and 40 SNPs and therefore we performed the same calculation for lengths 30, 32, 34, 36, 38 and 40 taking as starting positions all even positions. The haplotype length having the highest average informativity was 38 SNPs.

Haplotypes were also defined as in Jakobsson et al. (2008) by considering 20 haplotype clusters at each position and using fastPHASE (Scheet and Stephens 2006) to assign to each individual, at each SNP, a probability of haplotype cluster membership for each of the 20 possible haplotype clusters.

PCA analysis

PCA analysis was performed using EIGENSOFT software (Patterson et al. 2006). In order to control for the presence of correlated SNPs, LD correction was turned on. Outlier removal parameter was turned off. R package was used to draw the plots and to calculate the convex hull polygon for each population for the first two principal components (R Development CoreTeam 2009).

Structure analysis

The Bayesian clustering software STRUCTURE (Pritchard et al. 2000) was used to group individuals based on SNPs or on haplotypes. All runs used a burn-in period of 50,000 iterations followed by 50,000 iterations from which estimates were obtained. All runs were based on the admixture model in which each individual is assumed to have ancestry in multiple genetic clusters and using the F model of correlation in allele frequencies across clusters. The software Distruct (Rosenberg 2004) was used to create the images.

To select uncorrelated SNP for the analysis we used Haploview (Barrett et al. 2005) and took tag SNPs with a r^2 value lower than 0.8. We performed five replicas and considered the run wit the highest likelihood.

For the haplotype analysis, we took those five configurations of length 38 SNPs with different starting points that had the highest informativeness values; ran STRUCTURE on them as if it were multiallelic data and considered the run with the highest likelihood.

To perform the population structure analysis as in Jackobson et al. (2008), 10 different structure datasets were created based on the haplotype cluster membership probabilities for each individual and each SNP. We assigned, according to the corresponding probabilities, a specific haplotype to each individual at each position and this randomized process was performed 10 times to create 10 datasets. Each of the datasets was analyzed by the Structure software as if they were multiallelic markers and using the exact same parameters explained above and we considered the replica with the highest likelihood.

Shannon's diversity index was calculated for each individual and each K as follows:

$$H = -\sum_{i=1}^{K} p_i \ln p_i$$

where p_i is the probability of that individual (cluster membership) to belong to the ith cluster. Note the highest value of H will be achieved if all individuals belonged with equal probability to all populations, which would be a case with no structure whereas the highest value would be achieved when all individuals were assigned to one cluster with probability equal to 1.

References

Auton, A., K. Bryc, A.R. Boyko, K.E. Lohmueller, J. Novembre, A. Reynolds, A. Indap, M.H. Wright, J.D. Degenhardt, R.N. Gutenkunst, K.S. King, M.R. Nelson, and C.D. Bustamante. 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Research* 19: 795-803.

Barrett, J.C., B. Fry, J. Maller, and M.J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263-265.

Conrad, D.F., M. Jakobsson, G. Coop, X. Wen, J.D. Wall, N.A. Rosenberg, and J.K. Pritchard. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 38: 1251-1260.

Delfin, F., J.M. Salvador, G.C. Calacal, H.B. Perdigon, K.A. Tabbada, L.P. Villamor, S.C. Halos, E. Gunnarsdottir, S. Myles, D.A. Hughes, S. Xu, L. Jin, O. Lao, M. Kayser, M.E. Hurles, M. Stoneking, and M.C.A. De Ungria. 2011. The Y-chromosome landscape of the Philippines: extensive heterogeneity and varying genetic affinities of Negrito and non-Negrito groups. *Eur J Hum Genet* 19: 224-230.

Gunnarsdóttir, E.D., M. Li, M. Bauchet, K. Finstermeier, and M. Stoneking. 2011. High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Research* 21: 1-11.

Jakobsson, M., S.W. Scholz, P. Scheet, J.R. Gibbs, J.M. VanLiere, H.-C. Fung, Z.A. Szpiech, J.H. Degnan, K. Wang, R. Guerreiro, J.M. Bras, J.C. Schymick, D.G. Hernandez, B.J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H.M. Cann, J.A. Hardy, N.A. Rosenberg, and A.B. Singleton. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998-1003.

Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. 2008. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* 319: 1100-1104.

Lohmueller, K.E., C.D. Bustamante, and A.G. Clark. 2009. Methods for Human Demographic Inference Using Haplotype Patterns From Genome-wide SNP Data. *Genetics*: genetics.108.099275.

Melé, M., A. Javed, M. Pybus, F. Calafell, L. Parida, J. Bertranpetit, and C. The Genographic. 2010. A New Method to Reconstruct Recombination Events at a Genomic Scale. *PLoS Comput Biol* 6: e1001010.

Mendizabal, I., C. Valente, A. GusmÃ£o, C. Alves, V.n. Gomes, A. Goios, W. Parson, F. Calafell, L. Alvarez, A.n. Amorim, L. GusmÃ£o, D. Comas, and M.J.o. Prata. 2011. Reconstructing the Indian Origin and Dispersal of the European Roma: A Maternal Genetic Perspective. *PLoS ONE* 6: e15988.

Patterson, N., A.L. Price, and D. Reich. 2006. Population Structure and Eigenanalysis. *PLoS Genet* 2: e190.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155: 945-959.

R Development CoreTeam. 2009. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria.

Rosenberg, N.A. 2004. DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* 4: 137-138.

Rosenberg, N.A., J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. 2002. Genetic Structure of Human Populations. *Science* 298: 2381-2385.

Scheet, P. and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* 78: 629-644.

Stephens, M. and P. Scheet. 2005. Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. *American journal of human genetics* 76: 449-462.

Stephens, M., N.J. Smith, and P. Donnelly. 2001. A New Statistical Method for Haplotype Reconstruction from Population Data. *The American Journal of Human Genetics* 68: 978-989.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061-1073.

Xing, J., W.S. Watkins, Y. Hu, C. Huff, A. Sabo, D. Muzny, M. Bamshad, R. Gibbs, L. Jorde, and F. Yu. 2010. Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biology* 11: R113.

Figures and tables

Table 1. Information on the genotyped regions of the X chromosome. Positions are calculated based on Build 36. Quality control was for lvel of missing genotype and heterozygous in males. Final SNPs are the QC-passed SNPs that were polymorphic.

	Start (bp)	End (bp)	Length (bp)	Initial SNPs	QC-passed SNPs	Final SNPs
region 1	22509816	22728031	218215	250	234	206
region 2	39100654	39237964	137310	165	150	129
region 3	93525304	94555531	1030227	503	468	385
region 4	140885581	141035312	149731	197	170	158
region 5	144772688	145266246	493558	421	434	377
SUM			2029041	1536	1456	1255

Table 2. Information on the samples: name, acronym, continental group, population details (ancestry, sampling location, ethnicity), number of successfully genotyped males and females, final sample size.

Population Name	Pop Name	Continental group	sampling region	N males	N females	total N
Yoruba	YRI	Sub-saharan Africa	Ibadan, Nigeria	53	0	53
Maasai	MKK	Sub-saharan Africa	Kinyawa, Kenya	46	0	46
Luhya	LWK	Sub-saharan Africa	Webuye, Kenya	46	0	46
Chadian (Laal and Sara)	CHA	Sub-saharan Africa	Southern Chad	43	0	43
African American	ASW	Sub-saharan Africa	Southwest USA	45	0	45
Lebanese	LEB	Middle East and North Africa	Lebanon	42	0	42
Kuwaitis	KUW	Middle East and North Africa	Kuwait	43	0	43
Iranian	IRA	Middle East and North Africa	Egypt	32	0	32
Egyptian	EGY	Middle East and North Africa	Kordestan, Iran	46	0	46
Moroccan	MOR	Middle East and North Africa	Assa-zag, Morocco	20	0	20
N. and W. European	CEU	Europe	Utah, USA	44	1	46
British	BRI	Europe	Great Britain, UK	32	13	58
Dutch	DUT	Europe	Netherlands	29	0	29
Basque	BAS	Europe	Guipuzcoa, Spain	45	0	45
Gypsies	GYP	Europe	La Mina, Sant Adrià del Besòs, Spain	24	11	46
Toscans	TSI	Europe	Toscana, Italy	46	0	46
Romanian	ROM	Europe	Cluj-Napoca, Romania	33	0	33
Chechen	CHE	Europe	Chechnya, Ingushetia and Dagestan, Russia	38	0	38
Russian	RUS	Europe	Arkhangel, Kostroma and Pskov regions, Russia	42	0	42
Tatar	TAT	Central Eurasia	Tatarstan, Russia	46	0	46
Altaian	ALT	Central Eurasia	Gorniy Altay,Russia	30	0	30
Uigur	UIG	Central Eurasia	Xinjiam, China	45	0	45
Gujarati	GIH	Southern Asia	Houston, Texas, USA	46	0	46
Nadar	CAN	Southern Asia	Cape Comorin, Tamil Nadu, India	47	0	47
Parayar	NTN	Southern Asia	Villupuram, Northern Tamil Nadu, India	32	0	32
Kalita	KAL	Southern Asia	Guwahati, Assam, India	41	0	41
Adi	ADI	Southern Asia	Siang region, Arunachal Pradesh, India	32	0	32
Tibetan	TIB	East Asia	Tibetan from Tibet, China	47	0	47

Laotian	LAO	East Asia	Laos	44	0	44
Ati	ATI	East Asia	Phillipines	19	0	19
Han Chinese	CHB	East Asia	Beijing, China	22	12	46
Japanese	JPT	East Asia	Tokyo, Japan	23	12	47
Mexican	MEX	America	Los Angeles, California, USA	46	0	46

Table 3. Average individual Shannon diversity index based on cluster membership for each K. Note that the index decreases if individuals tend to be assigned to a single cluster.

K	haplotypes	snps	jackobsson
2	0.07	0.42	0.32
3	0.25	0.43	0.43
4	0.35	0.68	0.72

Table 4. Cluster memberships in populations at K=4 using haplotypes as genetic markers for the clustering analysis.

	cluster 1	cluster 2	cluster 3	cluster 5
YRI	98.8	0.5	0.3	0.3
MKK	85.5	11.4	2.4	0.7
LWK	98.2	0.8	0.5	0.5
CHA	97.9	0.9	0.8	0.4
ASW	87.1	7.7	2.7	2.5
LEB	3.3	84.9	7.3	4.4
KUW	12	78.6	6.8	2.6
IRA	1	86.4	8.6	4
EGY	15.9	76.8	3.9	3.4
MOR	27.6	63.8	7.2	1.4
CEU	1.4	90.5	4.1	4
BRI	0.5	93.1	5.3	1.1
DUT	0.5	95.3	3.4	0.7
BAS	0.6	95.5	2.5	1.4
GYP	0.6	63.2	33.9	2.3
TSI	0.8	93.7	3.6	1.9
ROM	0.7	88.7	7.4	3.2
CHE	0.7	85.1	8.1	6.1
RUS	0.6	88.1	6.2	5.1
TAT	0.5	79.6	6.9	13
ALT	0.5	26.4	10.9	62.1
UIG	1.3	33.6	11.8	53.2
GIH	0.9	33.7	60.3	5.1
CAN	0.9	9.7	80.3	9.1
NTN	1	8.7	78	12.2
KAL	0.8	19.8	55.9	23.4
ADI	0.7	1.7	6.1	91.6
TIB	0.4	2.1	3.9	93.6
LAO	0.7	0.8	11.5	87.1
ATI	0.7	2.6	16.6	80.1
CHB	1	2.1	4.2	92.7
JPT	0.9	2.1	1.8	95.2
MEX	4.7	28.2	11.6	55.5

Figure 1. Geographic distribution of the sampled populations

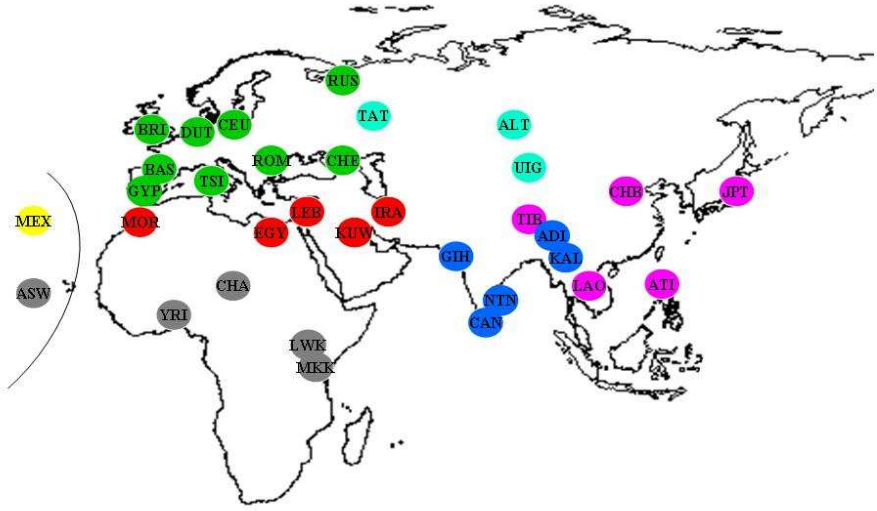


Figure 2 Informativeness versus haplotype length (L). Informativeness is defined as the product between average number of haplotypes across all windows and average number of sequences per haplotype across all windows. Standard deviations are calculated by changing the starting position in which haplotypes of length L start to be defined.

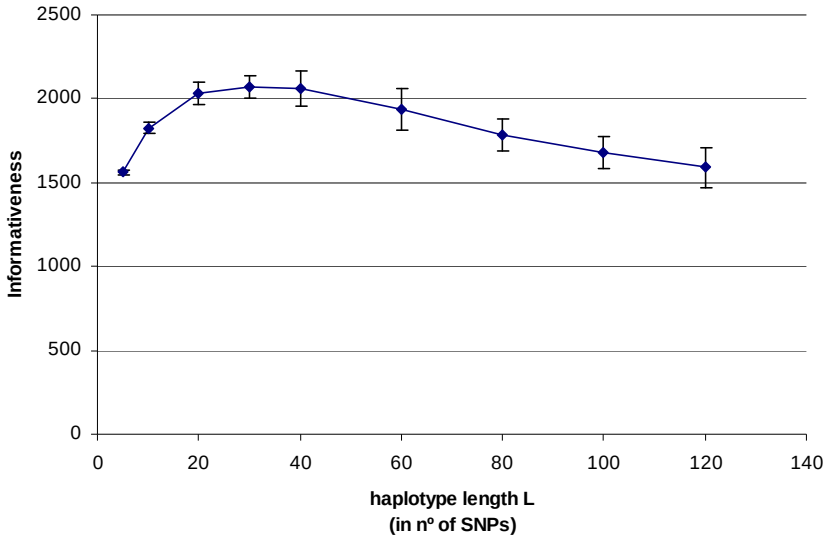


Figure 3a. Haplotype-based PCA

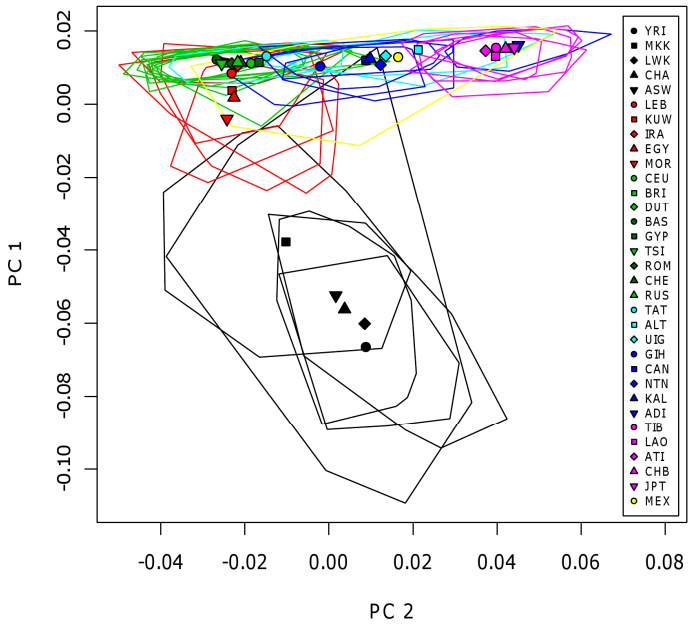


Figure 3b. SNP-based PCA analysis

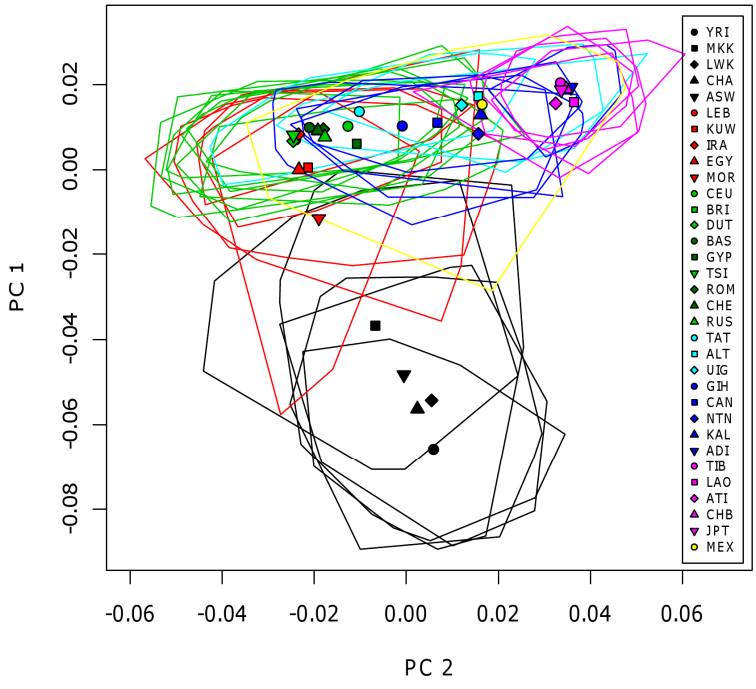


Figure 4a. haplotype-based STRUCTURE analysis.

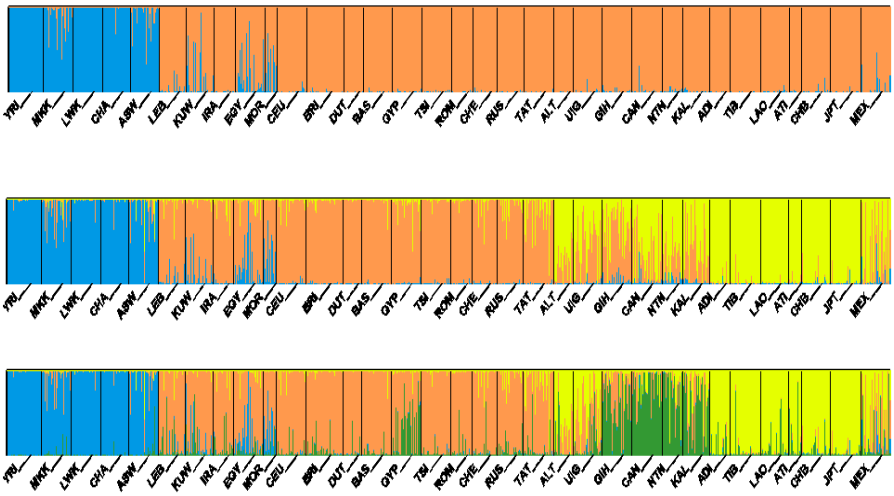


Figure 4b SNP - based STRUCTURE analysis

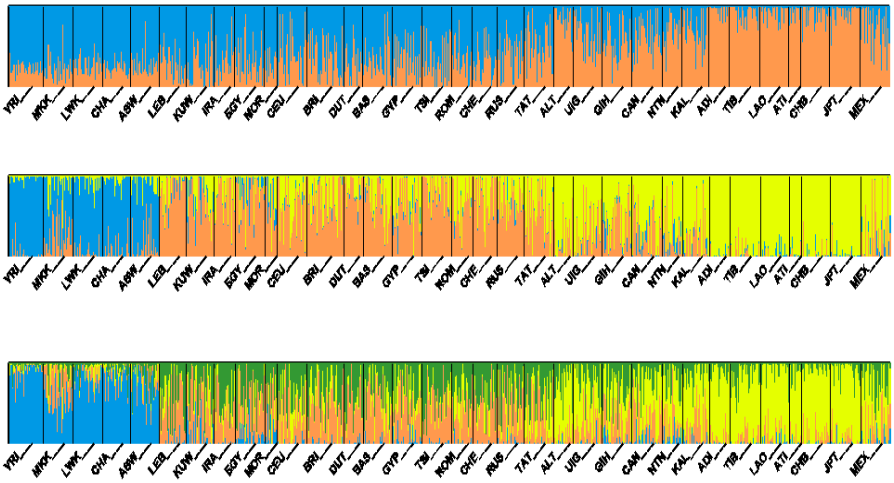
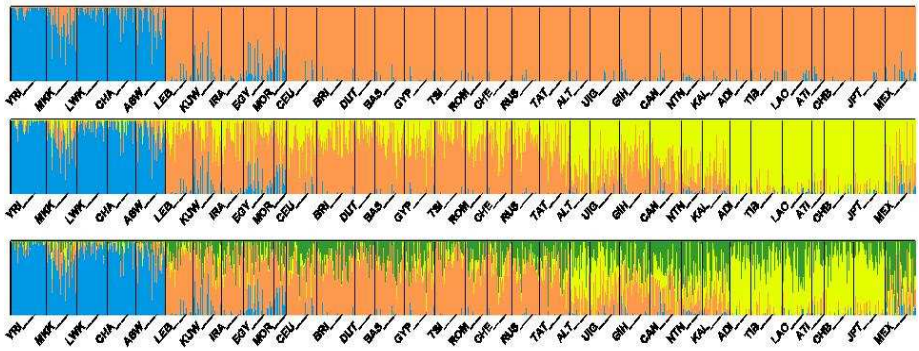


Figure 4c Structure analysis with haplotypes defined as in Jackobson et al. (2008)



4. DISCUSSION

4.1. Challenges of the study of recombination

Recombination is a very difficult process to detect, to model and to analyze; and this is the reason why recombinant regions of the genome have historically been eschewed in phylogeographic studies.

First of all, a network structure is needed to represent the phylogenetic relationships between sequences. This network or Ancestral Recombination Graph (ARG) is very difficult to infer, basically, because many possible networks would most likely produce the observed data and no information can be extracted to discern which one is best.

Secondly, whereas mutation generally leaves a visible footprint in the genome, recombination often draws on a palimpsest: recent recombination events overwrite past events and therefore the footprints of recombination quickly disappear from the genetic record. Therefore, those recombination events that can confidently be detected tend to be relatively recent and consequently, have a low frequency in the population. Moreover, recombination rates are highly variable along the genome, being very high in hotspots and low in the rest of the genome. This implies that, in hotspots, the signal of the past recombinations may be erased faster than in coldspots.

Third, the ancestral recombining sequences need to be different between them in order to leave a footprint; otherwise, recombination will be absolutely invisible. Consequently, a considerable amount of recombination events are invisible no matter which method we use to detect them.

Several frequent genetic events such as gene conversion and recurrent mutation, or technical issues such as genotyping errors or phasing errors, may lead to signals that will closely mimic those of recombination. This adds an additional burden to the study of

recombination since we need to tease apart those signals and separate them from the real recombination events.

In this thesis, we have faced all these problems, and more importantly, we have addressed them all. No doubt that recombination is a different process than mutation but we have shown that it can also be used as a genetic marker. In fact, these differences are the ones worth looking at in order to extract the complementary information hidden in our recombinant genomes.

4.2. Inferring the Ancestral Recombination Graph?

As mentioned in the previous section, inferring the true ARG from a set of extant sequences is a very complex problem. In Figure 17, an ARG inferred using coalescent theory for three human populations is shown and it may give an idea of the complexity that those networks reach.

Our approach in Parida et al. (2008), however, is aimed at constructing an ARG (or network) compatible with the data in order to detect true recombination events. In Melé et al. (2010), when performing the different runs of the algorithm with different starting positions, different window sizes and different directions, we do infer several different plausible ARGs compatible with the data and then we extract the recombination events that consistently show up on them. In this way, we use the ARG as a tool to detect recombinations. Further, we do infer small pieces of the true ARG (the recombinations), although we are not able to infer all of it.

Interestingly, in each of the ARGs generated by the basic algorithm, not only the recombinant sequences are inferred, but the sequences that carry the ancestral patterns are also pinpointed. Therefore, in a similar way as recombination events are detected by counting the number of detections, the corresponding sequences that carry the ancestral patterns could be extracted as well. This opens the door to

inferring some more pieces of this “unreachable” true ARG and the history of its recombinations.

4.3. Sensitivity, false discovery rate, accuracy in placing the breakpoint

In humans, and in most organisms, the recombination rate among adjacent basepairs is on the same order of magnitude as the mutation rate per basepair, something that makes the detection of recombination events harder. The basic algorithm (Parida et al. 2008) performed extremely well to analyze data in which mutation is much faster than recombination (see simulation section in Parida et al. (2008)) but needed to be optimized when run with simulations that mimicked human sequences.

To provide robust inferred recombination events, we ran the basic algorithm several times in a sliding window approach using different window sizes and directions. Our optimization parameters (Melé et al. 2010) were false discovery rate, sensitivity and accuracy in placing the breakpoint, which very often behaved in opposite directions. For example, false discovery rate increased with window sizes but accuracy in placing the breakpoint increased as well. Moreover, false discovery rate was given double weight compared to sensitivity and accuracy in placing the breakpoint. This seemed reasonable at the time of developing the method since first, all the applications we could think of were going to be based in the detected recombinations (and not in our not detected recombinations) and, secondly, we could study in depth which were those recombinations that were missed (see the following section). The optimal method used window sizes of 20, 10 and 5 and a threshold of 60%.

However, other optimization parameters could have been used and, in fact, other combinations of parameters performed very well and could even be more suitable in some circumstances. For example, to capture shared recombinations (recombinations that are shared between at least two individuals), it would be better to turn the

mergepats parameter on, whereas to have increased sensitivity, methods with lower threshold worked better. The accuracy in placing the breakpoint is good when the pattern sizes are small but the false discovery rate increases. In the supplementary figures of Melé et al. (2010), it can be seen how each of the different methods works depending on false discovery rate, sensitivity and accuracy with the idea in mind that different users could have different needs.

4.4. Which recombinations are detected?

It may seem at first sight that sensitivity estimates for IRiS are low, in Melé et al. (2010) the average estimate of sensitivity with the optimal method was around 20% and in Melé et al. (submitted), the estimated sensitivity for 5 specific chromosome X regions was estimated to be 7.3 %. In figure 17 it can be seen that most of the recombinations in the genealogy are not detected (grey) compared to those detected (red dots).

However, in any genealogy, many recombination events will be undetectable. First, the footprint of old recombinations is overwritten by newer recombinations and, as time goes by, the signals are blurred. Secondly, if the ancestral sequences that recombined were identical, the recombination would leave no footprint at all.

The specific task of IRiS is not only to detect presence or absence of recombination, but also to trace the whole history of each event. First, the two ancestral patterns that recombined in the past need to be present in the extant dataset for IRiS to detect the recombination event. This leads to low sensitivity estimates, and specifically, it leads to estimates which are biased towards recent recombination events (Figure 17).

When evaluated over recent recombination, IRiS sensitivity rises to 45% (Figures 5 and 7 in Melé et al. (2010). Still, we could ask which recombination events are missed. In Melé et al. (2010), we tried to answer this question by showing that for a recent

recombination event to be detected, the two parental sequences need to be different and, in fact, the higher the number of differences, the higher the probability of being detected (Figure 7 in Melé (2010)).

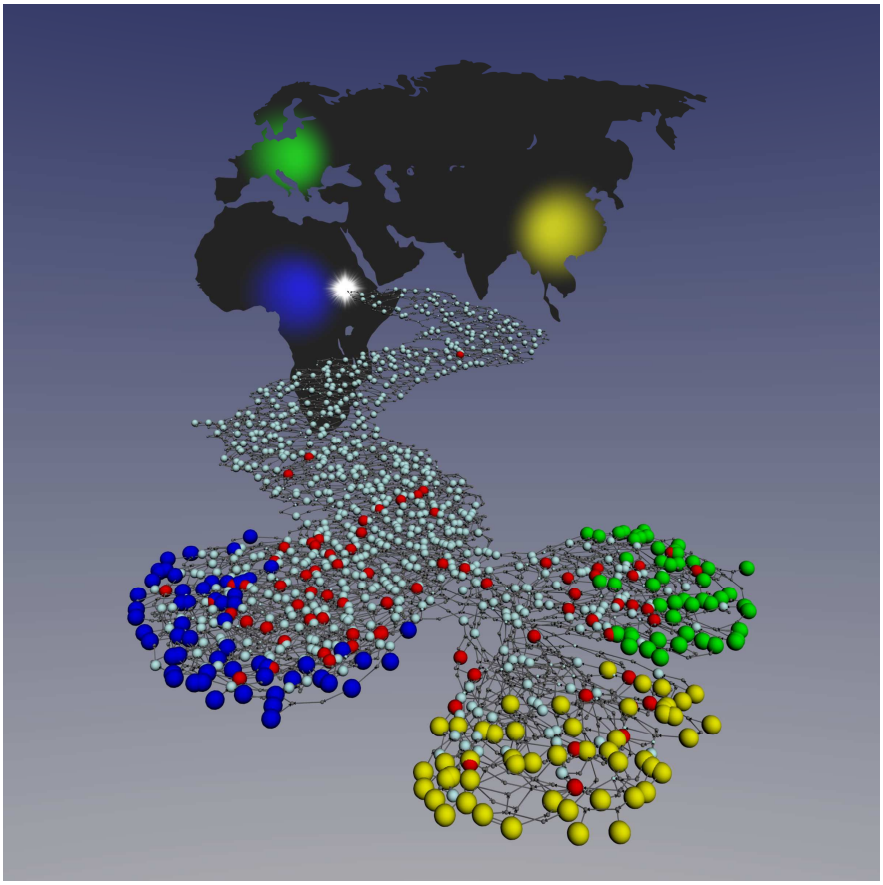


Figure 17. This picture shows a simulated coalescent network generated with *cosi* (Shaffner *et al*. 2005) that represents the history of humankind with an Out of Africa event and the emergence of the different human groups, Africans, Europeans and Asians. African individuals appear in blue, Asian individuals in yellow and Europeans in green. The cyan nodes represent past recombination events whereas the small gray nodes represent coalescent events. Red nodes are those past recombination events that were recovered by IRiS. These kind of coalescent networks were used to fine-tune and validate IRiS. Software used Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). Figure by Marc Pybus.

Moreover, a strong effect of the recombination rate can be observed on sensitivity (Figure 7 in Melé et al. (2010)). First, we showed that, within a hotspot, sensitivity clearly decreases, and that the number of recombinations present in a dataset is negatively correlated with the sensitivity of the method. In fact, IRiS saturates when a certain number of recombinations are present although methods that aim at estimating recombination rates such as LDhat behave similarly (Melé et al. 2010).

Sensitivity may also depend on the populations analyzed: the higher the diversity in a population the higher the sensitivity. This effect may be specifically strong in admixed populations because the signal left by recombinations between chromosomes of different populations may be clearer. In Melé et al. (submitted) explaining the observed higher recombinational diversity in all Indian populations due to admixture remains a possibility.

Finally, sensitivity will depend on the allele frequency and density of the selected SNPs with higher sensitivity if selecting SNPs with high allele frequency and at high density (Table S1 in Melé et al (2010)).

In Melé et al. (submitted), the sensitivity of the method for 5 regions of the X chromosome was inferred accurately. This allowed estimating the absolute number of recombinations that had occurred in the whole genealogy of the studied sequences, something that had not been done before.

The differences between the sensitivity estimates (20% vs 7.3%) of the two studies Melé et al. (2010) and Melé et al. (submitted) can be reconciled taking into account that the first estimate is an average value over thousands of simulations and the second one refers specifically to the sensitivity of the method when run on 5 regions of the X chromosome. These regions had higher recombination rates than average, the SNPs selected had allele frequencies not always being higher than 0.1, and finally, in the second study the number of chromosomes was the double than in the first.

4.5. Gene conversion, recurrent mutation, genotyping errors, and phasing errors

All these processes may, in some circumstances, mimic the footprint left by recombination (Figure 18). This issue was specifically addressed in Melé et al. (2010) when quantifying how often these events were confounded with recombination (Table S2) and was discussed there as well.

One of the most interesting findings was the mergepats parameter, which had a small effect on sensitivity and false discovery rate but showed to be very useful to increase IRiS robustness to such confounding factors. However, the accuracy in placing the breakpoint appears more affected if this parameter is turned on.

These results were specifically useful for the last paper (Melé et al. submitted). Since for the kind of analysis we wanted to perform, the breakpoint location was not necessary, we decided to turn on the mergepats parameter to make sure that our results were robust to these confounding factors.

Phasing errors may be the most dangerous since they behave exactly as recent recombination events; the only difference is that the complementary recombination event has to appear in the homologous chromosome. Although IRiS could potentially detect most of them, special care should be taken when dealing with unphased data. First, it is highly important to use an accurate phasing method such as PHASE, rather than fastPHASE (Table S3 in Melé et al. (2010)) because every gain phasing accurately is a singinficant improvement in false discovery rate. Secondly, IRiS may detect 30% of the phasing errors as such but a post-processing of the output is needed. This implies, first, that some recombinations could be removed from the sample even though they may be true and therefore, sensitivity will decrease. Therefore, in our analysis of the 30 Old World populations, we directly avoided phasing errors by selecting only males for the study.

However, the possible use of IRiS to specifically detect phasing errors should be taken into account for possible future applications.

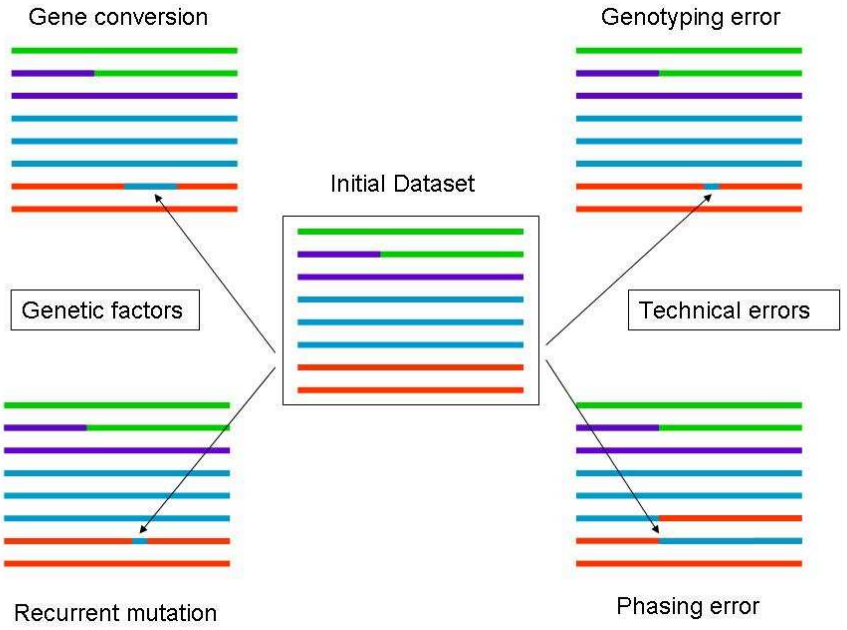


Figure 18. Different processes that may mimic the footprint of recombination.

4.6. Recombination in human population genetics

In order to introduce recombination in the study of human population genetics it may be necessary to think about what is new about using such a genetic marker. Again, we can claim that detected recombinations are recent, and this will allow us to restrict our studies in a very specific window time: right after the Out of Africa migration (Figure 17 for an example).

In the study by Melé et al. (submitted), the number of recombinations detected in a population is used as a proxy to infer the effective population size. The absolute values obtained are low

but this is justified by the fact that our diversity values represent the harmonic mean of the sizes for recent times, when it is known that anatomically modern humans went through a strong bottleneck. LD based measures of effective population size always give lower estimates than those based on heterozygosity again because LD estimates measure more recent times. In the case of the X chromosome, a sex bias may be present (Emery et al. 2010; Keinan and Reich 2010).

Further, the patterns of recombinational diversity of 30 populations of the Old World show significant differences among them. As expected, recombinational diversity was found to be much higher in African populations, consistent with their higher recent long term effective population size and older origin. In fact, diversity values will be a function of two factors: previous population sizes and the age of the population under study. Recombinational patterns being similar for East Asian and European populations point towards a similar time of settlement into their respective continental regions, something that is in agreement with the fossil record.

The significantly higher recombinational diversity in the Indian populations compared to Europeans and East Asian seems to suggest that the Indian subcontinent was settled very early. In fact, the most recent studies on the peopling of Eurasia by anatomically modern humans point towards a single migration event which took them fairly rapidly across southern and southeastern Asia with only a secondary and later dispersal into Europe (Mellars 2006). One of the most striking patterns regarding recombinational diversity is the significant correlation found between recombinational diversity and distance from Southern India, which raises the question on whether South Asia could have been a main source of Eurasian variation although there are other plausible explanations. The study clearly places Indian populations as having had a more central role in human population history than previously reported.

Moreover, patterns of recombinational diversity seem to be less affected by SNP ascertainment bias than other diversity measures such as nucleotide diversity: recombinational diversity in African

populations is much higher than in non-Africans, whereas, when looking at the patterns of nucleotide diversity of the exact same dataset, this pattern is not so clear. The same observation could be seen in Melé et al. (2010).

Young variants tend to be in low frequency in a population and, generally, the most informative variants are those at high frequency. Population genetics is generally based on the study of allele frequencies in a population. However, most of the recombinations detected by IRiS are singletons. From this observation, it follows that a very high number of sequences and SNPs at high density will be needed in order to use recombination as a marker of shared ancestry. We are however at the right time since this kind of data is currently being generated with projects such as The HapMap and the 1000 Genomes.

The intrinsic nature of recombination, which continuously overwrites previous events, makes tracing specific old recombination events nearly impossible, no matter how large our sample size will be. However, in order to use the complete recombinational history, we can always use haplotypes as genetic markers, since they provide more information than SNPs alone and carry information on both old and recent recombination events.

4.7. IRiS applied to other organisms

The study of recombination is not only restricted to humans, and it has been the focus of interest for several other species. In viruses for example, much effort has been devoted to study the specific mechanisms that underlie recombination, because this process is strongly related with their virulence and diseases caused by viruses.

Bacteria, on the other hand, do not undergo meiotic recombination and they do not have sexual reproduction. However, they have different mechanism to exchange DNA sequences either with other organisms (conjugation) or with the environment (transformation)

and therefore, they can incorporate foreign DNA to their own genetic material.

In eukaryotes, strong variation on recombination rates has been found. Wilfert et al (2007) collected information on recombination rates of several eukaryotic organisms and found that the highest recombination rates were found in fungi and protozoa whereas animals and plants had lower recombination rates in general (Figure 19). However, although yeasts are fungi, they have extremely low recombination rates (Zeyl and Otto 2007).

Wilfert et al. (2007) also detected exceptionally high recombination rates in social Hymenoptera compared to other higher eukaryotes and they hypothesize that it is the strong selection pressure in social insects that causes it. Another interesting observation is that insects with sex-restricted recombination such as *Drosophila* (in which only females recombine) show increased recombination rates compared to those where both sexes recombine.

Finally, recombination has been a focus of much study in mammals, especially in mice and humans. In mammals, the vast majority of meiotic recombination events are localized to hotspots and recently, the PRDM9 gene has been described as one of the major regulator of the distribution of recombination (Paigen and Petkov 2010).

Overall, recombination rates vary strongly in different species and organisms. Our method, however, has been calibrated to be run in human samples. In principle, any species that shows similar diversity patterns as humans can be analyzed by IRiS.

Humans have some specific characteristic that may differentiate them from other organisms. First of all, humans are much less diverse than other species. For example, they are more than three times genetically less diverse than chimpanzees, their closest relatives (Jobling et al. 2004). This would imply that, for a region of the same length and recombination rate, much more information would be present in chimpanzee sequences than in human

sequences. On the other hand, all human populations have experienced a recent expansion (Jobling et al. 2004) which affects the patterns of genetic diversity observed in the data. Specifically, an expansion would increase the length of the LD tracks, which makes the signal of recombination much clearer. Finally, the mutation and recombination rates in humans are, on average, of the same order of magnitude (10^{-8} per generation per locus).

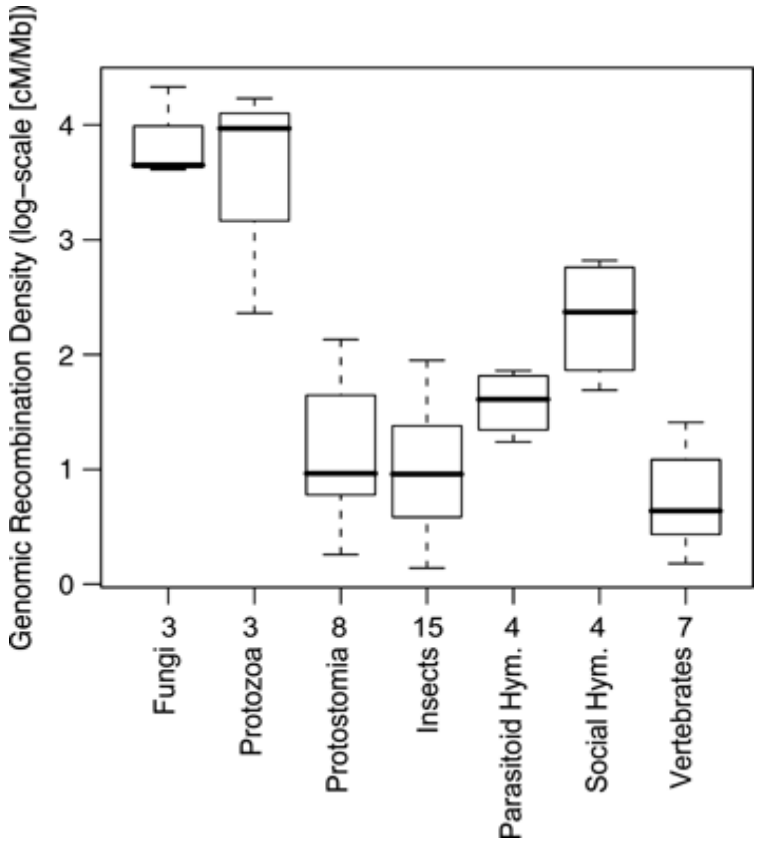


Fig 19. Average recombination densities (cM/Mb) across different large taxonomic groups. The social Hymenoptera stand out among the eukaryotes whereas the protozoa have the highest recorded values in all. The number of species per group (N) is indicated at the bottom. The horizontal line marks the median value, boxes indicate one quartile and vertical lines indicate the range of observations. Figure taken from Wilfert et. al 2007.

It is difficult to predict whether IRiS will be suitable for the analysis of other organisms rather than humans, since recombination occurs in most organisms and the dynamics and demographic history of each of them may differ strongly. In general terms, the performance of any method to detect recombination will be highly dependent on how strong the signal of recombination is in the sequences. IRiS is a pattern-based method and therefore, the correct detection of the recombinations in any organism will depend on the presence of such patterns in the sequences and how strong these patterns can show the footprint of recombination.

4.8. IRiS from SNPs to sequences and the genomic scale

IRiS was fine-tuned to be run in SNP data. Both studies in which IRiS was applied to study the recombinational diversity of human populations (Melé et al (2010) and Melé et al. (submitted)) spanned several megabases. In general terms, to analyze a moderate number of sequences and SNPs, IRiS can be perfectly run in a desktop computer and provide results in a short time.

Nowadays, however, more and more complete genomes are sequenced. SNP ascertainment biases are doomed to extinction and the huge amount of data produced by projects such as The 1000 Genomes Project demands new analysis tools. In light of this, IRiS will have to be adapted. First, it will need to be able to analyze data in which low variants dominate and this should not be difficult to do if using a similar pipeline as that used in Melé et al. 2010. Second, it will need to be scaled up in order to analyze complete genomes.

The limiting factors for IRiS in terms of speed and RAM are the number of sequences to be analyzed, the number of polymorphic sites (the length of those sequences) and the complexity of the region (the number of recombinations). Two changes would improve IRiS speed: first, moving from perl to C++ (perl is more generic but a little slower) and two, maintaining information mostly in RAM memory. Both are currently being implemented.

Being able to analyze complete genomes opens the door to a whole set of possible different applications for IRiS such as detecting positive selection or studying the dynamics of recombinations. In the following sections, these options will be explored further.

4.9. Recombination and selection

The role that positive selection has had in shaping our genomes has attracted notable attention. One of the possible future applications of IRiS could be to detect the footprint of positive selection because LD, recombination, and selection are tightly related. Under Wright Fisher equilibrium, the amount of linkage disequilibrium present between two genetic markers being in the same chromosome is proportional to the recombination rate between them. When a new variant arises, however, it does so in a specific haplotype and this variant will initially be completely linked with its neighbors until recombinations separates them until the equilibrium is reached.

Positive selection can change this pattern. When a new and favorable variant arises, it will increase much more rapidly in frequency in the population taking the linked markers with it (which are said to *be hitchhiked* because recombination will not have time to break their association). This will be translated into a fast increase of the frequency of the whole haplotype in the population, and a reduction of the diversity present in this specific area. The selected haplotype will show an increase in LD (or a lack of recombination) whereas the other haplotypes will not show such a signal. Trying to find specific regions of the genome in which a specific haplotype has a lower number of recombinations around it than expected may be an interesting approach to detect the signals of positive selection.

Specifically, in order to detect putative regions under positive selection using IRiS, the following could be done. First, IRiS should be run genome wide, and recotypes should be extracted as well as all the breakpoint positions. Then, recotype and haplotype information should be overlapped by analyzing the data in a sliding

window approach. Each window (of, for example, 1Mb size) should be centered in a combination of a small number of SNPs forming specific haplotypes. Then, the number of recombinations detected in those haplotypes looking at the whole region (delimited by the window size) should be counted. If the number of recombinations present in the sequences carrying the selected haplotype is much lower than the number of recombinations detected in other haplotypes, potentially, this particular haplotype may have undergone the effect of positive selection.

Tests specifically aimed at detecting this kind of signal have already been designed but they are based on the study of LD patterns. Our approach is based on specific counts of recombination and therefore, we may be able to detect different signals of positive selection than the other methods existing so far.

4.10. Recombinations detected by IRiS, recombination rates and the evolution of hotspots

The mechanisms that underlie the evolution of recombination are not fully understood. Recent research seems to point towards recombination being a very fast evolving system in which hotspots appear and disappear at a higher rate than sequence evolution (Hochwagen and Marais 2010). In light of this, it may be interesting to study the patterns of recombination genomewide using a method such as IRiS that is biased towards recent events. LD-based methods such as LDhat (McVean et al. 2004) or PHASE (Crawford et al. 2004; Li and Stephens 2003) are able to detect hotspots that have been active for a while and have broken linkage disequilibrium. IRiS, on the other hand, although it does detect ancient hotspots, will specially detect young hotspots which show an accumulation of recent recombinations which have left a clear footprint on the data (without necessarily having broken completely LD).

Looking at which features of the genome such as GC content, gene density and others, correlate with recombination rates has already been done, both at the megabase scale and at the kilobase scale (Crawford et al. 2004; Myers et al. 2005; Myers et al. 2006). However, it may be very interesting to study these patterns by only looking at the recent recombinational landscape at the genome wide scale.

Moreover, it has been shown that recombination rates between populations and individuals within populations are different. This seems to be directly related to the different PRDM9 variants that those individuals or populations may harbor. By looking at specific recombinational events detected by IRiS in several different population from which the frequency of PRDM9 alleles are known, we could have a better understanding of the role this gene has played in shaping the recombinational landscape at the population level.

4.11. Concluding remarks

The aim of this work was to use recombination to study human genetic variation by first developing a method to detect recombination events, and second, use it to study the recombinational patterns of several human populations. The question is therefore whether these objectives have been achieved.

First of all, we have developed a method named IRiS that can extensively detect recombination events in a set of extant sequences. One of the largest challenges of this project was to adjust our theoretical approach to the detection of recombinations on real human sequences. We initially developed an algorithm aimed at detecting recombination events but later a huge effort had to be devoted to largely evaluate and validate it before it could be applied to real sequences. A natural consequence of detecting recombination events was that only the very recent ones could accurately be extracted. Although this was expected, the bias that our detection method had towards recent events was higher than

expected. This questioned our initial expectations because we thought that recombinations used as genetic markers could potentially be more informative than SNPs. We have seen now that our recombinational analysis is not better than that based on SNPs but it is indeed complementary.

Next, we performed an extensive genotyping of more than one thousand SNPs in order to study the recombinational patterns of several human populations of the Old World. More importantly, we were able to make inferences based on the differential recombinational patterns between populations. Further, the number of recombinations could be used as a proxy of the recent effective population size of human populations within such a recent time frame that no one had looked at before. Specifically, our results stressed the higher diversity in Indian populations raising the question of whether India could have played a major role in the Out of Africa expansion of the Anatomically Modern Humans.

Finally, we explored further the use of the information left by recombination by taking haplotypes as genetic markers. This is an indirect way to incorporate recombination into the study of human populations because haplotypes include in their structure all the visible recombinational history of the sequences and allow us to go further in time than using specific events. We demonstrate the potential use of incorporating haplotypes into future analysis of human genetic variation.

In summary, we have provided some insights to incorporate recombination into the study of population genetics and even population genomics by first developing a method to detect recombinations and second, by extracting relevant information from both specific recombination events and haplotypes in an extended study of human populations.

MAIN REFERENCES FOR THE INTRODUCTION

These references have been used as a main guide for the introduction:

Campbell, M.C. and S.A. Tishkoff. 2010. The Evolution of Human Genetic and Phenotypic Variation in Africa. *Current Biology : CB* **20:** R166-R173.

Coop, G. and M. Przeworski. 2007. An evolutionary view of human recombination. *Nat Rev Genet* **8:** 23-34.

Garrigan, D. and M.F. Hammer. 2006. Reconstructing human origins in the genomic era. *Nature Reviews Genetics* **7:** 669-680.

Handley, L.J.L., A. Manica, J. Goudet, and F. Balloux. 2007. Going the distance: human population genetics in a clinal world. *Trends in Genetics : TIG* **23:** 432-439.

Hey, J. and C.A. Machado. 2003. The study of structured populations - new hope for a difficult and divided science. *Nat Rev Genet* **4:** 535-543.

Hochwagen, A. and Gabriel A.B. Marais. 2010. Meiosis: A PRDM9 Guide to the Hotspots of Recombination. *Current Biology : CB* **20:** R271-R274.

Jobling, M.A., M.E. Hurles, and C. Tyler-Smith. 2004. *Human Evolutionary Genetics*. Garland Science.

Paigen, K. and P. Petkov. 2010. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet* **11:** 221-233.

Pool, J.E., I. Hellmann, J.D. Jensen, and R. Nielsen. 2010. Population genetic inference from genomic sequence variation. *Genome Research* **20:** 291-300.

Relethford, J.H. 2008. Genetic evidence and the modern human origins debate. *Heredity* **100:** 555-563.

Rosenberg, N.A. and M. Nordborg. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* **3:** 380-390.

Stumpf, M.P.H. and G.A.T. McVean. 2003. Estimating recombination rates from population-genetic data. *Nat Rev Genet* **4:** 959-968

REFERENCES

Avise, J.C., J. Arnold, R.M. Ball, E. Bermingham, T. Lamb, J.E. Neigel, C.A. Reeb, and N.C. Saunders. 1987. Intraspecific Phylogeography - the Mitochondrial-DNA Bridge between Population-Genetics and Systematics. *Annual Review of Ecology and Systematics* **18:** 489-522.

Ballard, J.W.O. and M.C. Whitlock. 2004. The incomplete natural history of mitochondria. *Molecular Ecology* **13:** 729-744.

Balloux, F. 2009. The worm in the fruit of the mitochondrial DNA tree. *Heredity* **104:** 419-420.

Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. 2009. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327:** 836-840.

Beaumont, M.A. 2004. Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity* **92:** 365-379.

Behar, D.M., R. Villems, H. Soodyall, J. Blue-Smith, L. Pereira, E. Metspalu, R. Scozzari, H. Makkan, S. Tzur, D. Comas, J. Bertranpetit, L. Quintana-Murci, C. Tyler-Smith, R.S. Wells, and S. Rosset. 2008. The Dawn of Human Matrilineal Diversity. *The American Journal of Human Genetics* **82:** 1130-1140.

Berg, I.L., R. Neumann, K.-W.G. Lam, S. Sarbajna, L. Odenthal-Hesse, C.A. May, and A.J. Jeffreys. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet* **42:** 859-863.

Bosch, E., F. Calafell, D. Comas, P.J. Oefner, P.A. Underhill, and J. Bertranpetit. 2001. High-Resolution Analysis of Human Y-Chromosome Variation Shows a Sharp Discontinuity and Limited Gene Flow between Northwestern Africa and the Iberian Peninsula. *American journal of human genetics* **68:** 1019-1029.

Boulton, A., R.S. Myers, and R.J. Redfield. 1997. The hotspot conversion paradox and the evolution of meiotic recombination. *Proceedings of the National Academy of Sciences of the United States of America* **94:** 8058-8063.

Brakez, Z., E. Bosch, H. Izaabel, O. Akhayat, D. Comas, J. Bertranpetit, and F. Calafell. 2001. Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. *Annals of Human Biology* **28:** 295-307.

Broman, K.W., J.C. Murray, V.C. Sheffield, R.L. White, and J.L. Weber. 1998. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *American Journal of Human Genetics* **63:** 861-869.

Bustamante, C.D. and S. Ramachandran. 2009. Evaluating signatures of sex-specific processes in the human genome. *Nat Genet* **41:** 8-10.

Cann, R.L., M. Stoneking, and A.C. Wilson. 1987. Mitochondrial DNA and human evolution. *Nature* **325:** 31-36.

Cavalli-Sforza, L.L. and M.W. Feldman. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat Genet*.

Conrad, D.F., M. Jakobsson, G. Coop, X. Wen, J.D. Wall, N.A. Rosenberg, and J.K. Pritchard. 2006. A worldwide survey of

haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38:** 1251-1260.

Coop, G. and M. Przeworski. 2007. An evolutionary view of human recombination. *Nat Rev Genet* **8:** 23-34.

Coop, G., X. Wen, C. Ober, J.K. Pritchard, and M. Przeworski. 2008. High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. *Science* **319:** 1395-1398.

Cox, M., A. Woerner, J. Wall, and M. Hammer. 2008. Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. *BMC Genetics* **9:** 76.

Crawford, D.C., T. Bhangale, N. Li, G. Hellenthal, M.J. Rieder, D.A. Nickerson, and M. Stephens. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet* **36:** 700-706.

Davison, D., J.K. Pritchard, and G. Coop. 2009. An approximate likelihood for genetic data under a model with recombination and population splitting. *Theoretical Population Biology* **75:** 331-345.

DeGiorgio, M., M. Jakobsson, and N.A. Rosenberg. 2009. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proceedings of the National Academy of Sciences* **106:** 16057-16062.

Emery, L.S., J. Felsenstein, and J.M. Akey. 2010. Estimators of the Human Effective Sex Ratio Detect Sex Biases on Different Timescales. *American journal of human genetics* **87:** 848-856.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164:** 1567-1587.

Fisher, R.A. 1930. *The Genetical Theory of Natural Selection* Oxford Univesity Press.

Fisher, R.A. 1954. A fuller theory of "Junctions" in inbreeding. *Heredity* **8:** 187-197.

Forster, P. and S. Matsumura. 2005. Did Early Humans Go North or South? *Science* **308:** 965-966.

Fu, Y.X. 1997. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. *Genetics* **147:** 915-925.

Garrigan, D., S.B. Kingan, M. Metni Pilkington, J.A. Wilder, M.P. Cox, H. Soodyall, B. Strassmann, G. Destro-Bisol, P. de Knijff, and A. Novelletto. 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. *Genetics* **177:** 2195 - 2207.

Griffiths, R.C. and P. Marjoram. 1996. Ancestral Inference from Samples of DNA Sequences with Recombination. *Journal of Computational Biology* **3:** 479-502.

Gusfield, D., V. Bansal, V. Bafna, and Y.S. Song. 2007. A decomposition theory for phylogenetic networks and incompatible characters. *Journal of Computational Biology* **14:** 1247-1272.

Gusmão, L., P. Sánchez-Diz, F. Calafell, P. Martín, C.A. Alonso, F. Álvarez-Fernández, C. Alves, L. Borjas-Fajardo, W.R. Bozzo, M.L. Bravo, J.J. Builes, J. Capilla, M. Carvalho, C. Castillo, C.I. Catanesi, D. Corach, A.M. Di Lonardo, R. Espinheira, E. Fagundes

de Carvalho, M.J. Farfán, H.P. Figueiredo, I. Gomes, M.M. Lojo, M. Marino, M.F. Pinheiro, M.L. Pontes, V. Prieto, E. Ramos-Luis, J.A. Riancho, A.C. Souza Góes, O.A. Santapa, D.R. Sumita, G. Vallejo, L. Vidal Rioja, M.C. Vide, C.I. Vieira da Silva, M.R. Whittle, W. Zabala, M.T. Zarrabeitia, A. Alonso, A. Carracedo, and A. Amorim. 2005. Mutation rates at Y chromosome specific microsatellites. *Human Mutation* **26:** 520-528.

Gutenkunst, R.N., R.D. Hernandez, S.H. Williamson, and C.D. Bustamante. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet* **5:** e1000695.

Hammer, M.F., F.L. Mendez, M.P. Cox, A.E. Woerner, and J.D. Wall. 2008. Sex-Biased Evolutionary Forces Shape Genomic Patterns of Human Diversity. *PLoS Genet* **4:** e1000202.

Hartl, D.L. and A.G. Clark. 1997. *Principles in Population Genetics*. SinauerAssociates Inc.

Hayes, B.J., P.M. Visscher, H.C. McPartlan, and M.E. Goddard. 2003. Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. *Genome Research* **13:** 635-643.

Hellenthal, G., A. Auton, and D. Falush. 2008. Inferring Human Colonization History Using a Copying Model. *PLoS Genet* **4:** e1000078.

Hellenthal, G. and M. Stephens. 2006. Insights into recombination from population genetic variation. *Current Opinion in Genetics & Development* **16:** 565-572.

Hill, W.G. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8:** 117-126.

Hochwagen, A. and Gabriel A.B. Marais. 2010. Meiosis: A PRDM9 Guide to the Hotspots of Recombination. *Current biology : CB* **20:** R271-R274.

Holder, M. and P.O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet* **4:** 275-284.

Hudson, R.R. 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23:** 183-201.

Hudson, R.R. 1987. Estimating the Recombination Parameter of a Finite Population-Model without Selection. *Genetical Research* **50:** 245-250.

Hudson, R.R. and N.L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111:** 147 - 164.

Hudson, R.R. and N.L. Kaplan. 1988. The Coalescent Process in Models With Selection and Recombination. *Genetics* **120:** 831-840.

Huson, D.H. and D. Bryant. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* **23:** 254-267.

Jakobsson, M., S.W. Scholz, P. Scheet, J.R. Gibbs, J.M. VanLiere, H.-C. Fung, Z.A. Szpiech, J.H. Degnan, K. Wang, R. Guerreiro, J.M. Bras, J.C. Schymick, D.G. Hernandez, B.J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H.M. Cann, J.A. Hardy, N.A. Rosenberg, and A.B.

Singleton. 2008a. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451:** 998-1003.

Jakobsson, M., S.W. Scholz, P. Scheet, J.R. Gibbs, J.M. VanLiere, H.C. Fung, Z.A. Szpiech, J.H. Degnan, K. Wang, and R. Guerreiro. 2008b. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451:** 998 - 1003.

Jeffreys, A.J. and R. Neumann. 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* **31:** 267-271.

Jeffreys, A.J. and R. Neumann. 2005. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Human Molecular Genetics* **14:** 2277-2287.

Jobling, M.A., M.E. Hurles, and C. Tyler-Smith. 2004. *Human Evolutionary Genetics*. Garland Science.

Keinan, A., J.C. Mullikin, N. Patterson, and D. Reich. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat Genet* **41:** 66-70.

Keinan, A. and D. Reich. 2010. Can a Sex-Biased Human Demography Account for the Reduced Effective Population Size of Chromosome X in Non-Africans? *Molecular Biology and Evolution* **27:** 2312-2321.

Kim, H.L., T. Igawa, A. Kawashima, Y. Satta, and N. Takahata. 2010. Divergence, demography and gene loss along the human lineage. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365:** 2451-2457.

Kimura, M. and G.H. Weiss. 1964. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics* **49:** 561-576.

Kingman, J.F.C. 1982. The coalescent. *Stochastic Processes and their Applications* **13:** 235-248.

Kong, A., D.F. Gudbjartsson, J. Sainz, G.M. Jonsdottir, S.A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S.T. Palsson, M.L. Frigge, T.E. Thorgeirsson, J.R. Gulcher, and K. Stefansson. 2002. A high-resolution recombination map of the human genome. *Nature Genetics* **31:** 241-247.

Kong, A., G. Thorleifsson, D.F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G.B. Walters, A. Jonasdottir, A. Gylfason, K.T. Kristinsson, S.A. Gudjonsson, M.L. Frigge, A. Helgason, U. Thorsteinsdottir, and K. Stefansson. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467:** 1099-1103.

Kong, A., G. Thorleifsson, H. Stefansson, G. Masson, A. Helgason, D.F. Gudbjartsson, G.M. Jonsdottir, S.A. Gudjonsson, S. Sverrisson, T. Thorlacius, A. Jonasdottir, G.A. Hardarson, S.T. Palsson, M.L. Frigge, J.R. Gulcher, U. Thorsteinsdottir, and K. Stefansson. 2008. Sequence Variants in the RNF212 Gene Associate with Genome-Wide Recombination Rate. *Science* **319:** 1398-1401.

Krings, M., A.-e.H. Salem, K. Bauer, H. Geisert, A.K. Malek, L. Chaix, C. Simon, D. Welsby, A. Di Rienzo, G. Utermann, A. Sajantila, S. Pääbo, and M. Stoneking. 1999. mtDNA Analysis of Nile River Valley Populations: A Genetic Corridor or a Barrier to Migration? *American journal of human genetics* **64:** 1166-1176.

Lahr, M.M. and R. Foley. 1994. Multiple dispersals and modern human origins. *Evolutionary Anthropology: Issues, News, and Reviews* **3:** 48-60.

Lahr, M.M. and R.A. Foley. 1998. Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. *American journal of physical anthropology* **Suppl 27:** 137-176.

Lao, O., T.T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf, A. Caliebe, M. Balascakova, J. Bertranpetit, L.A. Bindoff, D. Comas, G. Holmlund, A. Kouvatsi, M. Macek, I. Mollet, W. Parson, J. Palo, R. Ploski, A. Sajantila, A. Tagliabracci, U. Gether, T. Werge, F. Rivadeneira, A. Hofman, A.G. Uitterlinden, C. Gieger, H.-E. Wichmann, A. Rüther, S. Schreiber, C. Becker, P. Nürnberg, M.R. Nelson, M. Krawczak, and M. Kayser. 2008. Correlation between Genetic and Geographic Structure in Europe. *Current biology : CB* **18:** 1241-1248.

Laval, G., E. Patin, L.B. Barreiro, and L. Quintana-Murci. 2010. Formulating a Historical and Demographic Model of Recent Human Evolution Based on Resequencing Data from Noncoding Regions. *PLoS ONE* **5:** e10284.

Lenzi, M.L., J. Smith, T. Snowden, M. Kim, R. Fishel, B.K. Poulos, and P.E. Cohen. 2005. Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis I in human oocytes. *American Journal of Human Genetics* **76:** 112-127.

Lewontin, R.C. and J. Krakauer. 1973. DISTRIBUTION OF GENE FREQUENCY AS A TEST OF THE THEORY OF THE SELECTIVE NEUTRALITY OF POLYMORPHISMS. *Genetics* **74:** 175-195.

Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-

Sforza, and R.M. Myers. 2008. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* **319:** 1100-1104.

Li, N. and M. Stephens. 2003. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165:** 2213-2233.

Liu, H., F. Prugnolle, A. Manica, and F. Balloux. 2006. A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *Am J Hum Genet* **79:** 230-237.

Lohmueller, K.E., C.D. Bustamante, and A.G. Clark. 2009. Methods for Human Demographic Inference Using Haplotype Patterns From Genome-wide SNP Data. *Genetics***:** genetics.108.099275.

Macaulay, V., C. Hill, A. Achilli, C. Rengo, D. Clarke, W. Meehan, J. Blackburn, O. Semino, R. Scozzari, F. Cruciani, A. Taha, N.K. Shaari, J.M. Raja, P. Ismail, Z. Zainuddin, W. Goodwin, D. Bulbeck, H.-J.r. Bandelt, S. Oppenheimer, A. Torroni, and M. Richards. 2005. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science* **308:** 1034-1036.

Makova, K.D. and W.-H. Li. 2002. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416:** 624-626.

Malecot, G. 1969. *The Mathematics of Heredity* Freeman, San Francisco.

Marjoram, P. and S. Tavare. 2006. Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet* **7:** 759-770.

Marth, G., G. Schuler, R. Yeh, R. Davenport, R. Agarwala, D. Church, S. Wheelan, J. Baker, M. Ward, M. Kholodov, L. Phan, E. Czabarka, J. Murvai, D. Cutler, S. Wooding, A. Rogers, A. Chakravarti, H.C. Harpending, P.-Y. Kwok, and S.T. Sherry. 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 376-381.

McDougall, I., F.H. Brown, and J.G. Fleagle. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433:** 733-736.

McVean, G.A.T. and N.J. Cardin. 2005. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360:** 1387-1393.

McVean, G.A.T., S.R. Myers, S. Hunt, P. Deloukas, D.R. Bentley, and P. Donnelly. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304:** 581-584.

Melé, M., A. Javed, M. Pybus, F. Calafell, L. Parida, J. Bertranpetit, and C. The Genographic. 2010. A New Method to Reconstruct Recombination Events at a Genomic Scale. *PLoS Comput Biol* **6:** e1001010.

Mellars, P. 2006. Going East: New Genetic and Archaeological Perspectives on the Modern Human Colonization of Eurasia. *Science* **313:** 796-800.

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310:** 321-324.

Myers, S., R. Bowden, A. Tumian, R.E. Bontrop, C. Freeman, T.S. Macfie, G. McVean, and P. Donnelly. 2009. Drive Against Hotspot

Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. *Science* **327:** 876-879.

Myers, S., C. Freeman, A. Auton, P. Donnelly, and G. McVean. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* **40:** 1124-1129.

Myers, S., C.C.A. Spencer, A. Auton, L. Bottolo, C. Freeman, P. Donnelly, and G. McVean. 2006. The distribution and causes of meiotic recombination in the human genome. *Biochem. Soc. Trans.* **34:** 526-530.

Myers, S.R. and R.C. Griffiths. 2003. Bounds on the Minimum Number of Recombination Events in a Sample History. *Genetics* **163:** 375-394.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annual Review of Genetics* **39:** 197-218.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante, and A.G. Clark. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet* **8:** 857-868.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, M. Stephens, and C.D. Bustamante. 2008. Genes mirror geography within Europe. *Nature* **456:** 98-101.

Novembre, J. and M. Stephens. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* **40:** 646-649.

Oliver, P.L., L. Goodstadt, J.J. Bayes, Z. Birtle, K.C. Roach, N. Phadnis, S.A. Beatson, G. Lunter, H.S. Malik, and C.P. Ponting.

2009. Accelerated Evolution of the *Prdm9* Speciation Gene across Diverse Metazoan Taxa. *PLoS Genet* **5:** e1000753.

Paigen, K. and P. Petkov. 2010. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet* **11:** 221-233.

Parida, L., A. Javed, M. Mele, F. Calafell, and J. Bertranpetit. 2009. Minimizing recombinations in consensus networks for phylogeographic studies. *BMC Bioinformatics* **10 Suppl 1:** S72.

Parida, L., M. Mele, F. Calafell, and J. Bertranpetit. 2008. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. *J Comput Biol* **15:** 1133-1154.

Parvanov, E.D., P.M. Petkov, and K. Paigen. 2009. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science* **327:** 835-.

Patterson, N., A.L. Price, and D. Reich. 2006. Population Structure and Eigenanalysis. *PLoS Genet* **2:** e190.

Plagnol, V. and J.D. Wall. 2006. Possible Ancestral Structure in Human Populations. *PLoS Genet* **2:** e105.

Posada, D. and K.A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc. Natl. Acad. Sci. U. S. A.* **98:** 13757-13762.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155:** 945-959.

Prugnolle, F., A. Manica, and F. Balloux. 2005a. Geography predicts neutral genetic diversity of human populations. *Current biology : CB* **15:** R159-R160.

Prugnolle, F., A. Manica, and F. Balloux. 2005b. Geography predicts neutral genetic diversity of human populations. *Curr Biol* **15:** R159 - 160.

Ptak, S.E., D.A. Hinds, K. Koehler, B. Nickel, N. Patil, D.G. Ballinger, M. Przeworski, K.A. Frazer, and S. Paabo. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet* **37:** 429-434.

Quintana-Murci, L., O. Semino, H.-J. Bandelt, G. Passarino, K. McElreavey, and A.S. Santachiara-Benerecetti. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet* **23:** 437-441.

Ramachandran, S., O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman, and L.L. Cavalli-Sforza. 2005a. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 15942-15947.

Ramachandran, S., O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman, and L.L. Cavalli-Sforza. 2005b. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* **102:** 15942 - 15947.

Reich, D.E., M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, S.F. Farhadian, R. Ward, and E.S. Lander. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199-204.

Relethford, J.H. 2004. Global patterns of isolation by distance based on genetic and morphological data. *Human Biology* **76:** 499-513.

Relethford, J.H. and L.B. Jorde. 1999. Genetic evidence for larger African population size during recent human evolution. *American Journal of Physical Anthropology* **108:** 251-260.

Roach, J.C., G. Glusman, A.F.A. Smit, C.D. Huff, R. Hubley, P.T. Shannon, L. Rowen, K.P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L.B. Jorde, L. Hood, and D.J. Galas. 2010. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328:** 636-639.

Rosenberg, N.A., J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. 2002. Genetic Structure of Human Populations. *Science* **298:** 2381-2385.

Sabeti, P.C., D.E. Reich, J.M. Higgins, H.Z.P. Levine, D.J. Richter, S.F. Schaffner, S.B. Gabriel, J.V. Platko, N.J. Patterson, G.J. McDonald, H.C. Ackerman, S.J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E.S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832-837.

Schaffner, S.F., C. Foo, S. Gabriel, D. Reich, M.J. Daly, and D. Altshuler. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15:** 1576-1583.

Schierup, M.H. and J. Hein. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156:** 879-891.

Sikora, M., H. Laayouni, F. Calafell, D. Comas, and J. Bertranpetit. 2010. A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. *Eur J Hum Genet* **19:** 84-88.

Slatkin, M. 1977. Gene Flow and Genetic Drift in a Species Subject to Frequent Local Extinctions. *Theoretical Population Biology* **12:** 253-262.

Soares, P., L. Ermini, N. Thomson, M. Mormina, T. Rito, A. Röhl, A. Salas, S. Oppenheimer, V. Macaulay, and M.B. Richards. 2009. Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. *The American Journal of Human Genetics* **84:** 740-759.

Song, Y.S. and J. Hein. 2005. Constructing Minimal Ancestral Recombination Graphs. *Journal of Computational Biology* **12:** 147-169.

Stephens, M. and P. Donnelly. 2000. Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62:** 605-635.

Stringer, C. 2003. Human evolution: Out of Ethiopia. *Nature* **423:** 692-695.

Stumpf, M.P.H. and G.A.T. McVean. 2003. Estimating recombination rates from population-genetic data. *Nat Rev Genet* **4:** 959-968.

Sturrock, K. and J. Rocha. 2000. A Multidimensional Scaling Stress Evaluation Table. *Field Methods* **12:** 49-60.

Tajima, F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123:** 585-595.

Tenesa, A., P. Navarro, B.J. Hayes, D.L. Duffy, G.M. Clarke, M.E. Goddard, and P.M. Visscher. 2007. Recent human effective

population size estimated from linkage disequilibrium. *Genome Research* **17:** 520-526.

Thangaraj, K., G. Chaubey, T. Kivisild, A.G. Reddy, V.K. Singh, A.A. Rasalkar, and L. Singh. 2005. Reconstructing the Origin of Andaman Islanders. *Science* **308:** 996.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061-1073.

The Hugo Pan-Asian SNP Consortium. 2009. Mapping Human Genetic Diversity in Asia. *Science* **326:** 1541-1545.

Thomas, J.H., R.O. Emerson, and J. Shendure. 2009. Extraordinary Molecular Evolution in the PRDM9 Fertility Gene. *PLoS ONE* **4:** e8505.

Tishkoff, S.A., E. Dietzsch, W. Speed, A.J. Pakstis, J.R. Kidd, K. Cheung, B. BonnÃ©-Tamir, A.S. Santachiara-Benerecetti, P. Moral, M. Krings, S. PÃ¤Ã¤bo, E. Watson, N. Risch, T. Jenkins, and K.K. Kidd. 1996. Global Patterns of Linkage Disequilibrium at the CD4 Locus and Modern Human Origins. *Science* **271:** 1380-1387.

Tishkoff, S.A., F.A. Reed, F.o.R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J.B. Hirbo, A.A. Awomoyi, J.-M. Bodo, O. Doumbo, M. Ibrahim, A.T. Juma, M.J. Kotze, G. Lema, J.H. Moore, H. Mortensen, T.B. Nyambo, S.A. Omar, K. Powell, G.S. Pretorius, M.W. Smith, M.A. Thera, C. Wambebe, J.L. Weber, and S.M. Williams. 2009. The Genetic Structure and History of Africans and African Americans. *Science* **324:** 1035-1044.

Torroni, A., A. Achilli, V. Macaulay, M. Richards, and H.-J. Bandelt. 2006. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics* **22:** 339-345.

Underhill, P.A. and T. Kivisild. 2007. Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. *Annual Review of Genetics* **41:** 539-564.

Voight, B.F., A.M. Adams, L.A. Frisse, Y. Qian, R.R. Hudson, and A. Di Rienzo. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 18508-18513.

Wade, M.J. and D.E. McCauley. 1988. Extinction and Recolonization - Their Effects on the Genetic Differentiation of Local-Populations. *Evolution* **42:** 995-1005.

Wall, J.D. and M. Przeworski. 2000. When Did the Human Population Size Start Increasing? *Genetics* **155:** 1865-1874.

Wang, Y. and B. Rannala. 2008. Bayesian inference of fine-scale recombination rates using population genomic data. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363:** 3921-3930.

Wang, Y. and B. Rannala. 2009. Population genomic inference of recombination rates and hotspots. *Proceedings of the National Academy of Sciences* **106:** 6215-6219.

Watkins, W.S., C.E. Ricker, M.J. Bamshad, M.L. Carroll, S.V. Nguyen, M.A. Batzer, H.C. Harpending, A.R. Rogers, and L.B. Jorde. 2001. Patterns of Ancestral Human Diversity: An Analysis of Alu-Insertion and Restriction-Site Polymorphisms. *American journal of human genetics* **68:** 738-752.

Weir, B.S. 1979. Inferences about Linkage Disequilibrium. *Biometrics* **35:** 235-254.

White, T.D., B. Asfaw, D. DeGusta, H. Gilbert, G.D. Richards, G. Suwa, and F. Clark Howell. 2003. Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature* **423:** 742-747.

Wilfert, L., J. Gadau, and P. Schmid-Hempel. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity* **98:** 189-197.

Winckler, W., S.R. Myers, D.J. Richter, R.C. Onofrio, G.J. McDonald, R.E. Bontrop, G.A.T. McVean, S.B. Gabriel, D. Reich, P. Donnelly, and D. Altshuler. 2005. Comparison of Fine-Scale Recombination Rates in Humans and Chimpanzees. *Science* **308:** 107-111.

Wiuf, C. 2002. On the minimum number of topologies explaining a sample of DNA sequences. *Theoretical Population Biology* **62:** 357-363.

Wiuf, C., T. Christensen, and J. Hein. 2001. A Simulation Study of the Reliability of Recombination Detection Methods. *Molecular Biology and Evolution* **18:** 1929-1939.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* **16:** 97-159.

Wright, S. 1940. Breeding Structure of Populations in Relation to Speciation. *The American Naturalist* **74:** 232-248.

Wright, S. 1943. Isolation by distance. *Genetics* **28:** 114-138.

Xing, J., W.S. Watkins, Y. Hu, C. Huff, A. Sabo, D. Muzny, M. Bamshad, R. Gibbs, L. Jorde, and F. Yu. 2010. Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biology* **11:** R113.

Zeyl, C.W. and S.P. Otto. 2007. A short history of recombination in yeast. *Trends in ecology & evolution (Personal edition)* **22:** 223-225.

Zhao, Z., N. Yu, Y.-X. Fu, and W.-H. Li. 2006. Nucleotide Variation and Haplotype Diversity in a 10-kb Noncoding Region in Three Continental Human Populations. *Genetics* **174:** 399-409.

APPENDIX A. Contributions to other articles

A1. Minimizing recombinations in consensus networks for phylogeographic studies

Laxmi Parida, Asif Javed, **Marta Melé**, Francesc Calafell, Jaume Bertranpetit and The Genographic Consortium.

BMC Bioniformatics 10Suppl 1:S72

http://www.biomedcentral.com/1471-2105/10/S1/S72

Parida L, Javed A, Melé M, Calafell F, Bertranpetit J, Genographic Consortium. Minimizing recombinations in consensus networks for phylogeographic studies. BMC Bioinformatics. 2009; 10(Suppl 1): S72.

A2. On recombination rates and with genetic differentiation in humans

Hafid Laayouni, Ludovica Montanucci, Martin Sikora, **Marta Melé**, Giovanni Dall'Olio, Belén Lorente-Galdos, Kate M McGee, Jan Graffelman, Philip Awadalla, Elena Bosch, David Comas, Arcadi Navarro, Francesc Calafell, Ferran Casals and Jaume Bertranpetit

Submitted

On recombination rates and with genetic differentiation in humans

Hafid Laayouni[1,2], Ludovica Montanucci[1], Martin Sikora[1], Marta Melé[1], Giovanni Dall'Olio[1], Belén Lorente-Galdos[1,2], Kate M McGee[4], Jan Graffelman[3], Philip Awadalla[5], Elena Bosch[1,2], David Comas[1,2], Arcadi Navarro[1,6,7], Francesc Calafell[1,2], Ferran Casals[1,5] and Jaume Bertranpetit[1,2, *]

[1]IBE, Institute of Evolutionary Biology (UPF-CSIC), CEXS-UPF-PRBB, Carrer Doctor Aiguader 88, 08003, Barcelona, Catalonia, Spain; [2] CIBER Epidemiología y Salud Pública (CIBERESP); [3] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Avinguda Diagonal 647, 08028, Barcelona, Spain; [4] NCI, National Institutes of Health, Frederick, MD 21072, USA; [5] Ste-Justine Hospital Research Centre, Faculty of Medicine, University of Montreal, Montreal, Quebec Canada; [6] Institució Catalana de Recerca i Estudis Avançats; [7] National Institute for Bioinformatics (INB), Barcelona (Spain)

Corresponding author: JB (jaume.bertranpetit@upf.edu)

Abstract

Recombination varies greatly among species, as illustrated by the poor conservation of the recombination landscape between humans and chimpanzees. Thus, shorter evolutionary time frames are needed to understand the evolution of recombination. Here, we analyze its recent evolution in humans. We calculated the recombination rates between adjacent pairs of 636,933 common single-nucleotide polymorphism loci in 28 worldwide human populations and analyzed them in relation to genetic distances between populations. We found a strong and highly significant correlation between similarity in the recombination rates and genetic differentiation between populations. This correlation is robustly maintained when considering presence/absence of recombination hotspots and after correcting for effective population size. A simulation analysis showed that the effect is not due to haplotype sharing. This result indicates a rapid pace of evolution of recombination, within the time span of differentiation of modern humans.

Introduction

The recombination rate is neither constant along chromosomes nor across species. The rate within genomes has been observed to vary at both the megabase level, with different chromosomal regions in the human genome showing differences in their recombination rates (Kong et al. 2002; Myers et al. 2005) and at a finer level, due to the existence of recombination hotspots (Myers et al. 2005; Crawford et al. 2004; McVean GA et al. 2004). A source of his variation could be to the existence of a 13-bp sequence motif recognized by a rapidly evolving zinc-finger protein, PRDM9 (Myers et al. 2010; Baudat et al. 2010). Comparisons of the human and the chimpanzee genomes have revealed poor conservation of recombination landscapes, likely due to these changes in PRDM9 (Myers et al. 2010), in contrast to the high level of DNA sequence conservation observed among these species (Ptak et al. 2005; Winckler et al. 2005). Recombination rates have also been compared among human populations, revealing large-scale conservation (Serre et al. 2005), while some differences in hotspot intensities and some population-specific hotspots have been described at a finer scale (Crawford et al. 2004; Bertranpetit et al. 2003; Conrad et al. 2006; Evans et al. 2005; Graffelman et al. 2007). Finally, different studies have shown the existence of individual variation in recombination (Coop and Przeworski 2007; Coop et al. 2008); and its heritability has been investigated, along with its biological consequences (Kong et al. 2004).

Measuring the fine-scale recombination rate is experimentally challenging and cannot be applied on a genome-wide scale; however good estimates can be obtained by applying population-genetic methods to DNA sequences (Stumpf and McVean 2003). Statistical methods have been developed to infer the fine-scale structure of recombination rate variation from genome-wide scale data (McVean GA et al. 2004). One of the widely used methods is implemented in the LDhat package, which is based on a composite-likelihood approach. Simulations have shown the LDhat produces largely unbiased rate estimates of the fine-scale genetic map. More recently, Khil and Camerini-Otero (2010) have shown that present-

day genetic crossovers are well predicted by a population averaged hotspot map computed from linkage disequilibrium data.

Differences in recombination rates among human populations provide an exceptional temporal framework to analyze the evolution of the recombination landscape, as they are well known and recent enough to capture fast evolutionary changes. The basal branches of the genetic diversification of human populations happened some 150,000 years ago, a much shorter time than the split between humans and chimpanzees (around 6.5 million years). The comparison of the recombination patterns among human populations provides a means to verify whether recombination landscapes evolve over time. To address this issue, we analyzed whether differences in recombination rates among human populations are correlated with their genetic differences computed as genetic distances. Whole genome estimations of recombination rates based on SNP data are already available for HapMap samples which, however, consist only of four populations for HapMap Phase I and II (International HapMap Consortium 2005; 2007) and 11 populations for HapMap Phase III. Here we computed the recombination rates using data for 660,918 SNPs on the Illumina HumanHap650K Beadchips genotyped in the full HGDP-CEPH panel samples (Li et al. 2008; Jakobsson et al. 2008) for 28 populations belonging to six continental groups representing worldwide human diversity (Cann et al. 2002).

Materials and methods

Recombination rate estimation

We considered the H971 subset of the Human Genome Diversity Cell Line Panel (HGDP-CEPH) recommended by Rosenberg (Rosenberg 2006). The 51 original HGDP-CEPH population samples (Cann et al. 2002) were re-grouped into 39 populations based on geographic and ethnic criteria as in Gardner et al. (Gardner et al. 2006). To avoid small sample sizes, the analysis was performed on genotypes from 28 populations belonging to six continental groups, with sample sizes over 19 individuals (a list of

used populations and their number of individuals is presented in Table 1). We used data for 660,918 SNPs on the Illumina HumanHap650K Beadchips successfully genotyped in the full HGDP-CEPH panel samples (Li et al. 2008; Jakobsson et al. 2008). SNPs are spaced 4.4 kb apart on average, an appropriate length given that hotspots occur every 200 Kb or less and their widths are 1-2 Kb (McVean GA et al. 2004; Jeffreys et al. 2005). Population recombination rates were calculated between neighboring SNPs according to the method implemented in the *rhomap* program (Auton and McVean 2007) within the LDhat package (Fearnhead and Donnelly 2001). LDhat methods have been demonstrated to give highly similar results to alternative approaches in human and chimpanzee datasets (Winckler et al. 2005; Jeffreys et al. 2005) and are computationally practicable for genome wide variation surveys. For a reliable estimation of the recombination rates, loci with more than 10% missing data in at least one population were discarded from the analysis (Auton and McVean 2007). After this cleaning procedure, the total number of SNPs included in the analysis was 636,933 (96% of all the SNPs in the HGDP). The number of SNPs for each chromosome is reported in Table 3. For each population, 5 independent runs of the *rhomap* program were carried out (with parameters: iterations=10.000.000, sampling=5.000, burnin=100.000). For each pair of adjacent SNPs we obtained 5 estimates of the population recombination rate ($4N_e r$ /kb) in each population and the median of these 5 estimates was used in the analysis.

Since population recombination rates (ρ) are dependent on the effective population size ($\rho = 4N_e r$), estimates of the population recombination rate in each population were normalized by $\theta = 4N_e \mu$, a scaled population mutation rate obtained from the same individuals and populations, where μ is the genome-wide average microsatellite mutation rate per locus and per generation (Graffelman et al. 2007). As there is no evidence of mutation rates varying among human groups, this correction produces values that are not biased by effective population size.

Correlation between genetic distance and recombination dissimilarity

We obtained a Spearman rank correlation matrix for the recombination rates among all pairs of populations. Each correlation value was obtained by comparing the values of corrected ρ (see above) for all pairs of adjacent typed SNPs between a population pair. In order to simplify the comparison with the genetic distance, the Spearman correlation values were turned into a dissimilarity measure by subtracting them from 1. The matrix obtained is then a measure of the dissimilarity of recombination rates between each pair of populations.

The differentiation among human populations was estimated through the F_{ST} measure (Weir and Cockerham 1984) among each pair of populations. F_{ST} values were calculated using a routine implemented in the PopGen module of BioPerl (Stajich and Hahn 2005) and stored in a 28×28 matrix

The matrix of recombination dissimilarity and that of genetic distance (F_{ST} matrix), were compared using a standardized Mantel test (Sokal and Rohlf 1995) by randomly permuting 9,999 times the rows and columns of one of the matrices. Statistical analyses were implemented using the R statistical software.

Simulation analysis

To further investigate the effect of the sharing of haplotypes and, hence of linkage disequilibrium patterns (which are at the base of the recombination rate estimates) on the relationship between genetic distance and recombination landscape, we designed a simulation study.

The simulations were carried out with the COSI program (Schaffner et al. 2005) which provide a simulation of the human demography under a three-population model based on the HapMap populations. This model was specifically designed to generate sequences that closely resemble empirical data of three human populations (African, European and Asian) by means of simulating a human-like demography and a variable recombination rate along the sequences,

allowing for presence and absence of hotspots. The simulator is already calibrated to obtain realistic F_{ST} values that mimic the divergence found among the three populations being simulated. We performed 1000 simulations using the best-fitting demographic model provided by COSI. We set the length of the simulated sequences to 1 Mb and adopted a sample size of 56 sequences for European and Asian populations and 42 for the African population with the aim of having the same amount of individuals as in a three chosen equivalent HGDP populations (Yoruba, French and Japanese). In each simulation, the recombination rate is exactly the same for the three simulated populations: this leads to simulated genotypes of different populations that share common haplotypes but do not have experienced differences in their recombination rate. Finally, in order to have a similar ascertainment bias in the simulations as in the observed data, we removed SNPs with MAF lower than 0.1 and performed a selection of tagSNPs with r^2 higher than 0.8 using Haploview software with the pairwise option (Barrett et al. 2005). In order to compare simulated data to a consistent empirical dataset, we randomly chose, along the whole genome, 1000 non-overlapping 1Mb long windows, and we analyzed them across the three populations of Yoruba, French and Japanese.

We then computed F_{ST} and recombination rates, following the same procedure as before, for real and simulated data. If the shared haplotypes were the main source of the high correlation found between recombination and genetic distance, we expect to observe this correlation also in the simulated data.

Results and discussion

Exploratory analysis of recombination rates

Population recombination rates were computed between 636,933 neighboring SNPs for 28 populations. As the recombination rate was estimated through several runs for each population, and to test for the agreement of estimates between runs of the same chromosome, 10 runs were performed for chromosome 22 for all populations. We carried out a repeated measure ANOVA testing

population and run as the main effects and pairs of adjacent SNPs as a covariate. No statistical significance of runs was found, but population and pairs of adjacent SNPs were highly significant (data not shown). This result reflects that the noise in the estimation procedure is low in relation to differences between populations.

Table 2 shows the mean estimated recombination rate for all populations, grouped according to geographical region**.** Results indicate considerable variation in recombination rates between populations, with small values for populations from East Asia. A repeated measure ANOVA show that differences between populations are highly significant ($F= 59479.8$ $p < 0.00001$). A Friedman ANOVA test shows similar results (ANOVA $\chi^2 = 2255369$ $p < 0.000001$). Post hoc analysis using a Bonferroni correction for the repeated ANOVA test show differences between populations remain significant, except for two homogenous groups from Central South Asia: Pathan, Burusho and Brahui; and Mozabite, Balochi and Makrani. Figure 1 shows estimated recombination rate, (scaled by the genome-wide average microsatellite mutation rate) along chromosome 22 for 6 populations (one from each continental region). The figure show similar pattern for all populations, however substantial variation could be detected by close observation. For example, South East China and Maya present by far less hotspots than the other populations. A hotspot located around 20 Mbp in all populations is absent (or much weaker) in Russian. A hotspots region around 32 Mbp is absent (or much weaker) in Brahui, Burusho, Hazara, Han, maya and Druze, but present in all other populations. This variation is consistent with previous reports in other genomic loci and genome-wide (Graffelman et al. 2007)

Genetic distance and recombination similarity between populations

Spearman rank correlation between populations recombination estimates were obtained by comparing the values of corrected recombination ρ for all pairs of adjacent typed SNPs between a population pair. The differentiation among human populations was

estimated through the F_{ST} measure (Weir and Cockerham 1984) among each pair of populations. The correlation values in recombination between population pair and F_{ST} measures were stored as a dissimilarity and distance matrices respectively and compared using a standardized Mantel test (Sokal and Rohlf 1995). A significant Mantel's *r* correlation of 0.894 ($P < 0.0001$) was observed, indicating that differences in recombination rates among populations increase with their genetic distance (Figure 2). In other words, genetic differentiation across human populations explains a considerable amount of recombination differences among them. This result also stands when the analysis is independently performed for each chromosome; then the Mantel test correlation ranges from 0.761 for chromosome 16 to 0.946 for chromosome 10 (Table 3).

It can be argued that these results could be explained by similar patterns of the recombination rate in the closest populations, due to the presence of common or shared haplotypes (see below), and not to a lower genetic differentiation among them. To test this hypothesis, we repeated the analysis considering only one population per continental group to avoid redundancy in the genetic composition of geographically close populations. In particular, the analysis was performed with data from Yoruba (Africa), French (Europe), Bedouin (Middle East / North Africa), Burusho (Central / South Asia), Han (East Asia) and Maya (America) populations. The observed correlation remained very high (Mantel's $r = 0.863$, $P = 0.002$) and was statistically significant even with the low number of pairwise comparisons.

To test for the impact of using the same data set for estimating recombination and genetic distance, we performed a mantel test between the Fst matrix calculated for one chromosome versus the recombination dissimilarity matrix computed on the other chromosomes. Results are presented in supplementary table 1 and show that this relationship remain and are highly significant in all cases ($p < 0.00001$) even when genetic distance and recombination dissimilarity are estimated from different parts of the genome. The maximum correlation is obtained when both matrices were

calculated for the same chromosome; however this is not always the case and may reflect inaccuracies in rates estimation.

Hotspots analysis

Alternatively, comparisons of recombination rates among populations can be evaluated by attending to the presence or absence of recombination hotspots. We defined a hotspot in each population as a recombination rate that exceeds 5 times the mean rate, producing a threshold of 10^{-6} ρ / θ. 22,413 hotspots have been detected at least in one population each. The number of hotspots vary from 2582 for South China to 8042 for Palestinian (no correlation between the number of hotspots and population sample size was observed, Pearson correlation test r = -0.08 p > 0.05; Spearman correlation test r = 0.34 p > 0.05). The proportion of shared hotspots between continental regions is maximum between EUR and MENA (0.34), EUR and CSASIA (0.31) and between MENA and CSASIA (0.29). These values are much lower when considering SSAFR or EASIA (Table 4).

We calculated the Jaccard distance between each pair of populations to measure the overall difference in presence/absence of hotspots (in this distance, the absence of a hotspot in a given position in two populations does not contribute to the similarity between them as would be in the case of a simple matching coefficient). Comparing this distance matrix with the F_{ST} matrix, highly significant results were obtained (Mantel's r = 0.866, P < 0.0001), suggesting that differences in the location of recombination hotspots increases with genetic differentiation between human populations.

Simulation analysis

With the mantel test analysis using only one population from each continent, we have shown that the effect of haplotype sharing in closely related populations does not explain the correlation between genetic differentiation and recombination. However, it is possible that the sharing of haplotypes and, hence of linkage disequilibrium

patterns, had a considerable effect also on distant populations, since its origin can be traced back to the Out of Africa origin of modern humans. To disentangle this point, we performed a simulation study designed to recognize the impact of using shared haplotypes on the estimates of recombination, rather than to represent a formal null model.

As the number of simulated populations is only three, the Mantel test cannot provide a robust comparison. To compare the relationship between recombination similarities and genetic differentiation in the three populations being simulated and in the three corresponding HGDP populations, we performed a Spearman correlation of the values of recombination between all neighboring SNPs in the 1Mbp and their FST values, for both simulated and empirical data. This is a more stringent test than the previous overall comparison between F_{ST} and recombination patterns, since, rather than general means, data points correspond now to 1000 windows of 1 Mb each. The correlation between recombination values and genetic distance for empirical data are 0.26, 0.25 and 0.27 for Yoruba-French, Yoruba-Japanese, and French-Japanese respectively (all significant). Conversely, these values were only 0.05, 0.06 and 0.09 for the simulated African-European, African-Asian and European-Asian (only the last comparison was marginally significant). This shows that, within the simulated populations, F_{ST} and recombination rate were not correlated despite sharing common haplotypes, whereas they are clearly correlated within the three studied populations. The common origin of haplotype structure, as illustrated in the simulation data, is unlikely to have contributed measurably to the correlation between genetic distances and structure of the recombination landscape. The low (although significant) correlation values between F_{ST} and recombination dissimilarity in the empirical data show that SNP variation captures a low amount of the variation of the recombination events distribution (r^2 = 0.07 on average). Presumably, differences in allele frequencies between populations correlate with recombination patterns through linkage disequilibrium with any motifs or genomic signals that induce recombination. That such correlation is small may imply that recombination patterns evolve faster than the relatively stable allele frequencies.

Concluding remarks

The results of this work reveal the footprint of the evolutionary history of human populations on the recombination rate, implying that differentiation in recombination rate estimates across human populations could be explained, in an important part, by their genetic differentiation. The large differences found in the comparison of the recombination landscapes among humans and chimpanzees (Ptak et al. 2005; Winckler et al. 2005) showed that recombination evolves quickly. Here, we give evidence that, even at the narrow timescale separating human populations, on the order of tens of thousands of years, differences appear to be detectable and to be correlated with genetic differentiation among populations. Recombination rate appears to be a rapidly changing parameter, indicating that the underlying factors shaping the likelihood of a recombination event, such as DNA sequences controlling recombination rate variation, also change. This is consistent with recent data showing that allelic variants of PRDM zinc fingers are significantly associated with variability in genome hotspots among humans (Baudat et al. 2010). The results obtained in this work contribute to the growing perception of recombination not as a genome-wide, cross-species fixed phenomenon, but as a fluctuating property well in accordance with its basic, molecular mechanism.

Acknowledgments

We are grateful to Brandon Invergo for reviewing the manuscript. This research was funded by grants BFU2007-63657, BFU2009-13409-C02-02 and SAF-2007-63171 awarded by Ministerio de Educación y Ciencia (Spain), by the Direcció General de Recerca of Generalitat de Catalunya (Grup de Recerca Consolidat 2005SGR/00608 and 2009 SGR 1101), and by the National Institute for Bioinformatics (www.inab.org), a platform of Genoma España. MM is supported by a PhD fellowship from the Programa de becas FPU del Ministerio de Educación y Ciencia, Spain (AP2006-03268v).

Author Contributions

Conceived and designed the experiments: JB FCas HL FCal DC EB AN. Performed statistical analysis and contributed to analysis tools: HL LM MS MM GD JG BLL KMM PA. Wrote the paper: HL FCas LM FCal JB. All authors read and approved the manuscript.

References

Auton A, McVean G. 2007. Recombination rate estimation in the presence of hotspots. Genome Res. 17: 1219-1227.

Barrett JC, Fry B, Maller J, Daly MJ, 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-265

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de

Massy B. 2010. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science 327: 836-840.

Bertranpetit J, Calafell F, Comas D, Gonzalez-Neira A, Navarro A. 2003. Structure of linkage disequilibrium in humans: genome factors and population stratification. Cold Spring Harb Symp Quant Biol 68: 79-88

Cann HM, de Toma C, Cazes L, et al. (39 co-authors). 2002. A human genome diversity cell line panel. Science 296: 261-262

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. 38: 1251-1260

Coop G, Przeworski M. 2007. An evolutionary view of human recombination. Nat. Rev. Genet. 8: 23-34

Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. Science 319: 1395-1398.

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens

M. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat. Genet. 36: 700-706

Evans DM, Cardon LR. 2005. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. Am. J. Hum. Genet. 76: 681-687

Fearnhead P, Donnelly P. 2001. Estimating recombination rates from population genetic data. Genetics 159: 1299-1318.

Gardner M, Gonzalez-Neira A, Lao O, Calafell F, Bertranpetit J, Comas D. 2006. Extreme population differences across Neuregulin 1 gene, with implications for association studies. Mol. Psychiatry 11: 66-75.

Graffelman J, Balding DJ, Gonzalez-Neira A, Bertranpetit J. 2007. Variation in estimated recombination rates across human populations. Hum. Genet. 122: 301-310

International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299-1320

International HapMap Consortium. 2007. A second generationhuman haplotype map of over 3.1 million SNPs. Nature. 449: 851-861.

Jakobsson M, Scholz SW, Scheet P, et al. (24 co-authors). 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998-1003

Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P. (2005) Human recombination hot spots hidden in regions of strong marker association. Nat. Genet. 37:601-606.

Khil PP, Camerini-Otero RD (2010) Genetic crossovers are predicted accurately by the computed human recombination map. PLoS Genet. 6(1):e1000831.

Kong A, Gudbjartsson DF, Sainz J, et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. Nat. Genet. 31: 241-247

Kong A, Barnard J, Gudbjartsson DF, et al. (14 co-authors). 2004. Recombination rate and reproductive success in humans. Nat. Genet. 36: 1203-1206

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-1104

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304: 581-584

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321-324

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G,

Donnelly P. 2010. Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. Science 327: 876-879

Parvanov ED, Petkov PM, Paigen K. 2010. Prdm9 controls activation of Mammalian recombination hotspots. Science 327: 835.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M,

Frazer KA, Pääbo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet. 37: 429-434

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann. Hum. Genet. 70: 841-847

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15: 1576-1583

Serre D, Nadon R, Hudson TJ. 2005. Large-scale recombination rate patterns are conserved among human populations. Genome Res 15: 1547-1552

Sokal RR, Rohlf FJ. 1995. Biometry. 3rd edition New York: Freeman

Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Mol. Biol. Evol. 22: 63-73.

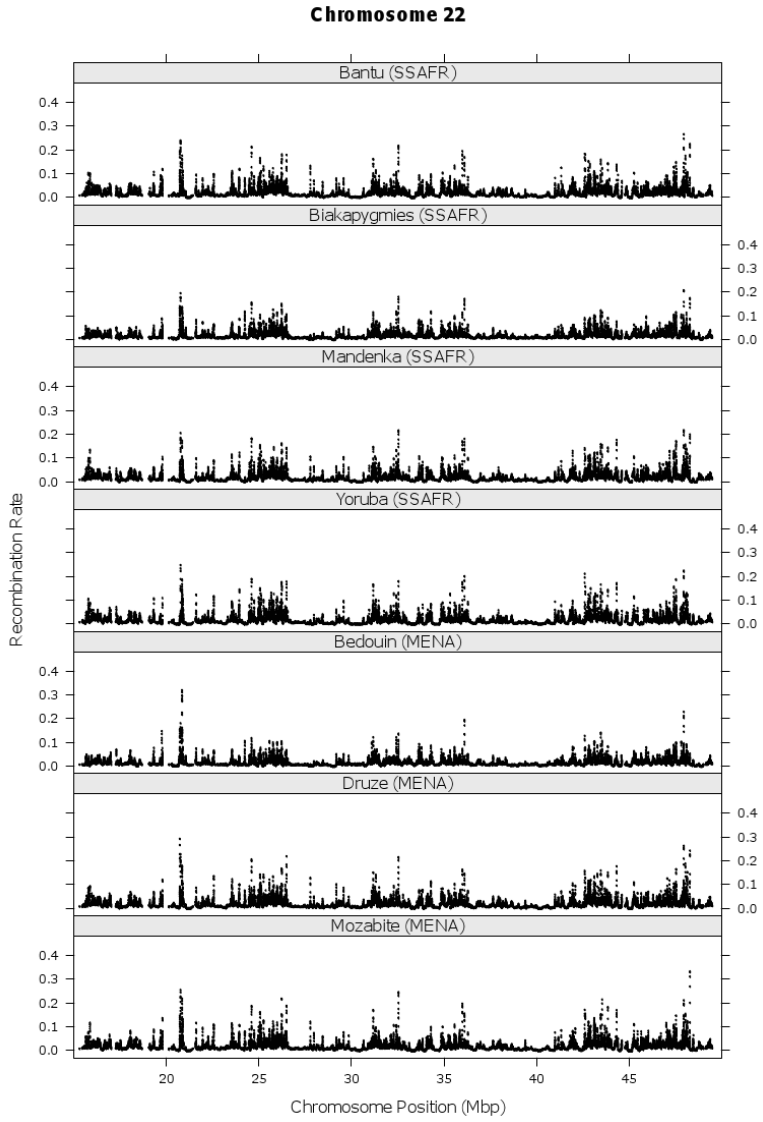
Stumpf MP, McVean GA. 2003. Estimating recombination rates from population-genetic data. Nat Rev Genet. 12: 959-968.

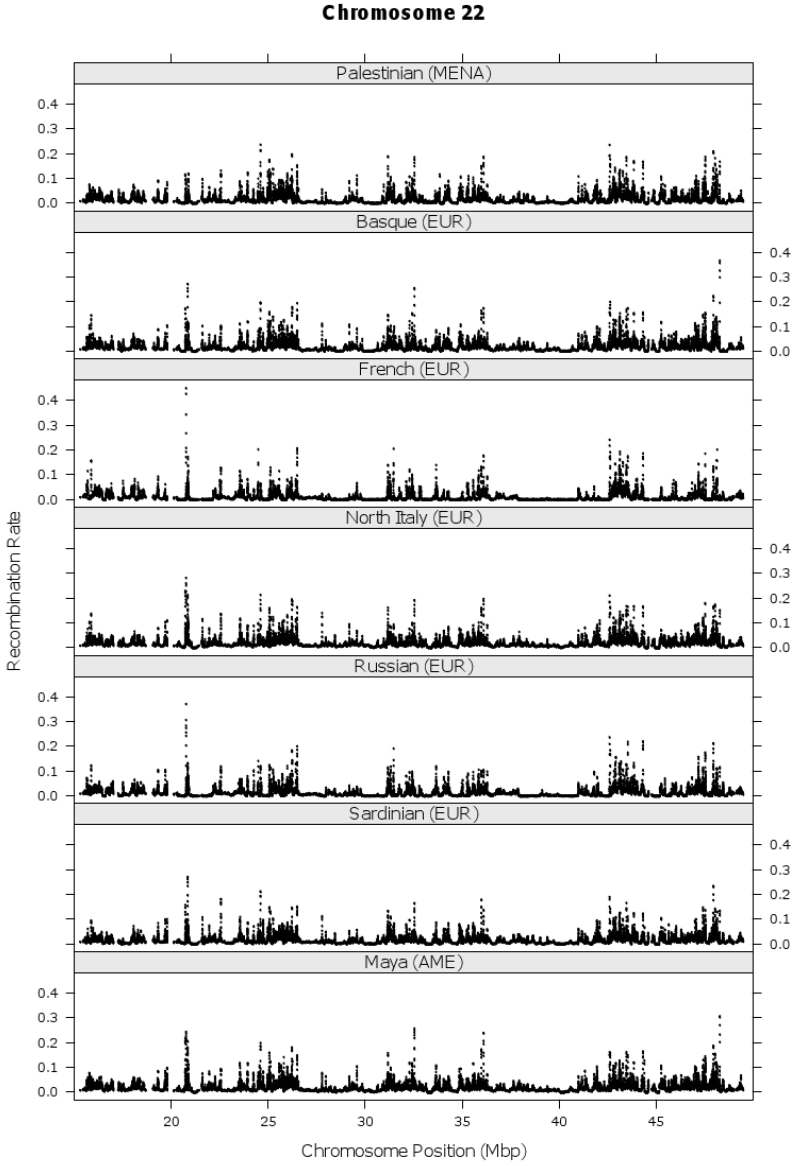
Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution Int J Org Evolution 38: 1358-1370

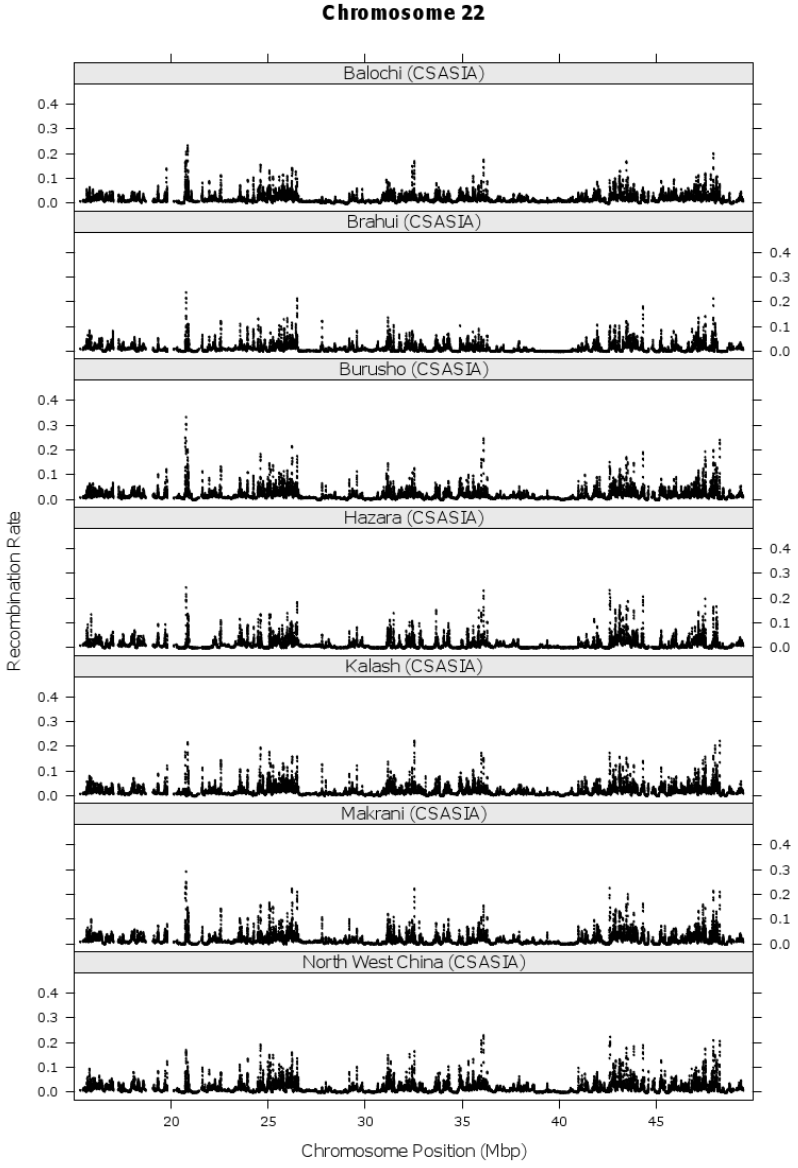
Winckler W, Myers SR, Richter DJ, et al. (11 co-authors). 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308: 107-111

Figure Legends

Figure 1. Recombination rate estimates (10^{-5}) for successive SNP-pairs for chromosome 22 and in each of 28 populations, grouped into geographical regions.




Chromosome 22
Palestinian (MENA)
Basque (EUR)
French (EUR)
North Italy (EUR)
Russian (EUR)
Sardinian (EUR)
Maya (AME)
Recombination Rate
0.0
0.1
0.2
0.3
0.4
20
25
30
35
40
45
Chromosome Position (Mbp)


Chromosome 22
Balochi (CSASIA)
Brahui (CSASIA)
Burusho (CSASIA)
Hazara (CSASIA)
Kalash (CSASIA)
Makrani (CSASIA)
North West China (CSASIA)
Recombination Rate
0.0
0.1
0.2
0.3
0.4
20
25
30
35
40
45
Chromosome Position (Mbp)

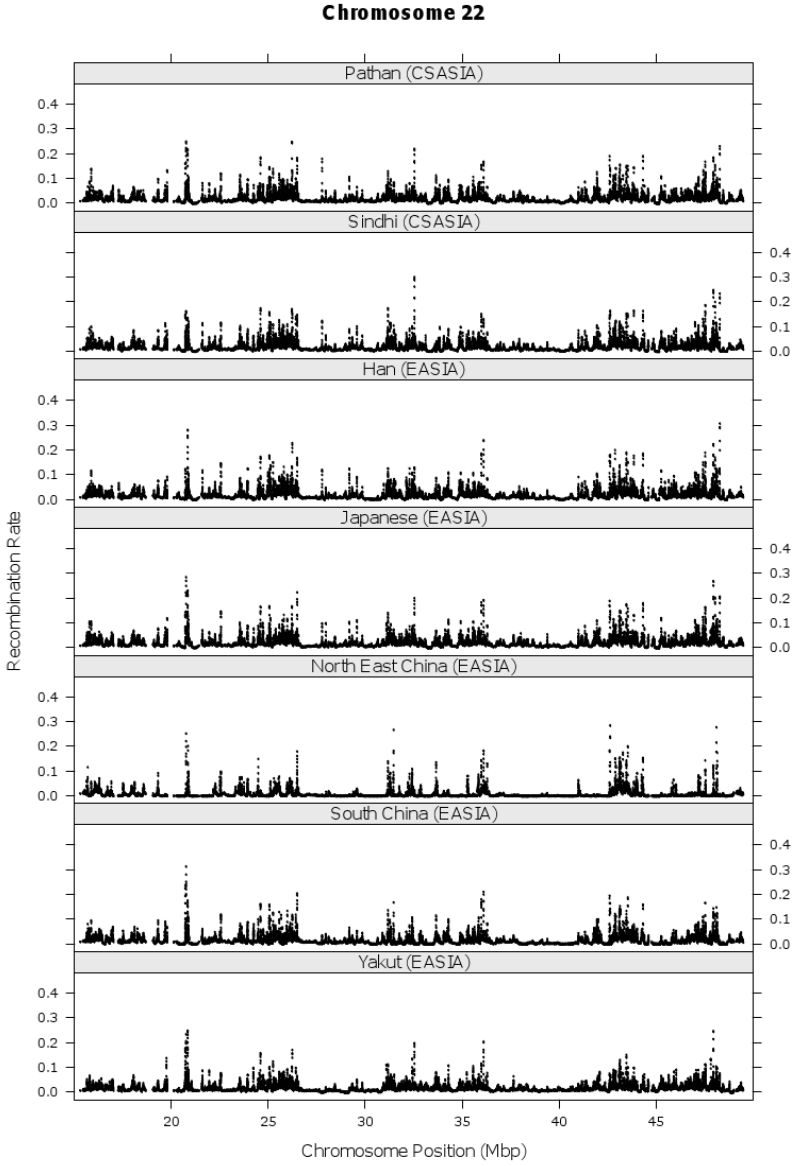
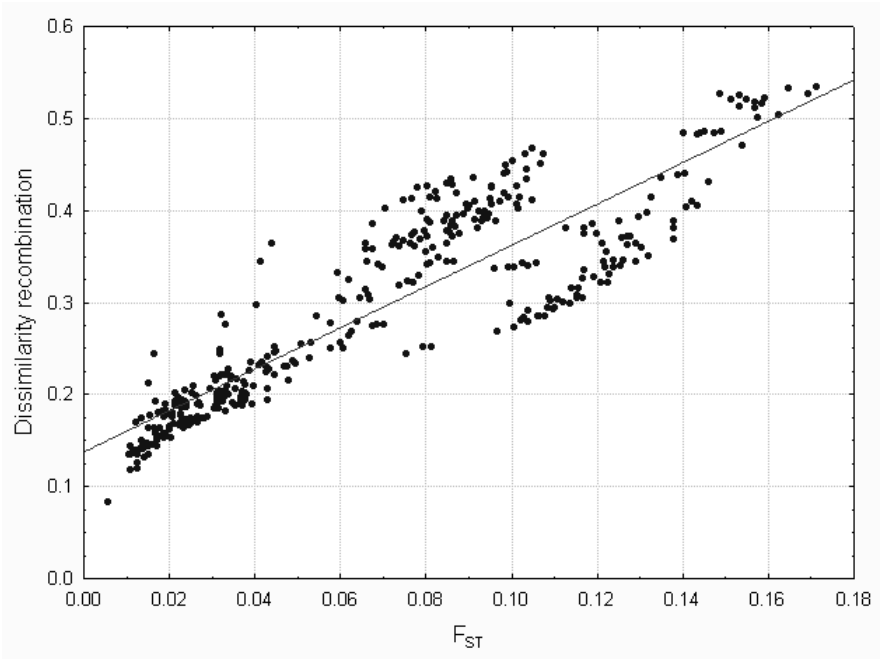

Chromosome 22
Pathan (CSASIA)
Sindhi (CSASIA)
Han (EASIA)
Japanese (EASIA)
North East China (EASIA)
South China (EASIA)
Yakut (EASIA)
Recombination Rate
0.0
0.1
0.2
0.3
0.4
20
25
30
35
40
45
Chromosome Position (Mbp)

Figure 2. Relationship between F_{ST} values and the recombination rate correlation based on 378 pairwise populations comparisons.



APPENDIX B. Contributions to other articles as a Genographic Consortium member

A Mitochondrial Revelation of Early Human Migrations to the Tibetan Plateau Before and After the Last Glacial Maximum

Zhendong Qin, Yajun Yang, Longli Kang, Shi Yan, Kelly Cho, Xiaoyun Cai, Yan Lu, Hongxiang Zheng, Dongchen Zhu, Dongmei Fei, Shilin Li, Li Jin1, Hui Li and The Genographic Consortium

American Journal of Physical Anthropology 143:555-569 (2010)

http://onlinelibrary.wiley.com/doi/10.1002/ajpa.21350/abstract;jsessionid=BB3A25E9887F949023594952A463576E.d02t02

Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities

Wolfgang Haak, Oleg Balanovsky, Juan J. Sanchez, Sergey Koshel, Valery Zaporozhchenko, Christina J. Adler, Clio S. I. Der Sarkissian, Guido Brandt, Carolin Schwarz, Nicole Nicklisch, Veit Dresely, Barbara Fritsch, Elena Balanovska, Richard Villems, Harald Meller, Kurt W. Alt, Alan Cooper, the Genographic Consortium

PLoS Biology 8(11): e1000536 (2010)

http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1000536

Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon

Marc Haber, Daniel E Platt, Danielle A Badro, Yali Xue, Mirvat El-Sibai, Maziar Ashrafian Bonab, Sonia C Youhanna, Stephanie Saade, David F Soria-Hernanz, Ajay Royyuru, R Spencer Wells, Chris Tyler-Smith and Pierre A Zalloua

European Journal of Human Genetics doi:10.1038/ejhg.2010.177 (1 December 2010)

http://www.nature.com/ejhg/journal/vaop/ncurrent/full/ejhg2010177a.html