Departament de Ciències

Experimentals i de la Salut

Universitat Pompeu Fabra (UPF)

Selection and linkage desequilibrium tests under complex demographies and ascertainment bias

Memòria presentada per Anna Ramírez Soriano per optar al grau de doctora per la Universitat Pompeu Fabra. Aquesta tesi ha estat realitzada sota la dirección del Dr. Francesc Calafell i Majó, a la Unitat de Biologia Evolutiva del Departament de Ciències Experimentals i de la Salut de la Universitat Pompeu Fabra, dins del programa de doctorat International PhD Programme in Basic Biomedical Research Health and Life Sciences (període 2004-2008).

Francesc Calafell i Majó

Anna Ramírez i Soriano

Barcelona, Juliol 2008

Com sempre, per tu, mama.
Tot i que ara camines per viaranys que ens són vetats
el teu record sempre ens acompanya, i et sento a prop
en tot allò que faig.

43
El sueño
de la razon
produce
monstruos.

Francisco de Goya y Lucientes
1799

Agraïments

INDEX

INTRODUCTION

"It's a dangerous business, Frodo, going out of the door.
You step into the Road, and if you don't keep your feet,
there is no knowing where you might be swept off to.
Do you realize that this is the very path that goes through
Mirkwood, and that if you let it, it might take you to
the Lonely Mountain or even further and to worse places?"

The Lord of The Rings
The Fellowship of the Ring
J.R.R. Tolkien

1 GENOMIC VARIATION

Earth is the only known planet that holds life. This life is extremely diverse and occupies a wide range of different ecological niches. However, despite all their diversity, all living organisms today share a common ancestor, named LUCA. LUCA, which stands for Last Unique Common Ancestor, lived between three and four billion years ago (Ridley. 2000). The force beyond the origin of the multiplicity of living forms present today from LUCA is evolution, which acts over the genomes creating variability.

The main forces driving evolution are mutation, selection, migration, and genetic drift. Mutation creates new variability, adding changes to the genome; while selection, migration and drift act over the present variation increasing or reducing it. Although the effect of those forces takes place on populations of a given species, in the end it may result in the split of one single species into two different ones.

Mutation occurs through different mechanisms involved in DNA duplication and repair. Consequently, it can create different types of polymorphism (see Figure 1), which take place at different rates. One of those are single nucleotide polymorphisms (SNPs), which consist in the substitution of one base for another in a single position. Nucleotide mutations occur at a rate between 10^{-8} and 10^{-10} per site and generation, depending on the genome context: for instance, they are more frequent in CpG islands. Another example are microsatellites or short tandem repeats (STRs), sequences of one to six bases in length that are repeated in tandem. In STRs, variation is found in the number of repeats between individuals, and the mutation rate is about 10^{-3}-10^{-4} per generation.

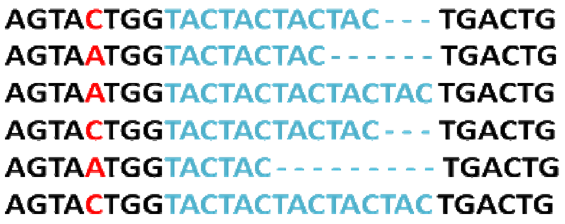

AGTACTGGTACTACTACTAC- - - TGACTG
AGTAATGGTACTACTAC- - - - - - TGACTG
AGTAATGGTACTACTACTACTACTGACTG
AGTACTGGTACTACTACTAC- - - TGACTG
AGTAATGGTACTAC- - - - - - - - - TGACTG
AGTACTGGTACTACTACTACTACTGACTG

Figure 1: fragment of DNA in six individuals. Red indicates a SNP, blue an STR.

As the genome is shaped by the evolutionary forces mentioned above, the study of its patterns of variation provides information on the processes that have acted over it, both globally (demographic factors) or locally (selective forces). In fact, a large number of papers are devoted to inferring the processes underlying population history from variation (Spurdle and Jenkins. 1992; Comas *et al.* 1998; Garrigan *et al.* 2007; Goebel *et al.* 2008; Friedlaender *et al.* 2008).

1.1 Demography

Population size, changes in population size, and population movements have the potential to shape variation simultaneously along the whole genome. Some of those factors, such as migration or bottlenecks, reduce variation; while others, like population splits or expansions, increase it.

However, the mechanisms through which they act are different. For instance, migration reduces variability due to the gene flow between two separated populations, while bottlenecks reduce it by a sudden decrease of the population size, in which a number of its individuals and the variation they carry get lost. The same happens with population subdivision and expansions. The former increases variability through the split of one population in two, which will accumulate different changes over time. Expansions, on the other hand, increase it through a rapid increase of population size, which leads to the rapid creation of new variation. This is more extensively explained in section 3.

Particular demographic histories have been extensively studied through the analysis of variation patterns and how they change over populations. In humans, this has been typically achieved genotyping SNPs and STRs in the Y chromosome, and resampling mitochondrial DNA (Spurdle and Jenkins. 1992; Stoneking. 1994; Castro *et al.* 1998; Cavalli-Sforza. 1998; Comas *et al.* 1998; Stumpf and Goldstein. 2001; Underhill. 2003; McElreavey and Quintana-Murci. 2005; Torroni *et al.* 2006; Goebel *et al.* 2008), although more recently some studies have been performed using polymorphism in the X chromosome (Schaffner. 2004; Garrigan and Hammer. 2006; Garrigan *et al.* 2007) or even the autosomes (Comas *et al.* 2000; Garrigan and Hammer. 2006; Friedlaender *et al.* 2008).

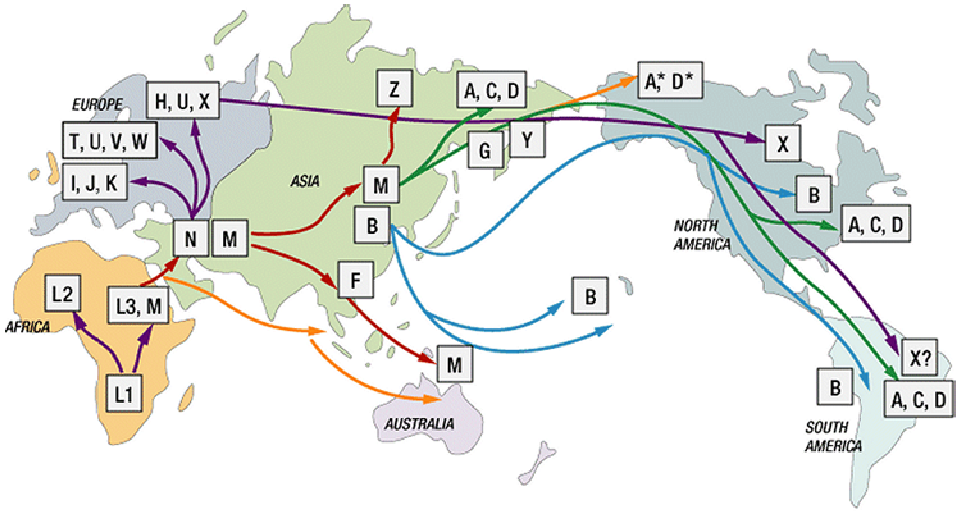


Figure 2: map of human expansion according to mitochondrial DNA. (http://www.familytreedna.com/)

Y chromosome and mitochondrial DNA (mtDNA) have been the favourite tools for population geneticists for nearly three decades (see Figure 2). This is mainly due to their particular characteristics, as they are haploid and uniparentally inherited. Being haploid means that a) their haplotypes are directly accessible and, thus, their phase does not need to be estimated, and b) their effective population size (N_e) is one quarter that of the autosomes, so they are more sensitive to genetic drift. The uniparental inheritance has the advantage than we can draw independent demographic maps for each sex and infer whether males and females have undergone the same processes.

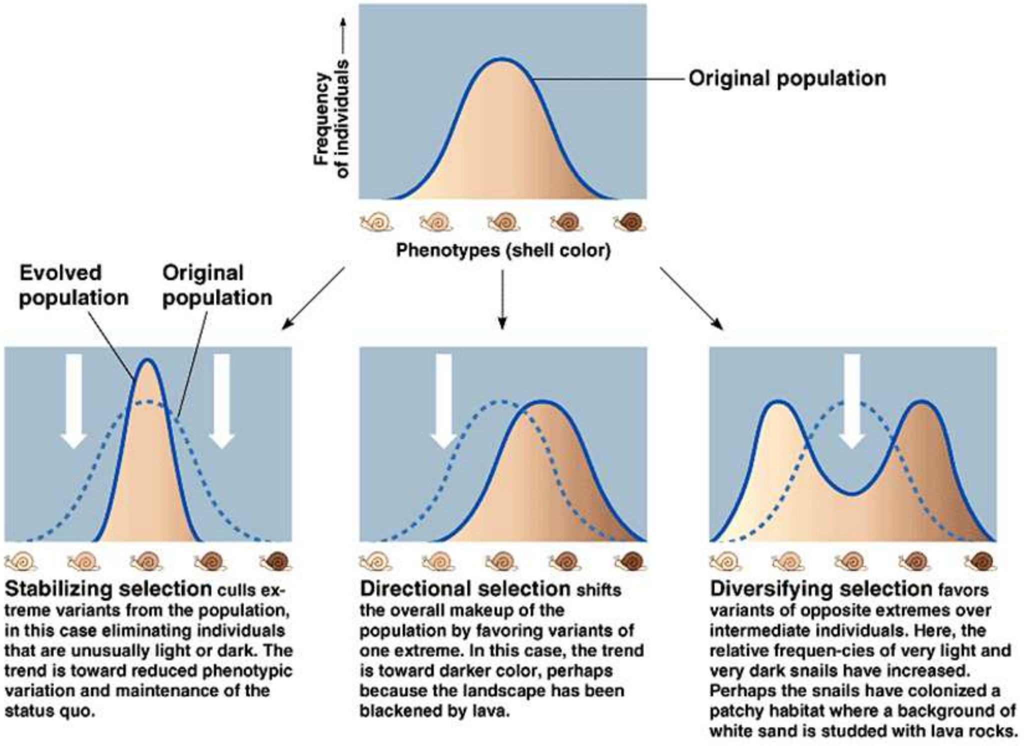
Too often Y chromosome and mtDNA studies realised over the same or very similar populations have been independently published, leading to confusing explanations about the demographic history of the peopling of a given area. As a consequence, several works aimed to reconcile both the male and female perspective have been written. This is the case of the work by Wood *et al.* (2005), in which they intended to establish the association between genetic markers for both sexes and linguistic and geographic variation in Africa. With this objective, they genotyped 50 SNPs in the Y chromosome in 40 African populations and compared their results with those of several previous mtDNA publications (Salas *et al.* 2002; Knight *et al.* 2003; Destro-Bisol *et al.* 2004; Coia *et al.* 2005). Their results suggested that the patterns of migration inside Africa have been different for males and females. In fact, genetic variation strongly correlates with linguistic differences in males but not in females, which points to sex-biased effects of the replacement of hunter-gatherer populations by Bantu farmers.

The X chromosome is a useful tool to disentangle human demography as it has several characteristics close to those of the Y chromosome and the mitochondrial DNA: a) it has low recombination rates, b) in the case of males its haplotypes can be directly obtained and c) as males have only one copy, its N_e is smaller than in autosomes. Moreover, as it is present in both males and females, it can yield a more general picture of the demographic factors affecting populations.

Demography was inferred from variation at the X chromosome by Laan *et al.* (2005). Their goal was to ascertain how the demographic history of populations affects linkage disequilibrium (LD) patterns and to which extent cross-over activity dilutes its effects. They focused in two different regions of the X chromosome, one with low cross-over activity (Xq13) and the other with high cross-over activity (Xp22). They genotyped eight microsatellites on Xq13 and six on Xp22 in 14 Eurasian populations, and calculated patterns of LD among them. Their results showed that Xq13, having low cross-over rates, maintained higher levels of LD, which are consistent with the demographic history of the populations involved in the study. Furthermore, demographic factors were shown to influence the haplotype distribution of the markers at Xq13 across populations. On the other hand, the haplotype and LD structure found gave insights of the demographic history of the populations included and of the gene flow among them.

1.2 Selection

Selection consists in the differential reproductive success of the distinct variants in the population. These forces can reduce variability, in the case of directional and stabilizing selection, or increase it in the case of disruptive selection, as shown in Figure 3.



Stabilizing selection culls extreme variants from the population, in this case eliminating individuals that are unusually light or dark. The trend is toward reduced phenotypic variation and maintenance of the status quo.

Directional selection shifts the overall makeup of the population by favoring variants of one extreme. In this case, the trend is toward darker color, perhaps because the landscape has been blackened by lava.

Diversifying selection favors variants of opposite extremes over intermediate individuals. Here, the relative frequen-cies of very light and very dark snails have increased. Perhaps the snails have colonized a patchy habitat where a background of white sand is studded with lava rocks.

Figure 3: kinds of selection which can act over phenotypes.

Directional selection reduces variability by favouring (positive selection) or hindering (negative selection) a variant, and shifts the distribution of a given trait in the population. Stabilizing selection does so by favouring intermediate phenotypes, and thus reduces the variability of the trait. On the contrary, disruptive selection increases the variance of a population by favouring extreme phenotypes, producing bimodal distributions of the trait involved.

As discussed in the previous paragraph, selection affects the genome locally because it acts over a single phenotypic trait, which translates to a gene or genomic region. Thus, the selective inferences made when analysing variation in a gene are limited to this gene and the adjacent area. This area of influence decreases as the recombination rates in the region increase, as recombination shuffles the variation present in the population.

One of the tools available to detect selective events is to compare the frequency of a genetic variant between populations. In fact, population-specific selection pressures can increase differences between them while strong, homogeneous selective pressure can lead to less diversity than expected.

F_{ST} is a measure of the proportion of the genetic variance explained by the differences among populations that counts the excess of heterozygotes found by pooling different populations. If many loci are studied simultaneously, the differences that will arise among populations will be due to demographic effects. However, if a single locus is compared to the rest of the genome, F_{ST} can be used to detect differential patterns of selection among populations.

Based on that property of F_{ST}, Akey *et al.* (2002) provided an empirical genome-wide distribution of F_{ST} values (see Figure 4). This distribution was created by comparing the frequency of 26,530 SNPs distributed along the whole genome in three populations. The idea behind the description of such a distribution is that the differences that will be observed among populations will be due to demographic events. Thus, it will provide a neutral distribution (see section 2) against which other SNPs could be compared. If the F_{ST} value of the SNPs of interest is extreme compared to the genome-wide distribution, we may infer that the difference between populations is not due to demography but to selective events.

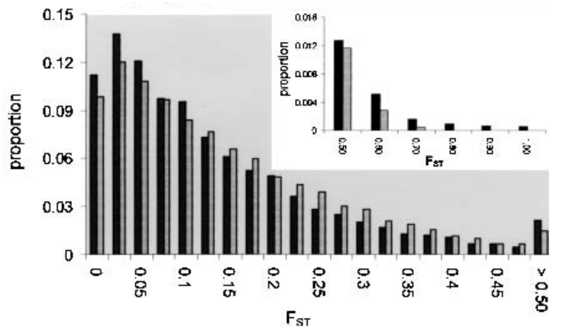


Figure 4: Akey's F_{ST} distribution. (Akey *et al.* 2002)

In fact, Akey *et al.* (2002) compared the F_{ST} values obtained for 8,862 SNPs located in gene-associated regions and identified 174 regions which extreme F_{ST}. Among these regions, they found 17 genes associated with Mendelian and complex diseases such as cystic fibrosis or type 2 diabetes, respectively.

2 THE NEUTRAL MODEL

As stated in the previous section, the detection of demographic or selective events in a population requires comparing the results obtained in this population against a model where these factors were absent. This is the neutral model, and it is based on the neutral theory of molecular evolution, which rests on the work developed by R.A. Fisher, J.B.S. Haldane and S. Wright in the early thirties.

2.1 The neutral theory of molecular evolution

Since Darwin's publication of *The Origin of Species by Means of Natural Selection* in 1859 and until the late nineteen sixties, it was widely believed that evolution was exclusively driven by positive selection. However, the next shock to the scientific community came in 1968, when Kimura published a paper entitled *Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles* (Kimura. 1968). In it, he proposed that most evolutionary change was not due to variants that affected the fitness of an individual but to neutral mutations. Five years later he published the book *The neutral theory of molecular evolution* (Kimura. 1984), that he started with the sentence: "This book represents my attempt to convince the scientific world that the main cause of evolutionary change at the molecular level […] is random fixation of selectively neutral or nearly neutral mutants rather than positive Darwinian selection." In the introduction, he stated that "I am convinced that no other existing theory can give a better and more consistent explanation of these facts [data from the new molecular revolution]", a sentence that still holds nowadays.

The neutral theory of molecular evolution states that the vast majority of mutations are neutral, that is, they are neither favourable nor unfavourable for the individual bearing them, and only a fraction will be exposed to natural selection. Under this scenario, the new variants arising in the genome increase and decrease in frequency at random due to the effect of genetic drift, as shown on Figure 5. Moreover, Kimura proposed that all mutations will eventually, given enough time, be fixed or disappear (see Figure 6). Thus, under this light, polymorphism is only an intermediate state between the appearance of a variant and its fixation or elimination. This theory, then, provided the necessary framework against which to compare the hypothesis of selection over genomic regions.

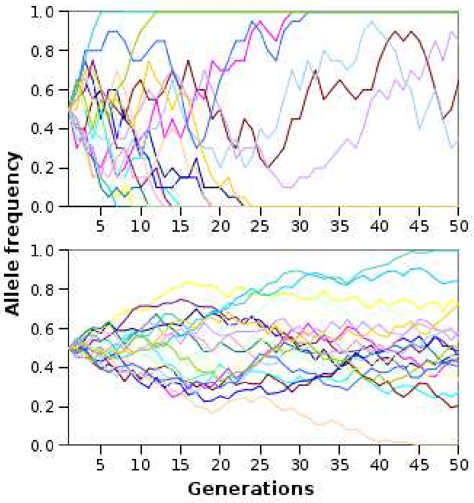


Figure 5: evolution of the frequency of 20 unlinked alleles with an initial frequency of 0.5 due to genetic drift on populations of 10 (above) and 100 (below) individuals (http://en.wikipedia.org/wiki/Evolution).

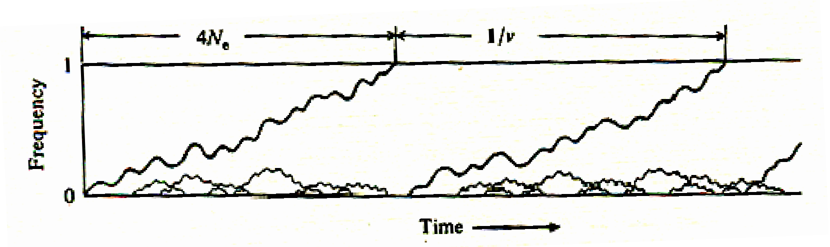


Figure 6: behaviour of new alleles arising in the population. ν corresponds to the mutation rate. (Kimura. 1984).

2.2 The Wright-Fisher's neutral model

The most widely used model to generate neutral genealogies is the Wright-Fisher model (Fisher. 1930; Wright. 1931; Hein *et al.* 2005). This model describes how an idealised population transmits its genes from one generation to the following and, thus, how it evolves forwards in time.

Note, however, that the Wright-Fisher model is not based on a real population but implies a number of simplifications. The model, then, follows six assumptions:

1. Constant size, that is, the number of individuals in the population does not change over generations.

2. Infinite sites. The Wright-Fisher model assumes that mutation may occur in an infinite number of sites. This implies that recurrent mutation is not allowed in this model, as the probability that two mutations occurs at the same site is zero.
3. Panmixia, which means that all individuals have the same probability to mate with any other individual of the opposite sex, without any internal subdivision.
4. Non-overlapping generations. All individuals in the population belonging to one generation reproduce and die at the same time, and they all mate only with members of their own generation.
5. No recombination: the genes involved in the model cannot recombine. This implies that the model as is can only be used on non-recombinant pieces of DNA, such as the non-recombining segment of the Y chromosome (NRY) or mitochondrial DNA.
6. Selection is absent: all individuals have the same probability of surviving and producing offspring, irrespectively of their genotypes.

Thus, if the population of interest does not match the Wright-Fisher model, it might mean that it is violating one or more of its assumptions. Finding which one can give clues of which processes underlie the history of the population.

The Wright-Fisher model can be used both to simulate haploid data (such as NRY chromosome and mitochondrial DNA) and diploid data, although there are some differences in the creation of a haploid or a diploid genealogy.

Assume a population of size 2*N* and two generations *t* and *t*+1. In the haploid model (Figure 7, above) each individual in generation *t*+1 is modelled taking randomly a gene from generation *t*, so every gene at generation *t*+1 will have an ancestor in generation *t* chosen at random with probability $\frac{1}{2N}$. The choice of one gene is independent of all other genes. As a consequence, an individual at generation *t* may have more than one descendant, and not all individuals at generation t will leave offspring in generation t+1.

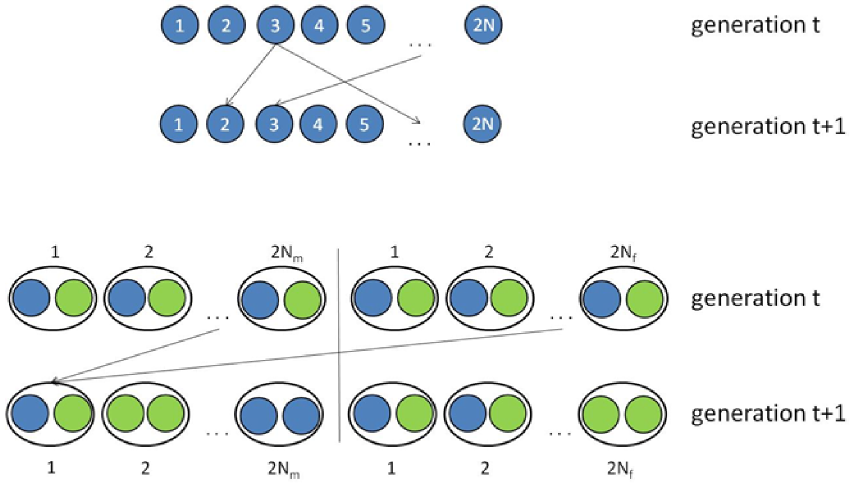


Figure 7: the Wright-Fisher model. The haploid model is shown above, and the diploid, below.

In the diploid model (Figure 7, below) there is an added difficulty, as each individual is formed by two genes, one coming from the father and the other from the mother. In this case, the population is subdivided into two subpopulations, males and females, with size N_m and N_f respectively. Together, they sum the 2*N* genes assumed in the haploid model. Now, each individual at generation *t*+1 has two ancestors at generation *t*, which are chosen among the males and the females with a probability of $\frac{0.5}{2N}$ in each case. Note that, in the diploid model, some restrictions apply: the choice of a gene in generation *t* to form an individual in generation *t*+1 is no longer independent, as the second gene chosen cannot belong to the same parent as the first one. However, if the time scales are corrected, for large values of *N*, N_f, and N_m, the two models are probabilistically similar.

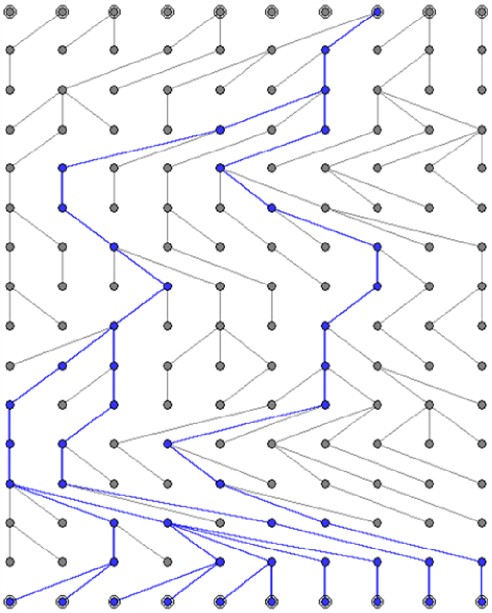


Figure 8: Wright-Fisher genealogy for a sample of 10 individuals. The present sample, their ancestors and their MRCA are shown in blue. (Genealogy built using http://www.coalescent.dk/).

The Wright-Fisher model can, then, simulate the genealogy of a given sample of size *n* over time. However, from this genealogy created forwards in time we can make inferences backwards in time. For instance, as not all individuals of a generation leave descendants in the next one, going backwards in time from the present generation to the past it will be seen that all genes existing in the present came from a single ancestor that lived a certain number of generations in the past (see Figure 8). This ancestor receives the name of most recent common ancestor (MRCA). This point will be later developed in more detail in section 3. For examples of applications of simulations based in the Wright-Fisher model, see Rendine *et al.* (1986) and Calafell and Bertranpetit (1993).

2.3 Moran's neutral model

One of the most problematic assumptions of the Wright-Fisher's model is that it does not allow overlapping generations. This was solved by Moran in the late fifties (Moran. 1958a; Moran. 1958b; Gale. 1990; Hein *et al.* 2005), when he proposed a new model with overlapping generations.

In Moran's model, assuming the same population of size 2*N* that in the Wright-Fisher's model, at each point of time, an individual is randomly chosen to die and is replaced by a newborn, who is a copy of one random, pre-existing individual, as shown in Figure 9. Depending on the formulation of the

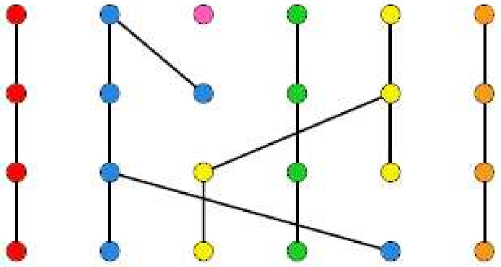


Figure 9: Moran's model. Colours indicate the same individual through time, and its descendant in case it reproduces (see blue and yellow). (Modified from Hein *et al.* 2005).

model, the individual who reproduces and the individual who dies can or cannot be the same. As the Wright-Fisher model, Moran's model also can be formulated for a diploid population.

Despite the differences between both models, they are equivalent if the time scale is corrected and *N* is large enough.

3 COALESCENT THEORY

Years after the publication of the Wright-Fisher model and based on it, J.F.C. Kingman developed what has been known since as the coalescent theory (Kingman. 1982a; Kingman. 1982b). Later on, in 2000, he published an amusing paper (Kingman. 2000) in which he describes the process from which such a theory came to light.

The main point of coalescent theory, which meant a breakthrough in the line of thinking that predominated at that moment, was that it was aimed to generate genealogies backwards in time. That meant that, on the contrary of what was being done with models such as Wright-Fisher's or Moran's, the genealogy was build starting on a population in the present whose genealogic relationships were modelled backwards until the MRCA was reached. This implies that, now, we would not know all the individuals existing in the previous generations but only those that had left descendants living in the present (see Figure 10).

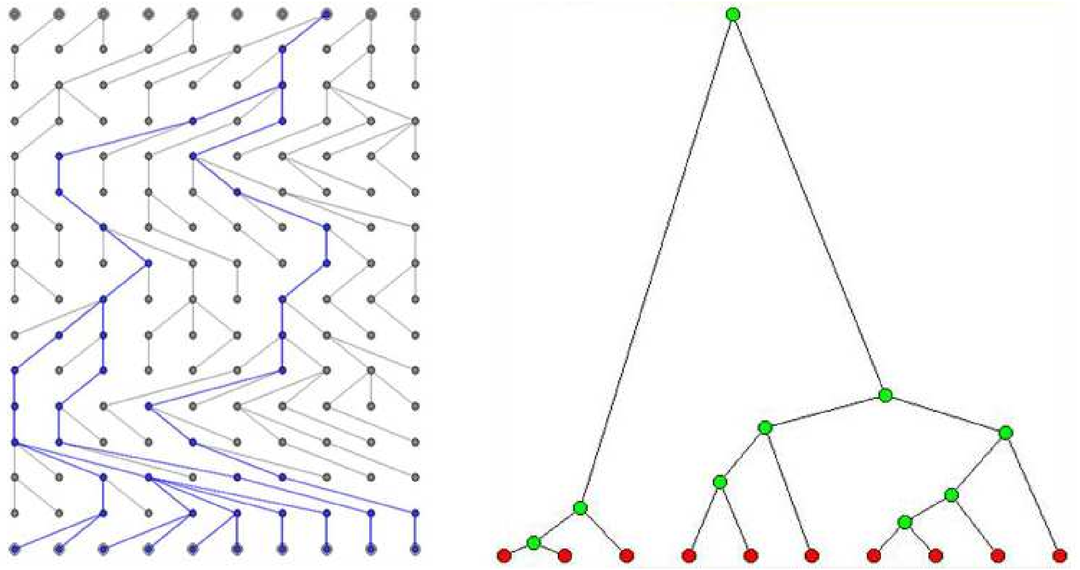


Figure 10: comparison between a Wright-Fisher (left) and a coalescent (right) tree. On the contrary than on the Wright-Fisher tree, in the coalescent tree, the only chromosomes known are those that are ancestors of the present sample. (Both genealogies are built using http://www.coalescent.dk/).

This theory, further developed by Hudson (1990), Donelly and Tavaré (1995), and Fu and Li (1999), provided a powerful framework to develop neutrality tests and further their study (see section 4), since it is computationally much more efficient than forward simulations.

3.1 Building the coalescent

The coalescent process includes two different parts: the topology of the tree and the mutation. Those parts cannot be simulated together. This is so because the topology of the tree is influenced by factors such as the sample size or its geographic structure. However, tree shape is not affected by neutral mutation since, by definition, neutral mutations do not alter the fitness of the individuals bearing them. Consequently, the coalescent must be build in two steps: the first step is to generate the topology of the tree and the second step to throw mutations over it.

Note that all the information that will be developed concerning the coalescent theory refers to the coalescent of a sample of *n* genes, what is known as the *n-coalescent*. Thus, this is not applicable to the coalescent process for an infinite or very large sample.

The topology of the tree is build upon the probability that two individuals share a common ancestor in the previous generation. This probability, on a population size of 2*N* individuals, is $\frac{1}{2N}$. This can be intuitively seen as that one individual has one chance over 2*N* possibilities to have as ancestor the same parent than another given individual of the population. Thus, given the probability that two individuals share an ancestor in the previous generation, it becomes obvious than the probability that they do not share it is $1-\frac{1}{2N}$.

As the Wright-Fisher's model, the coalescent is Markovian, that is, its probability only depends on the present state of the process. This means that the probability that two individuals share a common ancestor *t* generations before present is the probability that they do not share it in the first *t*-1 generation multiplied by that probability that they share it in the *t* generation. Numerically, this probability is expressed as

$$\left(1-\frac{1}{2N}\right)^{t-1}\frac{1}{2N}.$$

Furthermore, assuming that the lineages coalesce independently, that on each generation only one coalescent event is allowed and that the number of generations since the MRCA is large enough to be modelled as continuous time, the probability that *k* different sequences coalesce is $\frac{k(k-1)}{4N}$. This happens because, as said above, the probability of any pair of lineages coalescing is $\frac{1}{2N}$, and there are $\frac{k(k-1)}{2}$ possible pairs of sequences.

Based on these probabilities, it is possible to go from the present backwards in time establishing a probabilistic genealogic relationship between the individuals composing the sample. As stated before, on each stage of the tree the coalescent probability will only depend on the actual stage. However, each step backwards in time will contain one individual less than the previous one, as two individuals will have coalesced on a single ancestor (see Figure 11).

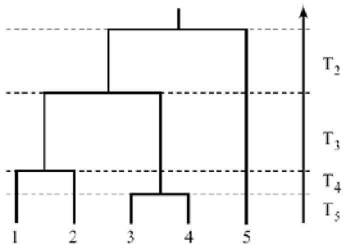


Figure 11: coalescent tree. Times are indicated as T_i, where *i* is the number of individuals in the sample at time T. At each point in time, there is an individual less than in the previous. (Hein *et al.* 2005).

Once the shape of the tree has been defined, mutation can be added on top of it (see Figure 12). Based on the standard neutral mutation model (Watterson. 1975; Kimura. 1984), the mutation process occurs independently in each individual and generation. Furthermore, the mutation rate μ is assumed to be constant and, thus, independent of the population size or the time.

Under the scenario described, mutation is thrown over the tree following a Poisson distribution with mean $2\mu t$, where t is the number of generations that two homologous sequences need to reach the MRCA. The consequence of this process is that the longest branches will accumulate a larger number of mutations than the shorter ones.

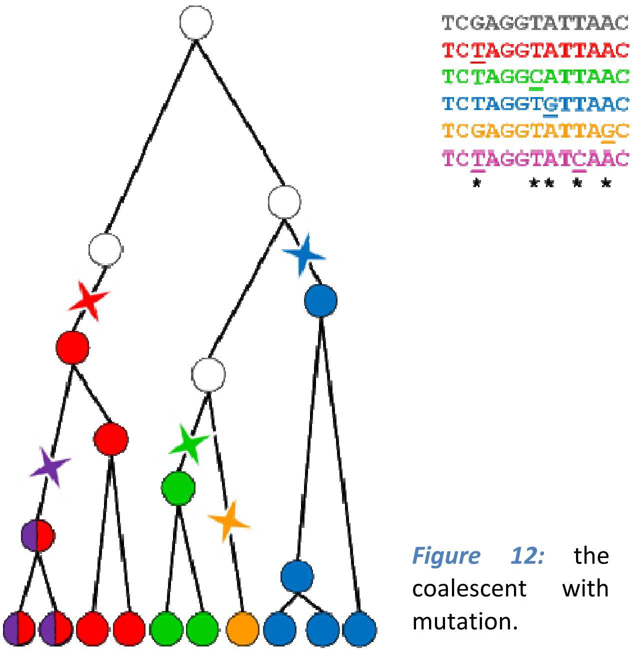


Figure 12: the coalescent with mutation.

3.2 Properties of the coalescent

Based on the probabilities stated in the previous point, some basic properties of the coalescent can be derived:

1. The number of coalescent events is directly proportional to sample size. As seen in equations above, if the number of lineages (k) increases, so does the number of possible pairs of sequences and, thus, the number of coalescent events until the MRCA.
2. The number of coalescent events is inversely proportional to population size, since in a larger population the probability that two individuals share a common ancestor is reduced (see equations above).
3. The expected time until the MRCA of a sample is
$$E[T_{MRCA}] = 4N\sum_{i=1}^{n-1}\frac{1}{i(i+1)}.$$
As the sample size grows, this time rapidly approaches 4N. That is, the MRCA for a sample rapidly approaches that for the whole population.
4. As stated above, in a neutral scenario the topology of the tree is independent of the mutational process.
5. The expected number of mutations between a pair of sequences is
$$2 \times E[T_{MRCA}] \times \mu = 4N\mu.$$

3.3 The coalescent with demography

The coalescent theory as stated above is strictly based on the Wright-Fisher model and follows all of its assumptions. However, the coalescent can be extended to more realistic models that include demographic events, selection, and even recombination. All these events can change the shape of the coalescent tree.

Among the demographic events, some of the most relevant for their effect on the tree topology are population expansions, bottlenecks, population subdivision, and migration.

3.3.1 Population expansions

In population expansions, each generation has a larger number of individuals than the preceding generation. For example, this is what could happen in a Neolithic population after several years of good harvests that allow the birth and survival of a larger amount of children.

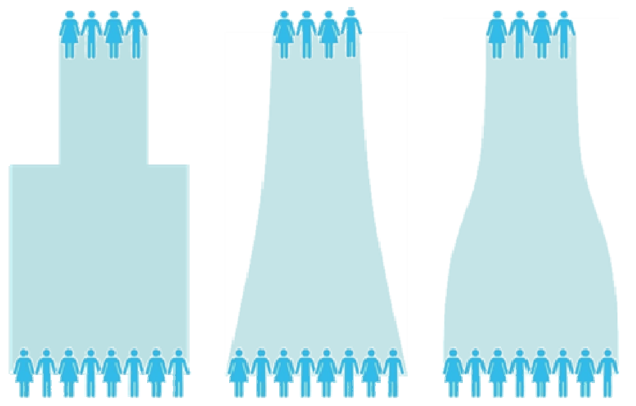


Figure 13: models of population expansion. From left to right are shown the sudden, the exponential, and the logistic model.

Despite this simple description of what a population expansion is, in coalescence it can be modelled in different ways. The simplest one is a size jump or sudden population expansion (Figure 13, left), in which the population changes its size in only one generation (Rogers and Harpending. 1992). In this model, an ancestral population in equilibrium of size N_0 experienced a sudden growth and reached maximum size $N_{\max}$ T_e generations before present. The strength of the expansions, that is, how much the population size increases, is defined by the degree of expansion $D_e = \frac{N_{max}}{N_0}$.

Another model of population increase is exponential growth (Slatkin and Hudson. 1991; Figure 13, middle). Exponential growth assumes a population in the present of size N_0 that has been growing at an exponential rate *r*. Then, its size *t* generations before present can be found by $N(t) = N_0 e^{-rt}$.

Finally, expansions can also be modelled as a logistic growth in population size (Fu. 1997; Figure 13, right). In this case a population has been growing at a logistic rate *r* from its initial size to its current size N_0. Now, the population size at time *t* is $\frac{dN}{dt} = r\left(1 - \frac{K}{N}\right)$.

As explained before, the probability that two individuals share a common ancestor in the generation before is $\frac{1}{2N}$ under neutrality. However, if the population size changes along time, this probability also changes with time because, as stated before, the probability of coalescence is smaller as *N* increases. Then, the new probability that two individuals coalesce in the previous generation will be a function of *t*: $p(t) = \frac{1}{2N(t)}$.

The simplest way to model a population expansion, then, is to create a neutral genealogy and, afterwards, compress or stretch the time before a coalescence occur according to the new *N*. That is, in those parts of the genealogy where *N* is smaller, the time to coalescence between two lineages will need to be shortened while, where *N* is larger, time will need to be extended.

Population expansions leave a particular footprint over genealogies. In fact, genealogies with underlying population growth are characterised by extremely long external branches and shortened internal ones, as shown in Figure 14. This happens because the most external branches represent the most recent times, where population size is large and, thus, the individuals in the sample need a long time to find a common ancestor in the populations. On the contrary, the branches near the MRCA represent generations with smaller *N*, where the probability of coalescence is larger.

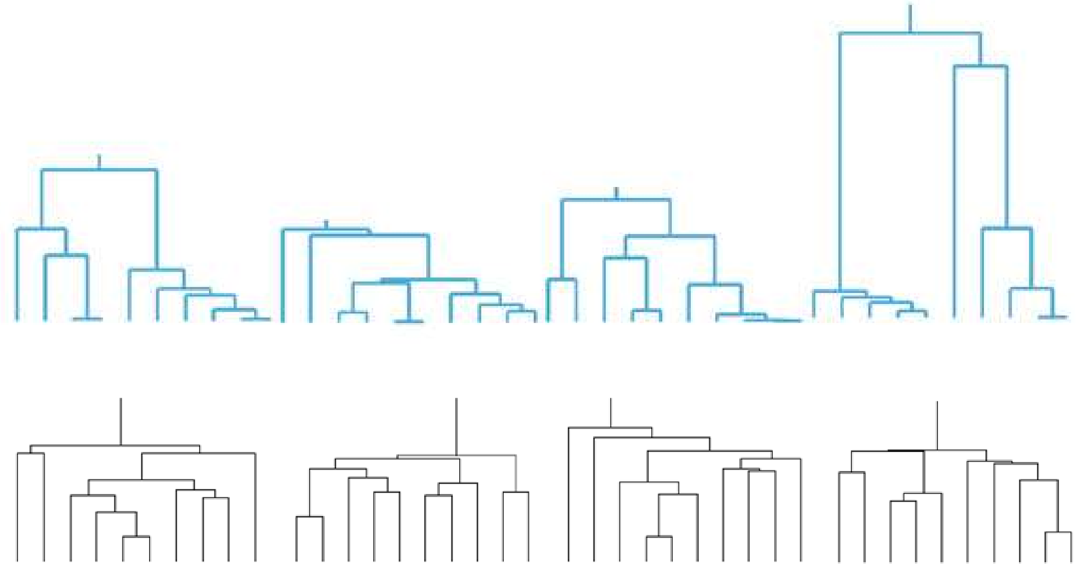


Figure 14: comparison between neutral genealogies (above, in blue) and genealogies with population expansions (below, in black). (Composed from Rosenberg. 2002 and Hein *et al.* 2005)

This effect on the genealogy will be more evident with larger increases in the number of individuals in the population, but it will also depend on sample size. For very large expansions the tree can even became star-like shaped, especially if the sample size is small.

This change in the topology of the tree will also affect the mutation pattern as mutation is placed on the tree following a Poisson distribution (see point 3.1). In this case, as the external branches will be much longer than the internal ones, mutations will tend to fall on them. This will result in an excess of singletons and mutations at low frequency compared with neutral genealogies.

3.3.2 Bottlenecks

Bottlenecks consist in a sudden decrease of a population size followed by a recovery or even an increase of the original population in a few generations. A typical example of a bottleneck is a plague such as the Black Death, which struck Asia and Europe in the fourteenth century. However, a founder effect in which a small subpopulation leaves its former habitat to establish a new one can also be considered a bottleneck. For instance, this is what happened when the first humans left Africa in what is known as the Out-of-Africa bottleneck (Reich *et al.* 2001).

The bottleneck can be modelled assuming a population of size N_A in equilibrium which, T_{start} generations ago, has been suddenly reduced by a factor $0<b<1$ to a second size bN_A. b is the bottleneck severity. The population maintains this second size bN_A for T_{dur} generations and, immediately afterwards, recovers instantaneously its original size N_A (Voight *et al.* 2005).

As in population growth, bottlenecks are a demographic event that changes population size. Thus, the same strategy used in expansions can be used to build demographies with underlying bottlenecks.

The effect of a bottleneck over the shape of the tree depends on T_{start} and b, that is, on the time since it started and on its severity. Using these two factors we can classify bottlenecks in severe and moderate, the former being older and causing a larger decrease in population size that the latter.

If the population is sampled just before the bottleneck, that is, before T_{start}, its genealogy follows neutrality. However, if it is sampled after the bottleneck, the topology of the tree will be different depending on the strength of the bottleneck and on the time since it finished, as shown in Figure 15.

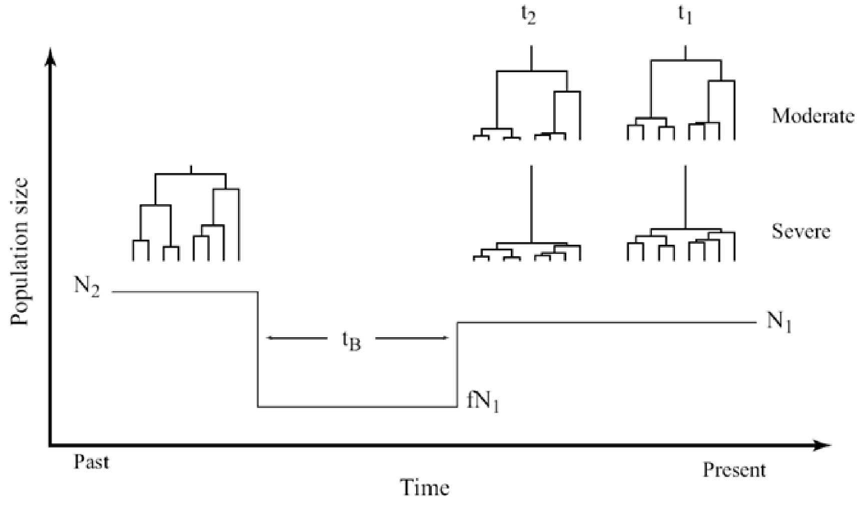


Figure 15: effect of a bottleneck on genealogies. (Hein *et al.* 2005)

It is expected that most lineages will die in a severe bottleneck. This implies that, just after the end of the bottleneck, the remaining individuals will presumably coalesce in the few surviving lineages and, thus, during the bottleneck. The result is a shortened genealogy whose individuals reach the MRCA much sooner than under neutrality. This genealogy, however, will change with time. A while after the end of the bottleneck, the time to the MRCA will increase, but all its lineages are expected to coalesce approximately at the time of the end of the bottleneck. At this moment, the genealogy is similar to that produced by a population expansion, with longer external branches compared to the internal ones. As in expansions, then, mutations will tend to accumulate in the external branches and will produce and excess of low-frequency variants.

On the other hand, in a moderate bottleneck it is expected that a number of lineages survives the bottleneck. This will result in longer internal branches, as the surviving lineages will tend to coalesce either just after the end of the bottleneck or before its beginning. At this time, mutations will tend to fall in the longest internal branches and will produce an excess of intermediate-frequency variants. After a while, the external branches will start to be longer, although most lineages will still coalesce at the start of the bottleneck. In this situation, mutations will be in excess both in the external branches and in those closer to the MRCA, what will produce an excess of singletons together with an excess of intermediate frequency variants.

3.3.3 Population subdivision and migration

The basic coalescent model assumes that two individuals of the population have the same probability to mate between them than with any other member of this population. However, this is not the case in most populations. Population subdivision can occur, for instance, due to geographic distance, such as in the case of individuals living in different continents. Nevertheless, it can also be found among individuals living in the same place but separated by social, linguistic or economical factors, as happens in India between the castes.

Barriers to mating, though, may not be absolute, with a number of migrants crossing between subpopulations each generation.

Population subdivision can be modelled in different ways, all of them including the possibility of migrants from one subpopulation to another.

One of those models is the finite island model (Li. 1976; Hudson. 1990; Hein *et al.* 2005), which assumes that the population is divided in *d* islands or demes of size 2*N* each which, together, sum 2*Nd* individuals. This model also assumes a given number of migrants *m* who can move to any other deme with equal probability. This implies that the parent of one individual will belong to his own deme with a probability of 1-*m* and to another deme with probability *m*.

Another group of models are the stepping-stone models (see Figure 16). The main difference between this and the finite island model is that in stepping stone models the demes are organized on a (one-dimensional or bidimensional) grid. This causes that the probability for migrants to move to one deme or another is no longer independent, as the individuals are allowed to move only to an adjacent deme. Moreover, the different demes can also have different probabilities to accept migrants. Another difference with the finite size model is that in this model the different demes do not need to have the same sample size.

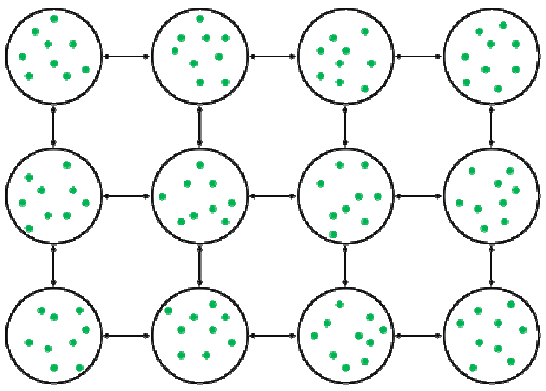


Figure 16: bidimensional stepping-stone model

Other models of population subdivision which have been formulated include continuous space models, in which individuals move through a continuous space, and non-equilibrium models, in which the demes have not reached a dynamic equilibrium and, thus, the number of total individuals they contain fluctuate through time.

The effect of population subdivision on the topology of the trees (Figure 17) is very similar to what happens just after a moderate bottleneck. As in moderate bottlenecks, the external branches will be short with respect to the internal ones. This happens because all the individuals in a deme coalesce very soon with respect to the coalescence time with members of another deme. With a smaller number of migrants this effect will be larger, and the time to the MRCA will also be longer. This scenario will produce an excess of mutations at intermediate frequencies.

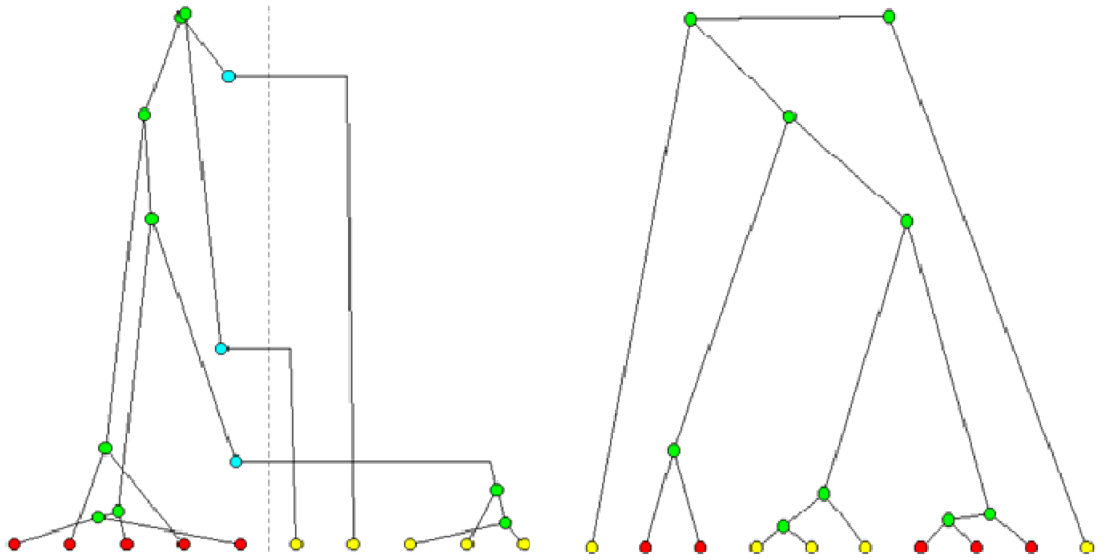


Figure 17: the coalescent with migration. Red and yellow represent the two populations, green are coalescent events and blue represent migration events. On the left it can be seen how the genealogy is produced, and on the right the same tree ordered to make it more readable. It can be seen that internal branches are long in respect to external ones. (Built using http://www.coalescent.dk/).

3.4 The coalescent with selection

Selection is another deviation of neutrality that affects the topology of the tree. In this section, it is described how balancing and positive selection modify the tree shape in coalescent theory.

3.4.1 Balancing selection

Balancing selection favours the presence of two or more alleles in the population, maintaining variation on it. One of the best known examples of this kind of selection is the case of sickle-cell anaemia and malaria. In populations where malaria is endemic, sickle-cell anaemia is more frequent than in other populations. This happens because the individuals with one sickle-cell anaemia allele are more resistant to malaria and, then, this allele, which is deleterious in homozygosity, is maintained in the population.

Balancing selection can be modelled considering biallelic or multiallelic models. The modelling and the consequences over the topology of the tree on each case are different, but I will focus on the first one.

In a biallelic situation (Kaplan *et al.* 1988; Hudson. 1990; Hein *et al.* 2005), in which two different alleles are maintained in the population, balancing selection can be modelled in a way analogous to the one used for population substructure, as shown on Figure 18. Consider two alleles, namely A and B, with a frequency of *p* and *q* respectively and a probability $\theta=4N\mu$ to mutate from one to the other. These two alleles, then, can be regarded as two independent demes of size 2*Np* and 2*Nq*, and θ can be assumed to be the migration rate between them.

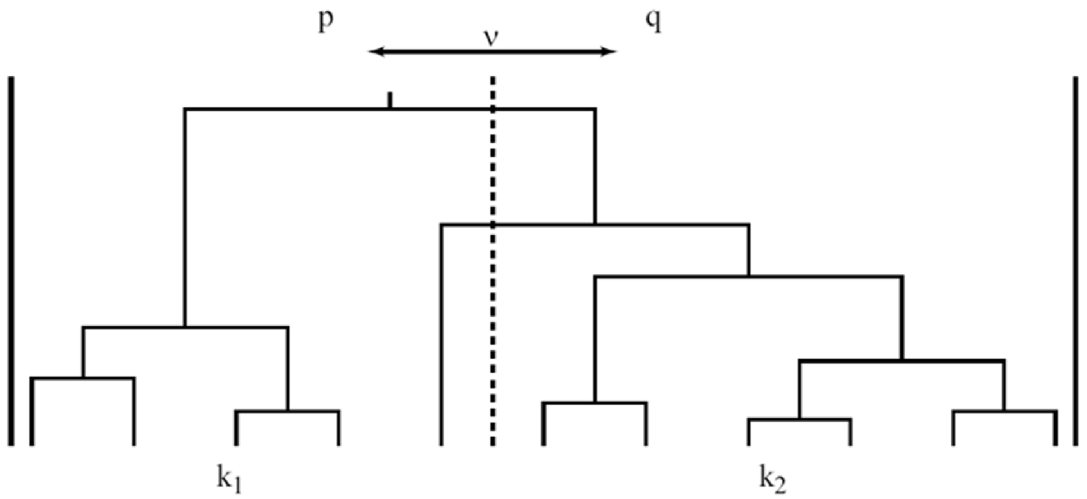


Figure 18: the coalescent with balancing selection can be modelled as the coalescent with migration. k_1 and k_2 are the two subpopulations, with alleles *p* and *q* respectively, and ν is the mutation rate from *p* to *q*. (Hein *et al.* 2005)

Under this scheme, it is assumed that two alleles will only be able to coalesce if they are of the same type. This means that two lineages with different alleles will only be allowed to coalesce after a mutation event transforming one allele into the other. Taking a sample of size n_A of the allele A and of size n_B of allele B, the probability of coalescence inside each type is

$$P_{CA} = \frac{n_A(n_A-1)}{2p}$$

and

$$P_{CB} = \frac{n_B(n_B-1)}{2q},$$

and the probability of mutation is

$$P_{AB} = n_A\frac{q\theta}{2p}$$

and

$$P_{BA} = n_B\frac{p\theta}{2q}.$$

Following these four probabilities it is possible to determine the kind of event –coalescence or mutation- that takes place at each point in time, and draw the genealogy accordingly.

As stated above, balancing selection can be modelled similar to population substructure, and thus will have the same effect on the genealogies. That is, the length of the tree will be larger than what is expected under neutrality and its internal branches will be longer. This will result in an excess of high-frequency variants.

The main difference between balancing selection and population substructure is that balancing selection will change the distribution of mutation in only one gene, while population substructure affects the whole genome.

3.4.2 Positive selection

Positive selection is a type of directional selection in which the selected allele is favoured, that is, the individuals carrying it will have more offspring. This would be the case of genes such as that coding for lactase, the enzyme that breaks lactose, the main sugar in milk. The ancestral lactase gene is active in babies to allow them to feed from their mother's milk, but it becomes inactivated in childhood. However, in these populations that relied in shepherding, the individuals that did not inactivate the lactase gene had an advantage as they could also feed from the milk of their cattle. In this population, then, positive selection acted favouring a gene that was active for all the lifespan of the individual.

When positive selection acts over a rare allele, it produces a rapid increase in frequency and, eventually, the fixation of this allele. However, this effect does not only take place on the selected variant but also on all mutations that are close to it in the genome. This way, a number of neutral alleles will also increase their frequency in the population. This is known as selective sweep or hitchhiking effect (Smith and Haigh. 1974).

Under a hitchhiking model (Kaplan *et al.* 1989; Braverman *et al.* 1995) two alleles are considered, namely b and B, where B has a selective advantage *s*. In continuous time, assuming a directional selection of strength $\alpha=2Ns$, for a large α the frequency of the allele will decrease backwards according to

$$x(t) = \frac{1-\epsilon}{\epsilon+(1-\epsilon)e^{-\alpha(t-\Delta t)}}$$

where $\epsilon = 5/\alpha$ (Kaplan *et al.* 1989). The process starts at $1-\epsilon$ and finishes at ϵ. If we call n_B the number of alleles in the sample linked to the favoured variant and n_b the number of alleles not linked to it, the probability of coalescence between two neutral alleles will be

$$P_C = \frac{\binom{i}{2}}{x(t)}\Delta t,$$

and the probability that they do not coalesce is

$$P_C = \frac{\binom{j}{2}}{1-x(t)}\Delta t.$$

The selective phase of the genealogy can be obtained following these probabilities. This phase will end if the genealogy reaches the MRCA or if x(t) becomes smaller than ϵ and it only remains one or any neutral allele in the sample not linked to the favoured allele. Once the selective phase is finished, the genealogy will keep building following the neutral coalescent.

After the end of a selective sweep in which the favoured allele has been fixed, all the individuals of the genealogy will coalesce soon before the beginning of the sweep. This will produce shortened trees with external branches longer than the internal ones. As in the case of population growth, this will result in the accumulation of mutation in those branches and, thus, an excess of singletons and low-frequency variants.

However, as shown in Figure 19, the effect of hitchhiking along the chromosome will not be homogeneous. If the sample is taken on the selected allele and the adjacent region, the tree will have a topology as the one described above. However, as we sample at longer distances from this point, the shape of the tree will change. At intermediate distances, trees will have very long internal branches, because most alleles will have as a common ancestor the selected allele but there will be a few that will be linked to the non-selected one. Both lineages, then, will need a long time to coalesce to the MRCA. At larger distances, the tree will become neutral.

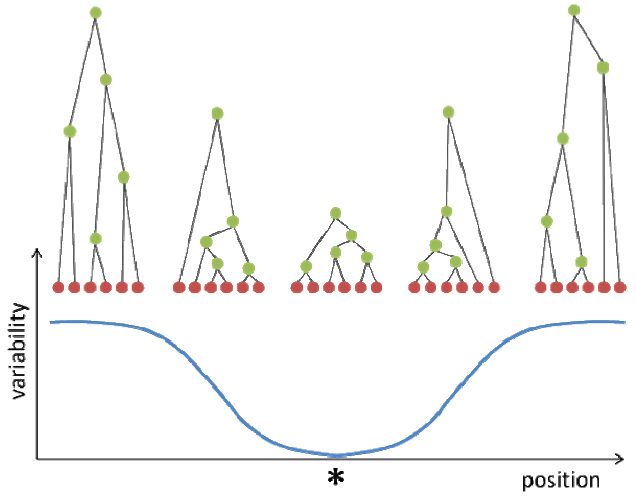


Figure 19: selective sweep. The selected position is denoted by *. Below is shown the variability along the chromosome, and above the coalescent trees for each part of the sequence.

3.5 The coalescent with recombination

Another assumption of the Wright-Fisher model that is highly unrealistic in most living organisms is the absence of recombination. In humans, for instance, all the genetic material recombines, with the only exception of the NRY chromosome and mitochondrial DNA.

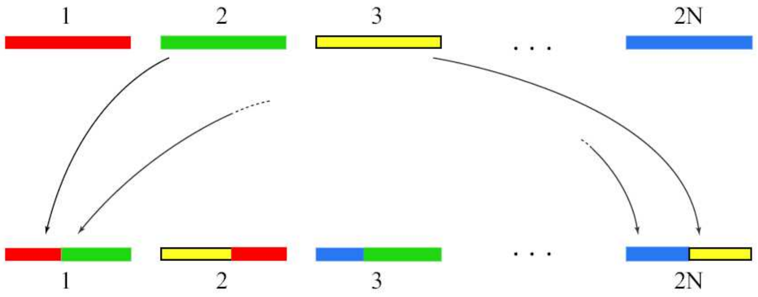


Figure 20: the haploid Wright-Fisher model with recombination (coloured from Hein *et al.* 2005)).

Recombination is a process that shuffles nucleotide variation among DNA sequences, thus creating new combinations. This implies that when recombination acts over a sample, a given sequence has not a single ancestor but two (or more) of them. If this is viewed forwards in time, as in the Wright-Fisher model, it will appear that each bit of the sequence will choose a different parent (Figure 20). To represent this situation, then, instead of using a genealogical tree, a group of local trees will be needed, one for each position in the sequence.

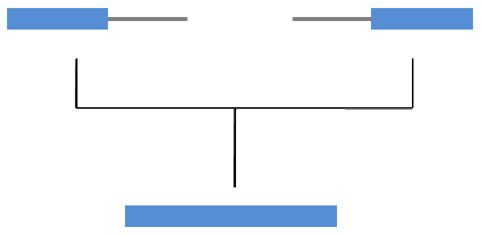


Figure 21: recombinant sequence and its ancestors. The fragment of the sequence found in the present sample and, thus, known, is shown in blue. The rest of the sequence, in grey, is not known.

In the coalescent, looking backwards in time, a sequence is formed by the combination of two ancestral sequences from which only the part that has left descendants is known (see Figure 21). In this situation it is assumed that coalescent and recombination events cannot happen at the same time. This is so because the probability that the two events occur together in the same sequence is $\frac{1}{8N^2}\rho$, that is, the probability of coalescence ($\frac{1}{2N}$) multiplied by the probability of recombination ($\frac{1}{4N}\rho$, where ρ is the recombination rate). In large populations, this probability is negligible. Again, this situation will produce different trees along the sequence.

The most widely used algorithm to construct the coalescent with recombination is the ancestral recombination graph (ARG, see Figure 22). This algorithm starts with a sample of *n* sequences and its first step is to determine the time to the next event. Once this time is set, the algorithm decides whether it is a coalescent event (with probability $\frac{n-1}{n-1+\rho}$) or a recombination event. If it is a recombination event, a sequence is chosen at random to be split at a random point into two ancestral sequences and the sample size *n* is increased by one. Otherwise, if the event is a coalescence it merges two randomly chosen ancestral sequences into a new sequence made of material from both of them. In this case, *n* is decreased by one. Once the new *n* has been obtained, the algorithm starts from the beginning until there is only one sequence left (the MRCA).

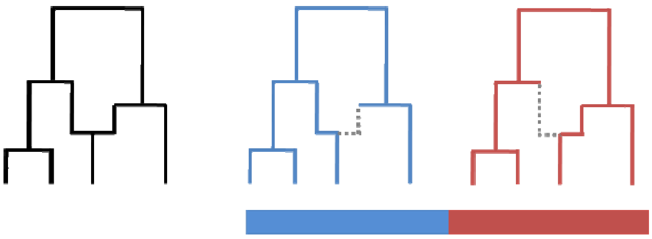


Figure 22: on the left, the ARG. On the right, the same ARG deconstructed, showing the particular genealogy of each fragment of the secuence.

4 NEUTRALITY TESTS

As discussed above, deviations from neutrality change the shape of the genealogies and, thus, their mutation pattern. This has led to the development of a number of statistical tests –the neutrality tests- aimed to explore different aspects of how, from the genetic diversity, it is possible to infer deviations from what is expected under a neutral model.

Coalescent theory has revealed itself to be a powerful tool in the development, study and use of neutrality tests. In fact, coalescent theory allows not only to know whether a sequence of DNA is neutral or non-neutral but also to obtain the direction of the deviation and the statistical significance of this deviation. This can be achieved through computer simulations, which make it possible to generate neutral distributions for a given neutrality test against which to compare the value obtained with the same test in an empirical sequence. Examples of this kind of work can be found in the following section.

The possibility to generate non-neutral genealogies has also allowed seeing how the different neutrality tests behave in the presence of deviations of neutrality. For instance, it is well known that the distribution of one of this statistics, Tajima's *D*, is centred on 0 under neutrality. However, under positive selection or population expansion, Tajima's *D* becomes negative, while under balancing selection or a substructured population it becomes positive. The causes of this effect will be explained in detail below, in section 4.1.1.

Although all neutrality statistics are based on genomic variation, not all of them rely on the same kind of information. This led Ramos-Onsins *et al.* (2002) to classify them into three classes, named Class I, II, and III, according to the information used. Class I tests are based on the frequency spectrum of mutations, Class II on the haplotype distribution, and Class III on the distribution of pairwise differences.

4.1 Class I tests

Class I statistics use information on the frequency of mutations in the sample. Most of them are based on the differences between two estimators of the population mutation rate $\theta=4N\mu$. From this class, the most relevant tests are Tajima's *D* (Tajima. 1989); Fu and Li's *D*, *F*, *D** and *F** (Fu and Li. 1993); Fay and Wu's *H* (Fay and Wu. 2000), and R_2 (Ramos-Onsins and Rozas. 2002).

4.1.1 Tajima's D

Tajima's *D* (Tajima. 1989) is the oldest neutrality test and one of the most widely used. This statistic is based on the standardized difference between the average pairwise difference, π, which takes into account the number of differences between two sequences; and the Watterson's estimator of θ, θ_W (Watterson. 1975), based on the number of segregating sites. Its equation, then, is

$$D = \frac{\theta_\pi - \theta_W}{\sqrt{Var\ (\theta_\pi - \theta_W)}},$$

where

$$\theta_{\pi = \sum_{ij} x_i x_j \pi_{ij}}$$

and

$$\theta_{W=\frac{S}{\sum_{i=1}^{n-1}\frac{1}{i}}},$$

being *n* the number of chromosomes in the sample and *S* the number of segregating sites.

Under neutrality, the two estimators of θ are equivalent, and both predict the true value of θ=4$N\mu$. This is the reason why, under neutrality, the distribution of values of Tajima's D is located around 0. Moreover, as it is normalized, its variance is 1, although it does not follow a normal but a beta distribution.

In case of positive selection or of population expansion, however, an excess of singletons and low frequency variants is found. If this happens, the number of segregating sites *S* will be too large compared to π and, thus, θ_W will be larger than θ_π. Therefore, the obtained values of Tajima's D will now be negative, and more negative as the larger is the deviation from neutrality. On the contrary, in case of balancing selection or population substructure an excess of intermediate frequency variants will be found and *S* will be too small compared to π, leading to positive Tajima's D values.

4.1.2 Fu and Li's tests

Fu and Li's tests (Fu and Li. 1993) are a group of tests based on the comparison between an estimator of θ and the number of derived unique mutations in external branches of the genealogy. Those tests are *D*, *F*, *D** and *F**, and the main difference between them is that the first two need an outgroup, while the latter two do not.

Fu and Li's *D* (or D^F) is computed from the normalised difference between θ_W and the expected number of derived mutations, that is,

$$D^F = \frac{S-a_n\eta_e}{\sqrt{Var(S-a_n\eta_e)}},$$

where

$$a_n = \sum_{i=1}^{n-1}\frac{1}{i}$$

and η_e is the number of derived singletons in the sample (that is, excluding singletons in the outgroup). The *F* statistic is very similar to *D* but, instead of θ_W, it uses π:

$$F = \frac{\theta_\pi-\eta_e}{\sqrt{Var(\theta_\pi-\eta_e)}}.$$

In both cases, the statistic is based on the idea that, under neutrality, the expected number of external mutations $E[\eta_e] = \theta_W = \theta_\pi = 4N\mu$.

However, it is not always possible to have an outgroup and, thus, to know whether a singleton is derived or ancestral. Taking all singletons as derived obviously overestimates the number of derived singletons. To solve that, Fu and Li provided two more statistics, *D** and *F**, which correct this overestimation:

$$D^* = \frac{\frac{n}{n-1}S-a_n\eta_s}{\sqrt{Var\left(\frac{n}{n-1}S-a_n\eta_s\right)}}$$

and

$$F^* = \frac{\frac{n}{n-1}\eta_s-\theta_\pi}{\sqrt{Var\left(\frac{n}{n-1}\eta_s-\theta_\pi\right)}},$$

where η_s is the total number of singletons in the sample.

4.1.3 Fay and Wu’s H

Fay and Wu’s *H* (Fay and Wu. 2000) is a neutrality test based on the standardised comparison between π and θ_H that is defined as

$$H = \frac{\theta_\pi - \theta_H}{\sqrt{Var\ (\theta_\pi - \theta_H)}}.$$

θ_H is a new unbiased estimator of θ developed by themselves in the same paper, that gives more weight to high-frequency derived variants. This estimator is a particular case of a more general class of estimators derived by Fu (Fu. 1995), based on the expected number of mutations with a derived frequency *i* in the sample. θ_H, then, was defined as

$$\theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 S_i,$$

where S_i is is the number of derived variants found *i* times in the sample.

As in Tajima’s D and Fu and Li’s tests, under neutrality those two estimators of θ are expected to be $4N\mu$. Furthermore, Fay and Wu’s *H* is specially indicated to detect selective sweeps, as they increase the frequency of derived variants.

4.1.4 R_2

The R_2 statistic (Ramos-Onsins and Rozas. 2002) behaves differently than other Class I tests, as it is not based on the difference between two different estimators of θ and, besides, it is not normalised, so its mean and variance are not 0 and 1 respectively.

R_2 is a test thought to be efficient to detect population expansions that is based on the difference between the number of singletons per sequence and the average number of nucleotide differences. It is computed as

$$R_2 = \frac{\left(\frac{\sum_{i=1}^{n}\left(U_i - \frac{k}{2}\right)^2}{n}\right)^{1/2}}{S},$$

where U_i is the number of singletons in sequence *i* and *k* is the average number of nucleotide differences between two sequences. After a population expansion R_2 is expected to decrease.

Furthermore, the authors developed another statistic R_{2E} which takes into account the derived variants and, thus, requires an outgroup; together with other statistics R_3, R_4, R_{3E} and R_{4E} which are equivalent to R_2 and R_{2E} but with exponents $\frac{1}{3}$ and $\frac{1}{4}$ respectively.

4.2 Class II tests

Class II includes statistics based on the haplotype distribution of the sample. Within this class the most relevant statistics are Fu’s F_S (Fu. 1997), *Dh* (Nei. 1987; equation 8.4 replacing 2n by n), Wall’s *B* and *Q* (Wall. 1999), Kelly’s Z_{nS} (Kelly. 1997), Rozas’ Z_A and *ZZ* (Rozas *et al.* 2001) and extended haplotype homozygosity *EHH* (Sabeti *et al.* 2002).

4.2.1 Fu’s F_S

Fu’s F_S (Fu. 1997) is a neutrality test based on the Ewens’ sampling distribution (Ewens. 1972) which takes into account the number of different haplotypes in the sample.

F_S is defined as

$$F_S = \ln\left(\frac{S\prime}{1-S\prime}\right),$$

where *S′* is the complementary of the Strobeck's statistic *S* (Strobeck. 1987) and corresponds to the probability of having the same or a higher number of mutations than expected under neutrality. *S′* is then computed as

$$S' = \sum_{k \geq k_0} \frac{|S_k|\theta_\pi^k}{S_n\theta_\pi},$$

where *k* is the number of alleles in the sample, k_0 is the number of alleles expected under neutrality and $S_n\theta_\pi$ and S_k are defined according to the Ewens' sampling distribution.

F_S, then, in expected to be negative if low frequency mutations (and, thus, haplotypes) are in excess in the sample. This also implies that it is a one sided test, which is expected to be efficient to detect positive selection and population expansions.

4.2.2 Dh

Haplotype diversity (*Dh*) or gene diversity (Nei. 1987, equation 8.4 replacing 2n by n) is a measure of the heterogeneity present in the sample. In fact, *Dh* is an unbiased estimate of the heterozigosity, and is defined as

$$Dh = \frac{n(1-\sum x_i^2)}{2n-1},$$

where x_i is the frequency of the allele *i* in the sample.

The expectation of *Dh* is not known beforehand, and its distribution must be simulated under neutrality before it can be used as a neutrality test.

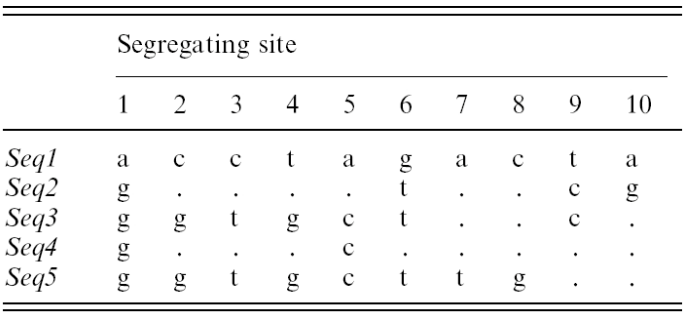
4.2.3 Wall's tests

Wall's tests (Wall. 1999), named *B* and *Q*, were developed to detect events that produce trees with relatively longer external branches, such as under balancing selection or population substructure.

Both *B* and *Q* are based on what is called *congruent* sites, that is, pairs of adjacent segregating sites which, if taken as a subset, form only two possible haplotypes (see Figure 23). *B*, which has been scaled between 0 and 1, is defined as

$$B = \frac{B\prime}{(S-1)},$$

where *B′* is the number of congruent pairs of adjacent segregating sites.



	Segregating site									
	1	2	3	4	5	6	7	8	9	10
Seq1	a	c	c	t	a	g	a	c	t	a
Seq2	g	.	.	.	.	t	.	.	c	g
Seq3	g	g	t	g	c	t	.	.	c	.
Seq4	g	.	.	.	c	.	.	.	.	.
Seq5	g	g	t	g	c	t	t	g	.	.

Figure 23: table showing ten SNPs in five sequences. This produce nine pairs of sites, from which three are congruent: 2-3, 3-4 and 7-8. The first two pairs induce the same partition, as they have the same combination of haplotypes. Instead, the pair 7-8 induces a diferent partition. (Wall. 1999)

The Q statistic adds a level of complexity including also the number of different partitions defined by congruent pairs. A partition is a subset of congruent SNP pairs, in which the two haplotypes defined by those SNPs are carried by exactly the same chromosomes (see Figure 23). The SNP pairs that define a partition need not be adjacent to each other. Considering A the total number of different partitions, Q is defined by

$$Q = \frac{B+|A|}{S}.$$

4.2.4 Z_{nS}, Z_A and ZZ

Z_{nS} (Kelly. 1997), Z_A, and ZZ (Rozas *et al.* 2001) are a group of statistics developed by two different authors that are based on the linkage disequilibrium (LD) measure r^2.

Z_{nS} uses information of the r^2 values between all pairs of polymorphic sites, and thus is defined as

$$Z_{nS} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} r_{ij}^2,$$

where r^2 among a pair of loci, namely i and j, is computed as

$$r_{ij}^2 = \frac{D_{ij}^2}{p_i(1-p_i)p_j(1-p_j)}.$$

p_i and p_j correspond to the frequencies of the mutant alleles i and j and D_{ij} is the mesure of LD between both loci, which is

$$D_{ij} = p_{ij} - p_i p_j.$$

Z_A is very similar to Z_{nS}, but only takes into account r^2 values between adjacent pairs. Is defined as

$$Z_A = \frac{1}{S-1} \sum_{i=1}^{S-1} r_{ij}^2.$$

Finally, ZZ is defined as

$$ZZ = Z_A - Z_{nS},$$

and thus is the difference between Z_A and Z_{nS}. Therefore, ZZ provides information about intragenic recombination, and is expected to become increasingly positive as recombination increases.

4.2.5 EHH and EHH-based tests

Unlike the other statistics explained, the extended haplotype homozigosity (*EHH*) test (Sabeti *et al.* 2002) is a heuristic test. Thus, its significance cannot be stated by means of computer simulations but by using as neutral model large amounts of empirical data. However, due to the relevance it has acquired to detect recent positive selection since it was first described, it will be included in this section.

EHH (see Figure 24) is built upon the realisation that, under neutrality, young mutations are found at low frequency and in areas with a long-range LD, while old mutations can be found at low or high frequency but surrounded by short-range LD. This happens because in young alleles, recombination has not had time to break LD, but it has had in old alleles. On the contrary, if a selective sweep has taken place, the selected variant will be found at high frequency and in a region with long-range LD.

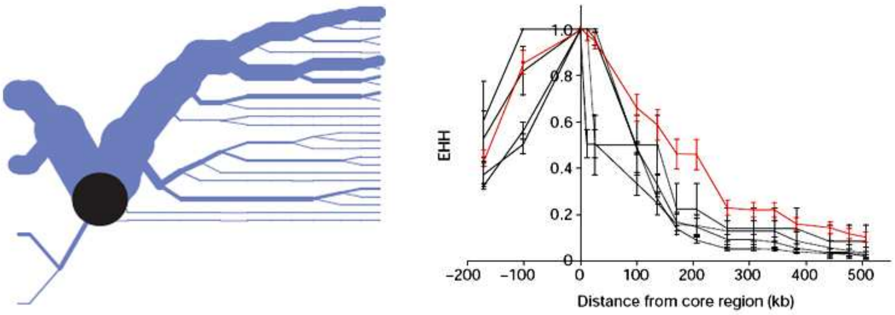


Figure 24: left, haplotype bifurcation diagram. Right, *EHH* values at different distances from the core region (at position 0). (Modified from Sabeti *et al.* 2002)

EHH is found by selecting a small region called "core haplotype" where to genotype a high density of SNPs, and adding other SNPs at lower densities and longer distances. This protocol, then, allows studying how LD decays along the region. The decay of the haplotype along distance goes from 1 (at the core) to 0. Under positive selection it is expected that the identity of the haplotype is maintained at a longer distance than under neutrality.

More recently, Voight *et al.* (2006) developed a normalised point estimator based on *EHH*, the *iHS*, defined as

$$iHS = \frac{ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]},$$

where E_p and SD_p are the expectation and standard deviation of $ln\left(\frac{iHH_A}{iHH_D}\right)$ conditioned to the frequency of the derived allele; and *iHH* is the area under curve of the decay of *EHH* until it reaches 0.5, that is, the integral of this curve. iHH_A corresponds to the integrated curve for the ancestral allele and iHH_D to the integrated curve for the derived allele.

4.3 Class III tests

Class III tests are those based on the distribution of pairwise differences or mismatch distribution. The most relevant among them are the raggedness (*rg*) statistic (Harpending *et al.* 1993; Harpending. 1994) and the mean absolute error (*MAE*) between the observed and the theoretical mismatch distribution (Rogers *et al.* 1996).

4.3.1 Raggedness statistic

The raggedness (*rg*) statistic (Harpending *et al.* 1993; Harpending. 1994) is a test developed to detect populations expansions that is based upon the realisation than, under neutrality, the mismatch distribution of a sample has ragged peaks. However, after population increase the distribution will be much smoother. The *rg* statistic, then, quantifies the smoothness of the mismatch distribution, and is defined as

$$rg = \sum_{i=0}^{d-1}(x_{i+1} - x_i)^2,$$

where *d* is the maximum number of differences between haplotypes and *x* is the observed relative frequencies of the mismatch classes.

4.3.2 Mean abolute error

The mean absolute error (*MAE*) between the observed and the theoretical mismatch distribution (Rogers *et al.* 1996) is a mismatch-distribution based test aimed to minimize the dependence that previously developed Class III tests had on the infinite-sites mutation model.

This test is based on the *MAE* function, which is described as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|t_i - o_i|,$$

where t_i is the value of the theoretical mismatch distribution and o_i is the value of the observed mismatch distribution, each at *i* pairwise differences.

5 EXAMPLES OF THE NEUTRALITY TESTS APPLICATIONS

Neutrality tests, together with the coalescent theory, have become the most useful tool available to detect departures from neutrality. Obtaining the value of a wide range of statistics from the empirical data and comparing them with their neutral distribution, build by means of the coalescent theory, provides a straightforward rationale to recover information of demographic and selective events. Furthermore, the possibility to do that in a simple and automated way using publicly available software such as DnaSP (Rozas *et al.* 2003) has made neutrality tests the tool of election in a large number of papers focused to disentangle the evolutionary history of populations (Canino and Bentzen. 2004; Nakajima *et al.* 2004; Verrelli and Tishkoff. 2004; Schmid *et al.* 2005; Civetta *et al.* 2006; Patin *et al.* 2006; Soejima *et al.* 2006; De Mita *et al.* 2007; Pinto *et al.* 2007; Sanchez-Gracia and Rozas. 2007; Derome *et al.* 2008; Wright *et al.* 2008; Alonso *et al.* 2008;). Below, I summarize a few of these applications.

5. 1 Unusual pattern of nucleotide sequence variation at the OS-E and OS-F genomic regions of *Drosophila simulans*

The sense of smell plays an important role in the interaction of most animals with their environment, and the genes coding for olfactory system proteins have been shown to be under positive selection in a number of organisms such as humans, rodents or insects.

As olfaction is crucial for invertebrates, Sánchez-Gracia and Rozas (2007) studied the evolution of two members of the odorant-binding protein (OBP) gene family, *OS-E* and *OS-F*, in *Drosophila simulans*. OBPs are in charge of transporting odor molecules to the odorant receptors. The results for *D. simulans* are further compared with the results of *D. melanogaster* found in a previous study.

In this work, the authors sequenced 11 European lines and 11 African lines of *D. simulans*, the latter from Rozas *et al.* (2001). Moreover, they used 14 lines of *D. melanogaster* and several lines of *D. mauritiana* and *D. erecta* from a previous work. They used Class I and Clas II statistics to detect departures from neutrality. From the former they used Tajima's *D*, Fu and Li's *D* and *F* and Fay and Wu's *H*, and from the latter Fu's F_S, *Dh*, Wall's *Q* and *ZZ*. Tests designed to detect departures from neutrality based on interespecific differences were also used.

In European samples, Tajima's *D*, Fu and Li's tests and Fay and Wu's *H* were non-significant. *ZZ* was significantly positive, suggesting that intragenic recombination has played an important role shaping variation in this region. F_S and *Dh* were also significant, reflecting a reduction in the number of haplotypes respect to the neutral model. Wall's *Q* was used to test whether the data was compatible with a bottleneck model. In the absence of recombination, *Q* was compatible with recent bottlenecks. However, when higher recombination rates were applied, the number of bottlenecks compatible with data was reduced to only the most recent ones. No test of neutrality was significant for African samples.

5.2 Multiple origins of knockdown resistance mutations in the Afrotropical mosquito vector *Anopheles gambiae*

Antibiotics and insecticides are useful to battle against plagues and pests. However, it is well known that populations subjected to these kinds of treatments quickly develop resistance to them, becoming immune. Besides the problems than this phenomena poses to the plague control, it also represents a great opportunity to study evolution at work, particularly directional selection.

Pinto *et al.* (2007) studied the emergence of the resistance to DDT and pyrethoid insecticides in *Anopheles gambiae sensu stricto*, the main vector of malaria in Africa. In this mosquito, resistance to insecticides can be reached through two point mutations, both of which inactivate a voltage-gated sodium channel. To unravel the history of these two mutations conferring resistance to insecticides they genotyped the *kdr* locus and the downstream region of intron 2, and sequenced intron-1 in 288 individuals from Western Africa, West-Central Africa and East Asia. With this information they (a) analyzed the frequency, distribution and genealogic relationship of the knock-down resistance (*kdr*) haplotypes and (b) used neutrality tests to detect traces of selection, especially of recent selective sweeps, in the sample.

The analysis of haplotypes suggested that the *kdr* alleles have four independent origins, two for each one of the two mutations causing resistance. To detect departures from neutrality they used Tajima's *D*, Fu and Li's *D** and *F**, and F_S statistics, and tested them separately in the three geographical groups of mosquitoes. Only F_S and Fu and Li's *F** showed any significance, the former in West and West-Central Africa and the latter in West Africa. The authors explain the lack of significance in other tests, together with the low genetic variation found, through a process of hitchhiking. Moreover, they found a higher significance of F_S in West Africa that they attribute to an intensive use of insecticide in the area in the last 20 years accompanying the increase of cotton production.

5.3 Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes

The human arylamine N-acetiltransferase genes are a genic family formed by the two genes *NAT1* and *NAT2*, and by the *NATP* pseudogene. Those genes code for two phase-II enzymes (*NATs*) involved in the metabolism of various drugs and carcinogens through their acetilation. Thus, variation in these genes is related to cancer susceptibility. However, some of the variants of the *NAT2* gene that have been shown to confer a higher risk for bladder cancer by coding for a "slow-acetilator" phenotype are at high frequencies in populations worldwide.

As those genes are involved in interaction with environmental factors and, thus, can easily be targets for selection, Patin *et al.* (2006) decided to unravel the demographic and selective history of *NAT* genes. They studied a sample of 80 humans on eight worldwide populations and resequenced and genotyped several fragments of the *NAT* loci, including the entire coding exon of *NAT1* and *NAT2*.

The analysis of the haplotypes in *NAT1* showed two clusters of haplotypes that share their *MRCA* 2.01±0.29 million years ago, an extreme estimated date compared to the rest of the genome, which significantly departs from neutrality. According to the authors, the most likely scenario to explain *NAT1* genealogy is an ancient population substructure demographic event. *NAT2* and *NATP* haplotypes genealogy coalesce at times that are in agreement with neutral expectations.

To find traces of selection over *NAT* genes they used three Class I neutrality tests: Tajima's *D*, Fu and Li's *F**, and Fay and Wu's *H*, as well as several interespecific tests. They performed independent tests for *NAT1* and *NAT2* as they are not in LD. Those tests were significantly negative in most of the populations showing variation for the exonic region of *NAT1*. However, when the flanking regions were included only two populations remained significant. The authors suggest that these values could be due to purifying selection acting over the *NAT1* gene. On the contrary, *NAT2* exonic region was significantly positive in three populations, but none remained if the flanking regions were included. These results point to the role of selection acting locally on *NAT2* locus, although interespecific tests do not depart significantly from neutrality. *NATP* was not significant in any population or test.

The authors also used the long range haplotype test (LRH), based on *EHH*, on genotyping data to detect traces of recent positive selection. One *NAT1* and one *NAT2* haplotype were found to depart from neutrality in Eurasian populations. This *NAT2* haplotype leads to the "slow-acetilator" phenotype and, in western and central Eurasians, was associated with the *NAT1* haplotypes in approximately 80% of cases. This suggests that both haplotypes could have been selected in a single selection event. Furthermore, the evidence of positive selection favouring an allele causing cancer nowadays points to the changes to the environmental carcinogens at which humans have been exposed along their evolution. These changes could be mainly due to the Neolithic and the Industrial revolutions, the two main events affecting human lifestyle.

In this work, then, neutrality tests have been used both to describe demographic events (in the *NAT1* gene) and to detect selection (*NAT2* gene).

5.4 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism

It is often difficult to distinguish between demographic and selective events when deviations from neutrality are found in a sample. This happens because some of them leave the same traces over the genome. However, their range of action is radically different, as demography affects the whole genome while selection acts locally. Thus, if similar deviations from neutrality are found in a number of different loci, this suggests that they fit in a model of demographic change.

Schmid *et al.* (2005) used this rationale to unravel the demographic history of *Arabidopsis thaliana*, a model organism of genetic variation in plants. With this goal, they sequenced 595 short genomic regions (STS) selected at random in 12 accessions. The STS were located both in coding and non-coding regions. They also used data from other species of *Arabidopsis* as outgroup.

The authors calculated Tajima's *D*, Fu and Li's *D*, *D**, *F* and *F** and Fay and Wu's *H* tests over the STS and compared the obtained distribution to different demographic scenarios, in order to detect the demographic events underlying *A. thaliana* history. When tested against neutrality, all statistics were significantly negative with the exception of Fay and Wu's *H*, which fitted the neutral model. This indicated that derived alleles are not in excess in the sample, which could have been interpreted as a signal of selection. However, neutrality statistics do not fit in logistic population growth, glacial refugia or bottleneck models. Thus, a more complex model should be built to explain *A. thaliana* demographic history.

Furthermore, they found 28 STS loci which were outliers in respect to the empirical distribution of the neutrality statistics, and thus they are candidates to have evolved under positive selection. When these loci were excluded from the analysis, the empirical distribution fitted the logistic growth model.

6 LIMITATIONS OF NEUTRALITY TESTS

Despite their advantages as a tool to detect departures from neutrality, neutrality tests have also several limitations that should be taken into account when using them. These limitations can be classified into two groups according to whether they relate to the neutral distribution or to the protocols designed to obtain information about the segregating sites in the sample.

6.1 The neutral distribution

As explained above, the rationale for using neutrality tests requires comparing them against a neutral distribution. However, as stated in section 2, the Wright-Fisher model and the basic coalescent make several assumptions that are unrealistic in most populations. Thus, comparing the value of a statistic against a purely neutral distribution will lead to erroneous conclusions. This implies that the neutral distribution to be used for assessing the significance of a neutrality statistic should not be neutral in the strict sense of the word, but it should take into account the demographic and recombinatory history of the sample under analysis. Nevertheless, building this more realistic neutral distribution is not trivial, as it requires an accurate knowledge of the forces shaping the genetic variation of the species of interest. However, this information is known, in the best of cases, only partially.

The demographic events that have taken place in the past of the population of interest are often a source of conflict when looking for selection. As seen above, demographic and selective events tend to leave very similar traces over sequences. This makes it difficult to interpret an extreme value of a neutrality statistic, as it often cannot be known whether it is due to demography or to selection. In the case of humans, for instance, it is well known that they have experienced several bottlenecks and expansions, among which the out-of-Africa bottleneck and their further expansion through the world (Jobling *et al.* 2004). An out-of-Africa bottleneck leading to extant worldwide populations has also been described in *Drosophila melanogaster* (Andolfatto. 2001; Baudry *et al.* 2004). As a consequence of these changes in population size, negative values of Tajima's *D* that are significant compared to a neutral distribution cannot be directly interpreted as the result of local selection. In the vast majority of cases, those significantly negative values will only reflect the demographic history of populations, and only the most extreme will point to positive selection.

This effect can be clearly seen looking at empirical distributions of the neutrality statistics, such as those publicly available at SeattleSNPs (http://pga.mbt.washington.edu/; Crawford *et al.* 2005), a database containing the sequences of more than 300 human genes related to inflammation in 24 African-American and 23 European-Americans individuals. The average Tajima's *D* over all genes in this database is -0.6, and not 0 as it would be expected under a neutral model. This was extensively studied by Stajich and Hahn in 2005 (2005) in the 151 genes from SeattleSNPs available at that moment. In their work, they calculated several neutrality tests and summary statistics -among which π (Tajima. 1983), θ (Watterson. 1975), Tajima's *D* (Tajima. 1989), Fu and Li's *D*, *D**, *F*, and *F** statistics (Fu and Li. 1993) - for European-Americans and African-Americans separately. They showed that the values of neutrality statistics for most European data can be explained by a population bottleneck, while African-American results fit a model considering admixture between subpopulations with an underlying bottleneck. However, they could also identify two loci which were extreme even when compared to a neutral model with the described demographic history and, thus, which have been under selection: *ABO* and *TRPV6*.

It becomes clear, then, that the solution for finding traces of selection acting over some genes is to take into account the demographic history of the populations. This can be done following two strategies: a) comparing the values of the statistical tests not against a coalescent build neutral model but against an empirical distribution, as done by Stajich and Hahn (2005), and b) simulating neutral models that incorporate all the information known about demography, with as much detail as possible. In humans, for instance, Schaffner *et al.* (2005) published a calibrated demography that, although it does not exactly fit the demographic processes shaping human evolution, it is consistent with a variety of statistics and measures as obtained from empirical data.

The second main assumption of the neutral model that is known to affect the power of neutrality tests is the lack of recombination (Wall. 1999). In fact, recombination breaks the existent haplotypes and shuffles them creating new haplotypes and, thus, increasing their number and causing decay in LD. Furthermore, it smoothes the mismatch distribution. For this reasons, recombination is expected to affect mainly the power of Class II and Class III tests, while it is not expected to have much effect on Class I ones.

As in the case of demographic events, the best solution to avoid errors when testing neutrality statistics against neutral distributions is to include accurate information about recombination rates in them. In experimental species this is not particularly difficult, as the possibility to perform directed crossings between individuals and to obtain large pedigrees has made possible that the recombination rates between classical mutants have long been known.

Thus, detailed recombination maps are available for organisms such as *Drosophila melanogaster* (Hoskins *et al.* 2001), *Drosophila pseudooscura* (Ortiz-Barrientos *et al.* 2006), *Arabidopsis* (Singer *et al.* 2006) or zebra fish (*Danio rerio*; Singer *et al.* 2002).

Although recombination rates cannot be estimated through directed crossings between individuals in humans, during the last years several recombination maps of the human genome have been produced. Kong *et al.* (2002) provided in 2002 the first high-resolution recombination map, based on the information of 1,257 meioses. This map used intervals of approximately 350 kb. The International HapMap Consortium (2005; 2007) has also provided fine-resolution recombination maps; the last one, build upon HapMap Phase II, identifying 32,996 recombination hotspots. However, their map is not based on meiotic counts but on the coalescent-based method of McVean *et al.* (2004), operating on LD patterns.

6.2 Ascertainment bias

Neutrality tests, mainly those based on the frequency spectrum of mutations and with the mismatch distribution (that is, Class I and III respectively), rely on an accurate description of the frequency at which segregation sites are found in the sample. This can only be achieved through an accurate resequencing of all the chromosomes in the sample. However, most researchers prefer to use genotyping technologies, as they are simpler, cheaper and much faster. This technique implies selecting a priori which SNPs will be genotyped, which means that information will not be obtained for all segregating sites. The bias produced by the choice of SNPs is named ascertainment bias.

Ascertainment bias can be produced by two mechanisms, although they are not mutually exclusive: (a) by not detecting all the possible SNPs in the sample or (b) by selecting only some of the SNPs present in the sample to genotype. A widely used strategy followed to produce genotyping data that produce the first kind of bias is to resequence only a small subsample, called the discovery sample, and afterwards genotype the SNPs found in a larger panel of similar ethnical composition (Picoult-Newberg *et al.* 1999; Altshuler *et al.* 2000). By using this procedure, it is more probable to detect alleles at intermediate or high frequencies, as the probability to identify a SNP is a function of its frequency, and thus common SNPs will be easier to detect than rare ones. Indeed, it has been shown that the frequency spectrum differs between the discovery panel and the genotyped sample (Nielsen and Signorovitch. 2003; Nielsen *et al.* 2004).

On the other hand, ascertainment bias can be caused by a selection of the SNPs to be genotyped from sources other than the actual sample. Usually these SNPs are selected from a public database such as HapMap (http://www.hapmap.org/; International HapMap Consortium. 2007) or Perlegen (http://www.perlegen.com/; Hinds *et al.* 2005), which in turn have incurred in an ascertainment bias of their own. A large fraction of SNPs provided by Perlegen, for instance, have been detected using a discovery panel. Protocols of SNP selection can vary in a number of ways, but usually all of them involve one or a combination of the following: (a) select SNPs by their minor allele frequency (MAF), e.g. only those with a frequency over 10%, (b) select by distance, one SNP every a given number of base pairs, (c) select by distance but not uniformly, e.g. genotyping a major density in genes, (d) selecting SNPs polymorphic in all the population of interest or (e) selecting SNPs that are polymorphic in only one of the populations of interest (see, for example, Moreno-Estrada *et al.* 2008). The effect of the SNPs selection in the frequency spectrum of mutations depends on the criteria followed, but in any case it will largely differ from the expected one.

As seen above, independently of how ascertainment bias is produced its final effect is always a distortion of the actual frequency spectrum of mutations. As a consequence, then, data provided by genotyping projects cannot be effectively analysed by means of neutrality tests. This problem has previously been reported by Kreitman and Di Rienzo (2004) and Soldevila *et al.*(2005), who showed that the putative effects of balancing selection detected in the *PRPN* by Mead *et al.* (2003) were due to ascertainment bias. In fact, they used a discovery panel plus genotyping, which led to the loss of the low-frequency variants that pointed to the presence of positive selection over the *PRPN* gene (Soldevila *et al.* 2005).

Although no neutrality test can be properly used on ascertained data, much work has been done to develop tools to detect departures from neutrality in these cases. The main efforts to solve this problem have been devoted to: (a) find new methods based on haplotype structure, such as the *EHH* statistic (Sabeti *et al.* 2002), (b) obtain critical values and confidence intervals for neutrality tests from distributions build upon simulated data which directly takes into account the same ascertainment bias than the empirical data (such as the work of Voight *et al.* (2006) or Carlson *et al.* (2004)), and (c) directly correct the statistical estimators and statistics for the ascertainment bias applied to the sample (e.g. Nielsen. 2000; Wakeley *et al.* 2001; Nielsen and Signorovitch. 2003; Polanski and Kimmel. 2003; Nielsen *et al.* 2004).

7 THIS THESIS

In the present thesis I pretend to define more clearly the properties of the neutrality tests and address some of their limitations. In order to do that, the results presented are organised in four sections that correspond to the four papers written during its development.

7.1 Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination (Ramirez-Soriano *et al.* 2008)

Departures from neutrality such as demographic events or recombination leave particular traces over the genome that can be detected by means of neutrality tests. However, although the power of neutrality tests to detect demographic events has been assessed in a number of papers (Ramos-Onsins and Rozas. 2002; Depaulis *et al.* 2003; Sano and Tachida. 2005), the effect of intragenic recombination on the power of the statistics has been much less explored, especially considered together with changes in population size.

In this paper, we have studied the power of the neutrality tests to detect sudden population expansions, population contractions and bottlenecks with and without recombination, using a wide range of parameters of length, time and strength of the event. Furthermore, we have focused on how the distributions of these tests are affected by different degrees of recombination. In this sense we have assessed how the inaccuracy in the estimate of the recombination in a genomic area, and thus in the levels of recombination used to simulate neutral models, affects reliability of the neutrality statistics.

As a conclusion of this work, we provide guidelines on which neutrality tests should be used to detect each of the demographic events explored. These guidelines take into account not only the time since the beginning of the event or its strength but also the recombination rate underlying the genomic area of interest.

7.2 Neutrality statistics in diploid sequences: estimating the loss of power due to statistical phasing (Ramirez-Soriano and Calafell. in preparation)

The significance of the neutrality tests as calculated from empirical data is ascertained by comparing it to a neutral, simulated distribution of values of the same statistic. However, both sets of data are produced with very different methods, as simulations provide haplotyes while resequencing provides genotypes. In empirical data, thus, haplotypes must be inferred from genotypes. Moreover, resequencing produces a number of missing genotypes, which can also be reconstructed.

In this work, we explore the amount of error produced by the algorithm implemented in fastPHASE, a program used to estimate haplotypes, and how it affects the power of neutrality statistics both in neutral models and under several demographic and selective scenarios. All models are tested assuming that the whole genotypes are known and including three different fractions of missing data.

7.3 Correcting Estimators of θ and Tajima's D for ascertainment biases caused by the SNP discovery process (Ramirez-Soriano and Nielsen. submitted)

As seen above, neutrality tests are most efficient in detecting departures from neutrality when data comes from resequencing projects, in which every single nucleotide variant harboured by a sampled individual is typed, including low frequency variants. However, much human data is currently generated by means of large-scale SNP genotyping projects, which cause that some of these variants might be overlooked. This loss of information, known as ascertainment bias, makes neutrality tests highly unreliable to detect departures from neutrality when analysing genotyping data.

Although a number of statistics and methods have been developed to deal with genotyping data (see above), there are currently no formal tests of neutrality that accurately take ascertainment biases into account. Our objective in this work was, therefore, to modify Tajima's D test to take ascertainment bias into account and to use the corrected statistic over a large ongoing genotyping project.

In this paper, then, we derived two corrected estimators of θ, θ_W and θ_π, together with their variances and co-variances, and provide a corrected version of Tajima's *D* statistic. We performed this correction assuming an ascertainment scheme in which a discovery subsample of size *d* has been resequenced and the SNPs found in it are afterwards genotyped in the whole sample. However, the equations given can be easily extended to other ascertainment schemes. Furthermore, we re-analyzed the Perlegen data set using the corrected Tajima's *D*, finding substantial differences in the results obtained with or without ascertainment bias correction.

7.4 FABSIM: a software for generating F_{ST} distributions with various ascertainment biases (Ramirez-Soriano and Calafell. submitted)

One of the methods to detect selection acting over populations besides from neutrality tests is, as seen in previous point 1.2, F_{ST}. F_{ST} measures the extent of genetic differentiation among populations and, thus, is a useful tool to detect local selective pressures acting on a single population, which would appear as more differentiated than with neutrally evolving genes. This is usually achieved by comparing the F_{ST} obtained from a single locus against an empirical distribution of F_{ST} values built upon a large number of SNPs, such as that provided by Akey *et al.* (2002). However, although widely used, this method has two main problems. One one hand, empirical distributions use a particular subset of SNPs, and so they have an underlying ascertainment bias that can modify the distribution. Moreover, this bias is usually different between the published empirical distributions and the genotyped gene of interest. On the other hand, a fraction of the SNPs included in the empirical distribution will indeed be under selection, thus producing broader confidence intervals.

In this work, we suggest to use simulated distributions of F_{ST} in order to look for selection in a genomic region of interest. Until now, simulated distributions of F_{ST} incorporated even more uncertainty than empirical ones mainly due to (a) the inaccuracy of the demographic models describing the populations and (b) the ascertainment bias. In humans, the first issue can currently be solved using the calibrated demographic model proposed by Schaffner *et al.* (2005).

This paper addresses the second problem by developing a software that generates F_{ST} distributions from simulations reconstructing the ascertainment bias of the sample. We have implemented seven biases classified into four categories which can be applied to the simulations independently or combined. Furthermore, we have explored the differences between simulated distributions and an empirical distribution built from the SNPs found by SeattleSNPs, and we provide several F_{ST} distributions for humans with different underlying biases.

MATERIALS AND METHODS

"And somewhere or other, quite anonymous,
there were the directing brains who co-ordinated
the whole effort and laid down the lines of policy
which made it necessary that this fragment
of the past should be preserved, that one falsified,
and the other rubbed out of existence."

1984
George Orwell

1 GENERAL METHODOLOGY

All the work presented in this thesis is based on computer simulations, and no empirical data has been produced.

Simulations have been performed using several programs designed to simulate neutral genealogies based on the coalescent theory (see Introduction, section 3), and analysed using a wide range of small programs and scripts especially developed to this purpose. The only exception has been the use of fastPHASE (Scheet and Stephens. 2006), a program used to infer haplotypes from genotype data.

Empirical data has been used in some cases to support, test or strengthen the conclusions reached using the coalescent-based methods. When empirical data has been needed it has been obtained from publicly available databases, especifically SeattleSNPs and Perlegen. SeattleSNPs (http://pga.gs.washington.edu/, Crawford *et al.* 2005) is a project devoted to resequencing genes belonging to pathways leading to inflammatory response in humans, which nowadays contains information for more than 300 genes. The vast majority of these genes are resequenced in 24 African-American and 23 European individuals. From SeattleSNPs we have used the genotypes of the 303 genes resequenced in African-American and European individuals, downloaded from http://pga.gs.washington.edu/data_download.html.

Perlegen (http://www.perlegen.com/indexNew.html?science/science.html, Hinds *et al.* 2005) is a private company created to discover patterns of genetic variation that are rellevant to clinical purposes. Perlegen has also a database containing 1.6 milion SNPs genotyped in 24 European-American, 23 African-American and 24 Han Chinese samples. Those SNPs have been obtained following three different protocols. 69% of the SNPs (class A) have been obtained by means of array-based genomic resequencing, that is, they have resequenced 20 to 50 chromosomes and have genotyped the polymorphic positions found on these individuals. Note that the number of resequenced chromosomes changes from site to site. Another 27% of the SNPs genotyped (class B) were segregating sites found in other public databases such as dbSNP. The rest of the SNPs (class C) were unvalidated polymorphic positions from dbSNP or low-confidence SNPs found by their resequencing protocols. From Perlegen, we have downloaded the information of all SNPs from http://genome.perlegen.com/browser/download.html but we have only used those belonging to class A.

2 COALESCENT SIMULATIONS

In the different papers included in this thesis, coalescent simulations have been performed using a wide range of parameters that include different sequence lengths, mutation rates, recombination rates and demographic and selective models. Moreover, these genealogies have been built using different programs.

2.1 Programs designed to run coalescent simulations

A number of programs have been developed in the last years to simulate neutral genealogies using coalescent alrogithms. However, besides the purely neutral model explained in section 3.1, they are also able to simulate departures from neutrality such as demographic events, recombination or selection. Indeed, each one of them includes particularities that make it more suitable for simulating coalescent trees under different assumptions. Some of the most representative of such programs are discussed below.

ms (http://home.uchicago.edu/~rhudson1/; Hudson. 2002) is one of the more widely used programs aimed to generate samples under neutral models. ms simulates neutral models, population size changes, migration, and recombination. Mutation can be simulated fixing θ or *S* and uses an infinite sites model, so no recurrent mutation is allowed. This program outputs the matrix of DNA sequences for each individual, and it can also include their genealogic tree.

SimCoal (http://cmpg.unibe.ch/software/simcoal/; Excoffier *et al.* 2000) is another program aimed to generate neutral genealogies with the possibility to include demographic events in them, but it does not take into account recombination. SimCoal allows running simulations conditioned only on θ, but it uses three different mutation models: (a) restriction fragment length polymorphisms (RFLP), that are modelled using a two-allele model with finite sites; (b) STRs, using an stepwise mutation model; and (c) nucleotide replacements, which can be modeled using different finite-sites models. Sequences for each individual are outputted as several Arlequin (Schneider *et al.* 2000; Excoffier *et al.* 2005) or Nexus compatible formats. A second version of the program, SimCoal2 (http://cmpg.unibe.ch/software/simcoal2/; Laval and Excoffier. 2004), also enables (a) to include recombination, (b) multiple coalescent events per generation, (c) to simulate SNP data with a given minimum frequency, (d) to output diploid genotypic data, and (e) to simulate different mutation models along the sequence.

cosi (http://www.broad.mit.edu/~sfs/cosi/; Schaffner *et al.* 2005) generates samples under stationarity and different demographic models, with or without recombination. The main difference with ms is that it allows applying different recombination rates along the sequence and that it can use either a finite or an infinite sites model. cosi outputs two files for each simulated population, a file with the haplotypes of each individual and another with the position and allele frequency of each segregating site. The cosi package includes two programs: coalescent, to run the simulations, and recosim, to produce random maps of recombination rates.

SelSim (http://www.stats.ox.ac.uk/~spencer/SelSim/Controlfile.html; Spencer and Coop. 2004) is a program to generate genealogies under neutrality and selection with recombination. It implements several mutation models, among which the possibility to produce SNPs or microsatellites, to fix *S* or θ, and to use finite or infinite models. Recombination rates can vary along the sequence. SelSim is able to simulate positive and overdominant selection and allows choosing a deterministic or a stochastic model for simulating the trajectory of the derived allele. The program outputs the haplotypes of all individuals in the population. Furthermore, if required, it can provide the genealogic trees of each sample and the location of the recombination events.

mlcoalsim (http://www.ub.es/softevol/mlcoalsim; Ramos-Onsins and Mitchell-Olds. 2007) is an ms-based program to generate samples under the stationary model, several demographic scenarios and strong positive selection, with recombination. Mutation can be simulated by fixing θ or S, or by using distributions of θ. This allows using different mutation rates along the sequence. Furthermore, mutation can be placed on the tree either following an infinite-sites model or allowing recurrent mutation. Recombination can also be incorporated in the model in different ways, including (a) a constant recombination rate, (b) a distribution of recombination values and (c) a fixed number of recombination events. Besides, a particularity of mlcoalsim is that it can output the sequences of every individual in the sample or calculate several neutrality statistics and output their value.

Of all the programs described above only ms, cosi and SelSim have been used. ms has been used to generate neutral genealogies and simple demographic models with and without recombination. Cosi has been used to run coalescent simulations under the best-fit model of human demography described in Schaffner *et al.* (2005). SelSim has been used to produce genealogies with positive selection.

2.2 Mutation models

Gene genealogies can be simulated according to several mutation models. As seen in the above section, these models can be classified according to whether they refere to (a) the kind of polymorphism produced, (b) the mutation parameter, and (c) the model of alleles used.

The most problematic of these points is how to include the mutation parameter, θ, in the simulated genealogy. The question arises because θ cannot be obtained directly from the observation of the data and, thus, it is usually unknown. To solve that, three main strategies have been proposed: (a) simulations can be produced using an estimate of θ, such as the Watterson's estimate θ_W (Watterson. 1975); however, it is difficult to ascertain the accuracy of an estimate of θ, especially if the region of interest is not under neutrality; and even if it was accurate, it produces broad confidence intervals, thus reducing the power of neutrality tests (Depaulis *et al.* 2005).

(b) simulations can be performed fixing the number of segregating sites, *S*, which can be directly counted in the sample after resequencing (Hudson. 1993). Nevertheless, this method is not accurate either, as it does not take into account the length of tree. Therefore, short trees will have relatively high mutation rates, while long trees will have lower mutation rates than expected (Tavare *et al.* 1997; Pritchard *et al.* 1999; Jakobsson *et al.* 2006).

(c) several strategies have been proposed which run simulations conditioned to *S* and taking into account the uncertainity of θ. In this case, the different trees produced for a given θ are weighted according to the probability that such a θ produce the *S* observed in the sample. This method, although more accurate, is also much more computationally intensive.

In the papers included in the Results section simulations have been run both fixing θ_W and *S*, depending on their purpose. In the first work, both methods have been used and differences between them have been discussed. For the second paper, devoted to see how phasing data modifies the power of neutrality tests, simulations have only been produced fixing *S*, as a consistent number of segregating sites along the samples was needed. In the third manuscript all simulations have been conditioned on θ_W, as we wanted to investigate the accuracy of the estimators of θ. On the fourth θ_W has been also used, as we used the parameters published by Schaffner *et al.* (2005). As for the other kinds of mutation models, DNA sequence data has been simulated in all cases, using an infinite-sites model.

2.3 General parameters used

In this section I will only discuss those parameters that have been used along all the works presented in the results. Therefore, I will not discuss the demographic or selective models used, as they only affect the two first (sudden population expansions, bottlenecks and sudden population contractions) and the last (Schaffner *et al.*'s (2005) calibrated genealogy for humans) manuscripts presented. I will neither discuss particular values of the parameters explained that only apply to specific cases.

Sample sizes have been generally assumed to be *n*=20, 50 and 100, as to compare results for small, medium and large sample sizes. Effective population size has been set to N_e=10,000, as this correponds to the value estimated for humans (Takahata *et al.* 1995).

In order to obtain realistic estimates of the number of segregating sites to simulate, we have used *S*=10, 100 and 400, which corresponds to the rounded minimum, average and maximum *S* found in the genes resequenced by SeattleSNPs (http://pga.gs.washington.edu/; Crawford *et al.* 2005). Sequence lenghts of 3,000, 21,000 and 72,000 base pairs (bp) have been assumed for each *S* respectively, as they also correspond to the rounded minimum, average and maximum lenghts resequenced in SeattleSNPs. θ_w values have been estimated from these *S* values when needed.

Recombination rates, when applied, have been set to r=0, r=10^{-10}, r=10^{-8}, and r=10^{-7} per bp. The non-null values correspond to the rounded minimum, average, and maximum values estimated by Kong *et al.* (2002) for the human genome.

3 NEUTRALITY TESTS

3.1 Neutrality tests used

As stated in the Introducton, neutrality tests can be classified in three classes: (a) Class I, based on the frequency spectrum of mutations, (b) Class II, based on the haplotype distribution, and (c) Class III, based on the distribution of pairwise differences.

As Class III tests have been shown to perform poorly in the presence of population expansions (Ramos-Onsins and Rozas. 2002), in the first manuscript presented only Class I and Class II tests have been used. Among each class all the statistics described above have been included; however, two point estimators of *EHH* have been developed in order to use it as a neutrality statistic. In order to do that, the first three SNPs of each sequence have been taken as the core haplotype, and the distance from each core at which *EHH* decays to or under 0.5 has been registered. The two point statistics, then, have been defined as: a) *EHH* Average, that is, the weighted average of the distance at which *EHH* decays to or under 0.5, for all core haplotypes, and b) *EHH* Maximum, the longest distance to which a core haplotype in the sample decays to or under 0.5. If a simulated segment finishes without *EHH* reaching a value under 0.5, *EHH* Maximum has been set to 2*L*, where *L* is the sequence length.

The second manuscript uses the same neutrality statistics than the first. On the third we have only used Tajima's *D*, as we have focused on correcting this statistic for its use under ascertainment bias. In the last manuscript we have not used any neutrality test but F_{ST}. However, neutrality tests have been included in the developed program.

3.2 Power of the tests

The power of the tests to detect an event has been estimated by comparing the distribution of the values of the statistic under the neutral model against the distribution of its values under the event of interest, as shown in Figure 25.

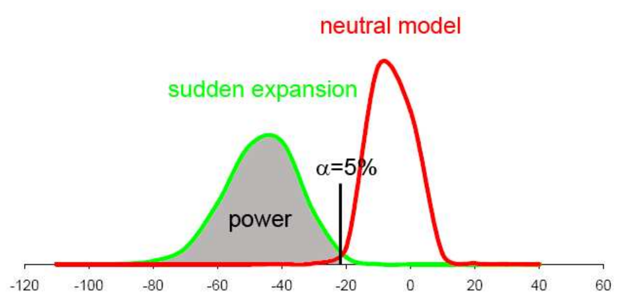


Figure 25: example of the calculation of the power of a test to detect population expansions.

In order to do that, 10,000 samples have been generated both for the null hypothesis, that is, the neutral model, and for the alternative hypothesis, that is, the non-neutral model. Two one-tailed tests have been performed, both with an α=5% significance level. That is, the 5% and the 95% percentiles have been obtained for each statistic.

Next, for the 10,000 non-neutral samples, the value of each statistic has been compared against its α, and all the values that were more extreme than α were counted. The power of a test, then, was the number of values more extreme than α divided by 10,000.

4 PHASE ESTIMATION

Nowadays, two main strategies are available to detect variation in genomic sequences. If the variation is known it can be characterised on the populations of interest by means of genotyping methods, while if it is unknown it can be found by resequencing a number of individuals.

However, given a polymorphism, both technologies are unable to specify which chromosome in a homologous pair each variant belongs to. Genotyping and resequencing, then, provide genotypes, from which the haplotypes, or the phase, needs to be inferred (see Figure 26).

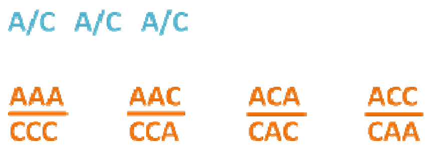


Figure 26: above, in blue, a genotype. Below, in orange, the four possible haplotype combinations.

4.1 Programs to estimate the phase

Several programs have been developed to estimate the phase of a set of genotypes, each one of them using different algorithms. The most important among them are Arlequin, PHASE and fastPHASE.

Arlequin (Schneider *et al.* 2000; Excoffier *et al.* 2005) is a program for genetic analysis that performs a number of different calculations among which Hardy-Weinberg (HW) equilibrium, measures of LD, some neutrality statistics or the analysis of variance (AMOVA). Moreover, it also implements phase reconstruction using an Expectation-Maximization (EM) algorithm. This algorithm provides the maximum-likelihood frequency of each haplotype working iteratively, starting with a random estimation of the frequencies. From these initial frequencies it follows two steps, namely E and M. In the E-step Arlequin calculates the most probable genotype frequencies from the current haplotype frequencies assuming HW equilibrium. In the M-step, it uses the new genotype frequencies as weight for producing new haplotype frequencies. These two steps are repeated until equilibrium is reached in haplotype frequencies, that is, until haplotype frequencies do not change more than a predefined value between iterations. However, this method is computationally very intensive and Arlequin has been progressively abandoned for this purpose in favour of other software. The kind of data accepted by Arlequin includes sequences, SNPs, STRs, RFLPs and allele frequencies. It can also deal with missing data, including the possibility to specify a maximum number of missing genotypes in a position to take it into account for analysis.

PHASE (Stephens *et al.* 2001; Stephens and Donnelly. 2003) is the most widely used method to estimate haplotypes nowadays. It uses a Bayesian algorithm, which is based on the prior distribution of haplotype frequencies and on the likelihood of this distribution. As a prior distribution PHASE uses a distribution approaching the coalescent, and implements a Markov Chain Monte Carlo (MCMC) method to estimate the posteriori distribution. This method reduces the error in the reconstruction and is more efficient than the EM algorithm, allowing reconstructing the haplotypes for larger numbers of SNPs (Stephens *et al.* 2001). PHASE is able to work on SNP data, STRs and multiallelic data such as triallelic SNPs, and to reconstruct missing positions.

Finally, fastPHASE (Scheet and Stephens. 2006) has been developed to deal with the huge numbers of SNPs produced by the new genotyping technologies, which make even PHASE largely inefficient. fastPHASE is based on a cluster method with two versions, which assume or not HW equilibrium. A cluster represents a combination of alleles at close SNPs, and each haplotype is composed by a mosaic of clusters of different size and distribution (see Figure 27). The algorithm implemented in fastPHASE, then, beggins with the assumption that an allele originates from one of the clusters, and calculates the cluster at which the next marker belongs using a Hidden Markov Model (HMM), which considers the probability of transition to another cluster. When the HW equilibrium is assumed, the two hapltoypes forming a genotype are assumed to be independent and use the same probability distribution to belong to a cluster. fastPHASE can deal only with biallelic SNP data, and also imputes which alleles correspond to a missing genotype.

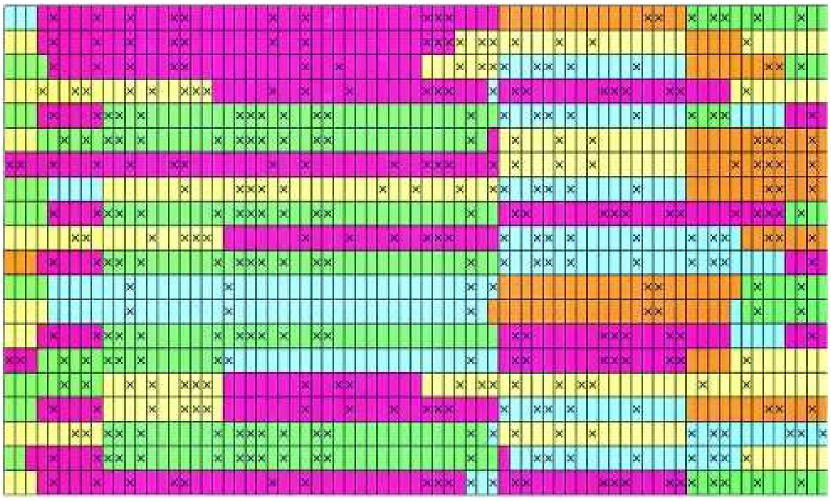


Figure 27:. example of clustering. Each line is a haplotype, and consecutive lines taken two by two represents individuals. Blanks and crosses represent the two possible states of each allele. Colours represent the different clusters. (Scheet and Stephens. 2006)

4.2 Methods used to assess the effect of phasing on neutrality statistics

In the second manuscript presented in the Results section we intended to ascertain the effect of phasing data on the power of neutrality statistics.

In order to do this we have simulated 10,000 sets of samples, formed by 100 chromosomes, under different demographic and recombination parameters. Afterwards, each set of samples have been split into 10,000 files, each one of them containing a sample formed by 50 individuals. To create the individuals, the 100 chromosomes of the sample have been randomly paired. Furthermore, these 10,000 files have been created five times each, assuming different fractions of missing genotypes. Missing values have been applied substituting the desired fraction of alleles by ‘?’. The alleles to be transformed to missing values have been selected at random through the sample.

Once the new files with one sample each have been created, their individuals have been phased. The phasing has been performed using fastPHASE as, even if each file did not contain many SNPs, the large number of files to phase made the use of PHASE computationally unfeasible.

Finally, the 10,000 phased files have been transformed again into a single file containing 10,000 samples. This new file has then been compared with the original simulations into two different ways, comparing (a) the error commited by fastPHASE and (b) the power of the neutrality statistics.

5 DATA ANALYSIS PROGRAMS

Besides from using programs publicly available through Internet, I have written several other programs and scripts to transform file formats as well as to analyse data.

5.1 Programming languages used

Thousands of programming languages have been developed since the first computers appeared, and still now new languages appear every year. Each one of them possesses characteristics that make it more suitable to particular uses, and can be classified in a number of ways according to their properties. Of all the available programming languages only C, Java, and Perl will be discussed, as they are the three I have used during the present work.

C was developed in 1972 at Bell Telephone Laboratories by Dennis Ritchie. It is an imperative language, that is, programs written in C are a sequence of commands for the computer to perform. Then, C is sequentially structured, and its code is organised in functions that receive parameters passed by value as well as memory positions, passed by means of pointers containing their addresses. Furthermore, C is a compiled language, which means that a compiled C program can run on any computer by itself, without the need of any supporting software. However, it must be compiled in the same operative system where it is to be used. As most programs designed to simulate neutral genealogies, such as ms or cosi, are programmed in C, I started using this language. However, lately I changed the programs I had written in C to Java, as it is quicker to program and powerful enough for the purposes of my work.

Java is a language developed by James Gosling at Sun Microsystems in 1991. Although its syntax derived from C and its further version C++, it is not an imperative language as the former but and object-oriented one as C++, even if simplified. As an object-oriented language its code is primarily organised in Objects, particular instances of code which own a set of variables and methods (functions) not shared with other parts of the code. These objects can be called from other objects, as well as inherit properties from them. Another difference between Java and C is that the former has been designed to be multiplatform, that is, it can be used on any operative system no matter in which one has been compiled. This is accomplished by compiling the font code to an intermediate language which can be further interpreted by the Java Virtual Machine (JVM). Furthermore, Java uses default methods that are continuously revised and sometimes eliminated or updated, so the JVM must be in the same version that the Java methods used. Most of the programs detailed below have been built or transformed to Java.

Perl was developed in 1987 by Larry Wall. As C, Perl is an imperative language, but also takes characteristics from the shell programming language developed for UNIX systems. However, unlike C, it is not compiled but interpreted, that is, it needs an interpreter that reads the code and executes it before every usage. During this work Perl has only been used to program some small scripts to perform file organisation tasks on Linux, and thus no Perl program is explained in detail nor included in the attached CD-rom.

5.2 Programs and scripts

In this section the different programs and scripts used in this thesis are described. Note that the programs developed in C and further translated to Java are included in their latest Java version.

The font code of all programs is provided in a CD-rom at the end of this thesis. For each program there is a folder with the program name, which coincides with which is presented below unless the contrary is stated, and a compiled .jar executable. Except for the programs with graphic environment, the programs are run using the command "java –jar program.jar" followed by the arguments required, and must be executed in the folder where their input files are placed. Programs with graphic environment are run by double-clicking the executable .jar file.

5.2.1 Scripts to modify simulation outfile formats

The first program to generate neutral genealogies that I used was ms. For this reason, the scripts written to analyse sample data were created to be able to read the standard ms outfile format. This format is characterized by a general header followed by the different simulated samples separated by a space and a double bar (//).

The header has two lines. The first includes the name of the program, the number of chromosomes per sample, the number of runs and the parameters indicating the mutation model. If any demographic event or recombination rate has been applied to the genealogy it is specified after the mutation model. The second line consists on a number specifying the seed used by ms to run the simulation. A simulation with 50 individuals per sample, 10,000 simulated samples, fixing 10 segregating sites and with no demography or recombination would look as follows:

```
ms 10 10000 -s 10
5294
```

After this header all the simulated samples (10,000 in the example) would follow. Samples start with the double bar. After that, the two subsequent lines contain the number of segregating sites in the sample and the position of each one of them. Positions are scaled between 0 and 1, so the absolute position on the sequence can be obtained multiplying the relative position by the total simulated sequence length. Finally the chromosomes are listed, one per line. The alleles are coded as 0 (ancestral) and 1 (derived). The first sample corresponding to the above header would be:

```
//
segsites: 10
positions:  0.0001 0.0193 0.0350 0.0442 0.0609 0.0864 0.0872 0.1004 0.1016 0.1071
1010000000
0010000101
0010001101
0010000010
0010000101
1010100000
0010000101
```

```
0010000000
0010000101
0010000000
```

The object of the scripts discussed in this section, then, is to transform other format files to the ms format. This transformation is basically aimed to use a second program, SampleStats, which calculates neutrality statistics for each sample. As of all the information that ms outputs in the header SampleStats only uses the number of chromosomes per sample and the number of runs, the transformation only includes this information on the first line of the header. On the second line, the seed number is always assigned to 111. Positions are assigned to 0, 1, 2 … n, as they are neither used by SampleStats. A version of these scripts which does include the actual position of the sample is implemented in the FABSIM program (see section 5.2.5).

CosiToMs transforms the outfile of cosi to a file with the ms format. It requires as arguments the name of the cosi's outfile, the number of individuals for sample, the number of simulated samples and a complete name, including the extension, for the outfile.

SelsimToMs transforms a Selsim outfile into the ms format. To do so, it requires the name of the SelSim's outfile, the selected segregating site, the number of individuals per sample, and the number of simulated samples. The resulting file is given the same name that the input file with the extension .out.

FastphaseToMs was developed to convert a number of fastPHASE output files to a single ms file format. In order to run, it requires the number of files to be included in the final file and a pattern, that is, a fragment of the name of the fastPHASE outfiles which is shared among all the files to be included. The FastphaseToMs output file is named as the entered pattern and has the extension .out.

ConvertPerlegen transforms the genotype files for each chromosome as downloaded from Perlegen (http://genome.perlegen.com/browser/download.html) into the ms format, as required for TajimaCorrection (see 5.2.5), conserving only Class A SNPs. It only requires as arguments the name of the file with the genotypes (g_chrxx.dat). However, a d_chrxx.dat file containing the information on the Class A SNPs (see Results, section 3, Theory and methods) must exist in the same folder. The output file is named c_chrxx.out.

5.2.2 SampleStats

SampleStats is a program developed to calculate neutrality statistics. The first version of this program was a modification of sample_stats, a program provided by Hudson together with the ms package that calculates the number of pairwise differences between sequences (π), Tajima's *D*, Fay and Wu's estimator of θ (θ_H), and the difference between θ_H and π. The first version of SampleStats, then, was written in C, as was Hudson's sample_stats. In that version I maintained the same structure for reading samples which was already implemented, but I changed the code to calculate the neutrality tests of my interest. In this sense I deleted θ_H and the difference between θ_H and π and implemented instead all the statistics cited in the Method's section 3. Afterwards, the program was translated to Java, optimizing the code and, thus, its speed.

SampleStats can calculate neutrality statistics over a number of samples in ms format, described above. To run, it only requires the infile name and an outfile name. For each sample, the outfile consists on a list of the different neutrality statistics included, classified according whether they belong to Class I or to Class II. An example of a SampleStats output for two samples would be as follows:

```
SAMPLE 1
Sequences: 46    Seg. sites: 42      Pi: 8.668598        Singletons: 10
Class I Statistics
Tajima's D: -0.320972
Fu and Li D*: -0.045210    Fu and Li F*: -0.170661
Fu and Li D: -0.099298     Fu and Li F: -0.223524
R2: 0.102917
Fay and Wu H: -9.773913
Class II Statistics
Fu's Fs: -2.411875
EHH average: 15.239130   EHH maximum: 34.000000
Dh: 0.88405797
Wall's B: 0.292682          Wall's Q: 0.428571
ZnS: 0.140208    Za: 0.344554      ZZ: 0.204346

SAMPLE 2
Sequences: 46    Seg. sites: 17      Pi: 6.558454        Singletons: 0
Class I Statistics
Tajima's D: 2.201818
Fu and Li D*: 1.616728     Fu and Li F*: 2.146899
Fu and Li D: 1.734385      Fu and Li F: 2.284009
R2: 0.192895
Fay and Wu H: 0.050242
Class II Statistics
Fu's Fs: 5.077704
EHH average: 34.000000   EHH maximum: 34.000000
Dh: 0.790338
Wall's B: 0.187500          Wall's Q: 0.352941
ZnS: 0.364722    Za: 0.380532      ZZ: 0.015810
```

SampleStats is accompanied by three other scripts: ExtractStats, Analyse, and Stats. ExtractStats takes the neutrality statistics from a SampleStats output file and displays them in a tabulated format, in columns. The name of the ExtractStats outfile is the same as the infile but with the extension .ext.

Stats has two associated main projects, Stats and Stats2. Stats calculates the value that defines α for a given confidence interval. It needs a file with the neutrality statistics tabulated, such as the produced by ExtractStats, and the confidence interval to calculate (e.g. 0.05). The outfile consists in a column with the α for every statistic without labels in the order they appear in the ExtractStats file. The outfile name is the same as the input finished with the confidence interval and with extension .stats (e.g. CSM_n100_s10_r0_0.05.stats). Stats2 calculates the mean and the variance of each statistic and displays them in two columns, with one column at the beginning with the labels. Its outfile name is the same as the infile but with the extension .stats.

Analyse was created to calculate the power of each statistic to detect a determinate event. It takes a SampleStats output file and, for each statistic, it compares its value against a Stats (version Stats) file which has the α of a given confidence interval calculated on another distribution. For example, if we want to see the power of neutrality test to detect population expansions with a confidence interval of 0.05, the SampleStats file would correspond to the expansion simulations and the Stats file to the neutral simulations. The parameters to run Analyse are the SampleStats file, the Stats file, the tail against which to compare (right or left, that is, under or over α), and the direction in which to display the results (line or columns). Results have no labels. The output file name is the same as the SampleStats input, finished with the tail and with extension .ana (e.g. CSM_n100_s10_r0_left.ana).

5.2.3 Scripts to create fastPHASE input files

Input fastPHASE files only need to include the number of diploid individuals in the sample, the number of segregating sites and the genotypes for each individual. However, it also accepts the PHASE infile format, which also includes the position of each segregating site and a row of 'S' characters, one for each segregating site. In order to create files which are compatible with both programs, the scripts described in this section produce the PHASE infile format. Thus, the output of these scripts will look as follows:

```
5
6
P    0    1    2    3    4    5
SSSSSS
DY01
?    ?    A    G    G    A    G
?    ?    A    G    G    A    G
DY02
?    ?    A    A    T    A    G
?    ?    A    A    T    A    G
DY03
?    T    A    A    T    A    G
?    T    A    A    T    A    G
DY04
?    ?    ?    A    G    A    G
?    ?    ?    G    T    A    G
DY05
?    T    A    A    T    A    G
?    T    A    A    T    A    G
```

The first two lines indicate that there are 5 individuals and 6 segregating sites. The third line gives the position of each segregating site, and the fourth is the row of 'S'. Next, it follows the genotype for each individual with its label. Missing genotypes are coded as '?'. Two scripts have been developed to convert files into PHASE format: msToPhase and SeattleSnpsToPhase.

msToPhase converts a file in ms format with a given number *n* of simulated samples into *n* input fastPHASE files with one sample each. Individuals in a sample are labelled #1, #2, …, #i, and their genotypes are coded as '0' and '1' as were in the simulations. Moreover, msToPhase allows including a number of missing genotypes in the sample, distributed randomly along all the genotypes. To run, msToPhase needs to be given the name of the file containing the simulations without extension, the extension preceded by a point (e.g. .out) and the desired number of missings to include (0 is allowed).

SeattleSnpsToPhase uses two of the files provided by SeattleSNPs for their resequenced genes: (i) the individual genotypes (gene.prettybase.txt) and (ii) the SNP alleles file (gene.alleles.txt) and converts them into two input fastPHASE file, one for the genotypes of African Americans (gene_AA.inp) and the other for the genotypes of Europeans (gene_EU.inp). Note that the program only works for those genes resequenced on, and exclusively on, African Americans and Europeans. The program looks for all the genes that have both an individual genotypes and an alleles files in the folder where it is placed and produces the two output files for each of them. Thus, SeattleSnpsToPhase does not require any argument to be run.

5.2.4 AscertainSample

AscertainSample is a program developed to bias simulated data as if its SNPs had been identified by using a discovery sample, a process consisting in resequencing a subsample of size *d* and genotype the SNPs found in the whole sample *n*. This is accomplished by selecting *d* random sequences and discarding from the sample all loci that are not polymorphic in this subsample.

The program has three versions, implemented as different packages inside the AscertainSample project, named ascertainsample_1, _2 and _3, which implement different ways to apply the bias. In all cases, the program only accepts as infile data in ms format, and the output consists on the same ms sample, also in ms format, but without the SNPs not found in the *d* sample.

Versions _1 and _2 ascertain data using the same subsample along the whole haplotype. The latter is an optimised version of the former, which also needs less parameters to run: version _2 requires the infile name and the *d* size, while version _1 also requires the *n* sample size as last argument. The output file of both versions is named as the infile, adding the word “_ascertained” after the infile name and with the extension .out. Furthermore, they can launch ms to produce new, non-ascertained simulations, with the same sample size than the original simulations and the same number of segregating sites after ascertainment. This new simulations are run for each sample and stored in a new file named infile_control.out. This option can be shut down by commenting the lines under the code //execute ms to generate a non-ascertained sample and //generates the final file with the non-ascertained sample.

Version _3 assumes a different *d* sample for each locus in the chromosome, that is, it selects randomly *d* different sequences for each SNP. As version _2*,* it only requires the infile name and the *d* size as arguments. The outfile name is the same as the infile followed by the *d* size and “_ascertained”, with the extension .out. If, for instance, the infile is named “test” and *d*=5, the outfile will be test_d5_ascertained.out.

5.2.5 TajimaCorrection

TajimaCorrection (Ramirez-Soriano and Nielsen. submitted) is a program with different versions that implements the corrected estimators of θ and their variances and covariances as derived in Results, section 3, and provides the corrected Tajima’s D value for data with underlying discovery sample ascertainment bias.

The paper is accompanied by a version with a graphic environment that can work both with simulation and empirical data, with or without changing the *d* size (see 5.2.4). This version, named Tajima’s D Corrector (in the TajimaCorrectionGraphic project), is built by adapting and putting together the versions of TajimaCorrection developed by the analysis of simulations and of Perlegen data plus the scripts developed to ascertain simulation data explained above. Details on how to use the graphic version are explained in the program’s documentation, included in the Apendices.

The TajimaCorrection project includes three different programmes implemented as different packages, each with their own main class: tajimacorrection, TajimaPerlegen and TajimaPerlegenUncorrected. Tajimacorrection was designed to calculate the corrected Tajima’s D using our corrected estimators for data with and without ascertainment bias. To run, it needs as arguments the input file and the *d* sample size. If *d*=0, our corrected formulas are used substituting the probability to find a SNP by 1. The output file is named after the input file with the extension .tcr and contains six columns, for the corrected estimators, their variances and covariances and for the Tajima’s *D*:

```
        Watterson's theta corrected        Tajima's theta corrected   Variance W_theta corrected Variance T_theta corrected       Covariance corrected       Tajima's D
corrected
        17.000000       15.642857       70.940928       109.347628      84.654057       -0.409558
        2.000000        0.542857        2.396624        2.952241        2.497456        -2.449229
        13.000000       11.300000       43.936709       66.437652       51.801591       -0.653306
        5.000000        4.771429        8.966245        12.336695       9.974563        -0.196445
        4.000000        1.814286        6.379747        8.547731        6.984738        -2.233110
        0.000000        0.000000        0.000000        0.000000        0.000000        0.000000
        1.000000        1.071429        1.000000        1.145714        1.000000        0.187122
        5.000000        4.571429        8.966245        12.336695       9.974563        -0.368335
        7.000000        6.142857        15.329114       21.897060       17.446582       -0.561171
        1.000000        1.071429        1.000000        1.145714        1.000000        0.187122
```

This example corresponds to the analysis of a sample with *d*>0. If *d*=0, the headers lack the word “corrected”.

TajimaPerlegen and TajimaPerlegenUncorrected were written to analyse Perlegen data, and thus take into account the two main particularities of this data: (a) the large region to analize, as each file contains a whole chromosome and (b) the different *d* size along the chromosome. To adress the first issue the program analyses the chromosomes using windows. To take into account the changing *d* size the programs need two files with the same name and extension, a c_file with the sample in the ms format and a d_file (see section 5.2.1 and Results, section3, Theory and methods) with the *d* size of each SNP. The d_file should be organized in eight columns separated by tabulators, the last two containing the *d* sample for each allele, and a row for each SNP plus another for the header. The arguments required for both programs are the name of the c_file, the size of the windows in kb and the step size between windows also in kb. The outfile is named after the c_file and contains 11 columns: (a) the window number (column 1), (b) the absolute position of the start and end SNPs, with 0 representing the first SNP in the sample (columns 2 and 3), (c) the average number of SNPs per window (column 4), (d) the average number of valid haplotypes per window, that is, the number of haplotypes without missings (column 5), (e) the corrected estimators of Watterson's and Tajima's θ and their variances and covariance (columns 6 to 10) and the corrected Tajima's *D* (column 11):

```
Window start_pos      end_possnp_num      average_n      W_theta       T_theta average_Theta  Var_W_theta    VarT_theta     Cov
Tajima's D
1      0      49      49      139      23.432733      17.535322      23.432733 79.241026      163.471632     106.042844     -0.192556
2      20     49      29      140      12.906103      8.235925       12.906103 25.571425      51.268497      33.403851      -0.465518
3      28     49      21      139      8.681747       6.103672       8.681747 12.306826       24.077158      15.766162      -0.531380
4      35     49      14      139      5.739914       4.164504       5.739914 6.053543        11.337225      7.525796       -0.673489
5      41     49      8       139      2.749935       2.359691       2.749935 1.745456        3.019952       2.061302       -0.607096
6      42     49      7       141      2.284971       1.744746       2.284971 1.270244        2.165306       1.477794       -1.125558
7      44     49      5       141      1.670762       1.352576       1.670762 0.826446        1.330367       0.931516       -1.083072
8      44     49      5       141      1.670762       1.352576       1.670762 0.826446        1.330367       0.931516       -1.083072
9      46     49      3       142      1.028643       0.651729       1.028643 0.439866        0.651925       0.473517       -2.603770
```

A script named AnalyseTajimaResults is included with the TajimaCorrection. AnalyseTajimaResults takes a TajimaPerlegen or a TajimaPerlegenUncorrected output file and selects only those windows that have a minimum number of SNPs and a Tajima's *D* value above a positive threshold or below a negative one. To run, this script requires the TajimaCorrection output file name, the minimum number of SNPs allowed for window and the threshold for Tajima's D. It returns two files named infile_xsnps.tcr and infile_tajimax.tcr, where x is the minimum number of SNPs per window and the Tajima's *D* threshold respectively.

5.2.6 FABSIM

FABSIM (Ramirez-Soriano and Calafell. submitted) is a program with a graphical interface developed to produce F_{ST} distributions with ascertainment bias. Furthermore, it also calculates minor (MAF) and derived (DAF) allele frequencies and neutrality statistics.

This program works on simulated data produced using ms, cosi or SelSim, or in any other software whose output is translated to one of those formats. It implements seven different ascertainment biases grouped into four categories: (a) related to the discovery sample, (b) related to the presence of polymorphism in a population, (c) related to the MAF, and (d) related to distance. Results can be outputted in two different formats, as a list of values per samples or as tabulated statistics.

More details on how to use it, its characteristics and the infile and output files can be found in Results, section 4, and on the program documentation, included in the Apendices.

5.2.7 Other scripts

PhaseEfficiency is a script that estimates the precision of fastPHASE to reconstruct haplotypes and impute missing alleles, using simulations (see Results, section 2). This program compares the true haplotypes with the reconstructed lines and gives four indicators of the precision: (a) the fraction of correctly estimated haplotypes, (b) the average number of incorrectly estimated positions per haplotype, (c) the average number of incorrectly estimated positions per incorrect haplotype and (d) the number of incorrectly estimated positions divided by total number of positions, the last understood as the number of chromosomes multiplied by the number of SNPs per chromosome. PhaseEfficiency compares the fastPhase input files (actual haplotypes) in a folder against the fastPHASE output files (reconstructed haplotypes) in the same folder. To run, the program only needs a file pattern, that is, a fragment of the input and output phase files. This pattern has to be shared among both files and to correspond to the beginning of the file names. PhaseEfficiency will compare two by two all the files sharing this pattern.

HistogramBuilder is a script that creates histograms from tabulated data. It needs seven parameters to run: (a) the infile name, (b) the minimum and maximum values of the distribution, (c) the range of classes (that is, the distance between them), (d) the symbol that codes invalid values (even if all values are valid in the input file, an invalid value character – such as any character absent in the file- must be added), (e) the column containing the values with which build the histogram and the number of header lines in the file. HistogramBuilder return a file named as the input file with the extension .hst that contains two columns separated by tabulators: the classes and the absolute frequency of values of each category, as shown below.

```
classes	frequency
0.05	21450
0.10	3305
0.15	2084
0.20	1180
0.25	837
0.30	512
0.35	403
```

```
0.40    368
0.45    248
0.50    335
0.55    324
0.60    260
0.65    259
0.70    199
0.75    168
0.80    262
0.85    302
0.90    295
0.95    299
1.00    381
```

This file can be directly imported to Excel or any other program that accepts tabulated formats to draw the graphic.

RESULTS

Un camí!
Quina cosa més curta de dir!
Quina cosa més llarga de seguir!

Poema de Nadal
Cant III
Josep Maria de Sagarra

1 STATISTICAL POWER ANALYSIS OF NEUTRALITY TESTS UNDER DEMOGRAPHIC EXPANSIONS, CONTRACTIONS AND BOTTLENECKS WITH RECOMBINATION (Ramirez-Soriano *et al.* 2008)

Ramírez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A.

Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination.

Genetics. 2008 May;179(1):555-67.

2 NEUTRALITY STATISTICS IN DIPLOID SEQUENCES: ESTIMATING THE LOSS OF POWER DUE TO STATISTICAL PHASING (Ramirez-Soriano and Calafell. in preparation)

Neutrality statistics in diploid sequences: estimating the loss of power due to statistical phasing

Anna Ramírez-Soriano[1] and Francesc Calafell[1,2,*]

[1] Departament de Ciències de la Salut i de la Vida. Universitat Pompeu Fabra. Doctor Aiguader, 88. 08003 Barcelona, Catalonia, Spain.

[2] CIBER en Epidemiologia y Salud Pública (CIBEREsp), Spain

Running title: Neutrality statistics in diploid sequences

Keywords: phase, neutrality statistics, haplotype

***Corresponding author:**

Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra

Doctor Aiguader 88, 08003 Barcelona (Spain)

Phone: +34 93 3160842

Fax: +34 93 3160901.

Email: francesc.calafell@upf.edu

ABSTRACT

The coalescent allows the estimation of the statistical significance of neutrality statistics through the simulation of neutral scenarios against which to compare the DNA data, and thus to infer whether the population deviates from the neutral model. However, there is a crucial difference between simulation and empirical, diploid data obtained from resequencing, as the former provide real haplotypes while the latter have to be phased. This could affect the power of neutrality statistics through (a) the loss of low-frequency variants and (b) the reduction in the number of different haplotypes. In this work, we have acknowledged how the reconstruction of haplotypes, performed using fastPHASE, affects the power of neutrality tests under several demographic models and different levels of recombination and missing genotypes. We have found that, although the accuracy of the phasing is not optimal, the power of neutrality statistics is not affected by it.

INTRODUCTION

In the last years, much interest has been devoted to unravel both the demographic processes and the selective forces that lay behind the populations. In this direction, a considerable effort has been put in the development of statistical tests (e.g. Tajima. 1989; Fu and Li. 1993; Fu. 1997; Kelly. 1997; Fay and Wu. 2000; Ramos-Onsins and Rozas. 2002) that can detect departures from the neutral theory of evolution, mainly modelled upon the Wright Fisher model (Fisher. 1930; Wright. 1931; Hein *et al.* 2005), which assumes neutral populations of constant size that are panmictic and non-recombining, in the DNA variability patterns.

These concerns have been usually approached from the coalescent theory (Kingman. 1982a; Kingman. 1982b; Hudson. 1990; Donnelly and Tavare. 1995; Fu and Li. 1999; Kingman. 2000), a theoretical framework that provides the mathematical basis for the development of neutrality statistics. The coalescent allows the estimation of the statistical significance of neutrality tests through the simulation of neutral scenarios against which to compare the DNA data, and thus to infer whether the population deviates from the neutral model (Wall. 1999; Ramos-Onsins and Rozas. 2002; Depaulis *et al.* 2003). However, this approach presents several limitations. In the first place, demographic and selective events leave similar traces in the genome, and thus it is often difficult to distinguish them when comparing against neutral models. That implies that the reliability of the results lies on the precision of the model against which empirical data is compared. Indeed, it has been shown that under- or over-estimation of recombination greatly affect the power of tests, mainly those based on haplotypes and linkage disequilibrium (LD; Ramirez-Soriano *et al.* 2008). Secondly, bias can be introduced in the detection or selection of SNPs as most tests, especially those based on the frequency spectrum of mutation, depend on low-frequency variants that are frequently lost if data is not carefully resequenced.

Besides those two well-known limitations of the use of coalescent methods to detect departures from neutrality, a third issue could affect the power of neutrality tests and that, to the best of our knowledge, has not yet been explored. When such methods are used there is a crucial difference between simulation data (used to build the distribution against which to compare) and empirical, diploid data obtained from resequencing, as the former provide real haplotypes while the latter have to be phased. Furthermore, as said above, resequencing often does not provide complete information on all loci, and a number of missing genotypes can be introduced. The most widely method to estimate haplotypes at this moment is Phase (Stephens *et al.* 2001; Stephens and Donnelly. 2003), which uses a Bayesian algorithm based on the known haplotypes (that is, on homozygote individuals); although fastPHASE (Scheet and Stephens. 2006), which uses an expectation-maximization (EM) algorithm based on dynamic haplotype clustering, is becoming increasingly popular as larger amounts of data are generated. Both methods are inconvenient as they can lead to the loss of rare variants. This can be more important if the data contains missing genotypes, which have to be inferred.

Moreover, high recombination introduces uncertainty in the estimation of haplotypes as it increases the number of different combinations.

In this work, we investigate how the reconstruction of haplotypes using the algorithm implemented in fastPhase affects the power of tests both in neutral models and under several demographic scenarios taking also into account the presence of missing genotypes.

METHODS

Statistics

We have tested the power of statistics based on a) the frequency spectrum of mutation (Class I), which use the differences between estimators of the population mutation rate $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate; and b) based on linkage disequilibrium and haplotype distribution (Class II), which are expected to be the most affected by recombination. From Class I, we present results for Tajima's D (Tajima. 1989), Fu and Li's D, F, $D*$ and $F*$ (Fu and Li. 1993), Fay and Wu's H (Fay and Wu. 2000), and R_2(Ramos-Onsins and Rozas. 2002), which is based on the difference between the number of singletons per sequence and the average number of nucleotide differences. Within Class II we have studied Fu's F_s (Fu. 1997), the unbiased haplotype diversity estimate *Dh* (Nei. 1987, equation 8.4 replacing 2n by n), Wall's B and Q (Wall. 1999), Kelly's Z_{nS} (Kelly. 1997), Rozas' Z_A and ZZ (Rozas *et al.* 2001), and two statistics based on the Extended Haplotype Homozygosity, *EHH* (Sabeti *et al.* 2002; Ramirez-Soriano *et al.* 2008) . Furthermore, we have also taken into account in all cases the number of segregating sites (S), the number of pairwise differences (π), and the number of singletons.

Coalescent Simulations

Simulations have been performed using the ms package (Hudson. 2002), a program that generates coalescent trees using the algorithm described by Hudson (1990). ms assumes an infinite-sites mutation model and can simulate any given population size, recombination rate, and demographic scenario.

Simulations have been run conditioned on the number of segregating sites (S), which have been fixed to 10 and 100. These values are representative for small and large S values, as S has been shown to affect the power of statistical tests (Ramos-Onsins and Rozas. 2002). These S values correspond to the rounded minimum and average number of segregating sites found in the genes resequenced by SeattleSNPs (http://pga.gs.washington.edu/; Crawford *et al.* 2005) and have been associated to sequence lengths of 3000 and 21000, which also corresponds to the rounded minimum and average lengths resequenced by them. One hundred chromosomes have been simulated per sample in order to create 50 individuals to phase. Simulations where performed with and without recombination. When applied, recombination rates were set to $r=10^{-10}$, $r=10^{-8}$, and $r=10^{-7}$ per nucleotide pair. These values correspond to the rounded minimum, average and maximum values found by Kong *et al.* (2002) in the human genome.

Scenarios

We have simulated two models: stationarity and sudden population growth, under the different conditions of S, n, and recombination listed above. For each scenario we have ran 10,000 simulations. As mutations were simulated under an infinite-sites model (which implies no recurrent mutation), for those statistics that require an outgroup it has been set to a string of 0's (the ancestral state as coded by ms).

The sudden population growth model (Rogers and Harpending. 1992) assumes that an initial population in equilibrium of size N_0 experienced a sudden growth and reached a size N_{max} T_e generations (scaled in units of 4N generations) before present. We have set two expansions at T_e = 0.05 and T_e = 0.2, as these times have been shown to have maximum and lowest power in a previous work (Ramirez-Soriano *et al.* 2008). The degree of expansion (D_e=N_{max}/N_0) has been set to D_e=10, that is, a 10-fold population increase.

Haplotype reconstruction

Haplotypes have been reconstructed using fastPHASE (http://www.stat.washington.edu/stephens/software.html; Scheet and Stephens. 2006), version 1.1. In order to reconstruct haplotypes we have generated one fastPhase input file for each of the 10,000 samples obtained with ms for each scenario. The 100 chromosomes of each sample have been paired randomly as to generate 50 individuals. For each condition, the 10,000 fastPHASE input infiles were created four times each, introducing in them either no missing genotypes or 5%, 10%, 15% missing genotypes.

Phase accuracy

The accuracy of fastPhase in the haplotype reconstruction has been estimated by comparing the original haplotypes with the reconstructed ones. To do that we have created a script that compares, locus by locus, the haplotypes before and after phasing. This scripts provides four measures of phase accuracy (a) the fraction of correctly estimated haplotypes, (b) the average number of incorrectly estimated positions per haplotype, (c) the average number of incorrectly estimated positions per incorrect haplotype, and (d) the number of incorrectly estimated positions divided by total number of positions, that is, the number of haplotypes multiplied by the number of segregating sites per haplotype.

Effects of phasing in neutrality statistics

How neutrality tests are affected by haplotype reconstruction is explored in three different ways, listed below. In all cases, a two-tailed α=0.95 has been used.

First, we have estimated the type I error committed as a result of phasing by comparing the neutrality statistics computed on (a) the simulated haplotypes as produced by ms against (b) the same haplotypes, grouped in pairs and phased. This procedure has been applied to the different demographic scenarios and to the different fractions of missing genotypes.

Secondly, we have compared each non-neutral scenario with its corresponding neutral model in order to estimate the power of each test to detect the event of interest. Afterwards, we have repeated this analysis but using the phased files. Again, this comparison has been repeated for each fraction of missing genotypes. Finally, in order to assess the effect of phasing on the power of the tests, we have calculated the difference in the power of the tests between both comparisons.

RESULTS

Phase reconstruction

Figure 1 shows the accuracy of fastPHASE in haplotype reconstruction. As expected, it is mainly dependant on the number of segregating sites, on the fraction of missing genotypes in the sample and on recombination. The larger effect on accuracy is seen for *S*=100 and for high levels of recombination (r=10^{-7}). However, the effect of high recombination is almost negligible for *S*=10. Furthermore, fastPhase is more accurate in the reconstruction for the constant size model than in sudden population expansions. This could be explained by the increase in the number of low frequency variants in the latter scenario.

Type I error analysis

Type I error is directly influenced by the accuracy of fastPhase reconstruction, increasing with the errors in haplotype estimation. In all simulation conditions the error committed is around 0.05, that is, what is expected using α=0.95; or lower (see Figure 2 for an example with an expansion at T_e=0.05). In all demographic scenarios the main exceptions to this are Fu's F_s and *Dh*, for *S*=10; and Fu's F_s, *Dh*, *EHH average* and *ZZ* for *S*=100. The largest errors are found for F_s, *Dh*, which in absence of missing genotypes reach errors above 0.60, and, with a 15% missing genotypes, of well above 0.90 (*S*=100 in both cases). The error in all those tests also increases with recombination, which is consistent with the fact that they are based on haplotypes. When a number of missing genotypes is included in the sample before phasing, the number of segregating sites after reconstruction shows also a high type I error (over 0.90 for *S*=100) when compared with the original *S*. However, this effect is not shared by singletons. The number of pairwise differences is also slightly affected by phasing in presence of intermediate and high fractions of missing genotypes (between 0.06 and 0.1).

Power of the tests after phasing

Figure 3 shows the difference in the power of the neutrality statistics due to statistical phasing for an expansion at T_e=0.05. It can be seen that the difference in the the power of the statistics if data has or has not been phased (computed as phased against neutral model minus unphased against neutral model) is mainly found between 0.06 and -0.06, which implies that most tests are not affected by phasing. This same pattern can be found for all the demographic scenarios tested (data not shown). The most widely affected statistics are Fu's F_s, *Dh*, Z_{nS}, Z_A, and *ZZ*, all belonging to Class II. However, their error is not the same under the different parameters. F_s is liberal in all conditions except in T_e=0.05, which could be explained by the fact that, at this time, the power of this test is nearly saturated (Ramirez-Soriano *et al.* 2008). *Dh* is conservative except in T_e=0.05 for *S*=100 and recombination values below r=10^{-7}, where it is liberal, and T_e=0.2 for *S*=10, where is not affected. Z_A is conservative in most conditions, while Z_{nS} is only affected for *S*=10 and *ZZ* for *S*=100.

DISCUSSION

We have explored in this work the possible effects of comparing empirical, phased diploid data against simulated data in order to see if the power of neutrality tests is affected by it. Although the most widely used tool to phase genotypes is PHASE (Stephens *et al.* 2001; Stephens and Donnelly. 2003), instead of it we have used fastPHASE (Scheet and Stephens. 2006), a faster algorithm developed to be used in large genotyping projects. In the framework of this work, the choice of the latter is justified by the large amounts of files to be phased, which made unfeasible the use of PHASE. However, given that fastPHASE has been shown to perform missing imputation better than PHASE but is slightly less efficient reconstructing haplotypes (Scheet and Stephens. 2006), we believe the conclusions of this work can be extrapolated to the use of PHASE.

Scheet and Stephens (2006) tested the accuracy of fastPHASE in haplotype reconstruction using a set of X-chromosome data and HapMap simulated and empirical data. They considered two types of error: (a) the proportion of ambiguous individuals whose haplotypes are not completely correct (individual error), and (b) the proportion of heterozygote genotypes that are not correctly phased relative to the previous heterozygote genotype (the switch error). For the X-chromosome data they obtained errors of 0.654 and 0.111, respectively, and of 0.879 and 0.055 respectively for HapMap empirical data. Those amount of errors are consistent with the values found in our data for S=100.

We have shown that the power of neutrality statistics to detect departures from neutrality when data is phased is mainly not affected in Class I statistics, while it is for most Class II, especially for F_S and *Dh*. When found, errors tend to be conservative, except in some statistics (Z_{nS}, Z_A and *ZZ*) that are mostly liberal. These results, then, show that although fastPHASE accuracy may not be perfect, that does not represent a problem to be taken into consideration when using neutrality statistics on diploid, phased data.

ACKNOWLEDGEMENTS

We would like to thank Arcadi Navarro for his comments on this manuscript. fastPHASE has been run in the supercomputers MareNostrum, at Barcelona Supercomputing Center - Centro Nacional de Supercomputación (The Spanish National Supercomputing Center), and Altamira, at the Universidad de Cantabria, both belonging to the Red Nacional de Supercomputación. We would like to thank especially Jorge Naranjo for his technical assistance in preparing the system for running fastPHASE in both supercomputers. Bioinformatic assistance was provided by the INB (National Bioinformatics Institute, Spain).

LITERATURE CITED

Crawford, D. C., D. T. Akey and D. A. Nickerson, 2005 The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet **6:** 287-312.

Depaulis, F., S. Mousset and M. Veuille, 2003 Power of neutrality tests to detect bottlenecks and hitchhiking. J. Mol. Evol. **57 Suppl 1:** 190-200.

Donnelly, P., and S. Tavare, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401-21.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive darwinian selection. Genetics **155:** 1405-13.

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

Fu, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915-25.

Fu, Y. X., and W. H. Li, 1999 Coalescing into the 21st century: An overview and prospects of coalescent theory. Theor. Popul. Biol. **56:** 1-10.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

Hein, J., M. H. Schierup and C. Wiuf, 2005 *Gene Genealogies,Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.

Hudson, R. R., 2002 Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18:** 337-8.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1-44 in *Oxford Surveys in Evolutionary Biology*, edited by J. Antonovics and D. Futuyama. Oxford University Press, Oxford.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197-206.

Kingman, J. F., 2000 Origins of the coalescent. 1974-1982. Genetics **156:** 1461-3.

Kingman, J. F. C., 1982a On the genealogy of large populations. J. Appl. Prob. **19A:** 27-43.

Kingman, J. F. C., 1982b The coalescent. Stoch.Proc.Applns. **13:** 235-248.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.* 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241-7.

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

Ramirez-Soriano, A., S. E. Ramos-Onsins, J. Rozas, F. Calafell and A. Navarro, 2008 Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics **179:** 555-567.

Ramos-Onsins, S. E., and J. Rozas, 2002 Statistical properties of new neutrality tests against population growth. Mol. Biol. Evol. **19:** 2092-100.

Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552-69.

Rozas, J., M. Gullaud, G. Blandin and M. Aguade, 2001 DNA variation at the rp49 gene region of drosophila simulans: Evolutionary inferences from an unusual haplotype structure. Genetics **158:** 1147-55.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.* 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832-7.

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. Am. J. Hum. Genet. **78:** 629-644.

Stephens, M., and P. Donnelly, 2003 A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. **73:** 1162-9.

Stephens, M., N. J. Smith and P. Donnelly, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68:** 978-989.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-95.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65-79.

Wright, S., 1931 Evolution in mendelian populations. Genetics 97-159.

FIGURE LEGENDS

Figure 1. Fraction of correctly estimated haplotypes in the neutral model (pale colours, CSM) and in the sudden expansion model at T_e = 0.05 (dark colours, SEM). At the left side of the graph are shown results for S=10, and at the right side for S=100. m0_CSM, m5_CSM, etc. indicate the fraction of missings.

Figure 2. Type I error for the sudden expansion model at T_e = 0.05. In blue are shown the values <0.04, in green 0.04<x<0.06 and in red the values >0.06. A. S=10. B. S=100.

Figure 3. Difference between the power of neutrality tests due to phasing for the sudden expansion model at T_e = 0.05. -0.04<x<0.04 is depicted in green, 0.04>x>0.06 and -0.04>x>-0.06 in blue and x>0.06 and x<-0.06 in red. Negative values are in light colours, and positive in dark. A. S=10. B. S=100.

Figure 1

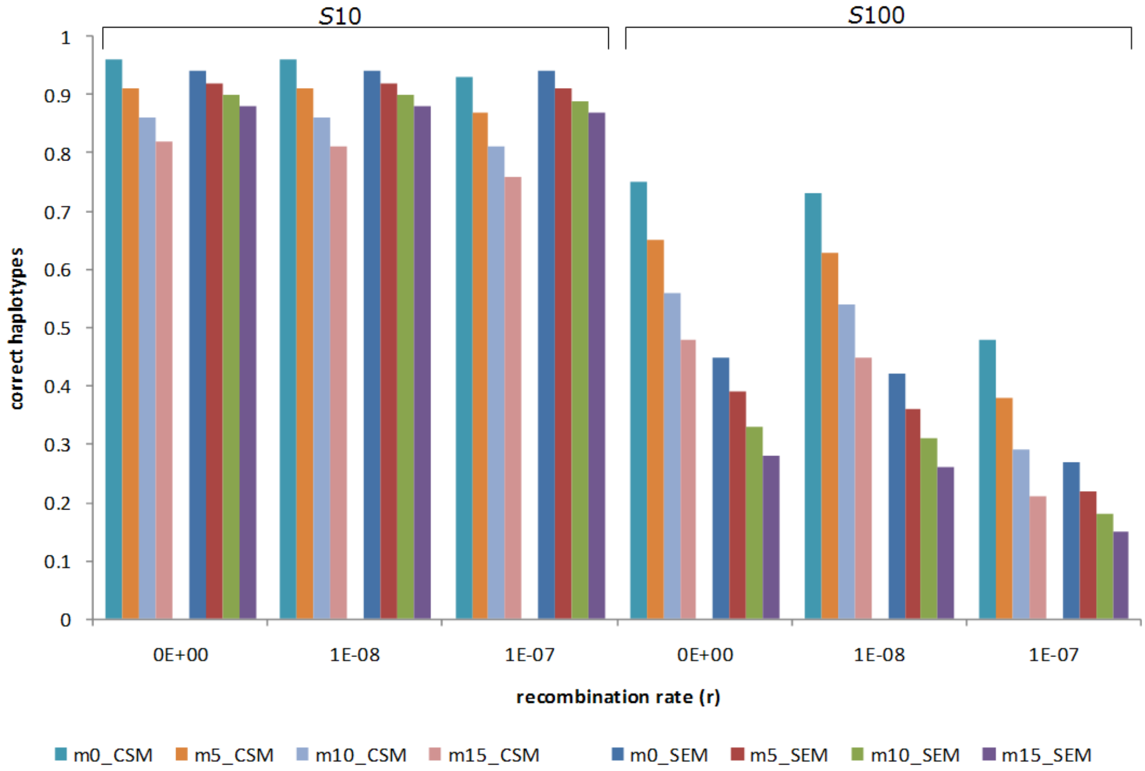
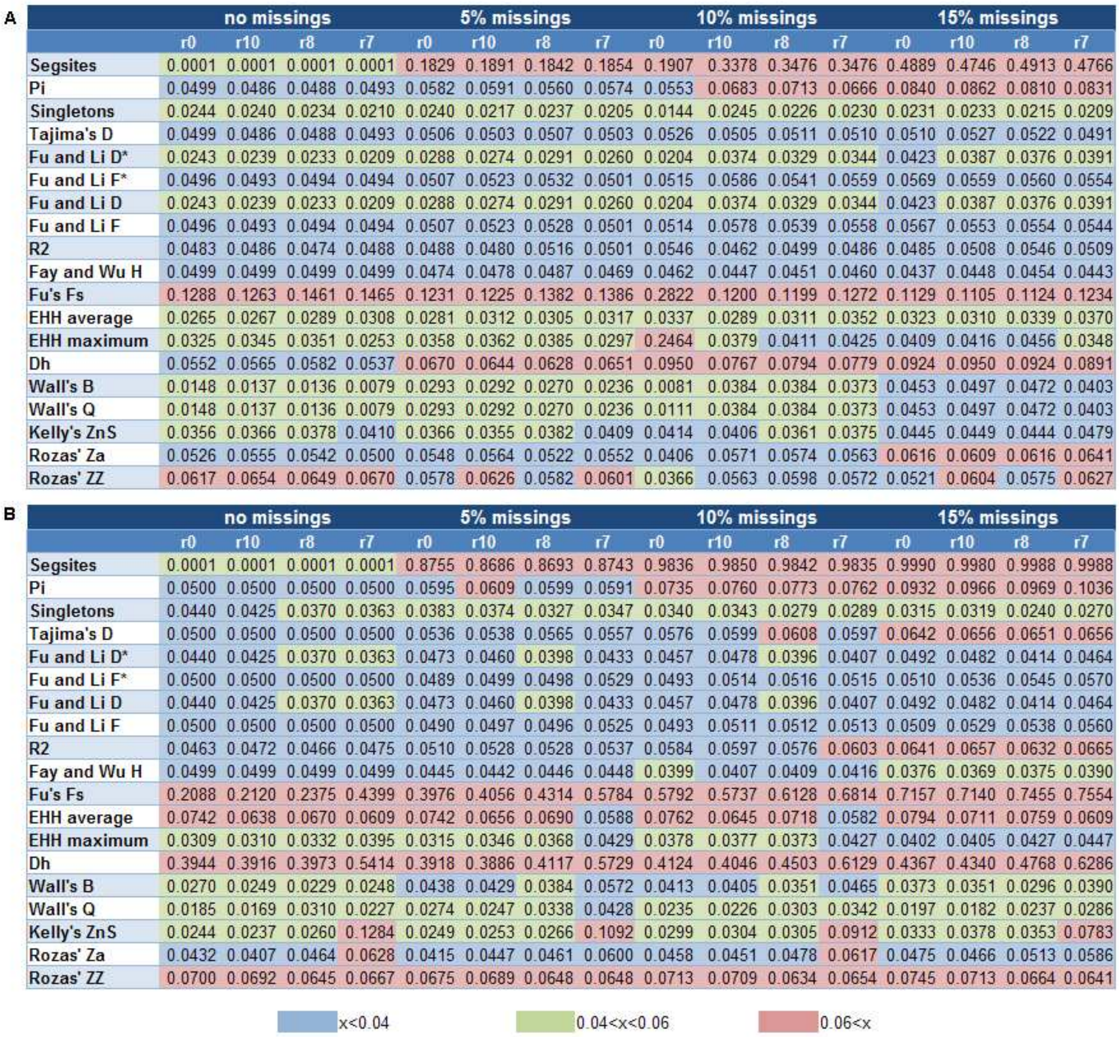

S10
S100
correct haplotypes
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0E+00
1E-08
1E-07
0E+00
1E-08
1E-07
recombination rate (r)
m0_CSM
m5_CSM
m10_CSM
m15_CSM
m0_SEM
m5_SEM
m10_SEM
m15_SEM

Figure 2



A

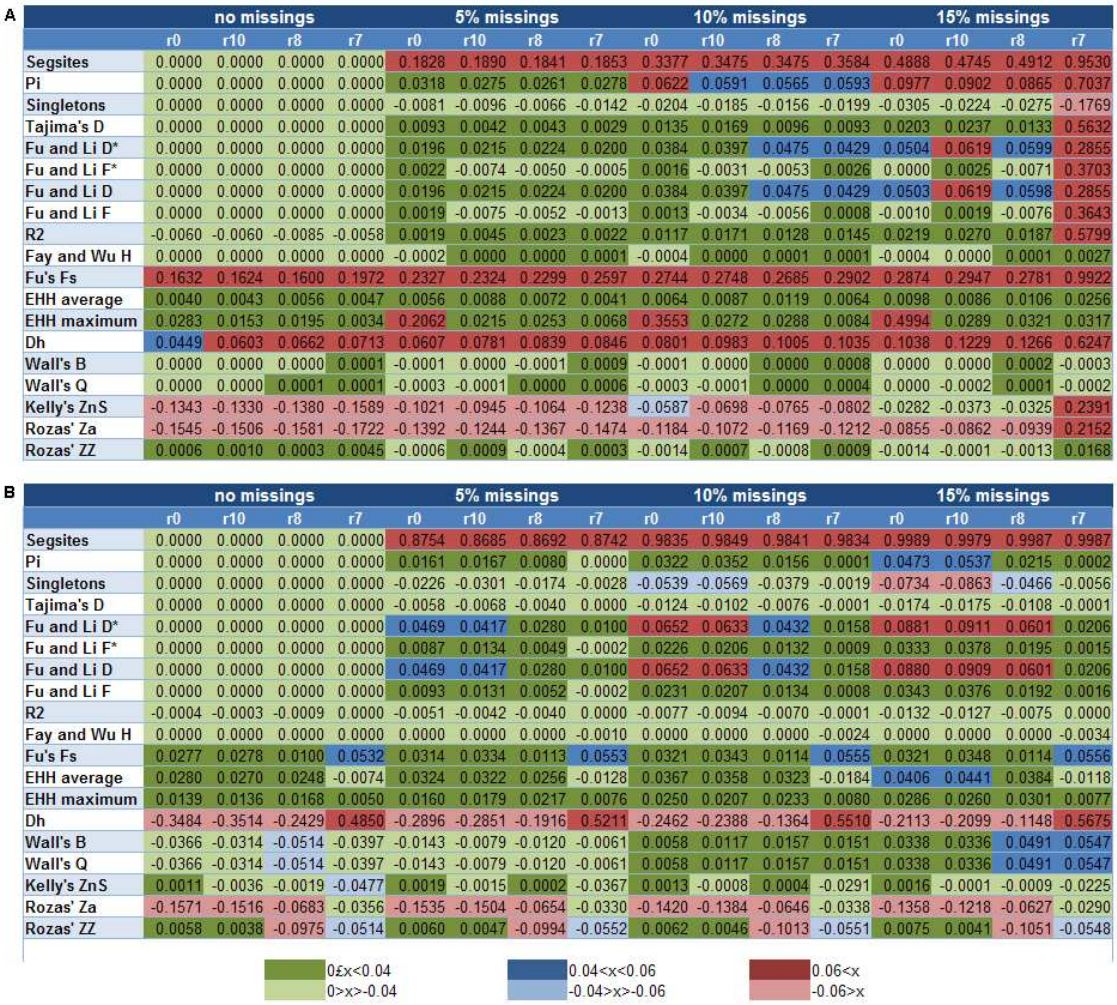
	no missings				5% missings				10% missings				15% missings			
	r0	r10	r8	r7	r0	r10	r8	r7	r0	r10	r8	r7	r0	r10	r8	r7
Segsites	0.0001	0.0001	0.0001	0.0001	0.1829	0.1891	0.1842	0.1854	0.1907	0.3378	0.3476	0.3476	0.4889	0.4746	0.4913	0.4766
Pi	0.0499	0.0486	0.0488	0.0493	0.0582	0.0591	0.0560	0.0574	0.0553	0.0683	0.0713	0.0666	0.0840	0.0862	0.0810	0.0831
Singletons	0.0244	0.0240	0.0234	0.0210	0.0240	0.0217	0.0237	0.0205	0.0144	0.0245	0.0226	0.0230	0.0231	0.0233	0.0215	0.0209
Tajima's D	0.0499	0.0486	0.0488	0.0493	0.0506	0.0503	0.0507	0.0503	0.0526	0.0505	0.0511	0.0510	0.0510	0.0527	0.0522	0.0491
Fu and Li D*	0.0243	0.0239	0.0233	0.0209	0.0288	0.0274	0.0291	0.0260	0.0204	0.0374	0.0329	0.0344	0.0423	0.0387	0.0376	0.0391
Fu and Li F*	0.0496	0.0493	0.0494	0.0494	0.0507	0.0523	0.0532	0.0501	0.0515	0.0586	0.0541	0.0559	0.0569	0.0559	0.0560	0.0554
Fu and Li D	0.0243	0.0239	0.0233	0.0209	0.0288	0.0274	0.0291	0.0260	0.0204	0.0374	0.0329	0.0344	0.0423	0.0387	0.0376	0.0391
Fu and Li F	0.0496	0.0493	0.0494	0.0494	0.0507	0.0523	0.0528	0.0501	0.0514	0.0578	0.0539	0.0558	0.0567	0.0553	0.0554	0.0544
R2	0.0483	0.0486	0.0474	0.0488	0.0488	0.0480	0.0516	0.0501	0.0546	0.0462	0.0499	0.0486	0.0485	0.0508	0.0546	0.0509
Fay and Wu H	0.0499	0.0499	0.0499	0.0499	0.0474	0.0478	0.0487	0.0469	0.0462	0.0447	0.0451	0.0460	0.0437	0.0448	0.0454	0.0443
Fu's Fs	0.1288	0.1263	0.1461	0.1465	0.1231	0.1225	0.1382	0.1386	0.2822	0.1200	0.1199	0.1272	0.1129	0.1105	0.1124	0.1234
EHH average	0.0265	0.0267	0.0289	0.0308	0.0281	0.0312	0.0305	0.0317	0.0337	0.0289	0.0311	0.0352	0.0323	0.0310	0.0339	0.0370
EHH maximum	0.0325	0.0345	0.0351	0.0253	0.0358	0.0362	0.0385	0.0297	0.2464	0.0379	0.0411	0.0425	0.0409	0.0416	0.0456	0.0348
Dh	0.0552	0.0565	0.0582	0.0537	0.0670	0.0644	0.0628	0.0651	0.0950	0.0767	0.0794	0.0779	0.0924	0.0950	0.0924	0.0891
Wall's B	0.0148	0.0137	0.0136	0.0079	0.0293	0.0292	0.0270	0.0236	0.0081	0.0384	0.0384	0.0373	0.0453	0.0497	0.0472	0.0403
Wall's Q	0.0148	0.0137	0.0136	0.0079	0.0293	0.0292	0.0270	0.0236	0.0111	0.0384	0.0384	0.0373	0.0453	0.0497	0.0472	0.0403
Kelly's ZnS	0.0356	0.0366	0.0378	0.0410	0.0366	0.0355	0.0382	0.0409	0.0414	0.0406	0.0361	0.0375	0.0445	0.0449	0.0444	0.0479
Rozas' Za	0.0526	0.0555	0.0542	0.0500	0.0548	0.0564	0.0522	0.0552	0.0406	0.0571	0.0574	0.0563	0.0616	0.0609	0.0616	0.0641
Rozas' ZZ	0.0617	0.0654	0.0649	0.0670	0.0578	0.0626	0.0582	0.0601	0.0366	0.0563	0.0598	0.0572	0.0521	0.0604	0.0575	0.0627

B

	no missings				5% missings				10% missings				15% missings			
	r0	r10	r8	r7	r0	r10	r8	r7	r0	r10	r8	r7	r0	r10	r8	r7
Segsites	0.0001	0.0001	0.0001	0.0001	0.8755	0.8686	0.8693	0.8743	0.9836	0.9850	0.9842	0.9835	0.9990	0.9980	0.9988	0.9988
Pi	0.0500	0.0500	0.0500	0.0500	0.0595	0.0609	0.0599	0.0591	0.0735	0.0760	0.0773	0.0762	0.0932	0.0966	0.0969	0.1036
Singletons	0.0440	0.0425	0.0370	0.0363	0.0383	0.0374	0.0327	0.0347	0.0340	0.0343	0.0279	0.0289	0.0315	0.0319	0.0240	0.0270
Tajima's D	0.0500	0.0500	0.0500	0.0500	0.0536	0.0538	0.0565	0.0557	0.0576	0.0599	0.0608	0.0597	0.0642	0.0656	0.0651	0.0656
Fu and Li D*	0.0440	0.0425	0.0370	0.0363	0.0473	0.0460	0.0398	0.0433	0.0457	0.0478	0.0396	0.0407	0.0492	0.0482	0.0414	0.0464
Fu and Li F*	0.0500	0.0500	0.0500	0.0500	0.0489	0.0499	0.0498	0.0529	0.0493	0.0514	0.0516	0.0515	0.0510	0.0536	0.0545	0.0570
Fu and Li D	0.0440	0.0425	0.0370	0.0363	0.0473	0.0460	0.0398	0.0433	0.0457	0.0478	0.0396	0.0407	0.0492	0.0482	0.0414	0.0464
Fu and Li F	0.0500	0.0500	0.0500	0.0500	0.0490	0.0497	0.0496	0.0525	0.0493	0.0511	0.0512	0.0513	0.0509	0.0529	0.0538	0.0560
R2	0.0463	0.0472	0.0466	0.0475	0.0510	0.0528	0.0528	0.0537	0.0584	0.0597	0.0576	0.0603	0.0641	0.0657	0.0632	0.0665
Fay and Wu H	0.0499	0.0499	0.0499	0.0499	0.0445	0.0442	0.0446	0.0448	0.0399	0.0407	0.0409	0.0416	0.0376	0.0369	0.0375	0.0390
Fu's Fs	0.2088	0.2120	0.2375	0.4399	0.3976	0.4056	0.4314	0.5784	0.5792	0.5737	0.6128	0.6814	0.7157	0.7140	0.7455	0.7554
EHH average	0.0742	0.0638	0.0670	0.0609	0.0742	0.0656	0.0690	0.0588	0.0762	0.0645	0.0718	0.0582	0.0794	0.0711	0.0759	0.0609
EHH maximum	0.0309	0.0310	0.0332	0.0395	0.0315	0.0346	0.0368	0.0429	0.0378	0.0377	0.0373	0.0427	0.0402	0.0405	0.0427	0.0447
Dh	0.3944	0.3916	0.3973	0.5414	0.3918	0.3886	0.4117	0.5729	0.4124	0.4046	0.4503	0.6129	0.4367	0.4340	0.4768	0.6286
Wall's B	0.0270	0.0249	0.0229	0.0248	0.0438	0.0429	0.0384	0.0572	0.0413	0.0405	0.0351	0.0465	0.0373	0.0351	0.0296	0.0390
Wall's Q	0.0185	0.0169	0.0310	0.0227	0.0274	0.0247	0.0338	0.0428	0.0235	0.0226	0.0303	0.0342	0.0197	0.0182	0.0237	0.0286
Kelly's ZnS	0.0244	0.0237	0.0260	0.1284	0.0249	0.0253	0.0266	0.1092	0.0299	0.0304	0.0305	0.0912	0.0333	0.0378	0.0353	0.0783
Rozas' Za	0.0432	0.0407	0.0464	0.0628	0.0415	0.0447	0.0461	0.0600	0.0458	0.0451	0.0478	0.0617	0.0475	0.0466	0.0513	0.0586
Rozas' ZZ	0.0700	0.0692	0.0645	0.0667	0.0675	0.0689	0.0648	0.0648	0.0713	0.0709	0.0634	0.0654	0.0745	0.0713	0.0664	0.0641

🟦 x<0.04 🟩 0.04<x<0.06 🟥 0.06<x

Figure 3



A

	no missings				5% missings				10% missings				15% missings			
	r0	r10	r8	r7	r0	r10	r8	r7	r0	r10	r8	r7	r0	r10	r8	r7
Segsites	0.0000	0.0000	0.0000	0.0000	0.1828	0.1890	0.1841	0.1853	0.3377	0.3475	0.3475	0.3584	0.4888	0.4745	0.4912	0.9530
Pi	0.0000	0.0000	0.0000	0.0000	0.0318	0.0275	0.0261	0.0278	0.0622	0.0591	0.0565	0.0593	0.0977	0.0902	0.0865	0.7037
Singletons	0.0000	0.0000	0.0000	0.0000	-0.0081	-0.0096	-0.0066	-0.0142	-0.0204	-0.0185	-0.0156	-0.0199	-0.0305	-0.0224	-0.0275	-0.1769
Tajima's D	0.0000	0.0000	0.0000	0.0000	0.0093	0.0042	0.0043	0.0029	0.0135	0.0169	0.0096	0.0093	0.0203	0.0237	0.0133	0.5632
Fu and Li D*	0.0000	0.0000	0.0000	0.0000	0.0196	0.0215	0.0224	0.0200	0.0384	0.0397	0.0475	0.0429	0.0504	0.0619	0.0599	0.2855
Fu and Li F*	0.0000	0.0000	0.0000	0.0000	0.0022	-0.0074	-0.0050	-0.0005	0.0016	-0.0031	-0.0053	0.0026	0.0000	0.0025	-0.0071	0.3703
Fu and Li D	0.0000	0.0000	0.0000	0.0000	0.0196	0.0215	0.0224	0.0200	0.0384	0.0397	0.0475	0.0429	0.0503	0.0619	0.0598	0.2855
Fu and Li F	0.0000	0.0000	0.0000	0.0000	0.0019	-0.0075	-0.0052	-0.0013	0.0013	-0.0034	-0.0056	0.0008	-0.0010	0.0019	-0.0076	0.3643
R2	-0.0060	-0.0060	-0.0085	-0.0058	0.0019	0.0045	0.0023	0.0022	0.0117	0.0171	0.0128	0.0145	0.0219	0.0270	0.0187	0.5799
Fay and Wu H	0.0000	0.0000	0.0000	0.0000	-0.0002	0.0000	0.0000	0.0001	-0.0004	0.0000	0.0001	0.0001	-0.0004	0.0000	0.0001	0.0027
Fu's Fs	0.1632	0.1624	0.1600	0.1972	0.2327	0.2324	0.2299	0.2597	0.2744	0.2748	0.2685	0.2902	0.2874	0.2947	0.2781	0.9922
EHH average	0.0040	0.0043	0.0056	0.0047	0.0056	0.0088	0.0072	0.0041	0.0064	0.0087	0.0119	0.0064	0.0098	0.0086	0.0106	0.0256
EHH maximum	0.0283	0.0153	0.0195	0.0034	0.2062	0.0215	0.0253	0.0068	0.3553	0.0272	0.0288	0.0084	0.4994	0.0289	0.0321	0.0317
Dh	0.0449	0.0603	0.0662	0.0713	0.0607	0.0781	0.0839	0.0846	0.0801	0.0983	0.1005	0.1035	0.1038	0.1229	0.1266	0.6247
Wall's B	0.0000	0.0000	0.0000	0.0001	-0.0001	0.0000	-0.0001	0.0009	-0.0001	0.0000	0.0000	0.0008	0.0000	0.0000	0.0002	-0.0003
Wall's Q	0.0000	0.0000	0.0001	0.0001	-0.0003	-0.0001	0.0000	0.0006	-0.0003	-0.0001	0.0000	0.0004	0.0000	-0.0002	0.0001	-0.0002
Kelly's ZnS	-0.1343	-0.1330	-0.1380	-0.1589	-0.1021	-0.0945	-0.1064	-0.1238	-0.0587	-0.0698	-0.0765	-0.0802	-0.0282	-0.0373	-0.0325	0.2391
Rozas' Za	-0.1545	-0.1506	-0.1581	-0.1722	-0.1392	-0.1244	-0.1367	-0.1474	-0.1184	-0.1072	-0.1169	-0.1212	-0.0855	-0.0862	-0.0939	0.2152
Rozas' ZZ	0.0006	0.0010	0.0003	0.0045	-0.0006	0.0009	-0.0004	0.0003	-0.0014	0.0007	-0.0008	0.0009	-0.0014	-0.0001	-0.0013	0.0168

B

	no missings				5% missings				10% missings				15% missings			
	r0	r10	r8	r7	r0	r10	r8	r7	r0	r10	r8	r7	r0	r10	r8	r7
Segsites	0.0000	0.0000	0.0000	0.0000	0.8754	0.8685	0.8692	0.8742	0.9835	0.9849	0.9841	0.9834	0.9989	0.9979	0.9987	0.9987
Pi	0.0000	0.0000	0.0000	0.0000	0.0161	0.0167	0.0080	0.0000	0.0322	0.0352	0.0156	0.0001	0.0473	0.0537	0.0215	0.0002
Singletons	0.0000	0.0000	0.0000	0.0000	-0.0226	-0.0301	-0.0174	-0.0028	-0.0539	-0.0569	-0.0379	-0.0019	-0.0734	-0.0863	-0.0466	-0.0056
Tajima's D	0.0000	0.0000	0.0000	0.0000	-0.0058	-0.0068	-0.0040	0.0000	-0.0124	-0.0102	-0.0076	-0.0001	-0.0174	-0.0175	-0.0108	-0.0001
Fu and Li D*	0.0000	0.0000	0.0000	0.0000	0.0469	0.0417	0.0280	0.0100	0.0652	0.0633	0.0432	0.0158	0.0881	0.0911	0.0601	0.0206
Fu and Li F*	0.0000	0.0000	0.0000	0.0000	0.0087	0.0134	0.0049	-0.0002	0.0226	0.0206	0.0132	0.0009	0.0333	0.0378	0.0195	0.0015
Fu and Li D	0.0000	0.0000	0.0000	0.0000	0.0469	0.0417	0.0280	0.0100	0.0652	0.0633	0.0432	0.0158	0.0880	0.0909	0.0601	0.0206
Fu and Li F	0.0000	0.0000	0.0000	0.0000	0.0093	0.0131	0.0052	-0.0002	0.0231	0.0207	0.0134	0.0008	0.0343	0.0376	0.0192	0.0016
R2	-0.0004	-0.0003	-0.0009	0.0000	-0.0051	-0.0042	-0.0040	0.0000	-0.0077	-0.0094	-0.0070	-0.0001	-0.0132	-0.0127	-0.0075	0.0000
Fay and Wu H	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0010	0.0000	0.0000	0.0000	-0.0024	0.0000	0.0000	0.0000	-0.0034
Fu's Fs	0.0277	0.0278	0.0100	0.0532	0.0314	0.0334	0.0113	0.0553	0.0321	0.0343	0.0114	0.0555	0.0321	0.0348	0.0114	0.0556
EHH average	0.0280	0.0270	0.0248	-0.0074	0.0324	0.0322	0.0256	-0.0128	0.0367	0.0358	0.0323	-0.0184	0.0406	0.0441	0.0384	-0.0118
EHH maximum	0.0139	0.0136	0.0168	0.0050	0.0160	0.0179	0.0217	0.0076	0.0250	0.0207	0.0233	0.0080	0.0286	0.0260	0.0301	0.0077
Dh	-0.3484	-0.3514	-0.2429	0.4850	-0.2896	-0.2851	-0.1916	0.5211	-0.2462	-0.2388	-0.1364	0.5510	-0.2113	-0.2099	-0.1148	0.5675
Wall's B	-0.0366	-0.0314	-0.0514	-0.0397	-0.0143	-0.0079	-0.0120	-0.0061	0.0058	0.0117	0.0157	0.0151	0.0338	0.0336	0.0491	0.0547
Wall's Q	-0.0366	-0.0314	-0.0514	-0.0397	-0.0143	-0.0079	-0.0120	-0.0061	0.0058	0.0117	0.0157	0.0151	0.0338	0.0336	0.0491	0.0547
Kelly's ZnS	0.0011	-0.0036	-0.0019	-0.0477	0.0019	-0.0015	0.0002	-0.0367	0.0013	-0.0008	0.0004	-0.0291	0.0016	-0.0001	-0.0009	-0.0225
Rozas' Za	-0.1571	-0.1516	-0.0683	-0.0356	-0.1535	-0.1504	-0.0654	-0.0330	-0.1420	-0.1384	-0.0646	-0.0338	-0.1358	-0.1218	-0.0627	-0.0290
Rozas' ZZ	0.0058	0.0038	-0.0975	-0.0514	0.0060	0.0047	-0.0994	-0.0552	0.0062	0.0046	-0.1013	-0.0551	0.0075	0.0041	-0.1051	-0.0548

🟩	0£x<0.04	🟦	0.04<x<0.06	🟥	0.06<x
🟩	0>x>-0.04	🟦	-0.04>x>-0.06	🟥	-0.06>x

3 CORRECTING ESTIMATORS OF THETA AND TAJIMA'S D FOR ASCERTAINMENT BIASES CAUSED BY THE SNP DISCOVERY PROCESS (Ramirez-Soriano and Nielsen. submitted)

Correcting Estimators of θ and Tajima's D for ascertainment biases caused by the SNP discovery process

Anna Ramírez-Soriano[1,*] and Rasmus Nielsen[2]

[1] Departament de Ciències de la Salut i de la Vida. Universitat Pompeu Fabra. Doctor Aiguader, 88. 08003 Barcelona, Catalonia, Spain.

[2] Departments of Integrative Biology and Statistics, 3060 Valley Life Science Building, University of California Berkeley, Berkeley, CA, 94720-3140, and Department of Biology, University of Copenhagen, Universitetsparken 15, 2100 Kbh Ø, Copenhagen, Denmark

Running title: Tajima's *D* under ascertainment bias

Keywords: Single Nucleotide Polymorphisms, ascertainment bias, Watterson's estimator, Tajima's estimator, Tajima's *D*

***Corresponding author:**

Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra

Doctor Aiguader 88, 08003 Barcelona (Spain)

Phone: +34 93 3160803

Fax: +34 93 3160901

E-mail: anna.ramirez@upf.edu

ABSTRACT

Most single nucleotide polymorphism (SNP) data suffer from an ascertainment bias caused by the process of SNP discovery followed by SNP genotyping. The final genotyped data are biased towards an excess of common alleles compared to directly sequenced data, making standard genetic methods of analysis inapplicable to this type of data. We here derive corrected estimators of the fundamental population genetic parameter $\theta = 4N_e\mu$ (N_e = effective population size, μ = mutation rate) based on the average number of pairwise differences and based on the number of segregating sites. We also derive the variances and co-variances of these estimators, and provide a corrected version of Tajima's *D* statistic. We re-analyze a human genome-wide SNP data set and find substantial differences in the results with or without ascertainment bias correction.

INTRODUCTION

The HapMap data (International HapMap Consortium. 2007) and other genome-wide SNP data sets provide a valuable resource for population genetic analysis. Much interest in the analysis of such data has focused on estimating demographic parameters or inferring natural selection (e.g. Bamshad and Wooding. 2003; Wooding. 2004; Voight *et al.* 2006; Wang *et al.* 2006; Carlson *et al.* 2006; Sabeti *et al.* 2006; Tang *et al.* 2007; Williamson *et al.* 2007). However, many of the studies of genome-wide SNP data have been challenged by the fact that the SNP genotyping data have been obtained by a process in which SNPs are first discovered in a small panel of individuals and subsequently typed in a much larger panel (e.g. Picoult-Newberg *et al.* 1999; Altshuler *et al.* 2000; Mead *et al.* 2003). Although this procedure provides a much faster and cheaper way of generating data than direct sequencing of the full panel, it also produces data with a relative excess of alleles of intermediate frequencies compared to directly sequenced data. Rare SNPs are more easily discovered in large panels than in small panels, so an initial discovery process based on a small panel produces an excess of high frequency alleles in the genotyped sample. As a consequence, the data will be different from what is assumed in standard population genetic models with respect to allele frequency distribution (e.g. Nielsen. 2000; Wakeley *et al.* 2001), patterns of linkage disequilibrium (Nielsen and Signorovitch. 2003) , and level of population subdivision (Nielsen. 2004). This ascertainment bias towards high-frequency alleles can have serious consequences when standard population genetic tools (e.g. Tajima. 1989; Fu and Li. 1993; Fay and Wu. 2000; Ramos-Onsins and Rozas. 2002) are used for the analysis of the data. For example, Kreitman & Di Rienzo (2004) and Soldevila *et al.* (2005) showed that the apparent effects of balancing selection detected in the *PRPN* by Mead *et al.* (2003) in fact were an artifact caused by this type of ascertainment bias.

Three different approaches have been used to address the problem of ascertainment biases in studies of real data: i) applying methods that may be more robust to the effect of ascertainment bias, such as methods based on haplotype structure (e.g. Sabeti *et al.* 2002), ii) simulating data under the ascertainment bias to derive appropriate critical values and confidence intervals using a distribution which directly takes ascertainment into account (e.g. Carlson *et al.* 2004; Voight *et al.* 2006), and iii) directly correcting the statistical estimators and statistics for the ascertainment bias (e.g. Nielsen. 2000; Wakeley *et al.* 2001; Nielsen and Signorovitch. 2003; Polanski and Kimmel. 2003; Nielsen *et al.* 2004) in specific models. However, hitherto there have been no ascertainment correction methods available for some of the most basic population genetic tools. Here we derive ascertainment corrected estimators of the fundamental population genetic parameter $\theta = 4N_e\mu$ (N_e = effective population size, μ = mutation rate) and an ascertainment corrected version of the popular statistic used for detecting selection: Tajima's *D*.

THEORY AND METHODS

Estimators of θ

Tajima's *D* (Tajima. 1989) is calculated as the difference between Tajima's estimator of θ, θ_T, (Tajima. 1989) and Watterson's estimator of θ, θ_w (Watterson. 1975). Tajima's estimator is based on the average number of pairwise differences (π), and is given by

$$\hat{\theta}_T = \pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \eta_i \big(i(n-i)\big), \qquad (1)$$

where η_i is the number of (arbitrarily labeled) alleles segregating at a frequency of *i/n*, in a sample of *n* chromosomes. Watterson's estimator is given by

$$\hat{\theta}_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}, \qquad (2)$$

where *S* is the number of segregating sites.

We will assume an ascertainment model in which a subset of *d* chromosomes have been chosen independently among the *n* chromosomes for ascertainment. We further assume that the chromosomes chosen for ascertainment are independent among SNPs. The probability of ascertainment of a SNP with alleles of frequencies *i/n* and (*n – i)/n*, is then (Nielsen. 2004):

$$P_A(i) = 1 - \frac{\binom{i}{d} + \binom{n-i}{d}}{\binom{n}{d}} \quad , \qquad (3)$$

where we use the definition $\binom{n}{k} = 0$ if *k*>*n*. The final sample after ascertainment is denoted the *genotyped sample*.

The expected number of segregating sites in the genotyped sample under this ascertainment scheme, $S^{(A)}$, is then simply the sum over all allelic classes of the expected number of segregating sites of that allelic class (E[η_i] = θ/I; Tajima. 1989; Fu. 1995) multiplied by the probability of ascertainment of the allelic class:

$$E_A[S^{(A)}] = \sum_{i=1}^{n-1} E[\eta_i] P_A(i) = \sum_{i=1}^{n-1} \frac{\theta}{i} P_A(i) \qquad (4)$$

An unbiased method of moments estimator of θ, similar to Watterson's estimator is then given by

$$\hat{\theta}_{W,C} = \frac{S^{(A)}}{\sum_{i=1}^{n-1} \frac{P_A(i)}{i}} \tag{5}$$

The expected number of pairwise differences in the genotyped sample is similarly given by the sum over all allelic classes of the expected contribution to the pairwise differences of the allelic class multiplied by the probability of ascertainment of the allelic class

$$E_A[\pi^{(A)}] = \sum_{i=1}^{n-1} \frac{\theta}{i} \left(\frac{2i(n-i)}{n(n-1)} P_A(i) \right) = \frac{2\theta}{n(n-1)} \sum_{i=1}^{n-1} (n-i) P_A(i) \,. \tag{6}$$

An unbiased ascertainment corrected method of moments estimator similar to Tajima's estimator is then given by

$$\hat{\theta}_{T,C} = \frac{\pi^{(A)} n(n-1)}{2\sum_{i=1}^{n-1} P_A(i)(n-i)} = \frac{\sum_{i=1}^{n-1} \eta_i i(n-i)}{\sum_{i=1}^{n-1} P_A(i)(n-i)} \,. \tag{7}$$

Notice that these estimators are identical to the traditional estimators, $\hat{\theta}_W$ and $\hat{\theta}_T$, when there is no ascertainment bias, i.e. $P_A(i) = 1$.

Variances of the estimators

We will use notation and some results from chapter 2 of Durrett (2008), to derive covariance and variances of these estimators. In the absence of any ascertainment bias (Fu. 1995; Durrett. 2008):

$$Var(\eta_i) = \frac{\theta}{i} + \theta^2 \sigma_{ii} \quad \text{and} \quad Cov(\eta_i, \eta_j) = \sigma_{ij}\theta^2 \text{ for } i \neq j$$

where σ_{ii} equals:

$$\begin{array}{ll} \beta_n(i+1) & i < n/2 \\ 2\frac{h_n - h_i}{n-i} - \frac{1}{i^2} & i = n/2 \\ \beta_n(i) - \frac{1}{i^2} & i > n/2 \end{array}$$

(8)

σ_{ij} equals:

$$\begin{array}{ll} \dfrac{\beta_n(i+1)-\beta_n(i)}{2} & i+j<n \\ \\ \dfrac{h_n-h_i}{n-i}+\dfrac{h_n-hj}{n-j}-\dfrac{\beta_n(i)-\beta_n(j+1)}{2}-\dfrac{1}{ij} & i+j=n \\ \\ \dfrac{\beta_n(j)-\beta_n(j-1)}{2}-\dfrac{1}{ij} & i+j>n \end{array}$$

and:

$$\beta_n(i)=\frac{2n}{(n-i+1)(n-i)}(h_{n+1}-h_i)-\frac{2}{n-i};\quad h_n=\sum_{k=1}^{n-1}\frac{1}{k}.$$

Using the conditional variance formula

$$Var(\eta_i^{(A)})=E[Var(\eta_i^{(A)}\,|\,\eta_i)]+Var[E(\eta_i^{(A)}\,|\,\eta_i)],\tag{9}$$

with $Var[E(\eta_i^{(A)}\,|\,\eta_i)]=\eta_i P_A(i)(1-P_A(i))$ and $E[E(\eta_i^{(A)}\,|\,\eta_i)]=\eta_i P_A(i)$, we get

$$Var(\eta_i^{(A)})=E(\eta_i)P_A(i)(1-P_A(i))+Var(\eta_i)P_A(i)=\frac{\theta}{i}P_A(i)+(P_A(i))^2\theta^2\sigma_{ii}.\tag{10}$$

A similar calculation leads to

$$Cov(\eta_i^{(A)},\eta_j^{(A)})=P_A(i)P_A(j)\sigma_{ij}\theta^2,\ i\neq j.\tag{11}$$

We can then easily get the variance of $S^{(A)}$:

$$Var[S^{(A)}]=\sum_{i=1}^{n-1}\left(\frac{\theta}{i}P_A(i)+(P_A(i))^2\theta^2\sigma_{ii}\right)+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n-1}\left(P_A(i)P_A(j)\sigma_{ij}\theta^2\right).\tag{12}$$

The variance of the ascertainment corrected estimator of θ based on the number of segregating sites is then given by

$$Var(\hat{\theta}_{W,C})=V\left[\frac{S^{(A)}}{\sum_{i=1}^{n-1}\frac{P_A(i)}{i}}\right]=\left(\sum_{i=1}^{n-1}\frac{P_A(i)}{i}\right)^{-2}Var[S^{(A)}].\tag{13}$$

The variance of the estimator based on the average number of pairwise differences becomes

$$Var(\hat{\theta}_{T,C}) = \frac{Var\left(\sum_{i=1}^{n-1}\eta_i i(n-i)\right)}{\left(\sum_{i=1}^{n-1}P_A(i)(n-i)\right)^2}$$

$$= \frac{\left[\sum_{i=1}^{n-1}(i(n-i))^2\left(\frac{\theta}{i}P_A(i)+P_A(i)^2\theta^2\sigma_{ii}\right)+2\theta^2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n-1}i(n-i)j(n-j)P_A(i)P_A(j)\sigma_{ij}\right]}{\left(\sum_{i=1}^{n-1}P_A(i)(n-i)\right)^2}$$

(14)

Covariances and Tajima's *D*

Defining the following coefficients

$C_W^{(A)} = \left(\sum_{i=1}^{n-1}\frac{P_A(i)}{i}\right)^{-1}$ and $C_{Ti}^{(A)} = i(n-i)\left(\sum_{l=1}^{n-1}P_A(l)(n-l)\right)^{-1}$, we have

$$Cov(\hat{\theta}_{TC},\hat{\theta}_{WC}) = Cov(C_W^{(A)}\sum_{i=1}^{n-1}\eta_i^{(A)},\sum_{i=1}^{n-1}\eta_i^{(A)}C_{Ti}^{(A)})$$

$$= C_W^{(A)}\sum_{i==1}^{n-1}C_{Ti}^{(A)}\left(\frac{\theta}{i}P_A(i)+\sum_{j=1}^{n-1}\sigma_{ij}\theta^2P_A(i)P_A(j)\right). \qquad (15)$$

Also

$$Var(\hat{\theta}_{WC}-\hat{\theta}_{TC}) = V(\hat{\theta}_{WC})+V(\hat{\theta}_{TC})-2Cov(\hat{\theta}_{WC},\hat{\theta}_{TC})\,. \qquad (16)$$

We now define an ascertainment corrected Tajima's *D* as

$$D_C = \frac{\hat{\theta}_{WC}-\hat{\theta}_{TC}}{\sqrt{Var(\hat{\theta}_{WC}-\hat{\theta}_{TC})}}\,. \qquad (17)$$

To calculate $Var(\hat{\theta}_{WC}-\hat{\theta}_{TC})$ for real data we need to know the value of θ and θ^2. We will estimate θ using $\hat{\theta}_{WC}$. We estimate θ^2 as

$$\hat{\theta}^2 = \frac{S^2-S}{\left(\sum_{i==1}^{n-1}\frac{1}{i}P_a(i)\right)^2+\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\sigma_{ij}P_a(i)P_a(j)} \qquad (18)$$

The D_C statistic is identical to the traditional Tajima's D in the absence of an ascertainment, i.e. when $P_A(i) = 1$.

Simulations

Simulated data were generated using the standard coalescent simulation program ms (Hudson. 2002) with 10.000 and /or 1.000.000 replicates. We explored three different values of θ: 2.23, 22.33 and 89.30, corresponding to the estimates of θ based on Watterson's estimator calculated from the minimum, average and maximum number of segregating sites found in the genes represented in the SeattleSNP database (http://pga.gs.washington.edu/, Crawford *et al.* 2005). We also explored results for an extreme value of theta θ =150. To generate ascertainment samples from the simulated data, we subsampled d (= 2, 5, or 10) gene-copies from each segregating site in the sample of size n (= 20 or 50). If the segregating site was polymorphic in the sub-sample, it was included in the final the sample, otherwise it was ignored.

Perlegen Data

Genotype data from Perlegen was obtained from http://genome.perlegen.com/browser/download.html, and we used information regarding the ascertainment protocol discussed in Clark *et al.* (2005) and Hinds *et al.* (2005). In the analysis of the Perlegen data we have only use those SNPs that were obtained using the ascertainment protocol described above, which in Perlegen nomenclature corresponds to ascertainment class A (array-based genomic resequencing). 69% of all SNPs of all the Perlegen SNPs were obtained using this protocol.

For Perlegen data, Tajima's D was calculated chromosome by chromosome through a sliding window of 100 and 500 kb, sliding by 10 kb at a time. Only those windows which contained at least 10 class A SNPs were included in the analysis. As Perlegen genotypes contain missing data we have corrected the sample size and the nucleotide diversity for each window. To correct the sample size we have used the average number of sequences at each segregating site in the window, and calculations of summary statistics have been done using the actual sample size in each site taking missing data into account. As the size of the discovery panel, d, also varies among SNPs, we calculate the ascertainment probability $P_A(i)$ by averaging over different values of d according to their relative contribution in the window. To examine the effect of the ascertainment bias, we have included results both for the uncorrected and the corrected values of Tajima's D.

RESULTS

Correction of the estimators of theta

Figure 1 shows the distribution of the estimates based on the uncorrected estimators $\hat{\theta}_W$ and $\hat{\theta}_T$ in the presence of an ascertainment bias (d = 5) and without an ascertainment bias, and the corresponding distributions of the corrected estimates, $\hat{\theta}_{WC}$ and $\hat{\theta}_{TC}$, in the presence of an ascertainment bias for n = 50 and θ =150. For $\hat{\theta}_W$, the average estimate of θ is 69.82 and 150.07 with and without an ascertainment bias, respectively. However, the ascertainment corrected estimate is $\hat{\theta}_{WC}$ = 150.11. For $\hat{\theta}_T$, the average estimate of θ is 104.18 and 150.11 with and without an ascertainment bias, respectively, and the ascertainment correct estimate is $\hat{\theta}_{TC}$ = 150.14. This shows that the traditional estimators, as expected, are biased in the presence of an ascertainment bias, but that the ascertainment corrected estimators derived here recover an unbiased estimate.

Correction of the variances and the covariance

As seen in Figure 1, the variance in the estimate is increased in the presence of an ascertainment bias when the number of SNPs in the data set is held constant. Formulas 13 and 14 quantify the variance in the estimate, and have been verified by simulations (not shown).

Figure 2 shows the relationship between d and the variance in the estimators. When the ascertainment sample size is small compared to the size of the sample, the variances and covariances are greatly increased (for d=2 the variance of Tajima's θ is nearly doubled, and the variance of $\hat{\theta}_W$ is nearly multiplied by four). However, when d approaches n/2 the difference between the real variance and the estimated variance is drastically reduced, especially for $\hat{\theta}_T$.

Correction of Tajima's D

Figure 3 shows the distribution of Tajima's D and D_C values. When there is no ascertainment bias, the distribution of Tajima's D values using Equation 17 is identical to the one obtained using the standard method, with mean=-0.1103 in both cases, while when there is ascertainment bias and we do not apply the corrected formula the distribution is greatly skewed towards positive values (mean=1.5170). If the correction is applied to the simulated data suffering from the ascertainment bias, the non-ascertained distribution is approximately recovered and its mean, -0.2497, gets closer to the non-ascertainment one. However, because the correction is non-linear it does not match the original distribution exactly but is slightly skewed towards negative values compared to the original distribution, and has a slightly larger variance.

Analysis of Perlegen data

To illustrate the use of the correction of Tajima's D, we applied it to a Perlegen data set (Hinds *et al.* 2005), previously analyzed by Clark *et al.* (2005) without correcting for ascertainment biases. The Perlegen data was analyzed chromosome by chromosome taking windows of 100 spanning 10 kb obtaining, on average, 12221 windows per chromosome. 74.47% of the windows have 10 or more SNPs and are, therefore, included for the comparison between the corrected and uncorrected Tajima's *D* values.

An example of the result, using windows of 500 kb on chromosome 1, is shown in Figure 4. Positive Tajima's *D* values (1.9) are found in the area containing the genes *TMEM57*, *MAN1C1* and *LDLRAP1*. The former is a transmembrane protein and the second a mannosidase. The latter encodes for a cytosolic protein that interacts with the LDL receptor, and mutations in it have cause hypercholesterolaemia, an autosomal recessive disorder (Mishra *et al.* 2005; Quagliarini *et al.* 2007). Negative Tajima's *D* values around -2 were found in windows containing *HIST2H**, *FCGR1A* and *PPIAL4*, a histone cluster, a fragment of the IgG receptor and the peptidylprolyl isomerase A, respectively. *D* values of -1.6 were found around the *SRGAP2* gene, whose mRNA has been found in melanoma, germ cell tumors, chondrosarcoma and retinoblastoma (Katoh and Katoh. 2003).

Figure 5 shows the correlation of Tajima's D results with and without correction for all chromosomes. As expected, the *D* values are higher than the D_C values. We examine windows with extreme values of Tajima's *D*, which we have arbitrarily defined as those with values lower than -2 or higher than 2, in more detail. While there are 210 windows with $D_C \leq -2$, there are only 17 windows with $D \leq -2$. Likewise, there are 99 windows with $D_C \geq 2$ and 8317 with $D \geq 2$. Table 1 summarizes the information about the 50 windows with the most extreme values of D_C (25 lowest and 25 highest). Of the 25 windows with lowest values of D_C, three would not be found among the 25 most significant windows using *D*, and eight, including the *GPC3* gene, would be excluded based on the $D \leq -2$ criterion. Among the 25 most significant windows with positive values of D_C, 10 of them are not included in the set of the 25 most extreme genes based on *D*. Among these windows there are genes such as *BRCA1* or *NF1*.

DISCUSSION

We have here derived estimators of the population genetic parameter θ, and the variances and covariances of the estimators, under a model with ascertainment bias. This leads us to an ascertainment correction of Tajima's *D*. We notice that similar corrections could easily be derived for other statistics as well, particularly if they can be written as functions of site frequency spectrum, i.e. η_I, $i = 1, 2,\ldots,n-1$. Statistics such as Fu and Li's *D* (Fu and Li. 1993) and Fay and Wu's H (Fay and Wu. 2000) are included in this category. We should also emphasize that while the ascertainment scheme here is quite specific, and the results may therefore not always apply to real data, all results are expressed in terms of the probability of ascertainment of a SNP as a function of its frequency, $P_A(i)$. It is, therefore, quite trivial to extend this work to other ascertainment schemes, including the ones considered in Nielsen *et al.* (2005), as long as appropriate ascertainment information is available.

The analysis of the Perlegen data illustrates that ascertainment bias correction is of great importance when analyzing SNP genotyping data. Even when just applying outlier approaches in studies of natural selection, the ranking of different genes is likely to change with and without ascertainment bias correction. Likewise, any study aimed at quantifying variability based on typical SNP data will be challenged by the ascertainment bias. It is, therefore, highly desirable that SNP genotyping projects keep close track of the SNP discovery/selection protocols used. Only when such detailed data regarding these protocols are available will it be possible to make accurate ascertainment bias corrections of the data.

A computer program implementing the ascertainment bias corrections discussed in this paper can be dowloaded from http://www.snpator.com/public/downloads/aRamirez/tajimasDCorrector/. A list of corrected Tajima's *D* values for different regions of the human genome can be found as Supplemental Data.

ACKNOWLEDGEMENTS

We would like to thank Marta Melé and Francesc Calafell for their comments on this manuscript. The work was supported by National Institutes of Health grant U01HL084706 and the Danish National Science Council (FSS).

LITERATURE CITED

Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin *et al.* 2000 An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature **407:** 513-6.

Bamshad, M., and S. P. Wooding, 2003 Signatures of natural selection in the human genome. Nat. Rev. Genet. **4:** 99-111.

Carlson, C. S., M. A. Eberle, L. Kruglyak and D. A. Nickerson, 2004 Mapping complex disease loci in whole-genome association studies. Nature **429:** 446-52.

Carlson, C. S., J. D. Smith, I. B. Stanaway, M. J. Rieder and D. A. Nickerson, 2006 Direct detection of null alleles in SNP genotyping data. Hum. Mol. Genet. **15:** 1931-1937.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson and R. Nielsen, 2005 Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. **15:** 1496-1502.

Crawford, D. C., D. T. Akey and D. A. Nickerson, 2005 The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet **6:** 287-312.

Durrett, R., 2008 *Probability Models for DNA Sequence Evolution (Probability and its Applications)*. Springer.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive darwinian selection. Genetics **155:** 1405-13.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172-197.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin *et al.* 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072-1079.

Hudson, R. R., 2002 Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18:** 337-8.

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

Katoh, M., and M. Katoh, 2003 FNBP2 gene on human chromosome 1q32.1 encodes ARHGAP family protein with FCH, FBH, RhoGAP and SH3 domains. Int. J. Mol. Med. **11:** 791-797.

Kreitman, M., and A. Di Rienzo, 2004 Balancing claims for balancing selection. Trends Genet. **20:** 300-4.

Mead, S., M. P. Stumpf, J. Whitfield, J. A. Beck, M. Poulter *et al.* 2003 Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. Science **300:** 640-643.

Mishra, S. K., P. A. Keyel, M. A. Edeling, A. L. Dupin, D. J. Owen *et al.* 2005 Functional dissection of an AP-2 beta2 appendage-binding sequence within the autosomal recessive hypercholesterolemia protein. J. Biol. Chem. **280:** 19270-19280.

Nielsen, R., 2004 Population genetic analysis of ascertained SNP data. Hum. Genomics **1:** 218-224.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931-942.

Nielsen, R., and J. Signorovitch, 2003 Correcting for ascertainment biases when analyzing SNP data: Applications to the estimation of linkage disequilibrium. Theor. Popul. Biol. **63:** 245-55.

Nielsen, R., M. J. Hubisz and A. G. Clark, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics **168:** 2373-2382.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.* 2005 Genomic scans for selective sweeps using SNP data. Genome Res. **15:** 1566-1575.

Picoult-Newberg, L., T. E. Ideker, M. G. Pohl, S. L. Taylor, M. A. Donaldson *et al.* 1999 Mining SNPs from EST databases. Genome Res. **9:** 167-174.

Polanski, A., and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics **165:** 427-36.

Quagliarini, F., J. C. Vallve, F. Campagna, A. Alvaro, F. J. Fuentes-Jimenez *et al.* 2007 Autosomal recessive hypercholesterolemia in spanish kindred due to a large deletion in the ARH gene. Mol. Genet. Metab. **92:** 243-248.

Ramos-Onsins, S. E., and J. Rozas, 2002 Statistical properties of new neutrality tests against population growth. Mol. Biol. Evol. **19:** 2092-100.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.* 2006 Positive natural selection in the human lineage. Science **312:** 1614-1620.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.* 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832-7.

Soldevila, M., F. Calafell, A. Helgason, K. Stefansson and J. Bertranpetit, 2005 Assessing the signatures of selection in PRNP from polymorphism data: Results support kreitman and di rienzo's opinion. Trends Genet. **21:** 389-391.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-95.

Tang, K., K. R. Thornton and M. Stoneking, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biology **5:** e171 OP.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biology **4:** e72 OP.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero and K. Ardlie, 2001 The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. Am. J. Hum. Genet. **69:** 1332-1347.

Wang, Y., L. P. Zhao and S. Dudoit, 2006 A fine-scale linkage-disequilibrium measure based on length of haplotype sharing. Am. J. Hum. Genet. **78:** 615-628.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-76.

Williamson, S. H., M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante *et al.* 2007 Localizing recent adaptive evolution in the human genome. PLoS Genet. **3:** e90.

Wooding, S., 2004 Natural selection: Sign, sign, everywhere a sign. Curr. Biol. **14:** R700-1.

Table 1: 50 windows with more extreme Tajima's D values for the corrected estimator

chromosome	**Window**	**first SNP**	**last SNP**	**gene containing first SNP**	**gene containing last SNP**	**corrected Tajima's D**	**uncorrected Tajima's D**
25 windows with lowest corrected Tajima's D							
19	1505	rs11883009	rs10775618	---	AKAP8L	-3.282039	-2.527051
22	1278	rs16986494	rs4035540	TTC28	CHEK2	-3.200966	-2.545256
X	1726	rs16980685	rs17320692	---	---	-3.164725	-2.499758
X	1722	rs10521677	rs17246666	---	---	-3.110302	-2.426383
19	1503	rs16980448	rs10775618	BRD4	AKAP8L	-3.049939	-2.356206
X	2088	rs16981582	rs6528025	CNKSR2	CNKSR2	-2.968399	-2.151076
02	13306	rs16849050	rs16849021	---	---	-2.958659	-2.765536
03	1909	rs10510486	rs17005761	KCNH8	KCNH8	-2.892485	-2.138216
X	10073	rs17331728	rs17342441	---	---	-2.867873	-2.051048
16	1463	rs17260976	rs16966953	PARN	NTAN1	-2.856588	-2.018186
X	13139	rs17251454	rs17000462	GPC3	GPC3	-2.831727	-1.991586
19	1497	rs16980438	rs4616406	---	---	-2.813864	-2.045236
X	13140	rs7061117	rs17000463	GPC3	GPC3	-2.809700	-1.991586
19	1171	rs17001730	rs10424893	ZNF700	---	-2.802120	-1.927546
06	6799	rs17446192	rs4710655	---	---	-2.795281	-2.372185
X	10074	rs16984144	rs10521499	BHLHB9	---	-2.794317	-2.070374
17	6325	rs16961696	rs2221741	---	---	-2.761039	-2.083343
X	1720	rs12845504	rs17246666	---	---	-2.754714	-2.094886
19	1499	rs16980438	rs16980462	---	---	-2.706380	-2.012035
07	7214	notfound	rs2353082	nf	BAZ1B	**-2.663340**	**-1.858293**

17	6324	rs16961697	rs2221741	---	---	-2.661940	-2.045867
01	5148	rs12094202	rs10489546	OSBPL9	OSBPL9	-2.659012	-1.964506
19	1501	rs8104223	rs10775618	BRD4	AKAP8L	-2.654787	-1.953696
08	9984	rs16897122	rs2029596	---	VPS13B	**-2.652911**	**-1.812917**
06	10935	rs17070142	rs351730	SESN1	---	**-2.645325**	**-1.770006**
25 windows with highest corrected Tajima's D							
01	11083	rs1774778	rs17026872	---	---	3.131594	5.174721
X	13405	rs5975710	rs6633822	MAP7D3	GPR112	3.069725	5.475525
01	11084	rs1774778	rs325910	---	---	**2.749263**	**4.683911**
X	13061	rs5975352	rs17324216	HS6ST2	HS6ST2	2.743219	4.817528
04	13648	rs7658327	rs13143611	---	---	2.723831	5.389725
05	7065	rs986217	rs1017225	---	BDP1	2.717593	4.800969
X	12501	rs203491	rs5931921	---	---	2.682832	4.922305
18	3626	rs2217945	rs7232770	---	---	2.656100	5.096458
04	12950	rs1870687	rs12510308	LARP2	---	2.645268	4.870015
10	12713	rs10794030	rs7918092	DHX32	FANK1	2.590828	4.981469
09	8180	rs7044691	rs9410888	GKAP1	KIF27	**2.577577**	**4.565764**
01	11082	rs1342353	rs17026872	---	---	2.553143	4.457859
X	13109	rs5975387	rs5977860	---	GPC4	**2.551131**	**4.648231**
03	4834	rs725310	rs734071	FBXW12	SCOTIN	2.534127	4.769015
04	5577	rs10434442	rs17085274	KDR	KDR	**2.531075**	**4.478948**
X	13062	rs17317147	rs5933229	HS6ST2	HS6ST2	2.523093	4.457818
09	8024	rs2788113	rs12686026	---	---	2.511495	4.827204
12	5603	rs537482	rs511752	---	ARHGAP9	**2.509088**	**4.637981**
17	2954	rs12948444	rs2952991	NF1	NF1	**2.501910**	**4.508276**

03	9662	rs6806361	rs1533148	---	---	2.484018	4.814285
17	4158	rs3950989	rs8070085	BRCA1	NBR1	**2.479324**	**4.643384**
X	13110	rs5975387	rs17317322	---	GPC4	**2.457019**	**4.491652**
06	7949	rs9352669	rs956550	IRAK1BP1	---	2.455673	5.427745
14	4683	rs9323475	rs17182817	GPHN	GPHN	**2.444739**	**4.476493**
03	9663	rs6806361	rs9833997	---	---	**2.428898**	**4.650462**

In bold, windows not found among the 25th most extreme for the uncorrected Tajima's D

FIGURE LEGENDS

Figure 1. The distribution of the estimates of θ assuming non-ascertained data (no asc), ascertained data with correction (asc|c) and ascertained data without correction (asc|nc). The mean and the variance of each set of data is shown next to the legend. Simulations have been performed for n = 50, θ = 150 and 1.000.000 replicates. (A) Watterson's estimator. (B) Tajima's estimator.

Figure 2. The variance of Watterson's estimator of θ, Tajima's estimator of θ and the covariance as a function of d calculated using estimated value of θ and θ^2 for a sample of size n=100. (A) θ = 150. (b) θ = 22.33.

Figure 3. The distribution of Tajima's D for data without ascertainment bias and without correction (no asc), ascertained data with correction (asc|c) and ascertained data without correction (asc|nc). The mean and the variance among estimates is shown next to the legend. A value of θ = 150 was used and 1.000.000 replicates were performed.

Figure 4. The distribution of the ascertainment bias corrected Tajima's D on chromosome 1 in the human genome based on the Perlegen data. The genes with the most extreme D values are also indicated on the Figure.

Figure 5. Correlation of Tajima's D results from Perlegen data with and without correction for all chromosomes.

Figure 1

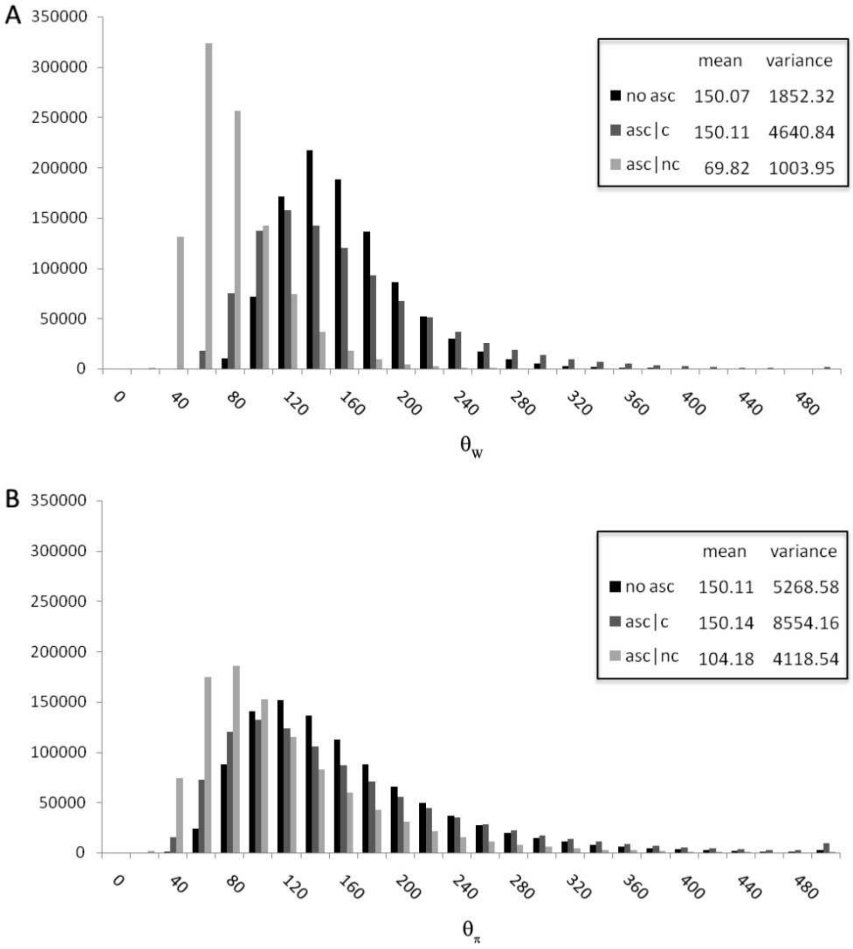

A
350000
300000
250000
200000
150000
100000
50000
0
0
40
80
120
160
200
240
280
320
360
400
440
480
θ_W
mean variance
■ no asc 150.07 1852.32
■ asc|c 150.11 4640.84
■ asc|nc 69.82 1003.95
B
350000
300000
250000
200000
150000
100000
50000
0
0
40
80
120
160
200
240
280
320
360
400
440
480
θ_π
mean variance
■ no asc 150.11 5268.58
■ asc|c 150.14 8554.16
■ asc|nc 104.18 4118.54

Figure 2

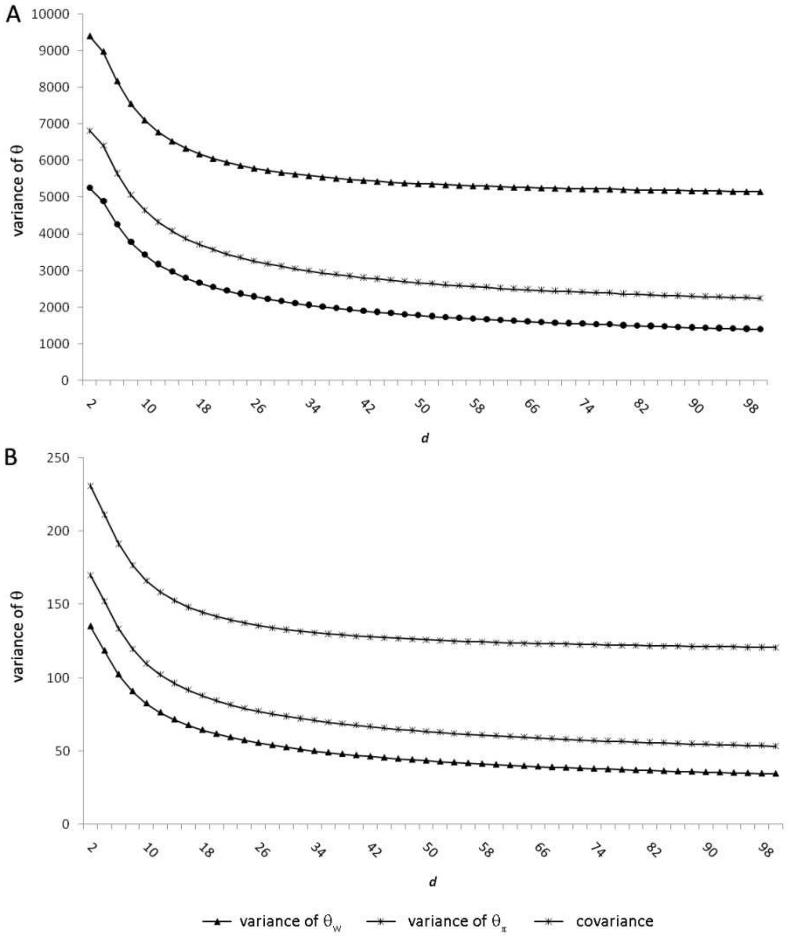
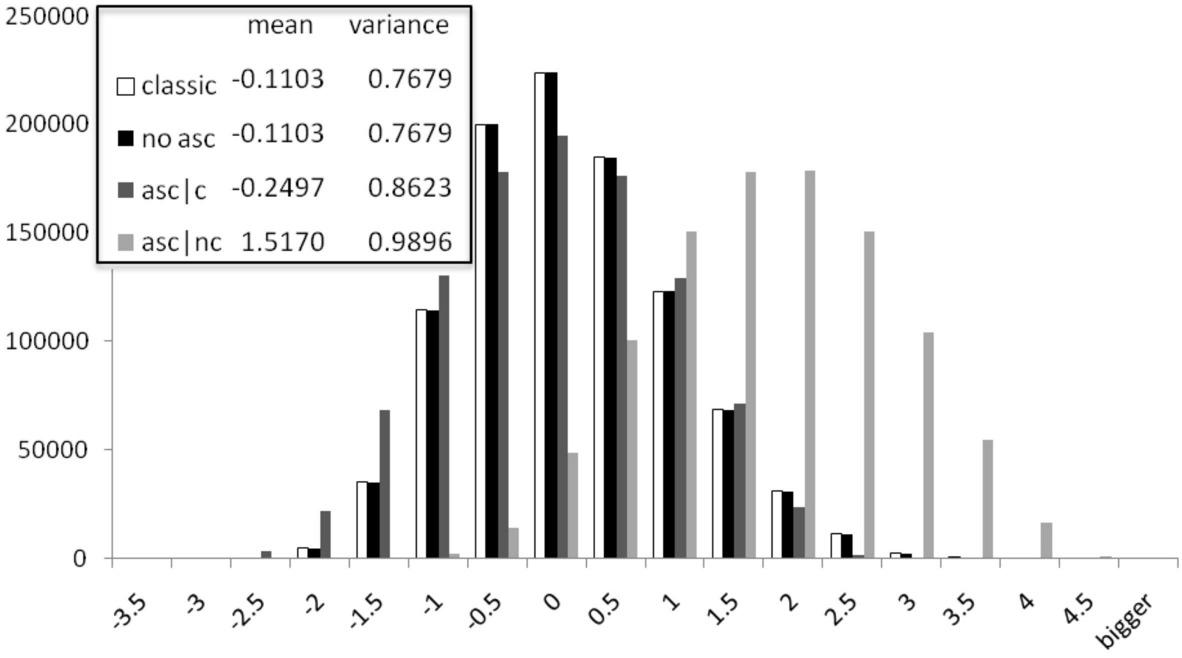

A
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
variance of θ
2
10
18
26
34
42
50
58
66
74
82
90
98
d
B
250
200
150
100
50
0
variance of θ
2
10
18
26
34
42
50
58
66
74
82
90
98
d
variance of θ_W
variance of θ_π
covariance

Figure 3



250000
200000
150000
100000
50000
0
mean
variance
□ classic -0.1103 0.7679
■ no asc -0.1103 0.7679
■ asc|c -0.2497 0.8623
■ asc|nc 1.5170 0.9896
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
bigger

Figure 4

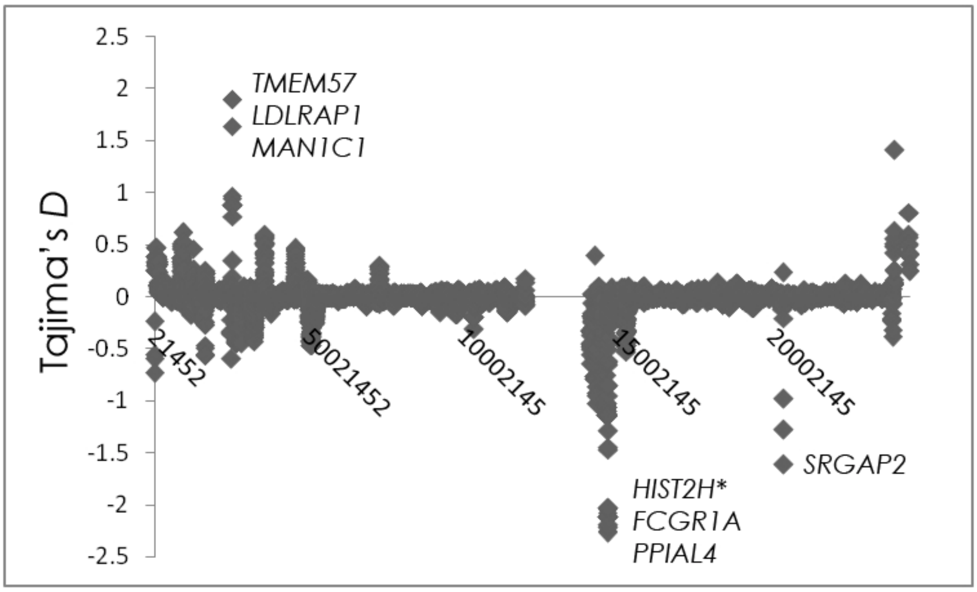
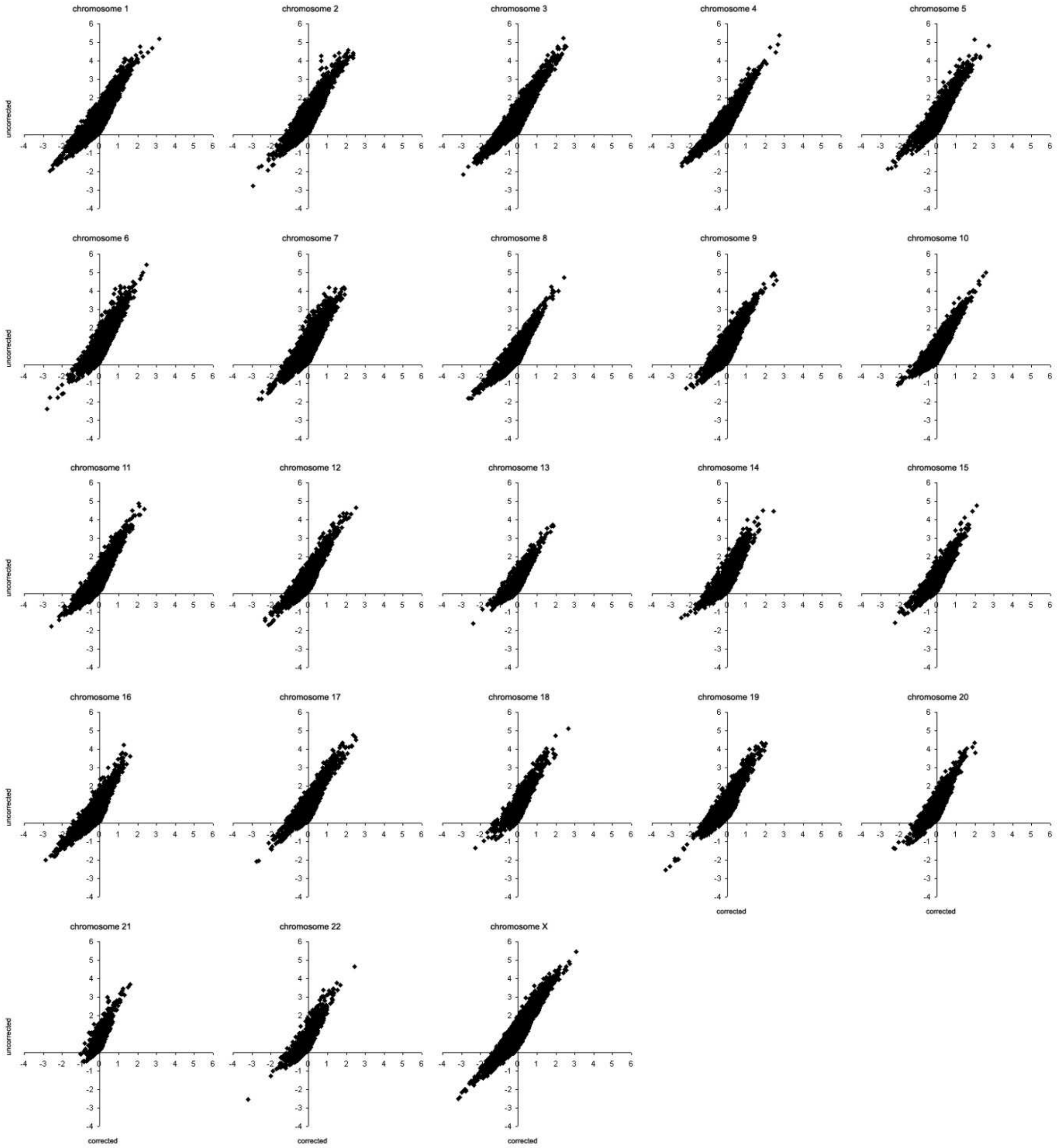

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Tajima's *D*
21452
50021452
10002145
15002145
20002145
TMEM57
LDLRAP1
MAN1C1
*HIST2H**
FCGR1A
PPIAL4
SRGAP2

Figure 5


chromosome 1
chromosome 2
chromosome 3
chromosome 4
chromosome 5
chromosome 6
chromosome 7
chromosome 8
chromosome 9
chromosome 10
chromosome 11
chromosome 12
chromosome 13
chromosome 14
chromosome 15
chromosome 16
chromosome 17
chromosome 18
chromosome 19
chromosome 20
chromosome 21
chromosome 22
chromosome X
uncorrected
corrected
-4 -3 -2 -1 0 1 2 3 4 5 6

4 FABSIM: A SOFTWARE FOR GENERATING F_{ST} DISTRIBUTIONS WITH VARIOUS ASCERTAINMENT BIASES (Ramirez-Soriano and Calafell. submitted)

FABSIM: a software for generating F_{ST} distributions with various ascertainment biases

Anna Ramírez-Soriano[1] and Francesc Calafell[1,2,*]

[1] Departament de Ciències de la Salut i de la Vida. Universitat Pompeu Fabra. Doctor Aiguader, 88. 08003 Barcelona, Catalonia, Spain.

[2] CIBER en Epidemiologia y Salud Pública (CIBEREsp), Spain

Running title: F_{ST} distributions with ascertainment bias

Keywords: F_{ST}, ascertainment bias, software

***Corresponding author:**

Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra

Doctor Aiguader 88, 08003 Barcelona (Spain)

Phone: +34 93 3160842

Fax: +34 93 3160901.

Email: francesc.calafell@upf.edu

ABSTRACT

F_{ST} is widely used to find genes under local selection by comparing the F_{ST} value of a single locus against genome-wide, empirical values. However, empirical distributions suffer from ascertainment bias caused by the protocol used to select SNPs, which affects the shape of the distribution. An alternative is working with simulated distributions, but this procedure also produces unreliable distributions as F_{ST} is highly dependant on the demographic history of the samples, and simulations do not take into account ascertainment bias. Provided that there is an increasing amount of information on the demographic history of populations, we have developed a software that applies ascertainment bias on simulated sequences and calculates F_{ST} on them. Moreover, we also used our program to generate several simulated F_{ST} distributions with different ascertainment biases and have compared them against the F_{ST} values found in an empirical database.

INTRODUCTION

Since the beginning of their journey out of Africa, approximately 60,000 years ago (Jobling *et al.* 2004), humans have colonized the entire world. In their expansion, they have adapted to a wide range of different habitats and, thus, of diverse ecological conditions. In these circumstances, human populations have been exposed to geographically localized selective pressures, leading to an increase in the genetic differences among them. Diseases such as sickle cell anemia (Motulsky *et al.* 1966) or cystic fibrosis (Lao *et al.* 2003), which show today a distinc geographical pattern across populations, are clear examples of the consequences of such local selective pressures.

One of the most widely used methods to detect differential selective pressures between populations is F_{ST}, a measure of the proportion of the genetic variance explained by differences among populations. F_{ST} can be used to find genes under local selection by comparing the F_{ST} value of a single locus against the genome-wide values. Allele frequency differences between populations are mainly caused by genetic drift, that is, by the random process driven by demographic history. Drift affects all of the genome and, thus, a genomewide F_{ST} distribution reflects primarily drift. Against this backdrop, a gene with extremely large F_{ST} values becomes suspect of having suffered local adaptation in a subset of the human populations. A large number of works have been published based on this principle, building genome-wide empirical distributions of F_{ST} based on increasing numbers of autosomical SNPs (Akey *et al.* 2002, Walsh *et al.* 2006, Xue *et al.* 2006, Myles *et al.* 2008, Savage *et al.* 2008).

However, this methodology has a few problems. Although empirical distributions are presumably neutral (since they report the differentiation due to demographic events), they are not built over the total variation found on the genome but on a particular subset of SNPs. The way the SNPs are ascertained may thus produce an underlying bias that affects the shape of the distribution (Hinds *et al.* 2005; International HapMap Consortium. 2007). Furthermore, when comparing the F_{ST} of the SNPs in a gene against a published empirical distribution, the biases applied to both sets of samples can be different (eg. Ferrer-Admetlla *et al.* 2008). An alternative is working with simulated distributions, as suggested by Beaumont and Nichols (1996). However, this procedure also produces unreliable distributions, as (a) F_{ST} is highly dependant on the demographic history of the samples, which may not be known with sufficient precision, and thus may not be accurately simulated, and (b) simulations do not take into account ascertainment biases. The former issue can now be addressed in humans using the calibrated demographic model proposed by Schaffner *et al.* (2005), which fits different empirical parameters, such as F_{ST} (under the ascertainment model used in Sachidanandam *et al.* (2001), minor allele frequencies (MAF) or linkage disequilibrium (LD).

In this work we adress the second issue, that is, how to produce an F_{ST} distribution with the same bias than the genotyped samples. In order to do that we have developed a program that applies several ascertainment biases on simulated sequences and calculates F_{ST} on them. This program, then, allows producing F_{ST} distributions with the same underlying ascertainment biases than a genotyped gene of interest against which to compare it. Moreover, we provide several simulated F_{ST} distributions with different ascertainment biases.

RESULTS

We have developed FABSIM, a Java software package that builds simulated F_{ST} distributions under different ascertainment biases. Moreover, as a complement, it can also calculate minor (MAF) and derived (DAF) allele frequencies and a number of neutrality statistics; although the last should only be calculated on simulations without applying any bias.

FABSIM works on coalescent-based simulation results, which may be generated using any of the available packages developed to simulate neutral genealogies. FABSIM supports only as input the output file formats of ms, cosi, and SelSim packages, or any other simulation translated to reproduce one of these formats. Different populations are introduced in the program in separate files.

FABSIM works on the data introduced by the user, to which four different bias categories (with a total of seven different bias types) can be applied, alone or combined with other biases; the input data can also be left unbiased. The four bias categories are: (a) related to the discovery sample, (b) related to the presence of polymorphism in a population, (c) related to the MAF, and (d) related to distance. If more than one population is introduced for analysis in the program, SNPs are selected over only one population (determined by the user), but the bias is applied over all populations. That is, the SNPs deleted from the chosen population are deleted from all populations. (a) Discovery sample biases imply that only some chromosomes (a subsample of size *d*) of a general sample of size *n* have been resequenced, and the segregating sites found on them have been genotyped on the whole sample *n*. If this bias is applied *d* sequences are randomly selected (where $0<d<n$ is specified by the user) over the total number of sequences in the sample, and keeps only the SNPs that are polymorphic in these *d* sequences. This procedure can be performed by gene, that is, the *d* selected sequences are the same over all the segregating sites; or by SNP, selecting a different *d* sample (but always of the same size) for each site. (b) In the bias related to the presence of polymorphism, only the SNPs that are polymorphic either in a given population or in all populations are kept. (c) In the MAF bias, all the SNPs that have a MAF below a threshold provided by the user are discarded. This procedure biases towards markers of high heterozigosity. (d) In physical distance biases SNPs are selected with a physical spacing specified by the user. To do so, FABSIM selects randomly one segregating site among the *x* first base pairs, where *x* is the spacing, in base pairs, selected by the user. From this first selected SNP, the position *x* base pairs downstream is determined. If in this new position a segregating site is found, it is selected; otherwise, the nearest one is selected, regardless of whether it is found upstream or downstream. FABSIM proceeds as explained until the new position is found outside the simulated fragment. This bias can be applied using the same distance along the gene or using different SNP densities. In the last case, a file must be provided stating which fragments are to have particular densities, and which densities these are.

Three different statistics can be calculated with FABSIM: F_{ST}, MAF and DAF, and 17 neutrality statistics. F_{ST} can be calculated correcting or not correcting by the different sample size of the populations involved; and by gene, by SNP, or both. The neutrality statistics included are the ones used in (Ramirez-Soriano *et al.* 2008), that is, Tajima's *D* (Tajima 1989); Fu and Li's *D*, *F*, *D** and *F** (Fu *et al.* 1993); Fay and Wu's *H* (Fay *et al.* 2000), R_2 (Ramos-Onsins *et al.* 2002), Fu's F_s (Fu 1997), *Dh* (Nei 1987, equation 8.4 replacing 2n by n), Wall's *B* and *Q* (Wall 1999), Kelly's Z_{nS} (Kelly, 1997), Rozas' Z_A and *ZZ* (Rozas *et al.* 2001) and extended haplotype homozygosity *EHH* (Sabeti *et al.* 2002, Ramirez-Soriano *et al.* 2008). The results can be provided in two different formats: as a list of the calculated parameters for sample or tabulated. In the former, each sample starts with a line stating the sample number (e.g. SAMPLE 1), and is followed by a list with the F_{ST}, MAF/DAF or neutrality statistic values. In the latter, FABSIM provides a file with as many columns as statistics plus one first colum with the sample number, separated by tabulators. The first line is a header stating what each column is.

As an example of a possible use of the program, we have used it to produce several F_{ST} distributions with and without ascertainment bias and have compared them with the empirical F_{ST} distribution of all the segregating sites found in the human genes resequenced by the SeattleSNPs project (http://pga.gs.washington.edu/, Crawford *et al.* 2005). We have run simulations using the parameters provided by Schaffner *et al.* (2005), which have been shown to fit empirical human data for several statistics. Only two populations, African-Americans and Europeans, have been simulated, as they are the populations resequenced in SeattleSNPs. To match SeattleSNPs data we have simulated 48 African-American and 46 European chromosomes. The F_{ST} values for all SNPs and simulations, together with the numerical results behind the histograms, are shown as Supplementary Data.

Figure 1 shows how F_{ST} distributions are affected by selecting SNPs according to their MAF, ascertaining either on European or on African-American samples. When SNPs are selected by MAF in European samples, the number of segregating sites with low F_{ST} (<0.05) decreases, while the number of SNPs with higher F_{ST} increases. This can be easily explained as low frequency SNPs tend to produce F_{ST} values near 0. Thus, if all these SNPs are eliminated from the analysis, the proportion of low F_{ST} is expected to decrease. However, if the selection of SNPs by MAF is performed over African-American samples, the fraction of low F_{ST} values is extremely reduced for MAF biases between 0.05 and 0.20 and increased for intermediate F_{ST} (between 0.05 and 0.2), in comparison with ascertainment on Europeans. This differential pattern can be due to the fact that African-Americans share a large fraction of low-frequency SNPs with Europeans due to asymmetrical gene flow.

Figure 2 shows the effect of selecting only those SNPs that are polymorphic in all populations or in a given population. It can be seen that, if we keep only the segregating sites that are polymorphic in all populations, F_{ST} tends to increase, as it can be expected that the SNPs shared among populations have originated before the Out-of-

Africa and, thus, that they have had more times to reach higher allelic frequencies, and so larger F_{ST} values. However, if SNPs are selected according to whether they are polymorphic in one of the populations irrespectively of whether they are polymorphic in the other, the effect is less pronounced, and a high fraction of SNPs at low frequency is maintained, especially if SNPs are selected in African-Americans. This is consistent with a larger fraction of African-American's low-frequency SNPs being also present in Europeans than the other way round.

The result of ascertaining SNPs by discovery sample is shown in Figure 3. As the *d* sample size increases the proportion of SNPs with low F_{ST} also increases. This happens because with larger d samples it is more probable to find SNPs at low frequency. The effect of using an ascertainment subsample on data is nearly identical independently on which population is ascertained.

We have also compared the distribution of F_{ST} obtained by simulation against the empirical distribution of the SNPs in the genes resequenced by SeattleSNPs, as shown in Figure 4. The number of SNPs with low F_{ST} is higher in simulations than in SeattleSNPs data. Furthermore, the distribution of F_{ST} in SeattleSNPs data has a larger tail of SNPs at high frequencies. This could be explained by (a) the effect of imputing missing genotypes or by (b) the presence of genes affected by natural selection. In order to ascertain the weight of these two possible explanations, the analysis was repeated by dropping the sites with missing data instead of imputing their alleles. Although deleting the sites that contain missing increases the number of low F_{ST} values (<0.05), it also increases the fraction of values with high F_{ST}, mainly those SNPs with F_{ST} >0.95. However, removing SNPs with missing genotypes does not make the empirical F_{ST} distribution significantly closer to the simulated distribution. This result points to positive selection as a significant force in shaping the F_{ST} distribution for SeattleSNPs genes; a plausible explanation given that genes in this database have been chosen for their relationship with human inflammatory response.

DISCUSSION

In this work we have presented FABSIM, a program that can generate F_{ST} distributions under different ascertainment schemes from simulations. FABSIM accepts simulated sequences to be inputted in three different formats and implements four different ascertainment bias categories. Furthermore, it can also calculate minor and derived allele frequencies and neutrality statistics.

Besides, we have also presented several F_{ST} distributions of simulated samples with different ascertainment biases matching the calibrated demographic history for humans described by Schaffner *et al.* (2005) in African-Americans and Europeans, which can be directly used to compare empirical F_{ST} values from human genes genotyped in those populations.

Until now, F_{ST} methods were always used over empirical distributions obtained from large numbers of genotyped SNPs. However, this method can be influenced by local selection acting on the SNPs from which the distribution has been built. The main problem arises as the number of SNPs under selection cannot be known and, thus, the empirical significance obtained from the distribution can be misleading. This is the case, for example, of the SeattleSNPs F_{ST} distribution presented here. On the other hand, F_{ST} is usually applied not on resequencing data but on genotyping projects, which often have different ascertainment biases than those affecting the F_{ST} distribution against which results are compared. In fact, from the distributions with ascertainment we provide, it can be clearly seen that ascertainment bias affects the distribution, mainly reducing the number of SNPs with low F_{ST} and, thus, increasing the estimated differences between populations. Moreover, they show that the different biases have discernible effects on the F_{ST} distribution, particularly on the fraction of F_{ST} values found in any histogram category.

FABSIM is a program designed to solve both problems at the same time. With FABSIM, the F_{ST} resulting from the analysis of data from the genotyping process can be compared not against an empirical distribution, which can have underlying selective processes shifting it, but against a neutral distribution, that is, a distribution that takes into account the demographic history of the sample but that does not include any selective events. Furthermore, FABSIM can reconstruct the ascertainment bias introduced in the SNP selection on the simulated data. Then, provided we have a valid description of the demographic events shaping the sample, FABSIM allows building simulated F_{ST} distributions that match with precision both the demographic history and the ascertainment process of the sample. Those simulated distributions became, then, a much more powerful framework against which to compare the F_{ST} of the samples than any empirical distribution, as the user can be sure that any of its SNPs is under selection.

METHODS

FABSIM programmation

FABSIM has been programed in Java using NetBeans IDE 6.0. It has been released as an executable file FABSIM.jar that can be run in any platform provided that a Java Runtime Environment (JRE) 6 has been installed (see the Java web page http://java.sun.com/javase/downloads/index.jsp). Both the executable file and the font code can be downloaded from http://www.snpator.com/public/downloads/aRamirez/FABSIM/, together with the Help.pdf file.

Simulation data

Simulations have been performed using the cosi package with the parameters provided by Schaffner *et al.* (2005) to reproduce their best-fit demographic model of human data. These parameters assume a sequence length of 250 kb, a mutation rate of 1.5×10^{-8} per site and per generation, a gene conversion rate of 4.5×10^{-9} and the recombination map provided by Kong *et al.* (2002). We have simulated only two populations, African-Americans and Europeans, with 48 and 46 chromosomes each. Asians and Africans have been considered in the demography but their number of individuals has been set to 0.

SeattleSNPs data

The genotypes of all the genes resequenced by SeattleSNPs have been obtained from http://pga.gs.washington.edu/data_download.html. We have dowloaded all variation data files as provided by the Bulk Dowload link and, from them, we have selected the individual genotypes (gene.prettybase.txt) and the SNP alleles file (gene. alleles.txt). Of the 319 genes dowloaded we have kept only the 303 that were resequenced exactly in African-Americans and Europeans. For each gene we have discarded the indel polymorphisms and the triallelic SNPs.

Missing genotypes reconstruction

The missing genotypes in SeattleSNPs genes have been estimated using fastPHASE (Scheet *et al.* 2006), a program that reconstructs haplotypes and estimates missing genotypes using a cluster method. The input fastPHASE files have been created using the SeattleSnpsToPhase.jar script. SeattleSnpsToPhase.jar uses the individual genotypes and SNP alleles files from SeattleSNPs and produces two input fastPHASE files, one for each population (African-Americans and Europeans), discarding the indel polymorphisms and triallelic positions. SeattleSnpsToPhase.jar can also be downloaded from http://www.snpator.com/public/downloads/aRamirez/toPhaseFormat.

ACKNOWLEDGEMENTS

We would like to thank Ignacio Guerra for his support in developing FABSIM; Arcadi Navarro and Ferran Casals for their helpful comments on this manuscript and Debborah Nickerson for making Seattle SNPs sequences widely available.

REFERENCES

Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12:** 1805-14.

Beaumont, M. A. and Nichols, R. A. 1996. Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proceedings of the Royal Society B: Biological Sciences* **263:** 1619-1626.

Crawford, D. C., Akey, D. T., and Nickerson, D. A. 2005. The patterns of natural variation in human genes. *Annu Rev Genomics Hum Genet* **6:** 287-312.

Fay, J. C. and Wu, C. I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155:** 1405-13.

Ferrer-Admetlla, A., Bosch, E., Sikora, M., Marques-Bonet, T., Ramirez-Soriano, A., Muntasell, A., Navarro, A., Lazarus, R., Calafell, F., Bertranpetit, J., *et al.* 2008. Balancing selection is the main force shaping the evolution of innate immunity genes. *J. Immunol.* **181:** 1315-1322.

Fu, Y. X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147:** 915-25.

Fu, Y. X. and Li, W. H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133:** 693-709.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072-1079.

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

Jobling, M. A., Hurles, M. E., and Tyler-Smith, C. 2004. *Human evolutionary genetics: origins, peoples & disease.* Garland Science, New York.

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. *Genetics* **146:** 1197-206.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., *et al.* 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241-7.

Lao, O., Andres, A. M., Mateu, E., Bertranpetit, J., and Calafell, F. 2003. Spatial patterns of cystic fibrosis mutation spectra in European populations. *Eur. J. Hum. Genet.* **11:** 385-394.

Motulsky, A. G., Vandepitte, J., and Fraser, G. R. 1966. Population genetic studies in the Congo. I. Glucose-6-phosphate dehydrogenase deficiency, hemoglobin S, and malaria. *Am. J. Hum. Genet.* **18:** 514-537.

Myles, S., Tang, K., Somel, M., Green, R. E., Kelso, J., and Stoneking, M. 2008. Identification and analysis of genomic regions with large between-population differentiation in humans. *Ann. Hum. Genet.* **72:** 99-110.

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

Ramirez-Soriano, A., Ramos-Onsins, S. E., Rozas, J., Calafell, F., and Navarro, A. 2008. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics* **179:** 555-567.

Ramos-Onsins, S. E. and Rozas, J. 2002. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* **19:** 2092-100.

Rozas, J., Gullaud, M., Blandin, G., and Aguade, M. 2001. DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. *Genetics* **158:** 1147-55.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., *et al.* 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832-7.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., *et al.* 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928-933.

Savage, S. A., Gerstenblith, M. R., Goldstein, A. M., Mirabello, L., Fargnoli, M. C., Peris, K., and Landi, M. T. 2008. Nucleotide diversity and population differentiation of the melanocortin 1 receptor gene, MC1R. *BMC Genet.* **9:** 31.

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

Scheet, P. and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78:** 629-644.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585-95.

Wall, J. D. 1999. Recombination and the power of statistical tests of neutrality. *Genet. Res.* **74:** 65-79.

Walsh, E. C., Sabeti, P., Hutcheson, H. B., Fry, B., Schaffner, S. F., de Bakker, P. I., Varilly, P., Palma, A. A., Roy, J., Cooper, R., *et al.* 2006. Searching for signals of evolutionary selection in 168 genes related to immune function. *Hum. Genet.* **119:** 92-102.

Xue, Y., Daly, A., Yngvadottir, B., Liu, M., Coop, G., Kim, Y., Sabeti, P., Chen, Y., Stalker, J., Huckle, E., *et al.* 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am. J. Hum. Genet.* **78:** 659-670.

FIGURE LEGENDS

Figure 1: F_{ST} distributions for simulations ascertained by MAF. Sims_maf05, Sims_maf10, etc. indicate that the thresold MAF has been set to 0.05, 0.10, etc. Sims_maf0 are the simulations without ascertainment. A) Ascertainment applied over European populations. B) Ascertainment applied over African-American populations.

Figure 2: F_{ST} distributions for simulations ascertained according to whether SNPs are polymorphic in all populations (Sims_polymorphicAll), on Europeans (Sims_polymorphicEU) or on African-Americans (Sims_polymorphicAA). Sims_allSNPs stands for the non-ascertaiend distribution.

Figure 3: F_{ST} distributions for simulations ascertained using a discovery sample. Sims_d8, Sims_d16, etc. indicate *d* sample sizes of 8, 16, etc. Sims_all are the simulations without ascertainment. A) Ascertainment applied over European populations. B) Ascertainment applied over African-American populations.

Figure 4: comparison between simulated (Sims) and empirical (SeattleSNPs and SeattleSNPs_nm) F_{ST} distributions. SeattleSNPs shows the empirical distribution reconstructing missing data and SeattleSNPs_nm the empirical distribution deleting all the loci with missing genotypes.

Figure 1

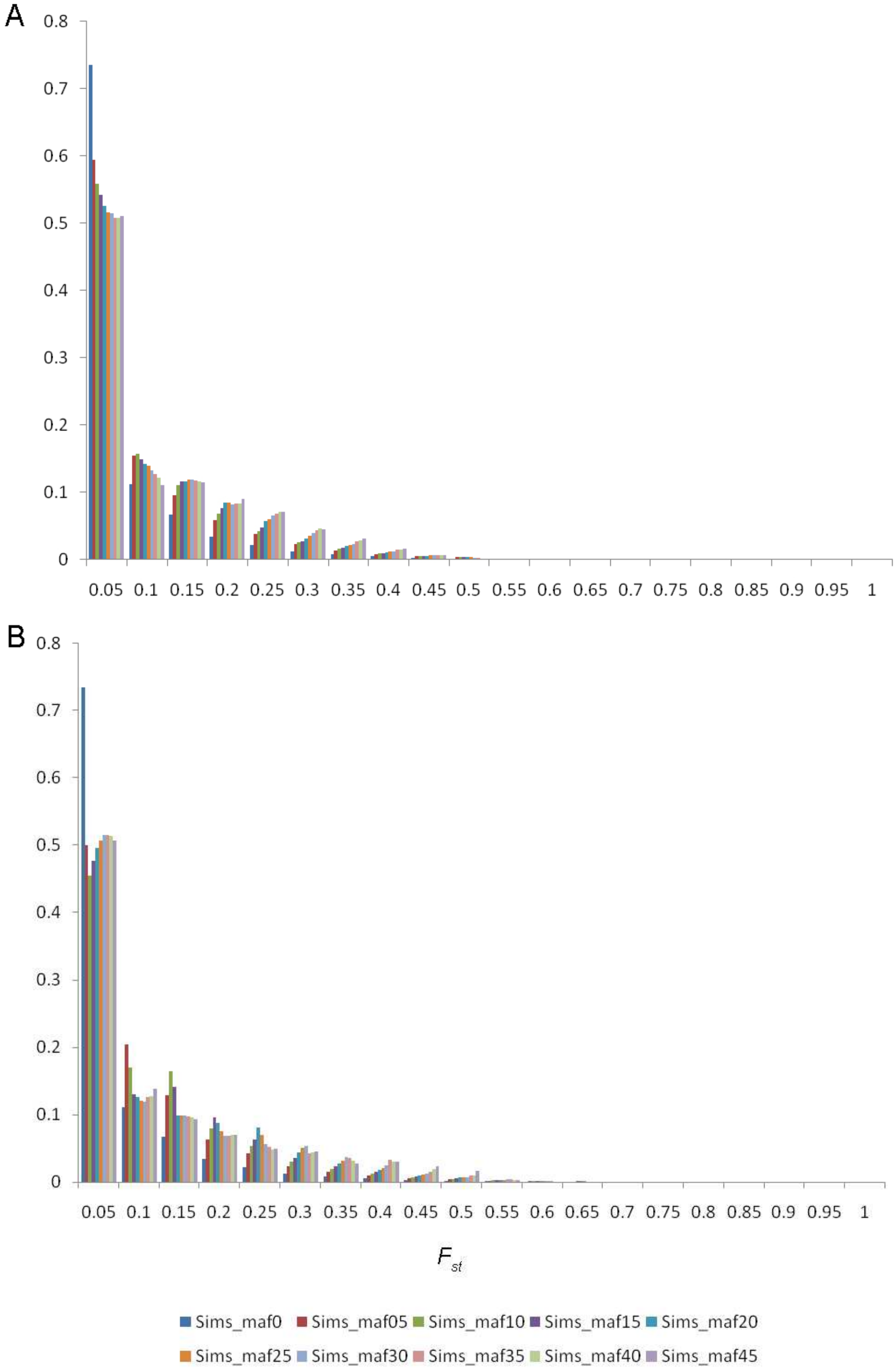

A
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
B
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
F_{st}
Sims_maf0
Sims_maf05
Sims_maf10
Sims_maf15
Sims_maf20
Sims_maf25
Sims_maf30
Sims_maf35
Sims_maf40
Sims_maf45

Figure 2

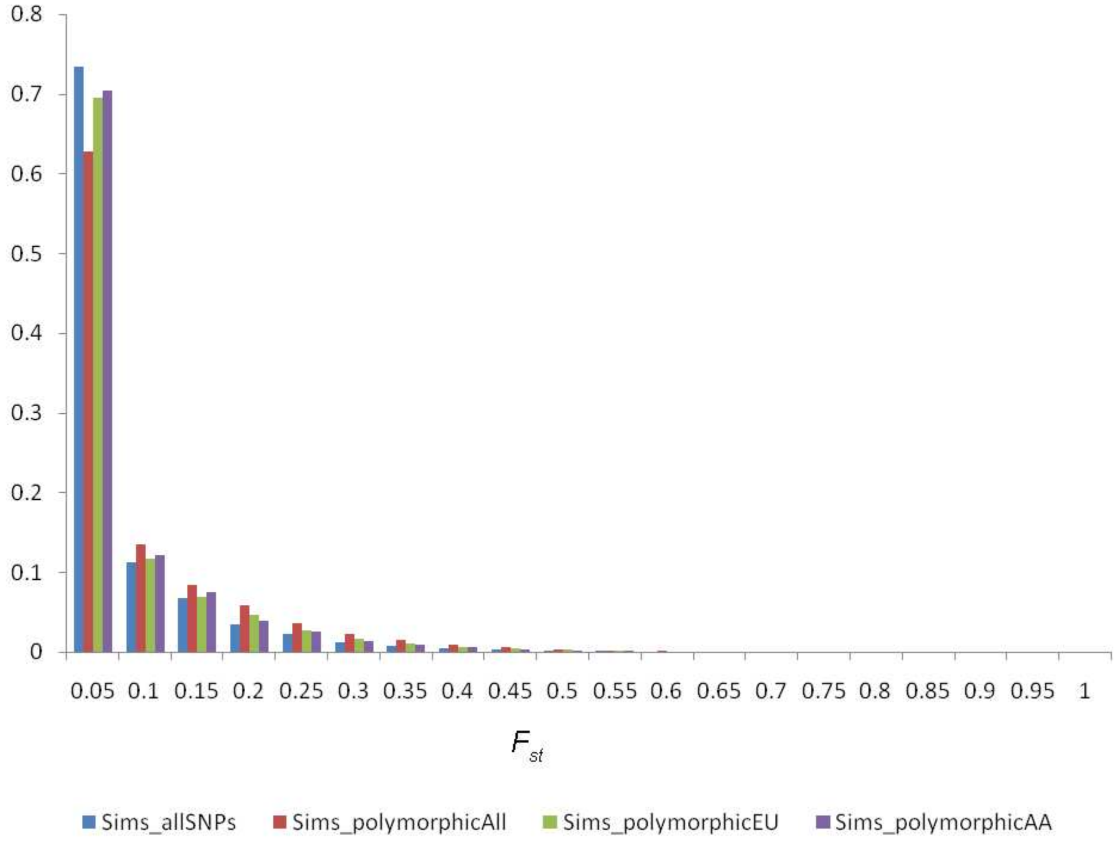

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
F_{st}
Sims_allSNPs
Sims_polymorphicAll
Sims_polymorphicEU
Sims_polymorphicAA

Figure 3

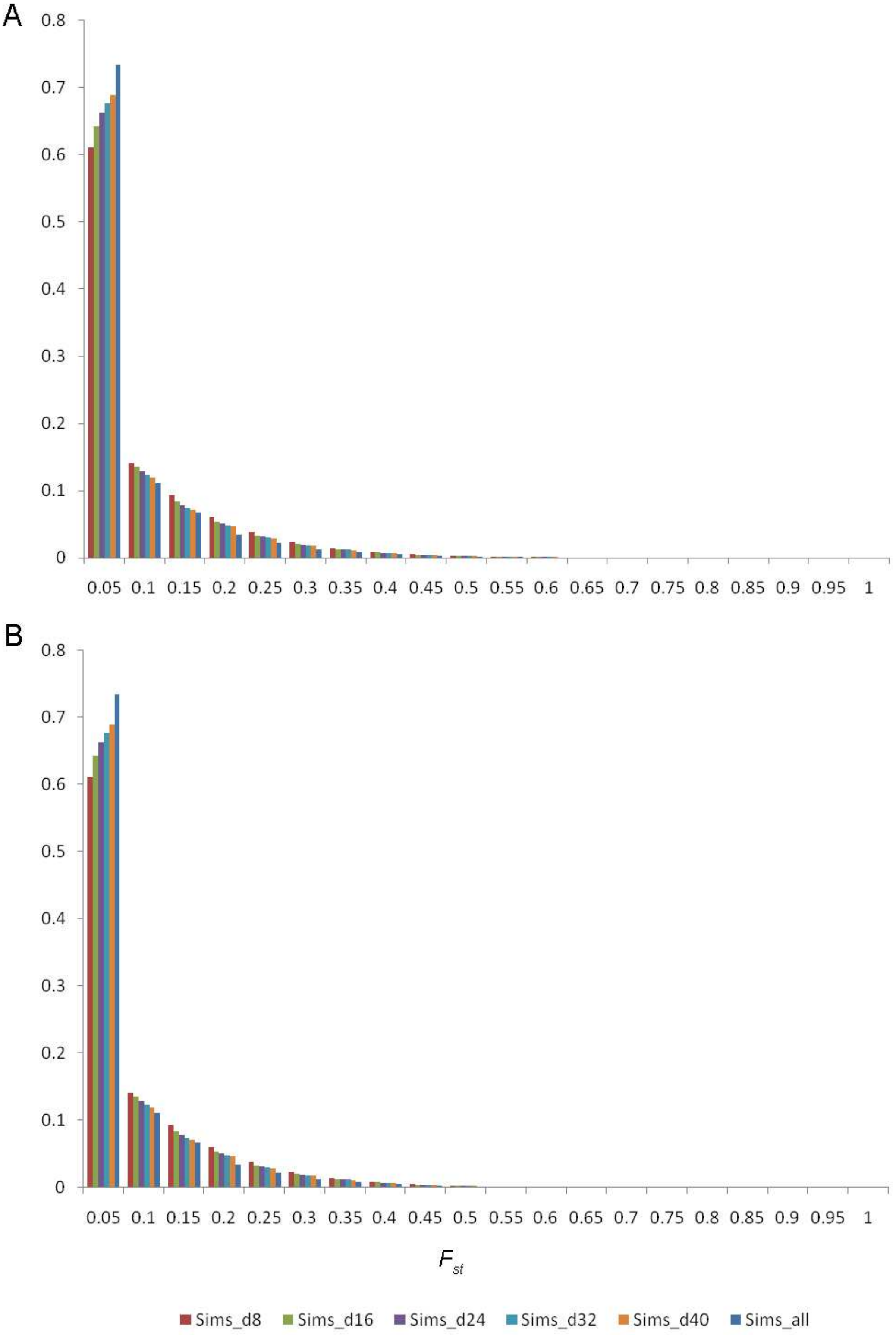
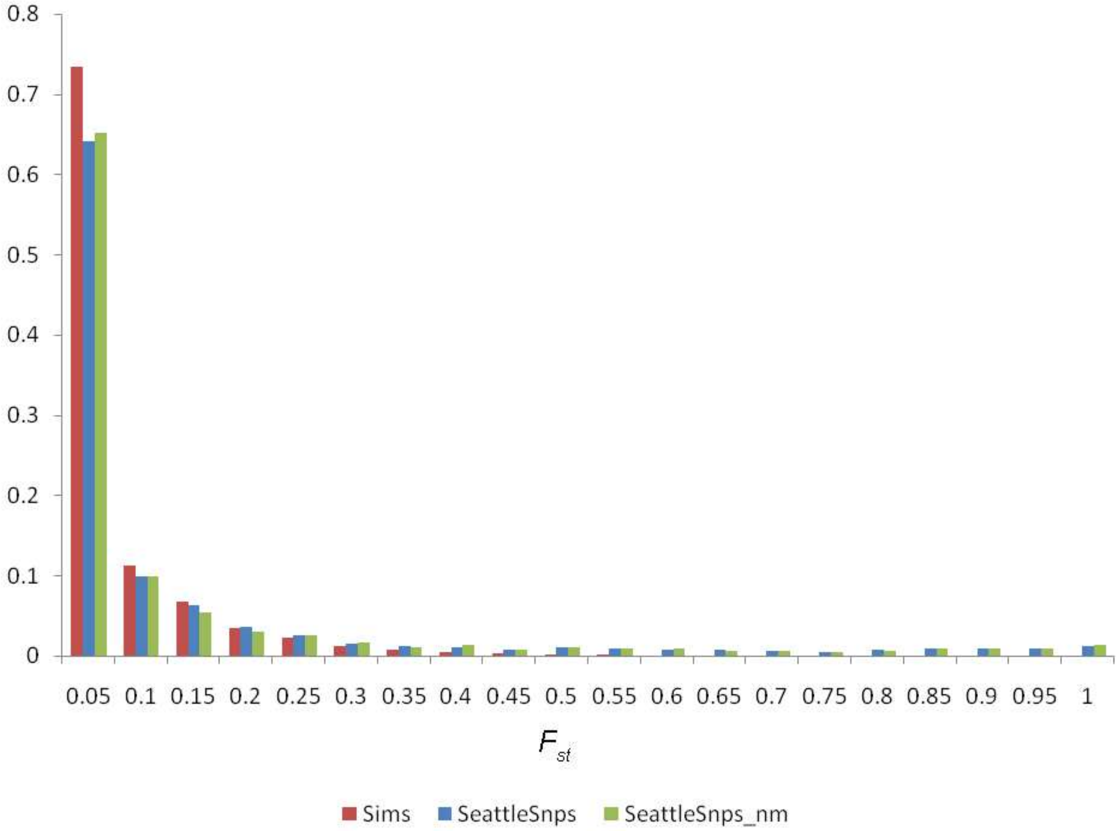

A
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
B
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
F_{st}
Sims_d8
Sims_d16
Sims_d24
Sims_d32
Sims_d40
Sims_all

Figure 4


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
F_{st}
Sims
SeattleSnps
SeattleSnps_nm

DISCUSSION

“ Ítaca t'ha donat el bell viatge,
sense ella no hauries sortit.
I si la trobes pobra, no és que Ítaca
t'hagi enganyat. Savi, com bé t'has fet,
sabràs el que volen dir les Ítaques.”

Viatge a Ítaca
Konstantinos P. Kavafis
(Adaptació de Lluís Llach sobre la versió de Carles Riba)

1 THE FINAL FRONTIER

The papers presented in the Results section tackle two different issues, both directed to detect departures from neutrality based in genomic data.

The first two works are aimed to ascertain how the power of the neutrality tests is affected by different parameters such as recombination and phasing of the sequences in samples with underlying demographic histories. In the first paper (Results, section 1) we have tested the power of neutrality statistics under different demographic events and recombination rates. We have shown that recombination has a large impact on the power of the neutrality statistics, mainly on those that are based on the haplotype distribution. This effect is mostly relevant if the recombination rate has not been correctly estimated, especially if it has been overestimated in the neutral model. Based on these results, we also provide some basic guidelines for the election of the neutrality tests to be used in empirical studies, according to the characteristics of the resequenced sample.

In the second work (Results, section 2), we were concerned about the possible effect of comparing empirical, phased sequences, against simulated distributions, of known phase that, thus, do not need to be estimated. Our results show that, although fastPHASE is not very efficient to reconstruct individual haplotypes, especially for large numbers of segregating sites and high recombination rates, in most cases the power of neutrality tests is only slightly affected. Thus, neutrality statistics can be used without any concern over phased data.

In the last part of the thesis we have explored different methods to deal with ascertainment bias. First, we have developed corrected estimators of θ and their variances and covariances and we have presented a Tajima's *D* corrected to work with data that has been ascertained finding the SNPs in only a subset of samples (Results section 3). Using coalescent simulations, this new, corrected Tajima's *D* can reconstruct the distribution of the usual *D* without ascertainment with considerable accuracy, although it slightly tends towards negative values. We have applied the corrected *D* over Perlegen data, which has mostly been genotyped using this kind of bias, and have found some genes with extreme values of Tajima's *D*. We also provide a program to calculate the corrected Tajima's *D* over sequences.

Finally, we have developed a software that allows comparing F_{ST} from genotyping data against coalescent simulations (Results, section 4). This program is based upon the idea that, if we have good information on the demographic history of the populations, we can simulate an F_{ST} distribution that matches the bias applied to the samples. We have applied our program to build that simulated F_{ST} distributions with different biases.

2 RAIDERS OF THE LOST ARK

In a world in which, in the best of cases, empirical data is obtained through resequencing, but in which the information we have about recombination and demography is only partial and under constant change, is there any hope to detect natural selection?

Although the tools we currently possess to detect departures of neutrality are powerful enough, we are well aware that their blind application over genomic data does not provide information on selection but on a number of factors that also leave their trace on sequences. This 'genetic noise' is mainly produced by demographic events and recombination, which, in fact, are also deviations from strict neutrality.

Along the first two papers presented in this thesis (Results sections 1 and 2) we have explored intensely these two forces shaping genetic variation and how they affect the power of neutrality tests. For the former we have stated that, as previously known, neutrality statistics can detect a large fraction of demographic events when compared against a neutral distribution. Considering that, in addition, selection and demography leave similar traces on the genomes, how can we distinguish them using such tests? The clue to solve this problem lies in the fact that, while demographic events affect the whole genome, selection has a local effect, acting over particular genetic variants. Thus, two strategies can be employed: (a) to use a neutral model that assumes an underlying demography or (b) to compare against an empirical distribution.

Both strategies were used by Soldevila *et al.* (2006) in a resequecing study aimed to describe the patterns of variation on the *PRNP* gene in humans and to unravel the selective pressures that have acted on it. In this work, we sequenced 348 chromosomes from populations worldwide and calculated Tajima's *D*, Fay and Wu's *D** and *F**, Fu's F_S and Fay and Wu's *H* for the eight population groups defined and taking together the whole world sample. Significant negative values were found for each statistic except for Fay and Wu's *H* in the world sample. The significance of the tests was first estimated by means of comparison against neutral coalescent simulations. However, taking into account the confounding effect of demographic events, confidence intervals were recalculated from a series of simulations assuming population histories with sudden expansions, as done by Wooding *et al.* (2004). Furthermore, Tajima's *D* values were also compared to their distribution in the 132 genes resequenced at that moment by SeattleSNPs.

After the publication of the calibrated human demographic history by Schaffner *et al.* (2005) in late 2005, we changed our strategy to ascertain the significance of the neutrality statistics obtained from resequencing projects. In the subsequent papers (Ogorelkova *et al.* in preparation; Calafell *et al.* in preparation; Casals *et al.* in preparation, *a*; Casals *et al.* in preparation, *b*), then, we compared our results against Schaffner's demographic model (Schaffner *et al.* 2005). In order to do that, different models were generated for each gene or genes resequenced, all with the same underlying demography but with the sequence length, and *S* or θ, of the region of interest.

In humans, this confounding effect of demographic events is particularly important when looking for positive selection, as was done in the papers explained above. This is due to the fact that humans have experienced several bottlenecks and population expansions, which shape the genome in an analogous way as selection does and, thus, made the use of neutral models liberal to detect selective sweeps. However, as balancing selection leaves the opposite traces over the genome, in this particular case using pure neutral models would be conservative. Based on that, in Calafell *et al.* (2008) Tajima's *D* and Fu and Li's *F* were used to detect traces of balancing selection on the *ABO* gene by comparing them against stationary distributions. Moreover, both tests were also computed in sliding windows along the sequence, and its statistical significance was tested using parametric bootstrapping.

As for the second issue, in the first paper included in the Results section we have shown that recombination affects the power of neutrality statistics, and particularly of those based on haplotypes. Using incorrect estimates of the recombination rate in the neutral model, that is, comparing against a neutral distribution with an underlying recombination rate larger or smaller than the one of the region of interest, leads to increased type I and II errors, depending on the neutrality statistic. For this reason, in all the papers discussed regarding neutral distributions, with or without underlying demography, simulations have been run assuming the same recombination rate than the resequenced gene. In Soldevila *et al.* (2006) and Calafell *et al.* (2008; in preparation) we have used the recombination rate provided by Kong *et al.* (2002) for STR D20597, D9S754 and D5S469, respectively. In Ogorelkova *et al.* (in preparation) we have used no recombination and the minimum, maximum, and average recombination in Kong *et al.*'s (2002) recombination map. In Casals *et al.* (in preparation, *a;* in preparation, *b*), the recombination rate for each region was obtained from HapMap (http://www.hapmap.org).

3 SNPs: THE FATAL ATTRACTION

Although resequencing data is the ideal base upon which to look for traces of selection by means of neutrality statistics, it a slow and expensive procedure. On the other hand, genotyping technologies have experienced a quick development that has made them the most attractive way to characterize genetic variation. SNPs are not only quick and cheap to obtain for anyone's region of interest, but huge amounts of them are available in public databases. Genotyping platforms such as SNPlexTM by AppliedBiosystems or the Illumina BeadArray 650Y, that allow whole-genome genotyping of over 655,000 tag SNPs described by the International HapMap Project (www.hapmap.org), have popularized the use of genomewide SNPs. Furthermore, public projects and databases such as HapMap, Perlegen (http://www.perlegen.com/) or the Stanford HGDP SNP Genotyping Data (http://shgc.stanford.edu/hgdp/files.html) provide large amounts of genotypes ready to use.

It is thus tempting to use all this available information and to analyse it in terms of selection. However, the use of neutrality tests is not recommended on SNP data as the process of selecting them produces a bias that often shifts the statistics' distributions, making their results unreliable if not completely misleading. Some tools, then, have been developed to make it possible to find traces of selection on genotyping data. One of the most popular of such methods is Sabeti's *EHH* statistic (Sabeti *et al.* 2002), designed to detect recent selective sweeps based on haplotypes. Another widely used method is to use F_{ST}, comparing its value for each SNP of interest against an empirical distribution (Akey *et al.* 2002), to detect geographically localized selection events.

EHH, the first of these methods, has been used in Moreno-Estrada *et al.* (2008), and Casals *et al.* (in preparation, *a*). In Moreno-Estrada *et al.* (2008) we used *EHH* over HapMap data, as the analysis of our genotypes was not feasible due to their low recombination rate. In Casals *et al.* (in preparation, *a*) we calculated *EHH* both over our genotypes and over HapMap data. Moreover, the iHS test (Voight *et al.* 2006), a derivation of *EHH*, was also applied. In neither of the works we could find traces of recent selective sweeps.

F_{ST} was used to detect population specific selection on 15 innate immunity genes by Ferrer-Admetlla *et al.* (2008) in the 1,049 samples included in the HGDP-CEPH Human Genome Diversity Panel (Cann *et al.* 2002). The F_{ST} distribution for all the SNPs in the selected genes were compared against empirical distributions obtained from three different sets of data composed by equivalent samples: (a) the SNPs in the Alfred database (http://alfred.med.yale.edu/alfred/, Rajeevan *et al.* 2003), (b) the SNPs from a gene free region on chromosome 22 (Gonzalez-Neira *et al.* 2004) and (c) an insertion/deletion set (Weber *et al.* 2002). Besides from the distributions comparison, single SNPs were also checked for extreme F_{ST} values. However, F_{ST} did not show in any case consistent traces of local selection.

In the three works, moreover, neutrality tests were also used on resequencing data, obtained (a) by directly sequencing (Casals,F. *et al.* in preparation, *a*), (b) from the sequences published by Nielsen *et al.* (2005a) (Moreno-Estrada *et al.* 2008), and (c) by dowloading the sequences from SeattleSNPs (Ferrer-Admetlla *et al.* 2008). The significance of neutrality statistics was estimated by using Schaffner's calibrated demography (Schaffner *et al.* 2005) as neutral model.

In this thesis, we have adressed the SNP problem from two different perspectives. One one hand, we have adapted neutrality statistics to their use on SNP data (Results section 3). We have provided two corrected estimators of the mutation parameter θ, $\theta_{W|C}$ and $\theta_{T|C}$, their variances and covariances, together with a corrected Tajima's *D*. Our equations fit a discovery sample bias, that is, assume that only a subsample has been resequenced and the SNPs found have been genotyped in the whole sample. However, as the results are expressed in terms of the probability of ascertainment of a SNP as a function of its frequency, this work can be extended to other ascertainment schemes, such as those considered in Nielsen *et al.* (2005b).

On the other hand, we have focused on the F_{ST}-based methods to detect local selection (Results section 4). In this sense, we have developed a program, FABSIM, that allows working with F_{ST} by comparing it against simulated distributions. Those simulations, built using demographic models close to what we know from the sample, can be biased using different ascertainment schemes, thus reproducing the process by which genotypes have been obtained.

4 BACK TO THE FUTURE

Although young, large-scale genotyping technologies are starting to see the beginning of their end. While the Human Genome Project, released on 2003, cost 13 years and 3 billion dollars, on 2007 James Watson had his genome resequenced in two months for 1 milion dollars (Check. 2007) using the 454 pyrosequencing platform acquired by Roche (Margulies *et al.* 2005). But 454 is not the only platform for large-scale resequencing. Illumina and AppliedBiosystems have also developed or acquired their own sequencing platforms, Solexa and SOLiD, respectively. All this new technologies are generally named ultrasequencing, in front of the traditional, Sanger sequencing.

Besides the existing platforms, companies and institutions are pushing hard for the improvement of these new sequencing technologies, mainly looking to saving money and time with each sequence. Companies are on the race to build the 'thousand-dollar genome', the goal of the new genomics era (Mardis. 2006). In the meantime, on 2006, Archon Genomics announced that they will give 10 million dollars to the first group that can sequence 100 human genomes in 10 days with less than 10,000 dollars per genome (http://genomics.xprize.org/).

The question that arises in this context is what impact will have the availability of hundreds, or even thousands, of resequenced genomes on the search for selection? This can move from science fiction into reality in only three years with the 1,000 Genomes Project (Kaiser. 2008), which plans to resequence 1,000 genomes in this time. The project, announced in January 2008, will be developed by an international consortium and has an estimate cost of between 30 and 50 million dollars. And they are not the only ones. The BGI Shenzhen, in China, will resequence 99 individuals in the Yanhuang Project, which also has an expected length of three years and will overlap and share some samples with the 1,000 Genomes Project. Furthermore, J. Craig Venter has vowed to provide complete diploid sequences for 10,000 humans in 10 years.

One of the first implications of the new, and mainly the future, ultrasequencing technologies will be the progressive abandonment of genotyping platforms to generate data aimed to detecting selection, especially when the $1,000-genome will be available. As a consequence, all the methods developed to find selection based on SNP data, such as F_{ST} (Akey *et al.* 2002), *EHH* (Sabeti *et al.* 2002) or the ones presented in this thesis, will became obsolete. However, even given that those tools are being produced with a short expiring date, the effort is worthwhile. Nowadays, only a very short fraction of all the scientific institutions worldwide can afford the price of large-scale resequencing, even with ultrasequencing, and even they cannot do that alone. Although we are glimpsing the future, this future will still be unavailable for most researchers for some years, and during this time, as well as in the present and recent past, there will be a great need for such methods. Moreover, even when the price of ultrasequencing reaches $1,000 or even less, a considerable large number of poor countries and laboratories will still not have the resources to pay for their own resequencing and will need to keep using SNPs or to work with publicly available sequences.

On the other hand, ultrasequencing will represent the reemergence of classical tools such as neutrality tests. In this scenario, on the long run genotyping technologies and the methods developed for their analysis will look as a small oasis, a short transition period or parenthesis between resequencing technologies. However, how will we use these tools? Nowadays, neutrality statistics are usually applied on single genes or genomic regions. But when ultrasequencing spreads, we will be able to find traces of selection on full genomes. Under these circumstances, the efforts will surely be directed to apply neutrality statistics by overlapping windows in order to identify those regions under selection. Another problem that will arise with ultrasequencing is that it does not provide the complete diploid sequence for each individual and, thus, the sample size is not the same in all loci. This issue, also found in some Sanger resequencing projects that are performed assembling small reads of DNA, such as the genome projects, is currently been addressed using Composite Likelihood Estimators (Hellmann *et al.* 2008).

Finding how humans have adapted genetically to new environments provides a window to our evolutionary past. However, we are opening this window from our present genomes, and both the key to open the window (neutrality tests) and the lock in the window (the genome itself) have gathered rust in the process. We have attempted to polish both lock and key, probably just before new technologies will pick that lock by brute force.

BIBLIOGRAPHY

"Sí, hablo de un lugar donde leer te puede llevar a la locura.
Donde los libros pueden herir, envenenar, incluso matar."

La ciudad de los libros soñadores
Walter Moers

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. **12:** 1805-14.

Alonso, S., N. Izagirre, I. Smith-Zubiaga, J. Gardeazabal, J. L. Diaz-Ramon *et al.* 2008 Complex signatures of selection for the melanogenic loci TYR, TYRP1 and DCT in humans. BMC Evol. Biol. **8:** 74.

Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin *et al.* 2000 An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature **407:** 513-6.

Andolfatto, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in drosophila melanogaster and drosophila simulans. Mol. Biol. Evol. **18:** 279-290.

Baudry, E., B. Viginier and M. Veuille, 2004 Non-african populations of drosophila melanogaster have a unique origin. Mol. Biol. Evol. **21:** 1482-1491.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783-796.

Calafell, F., F. Roubinet, A. Ramirez-Soriano, N. Saitou, J. Bertranpetit *et al.* 2008 Evolutionary dynamics of the human ABO gene. Hum Genet. DOI: 10.1007/s00439-008-0530-8

Calafell, F., and J. Bertranpetit, 1993 The genetic history of the iberian peninsula: A simulation. Curr. Anthropol. **34:** 735.

Canino, M. F., and P. Bentzen, 2004 Evidence for positive selection at the pantophysin (pan I) locus in walleye pollock, theragra chalcogramma. Mol. Biol. Evol. **21:** 1391-1400.

Cann, H. M., C. de Toma, L. Cazes, M. F. Legrand, V. Morel *et al.* 2002 A human genome diversity cell line panel. Science **296:** 261-2.

Carlson, C. S., M. A. Eberle, L. Kruglyak and D. A. Nickerson, 2004 Mapping complex disease loci in whole-genome association studies. Nature **429:** 446-52.

Castro, J. A., A. Picornell and M. Ramon, 1998 Mitochondrial DNA: A tool for populational genetics studies. Int. Microbiol. **1:** 327-332.

Cavalli-Sforza, L. L., 1998 The DNA revolution in population genetics. Trends Genet. **14:** 60-65.

Check, E., 2007 *James Watson's Genome Sequenced*. DOI: 10.1038/news070528-10

Civetta, A., S. A. Rajakumar, B. Brouwers and J. P. Bacik, 2006 Rapid evolution and gene-specific patterns of selection for three genes of spermatogenesis in drosophila. Mol. Biol. Evol. **23:** 655-662.

Coia, V., G. Destro-Bisol, F. Verginelli, C. Battaggia, I. Boschi *et al.* 2005 Brief communication: MtDNA variation in north cameroon: Lack of asian lineages and implications for back migration from asia to sub-saharan africa. Am. J. Phys. Anthropol. **128:** 678-681.

Comas, D., F. Calafell, N. Benchemsi, A. Helal, G. Lefranc *et al.* 2000 Alu insertion polymorphisms in NW africa and the iberian peninsula: Evidence for a strong genetic boundary through the gibraltar straits. Hum. Genet. **107:** 312-319.

Comas, D., F. Calafell, E. Mateu, A. Perez-Lezaun, E. Bosch *et al.* 1998 Trading genes along the silk road: MtDNA sequences and the origin of central asian populations. Am. J. Hum. Genet. **63:** 1824-1838.

Crawford, D. C., D. T. Akey and D. A. Nickerson, 2005 The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet **6:** 287-312.

De Mita, S., J. Ronfort, H. I. McKhann, C. Poncet, R. El Malki *et al.* 2007 Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in medicago truncatula. Genetics **177:** 2123-2133.

Depaulis, F., S. Mousset and M. Veuille, 2005 Detecting selective sweeps with haplotype tests in *Selective Sweep*, edited by D. Nurminsky. Landes Bioscience, Georgetown.

Depaulis, F., S. Mousset and M. Veuille, 2003 Power of neutrality tests to detect bottlenecks and hitchhiking. J. Mol. Evol. **57 Suppl 1:** 190-200.

Derome, N., E. Baudry, D. Ogereau, M. Veuille and C. Montchamp-Moreau, 2008 Selective sweeps in a 2-locus model for sex-ratio meiotic drive in drosophila simulans. Mol. Biol. Evol. **25:** 409-416.

Destro-Bisol, G., F. Donati, V. Coia, I. Boschi, F. Verginelli *et al.* 2004 Variation of female and male lineages in sub-saharan populations: The importance of sociocultural factors. Mol. Biol. Evol. **21:** 1673-1682.

Donnelly, P., and S. Tavare, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401-21.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87-112.

Excoffier, L., G. Laval and S. Schneider, 2005 Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online **1**:47-50.

Excoffier, L., J. Novembre and S. Schneider, 2000 SIMCOAL: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J. Hered. **91:** 506-509.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive darwinian selection. Genetics **155:** 1405-13.

Ferrer-Admetlla, A., E. Bosch, M. Sikora, T. Marques-Bonet, A. Ramirez-Soriano *et al.* 2008 Balancing selection is the main force shaping the evolution of innate immunity genes. J. Immunol. **181:** 1315-1322.

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

Friedlaender, J. S., F. R. Friedlaender, F. A. Reed, K. K. Kidd, J. R. Kidd *et al.* 2008 The genetic structure of pacific islanders. PLoS Genet. **4:** e19.

Fu, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915-25.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172-197.

Fu, Y. X., and W. H. Li, 1999 Coalescing into the 21st century: An overview and prospects of coalescent theory. Theor. Popul. Biol. **56:** 1-10.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

Gale, J. S., 1990 *Theoretical Population Genetics*. Unwin Hyman Inc., UK.

Garrigan, D., and M. F. Hammer, 2006 Reconstructing human origins in the genomic era. Nat. Rev. Genet. **7:** 669-680.

Garrigan, D., S. B. Kingan, M. M. Pilkington, J. A. Wilder, M. P. Cox *et al.* 2007 Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics **177:** 2195-2207.

Goebel, T., M. R. Waters and D. H. O'Rourke, 2008 The late pleistocene dispersal of modern humans in the americas. Science **319:** 1497-1502.

Gonzalez-Neira, A., F. Calafell, A. Navarro, O. Lao, H. Cann *et al.* 2004 Geographic stratification of linkage disequilibrium: A worldwide population study in a region of chromosome 22. Hum. Genomics **1:** 399-409.

Harpending, H. C., 1994 Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum. Biol. **66:** 591-600.

Harpending, H. C., S. T. Sherry, A. R. Rogers and M. Stoneking, 1993 The genetic structure of ancient human populations. Curr. Anthropol. **34:** 483.

Hein, J., M. H. Schierup and C. Wiuf, 2005 *Gene Genealogies,Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.

Hellmann, I., Y. Mang, Z. Gu, P. Li, F. M. de la Vega *et al.* 2008 Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res. DOI: 10.1101/gr.074187.107

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin *et al.* 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072-1079.

Hoskins, R. A., A. C. Phan, M. Naeemuddin, F. A. Mapa, D. A. Ruddy *et al.* 2001 Single nucleotide polymorphism markers for genetic mapping in drosophila melanogaster. Genome Res. **11:** 1100-1113.

Hudson, R. R., 2002 Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18:** 337-8.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23-36 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1-44 in *Oxford Surveys in Evolutionary Biology*, edited by J. Antonovics and D. Futuyama. Oxford University Press, Oxford.

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

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

Jakobsson, M., J. Hagenblad, S. Tavare, T. Sall, C. Hallden *et al.* 2006 A unique recent origin of the allotetraploid species arabidopsis suecica: Evidence from nuclear DNA markers. Mol. Biol. Evol. **23:** 1217-1231.

Jobling, M. A., M. E. Hurles and C. Tyler-Smith, 2004 *Human Evolutionary Genetics: Origins, Peoples & Disease*. Garland Science, New York.

Kaiser, J., 2008 DNA SEQUENCING: A plan to capture human diversity in 1000 genomes. Science **319:** 395.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887-899.

Kaplan, N. L., T. Darden and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics **120:** 819-829.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197-206.

Kimura, M., 1984 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kimura, M., 1968 Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. **11:** 247-269.

Kingman, J. F., 2000 Origins of the coalescent. 1974-1982. Genetics **156:** 1461-3.

Kingman, J. F. C., 1982a On the genealogy of large populations. J. Appl. Prob. **19A:** 27-43.

Kingman, J. F. C., 1982b The coalescent. Stoch.Proc.Applns. **13:** 235-248.

Knight, A., P. A. Underhill, H. M. Mortensen, L. A. Zhivotovsky, A. A. Lin *et al.* 2003 African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr. Biol. **13:** 464-473.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.* 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241-7.

Kreitman, M., and A. Di Rienzo, 2004 Balancing claims for balancing selection. Trends Genet. **20:** 300-4.

Laan, M., V. Wiebe, E. Khusnutdinova, M. Remm and S. Paabo, 2005 X-chromosome as a marker for population history: Linkage disequilibrium and haplotype study in eurasian populations. Eur. J. Hum. Genet. **13:** 452-462.

Laval, G., and L. Excoffier, 2004 SIMCOAL 2.0: A program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics **20:** 2485-2487.

Li, W. H., 1976 Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: The finite island model. Theor. Popul. Biol. **10:** 303-308.

Mardis, E. R., 2006 Anticipating the 1,000 dollar genome. Genome Biol. **7:** 112.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader *et al.* 2005 Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:** 376-380.

McElreavey, K., and L. Quintana-Murci, 2005 A population genetics perspective of the indus valley through uniparentally-inherited markers. Ann. Hum. Biol. **32:** 154-162.

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

Mead, S., M. P. Stumpf, J. Whitfield, J. A. Beck, M. Poulter *et al.* 2003 Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. Science **300:** 640-643.

Moran, P. A., 1958a A general theory of the distribution of gene frequencies. II. non-overlapping generations. Proc. R. Soc. Lond. B. Biol. Sci. **149:** 113-116.

Moran, P. A., 1958b A general theory of the distribution of gene frequencies. I. overlapping generations. Proc. R. Soc. Lond. B. Biol. Sci. **149:** 102-112.

Moreno-Estrada, A., F. Casals, A. Ramirez-Soriano, B. Oliva, F. Calafell *et al.* 2008 Signatures of selection in the human olfactory receptor OR5I1 gene. Mol. Biol. Evol. **25:** 144-154.

Nakajima, T., S. Wooding, T. Sakagami, M. Emi, K. Tokunaga *et al.* 2004 Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. Am. J. Hum. Genet. **74:** 898-916.

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

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931-942.

Nielsen, R., and J. Signorovitch, 2003 Correcting for ascertainment biases when analyzing SNP data: Applications to the estimation of linkage disequilibrium. Theor. Popul. Biol. **63:** 245-55.

Nielsen, R., M. J. Hubisz and A. G. Clark, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics **168:** 2373-2382.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton *et al.* 2005a A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. **3:** e170.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.* 2005b Genomic scans for selective sweeps using SNP data. Genome Res. **15:** 1566-1575.

Ortiz-Barrientos, D., A. S. Chang and M. A. Noor, 2006 A recombinational portrait of the drosophila pseudoobscura genome. Genet. Res. **87:** 23-31.

Patin, E., L. B. Barreiro, P. C. Sabeti, F. Austerlitz, F. Luca *et al.* 2006 Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. Am. J. Hum. Genet. **78:** 423-436.

Picoult-Newberg, L., T. E. Ideker, M. G. Pohl, S. L. Taylor, M. A. Donaldson *et al.* 1999 Mining SNPs from EST databases. Genome Res. **9:** 167-174.

Pinto, J., A. Lynd, J. L. Vicente, F. Santolamazza, N. P. Randle *et al.* 2007 Multiple origins of knockdown resistance mutations in the afrotropical mosquito vector anopheles gambiae. PLoS ONE **2:** e1243.

Polanski, A., and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics **165:** 427-36.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: A study of Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791-1798.

Rajeevan, H., M. V. Osier, K. H. Cheung, H. Deng, L. Druskin *et al.* 2003 ALFRED: The ALelle FREquency database. update. Nucleic Acids Res. **31:** 270-271.

Ramirez-Soriano, A., and R. Nielsen, Correcting estimators of theta and Tajima's D for ascertainment biases caused by the SNP discovery process. Submitted.

Ramirez-Soriano, A., and F. Calafell, FABSIM: A software for generating fst distributions with various ascertainment biases. Submitted.

Ramirez-Soriano, A., S. E. Ramos-Onsins, J. Rozas, F. Calafell and A. Navarro, 2008 Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics **179:** 555-567.

Ramos-Onsins, S. E., and T. Mitchell-Olds, 2007 Mlcoalsim: Multilocus coalescent simulations. Evol Bioinform Online **3:** 41-44.

Ramos-Onsins, S. E., and J. Rozas, 2002 Statistical properties of new neutrality tests against population growth. Mol. Biol. Evol. **19:** 2092-100.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.* 2001 Linkage disequilibrium in the human genome. Nature **411:** 199-204.

Rendine, S., A. Piazza and L. L. Cavalli-Sforza, 1986 Simulation and separation by principal components of multiple demic expansions in europe. Am. Nat. **128:** 681.

Ridley, M., 2000 *The Search for LUCA - Last Universal Common Ancestor*.

Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552-69.

Rogers, A. R., A. E. Fraley, M. J. Bamshad, W. S. Watkins and L. B. Jorde, 1996 Mitochondrial mismatch analysis is insensitive to the mutational process. Mol. Biol. Evol. **13:** 895-902.

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

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496-2497.

Rozas, J., M. Gullaud, G. Blandin and M. Aguade, 2001 DNA variation at the rp49 gene region of drosophila simulans: Evolutionary inferences from an unusual haplotype structure. Genetics **158:** 1147-55.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.* 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832-7.

Salas, A., M. Richards, T. De la Fe, M. V. Lareu, B. Sobrino *et al.* 2002 The making of the african mtDNA landscape. Am. J. Hum. Genet. **71:** 1082-1111.

Sanchez-Gracia, A., and J. Rozas, 2007 Unusual pattern of nucleotide sequence variation at the OS-E and OS-F genomic regions of drosophila simulans. Genetics **175:** 1923-1935.

Sano, A., and H. Tachida, 2005 Gene genealogy and properties of test statistics of neutrality under population growth. Genetics **169:** 1687-1697.

Schaffner, S. F., 2004 The X chromosome in population genetics. Nat. Rev. Genet. **5:** 43-51.

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

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. Am. J. Hum. Genet. **78:** 629-644.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar and T. Mitchell-Olds, 2005 A multilocus sequence survey in arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169:** 1601-1615.

Schneider, S., D. Roessli and L. Excoffier, 2000 Arlequin: A software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

Singer, A., H. Perlman, Y. Yan, C. Walker, G. Corley-Smith *et al.* 2002 Sex-specific recombination rates in zebrafish (danio rerio). Genetics **160:** 649-657.

Singer, T., Y. Fan, H. S. Chang, T. Zhu, S. P. Hazen *et al.* 2006 A high-resolution map of arabidopsis recombinant inbred lines by whole-genome exon array hybridization. PLoS Genet. **2:** e144.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555-62.

Smith, J. M., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23-35.

Soejima, M., H. Tachida, T. Ishida, A. Sano and Y. Koda, 2006 Evidence for recent positive selection at the human AIM1 locus in a european population. Mol. Biol. Evol. **23:** 179-188.

Soldevila, M., F. Calafell, A. Helgason, K. Stefansson and J. Bertranpetit, 2005 Assessing the signatures of selection in PRNP from polymorphism data: Results support kreitman and di rienzo's opinion. Trends Genet. **21:** 389-391.

Soldevila, M., A. M. Andres, A. Ramirez-Soriano, T. Marques-Bonet, F. Calafell *et al.* 2006 The prion protein gene in humans revisited: Lessons from a worldwide resequencing study. Genome Res. **16:** 231-239.

Spencer, C. C. A., and G. Coop, 2004 SelSim: A program to simulate population genetic data with natural selection and recombination. Bioinformatics **20:** 3673-3675.

Spurdle, A. B., and T. Jenkins, 1992 The Y chromosome as a tool for studying human evolution. Curr. Opin. Genet. Dev. **2:** 487-491.

Stajich, J. E., and M. W. Hahn, 2005 Disentangling the effects of demography and selection in human history. Mol. Biol. Evol. **22:** 63-73.

Stephens, M., and P. Donnelly, 2003 A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. **73:** 1162-9.

Stephens, M., N. J. Smith and P. Donnelly, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68:** 978-989.

Stoneking, M., 1994 Mitochondrial DNA and human evolution. J. Bioenerg. Biomembr. **26:** 251-259.

Strobeck, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: A test for population subdivision. Genetics **117:** 149-153.

Stumpf, M. P., and D. B. Goldstein, 2001 Genealogical and evolutionary inference with the human Y chromosome. Science **291:** 1738-1742.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-95.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437-460.

Takahata, N., Y. Satta and J. Klein, 1995 Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. **48:** 198-221.

Tavare, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505-518.

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

Underhill, P. A., 2003 Inferring human history: Clues from Y-chromosome haplotypes. Cold Spring Harb. Symp. Quant. Biol. **68:** 487-493.

Verrelli, B. C., and S. A. Tishkoff, 2004 Signatures of selection and gene conversion associated with human color vision variation. Am. J. Hum. Genet. **75:** 363-375.

Voight, B. F., A. M. Adams, L. A. Frisse, Y. Qian, R. R. Hudson *et al.* 2005 Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Natl. Acad. Sci. U. S. A. **102:** 18508-18513.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biology **4:** e72 OP.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero and K. Ardlie, 2001 The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. Am. J. Hum. Genet. **69:** 1332-1347.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65-79.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-76.

Weber, J. L., D. David, J. Heil, Y. Fan, C. Zhao *et al.* 2002 Human diallelic insertion/deletion polymorphisms. Am. J. Hum. Genet. **71:** 854-862.

Wood, E. T., D. A. Stover, C. Ehret, G. Destro-Bisol, G. Spedini *et al.* 2005 Contrasting patterns of Y chromosome and mtDNA variation in africa: Evidence for sex-biased demographic processes. Eur. J. Hum. Genet. **13:** 867-876.

Wooding, S., U. K. Kim, M. J. Bamshad, J. Larsen, L. B. Jorde *et al.* 2004 Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. Am. J. Hum. Genet. **74:** 637-646.

Wright, S., 1931 Evolution in mendelian populations. Genetics 97-159.

Wright, S. I., N. Nano, J. P. Foxe and V. U. Dar, 2008 Effective population size and tests of neutrality at cytoplasmic genes in arabidopsis. Genet. Res. **90:** 119-128.

LIST OF ABBREVIATIONS

"The Spacer said, "I shall introduce myself. I am R. Daneel Olivaw."
"Yes? Am I making a mistake? I thought the first initial--"
"Quite so. I am a Robot. Were you not told?""

The Caves of Steel
Isaac Asimov

ARG	Ancestral Recombination Graph
DAF	Derived Allele Frequency
EHH	Extended HaplotypeHomozigosity
HW	Hardy-Weinberg
LD	Linkage Disequilibrium
LUCA	Last Unique Common Ancestor
MAE	Mean Absolute Error
MAF	Minor Allele Frequency
MRCA	Most Recent Common Ancestor
mtDNA	mitochondrial DNA
NRY	Non-Recombining segment of the Y chromosome
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat

APPENDICES

"The Wheel of Times turns, and Ages come and pass,
leaving memories that become legend. Legend fades to myth,
and even myth is long forgotten when the Age that gave it birth
comes again. […] There are neither beginnings nor endings
to the running of the Wheel of Time."

The Wheel of Time
Robert Jordan

1 STATISTICAL POWER ANALYSIS OF NEUTRALITY TESTS UNDER DEMOGRAPHIC EXPANSIONS, CONTRACTIONS AND BOTTLENECKS WITH RECOMBINATION (Ramirez-Soriano *et al.* 2008): SUPPLEMENTAL DATA

This section includes Supplemental Results and Supplemental Figures from Ramirez-Soriano *et al.* 2008. Supplemental Data A-F is included in the CD-ROM attached.

CONTRACTION RESULTS

In contrast to the pattern described in sudden expansions, in the sudden contraction model (Supplemental Figure 1) neutrality tests show power for the right tail of the distribution with two exceptions: *Dh* and Fay and Wu's *H*. This two tests not only show power at the right tail but also at the left tail, *Dh* for *S*=100 and *H* for all *S* values. As expected, power increases with sample size, number of segregating sites and degree of expansion. The most powerful tests for *S*=10 are Fu and Li's *D** for population contractions occurred before T_c=0.2 (except for *n*=100, which makes Fu and Li's *D* more powerful) and F_s for older contractions. For larger *S* values, F_s remains the most powerful statistic for expansions more recent than T_c=0.2, while for older generations other tests reach similar or greater powers, mainly Z_{nS} and Z_A. Maximum power to detect population contractions is extremely different than in the case of expansions: it is mainly influenced by the number of segregating sites and can be found between T_c=0.1 (*S*=100) and T_c=0.4 (*S*=10).

The effects of recombination over the power of tests in the sudden contraction model can be seen in Supplemental Figure 3. As in population expansions, Class I tests generally maintain or improve their power under recombination, while Class II tests perform worse (with the exception of Dh for *S*=10).

In the contraction model, the errors induced by erroneous recombination rates (Supplemental Figure 5) differ from what is seen in population expansions. If recombination is underestimated (Supplemental Figure 5A) all tests, with exception of *Dh* and *ZZ*, experience an increased type I error. In contrast, the effect of overestimating recombination rates (Supplemental Figure 5B) is an increase of type I error for all tests (except *Dh*), Class II tests behaving opposite than in expansions.

BOTTLENECK RESULTS

For bottlenecks (Supplemental Figure 2) the power of tests increases with the number of segregating sites with the exception of *Dh*, which has greater power for *S*=10. Regarding bottleneck properties, statistics have maximum power at their left tails to detect old (T_{start}=0.04-0.08), strong (b=0.05 or stronger), long-lasting bottlenecks, and they have no power to detect bottlenecks that have just finished. In contrast, when testing at the right tail, most tests show power for weak, recent (T_{start}=0.02-0.04) and just finished bottlenecks. This effect is most clear for F_s, Z_{nS} and Class I tests (except Fay and Wu's *H*). The most powerful tests are, as a general rule, Class I tests (except *H*) and Fu's F_s, although Fu and Li's tests lose power for older and short-lasting expansions. Tajima's *D* also loses power with ancient bottlenecks, especially for small sample sizes. For *S*=100, Z_{nS} and to a lesser degree $Z_{A,}$ are also among the most powerful statistics. *Dh* follows a particular pattern, being similar to other well-performing tests for *S*=10 but not for *S*=100. In the latter case, it is similar to other tests but for the right tail of the distribution, while for the left tail shows power only for recent bottlenecks (T_{start}=0.02-0.04), mainly when they have just finished.

In the case of bottlenecks power generally increases with recombination (Supplemental Figure 4), especially for *Dh* and Wall's *B* and *Q*, although the latter two tend to loose power again for higher recombination rates (from r=10^{-8} to r=10^{-7}). In contrast, F_s, Z_{nS} and Z_A decrease their power in the presence of medium-high recombination levels.

A comparison of the apparent and the true power of tests when using misestimates of recombination (Supplemental Figure 6) shows that doing so can lead to serious errors. When recombination is underestimated (comparing the alternative hypothesis with a neutral model without recombination) (Supplemental Figure 6A) most test became conservative with the exception of F_s, EHH and Z_{nS}, in which there is an increase of type I error. In this case, errors are larger than when testing for population expansions, and *Dh*, which was the most liberal test for expansions, becomes greatly conservative in bottlenecks. Recombination overestimates (Supplemental Figure 6B) produce the opposite pattern, Class II F_s, EHH and Z_{nS} being conservative and the rest of tests liberal.

SUPPLEMENTARY FIGURE LEGENDS

Supplemental Figure 1. Power of the test depending on the time elapsed since the contraction, without recombination. n=100 D_c=0.1. (A) S=10. (B) S=100.

Supplemental Figure 2. Power of the test depending on duration of the bottleneck, without recombination. n=100 $T_{start}=0.04$ b=0.01. (A) S=10. (B) S=100.

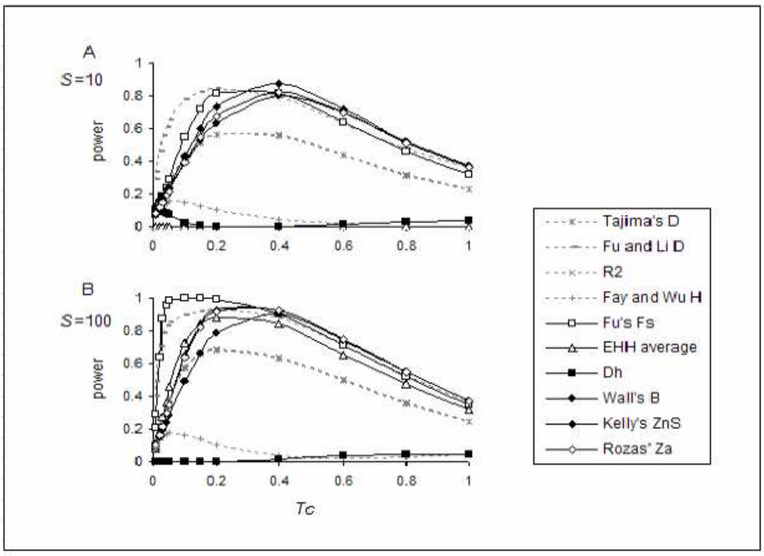
Supplemental Figure 3. Power of the test in the contraction model depending on recombination rates. n=100 De=10. (A) S=10, T_c=0.02. (B) S=100, T_c=0.02. (C) S=10, T_c=0.15. (D) S=100, T_c=0.15.

Supplemental Figure 4. Power of the test in the bottleneck model depending on recombination rates. n=100 T_{start}=0.04 T_{dur}=0.02 b=0.01. (A) S=10. (B) S=100.

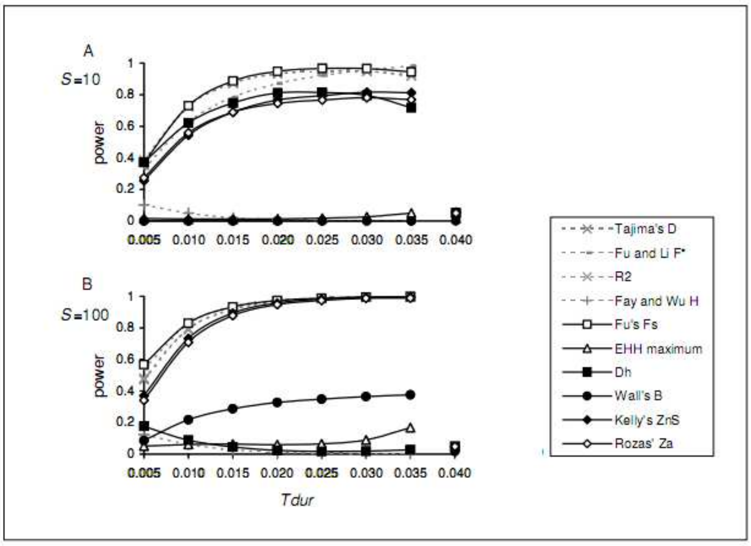
Supplemental Figure 5. Error made by tests in population contractions when recombination is under- or overestimated in the null model. S=100 n=100 D_c=0.1 T_{c}=0.15. (A) the apparent power of the null hypothesis was produced without recombination. (B) the apparent power of the null hypothesis has a recombination rate of 10^{-7}.

Supplemental Figure 6. Error made by tests in bottlenecks when recombination is under- or overestimated in the null model. S=100 n=100 b=0.1 T_{start}=0.04 T_{dur}=0.02. (A) the apparent power of the null hypothesis was produced without recombination. (B) the apparent power of the null hypothesis has a recombination rate of 10^{-7}.

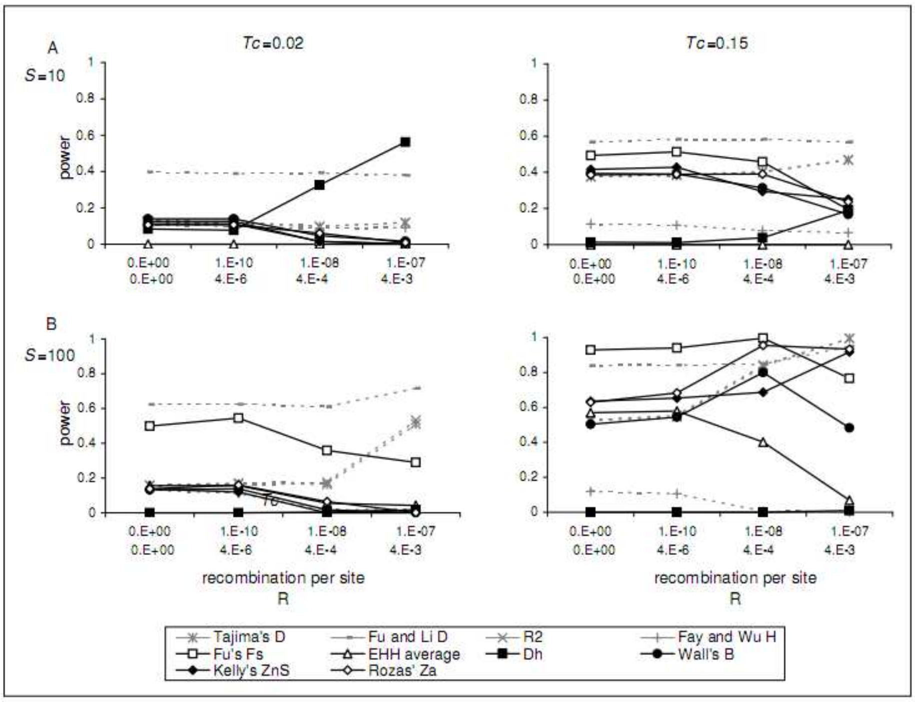
Supplemental Figure 1



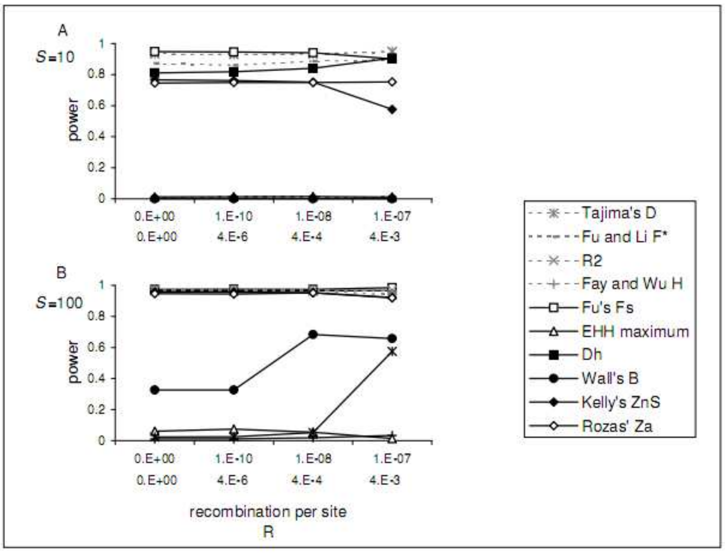
Supplemental Figure 2



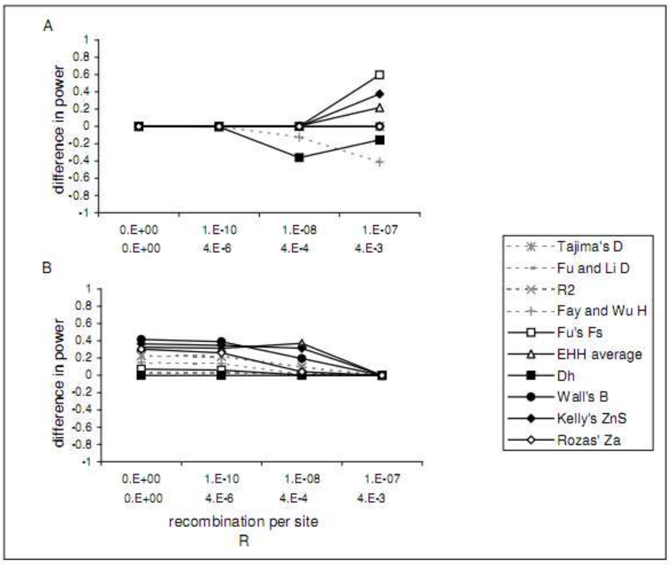
Supplemental Figure 3



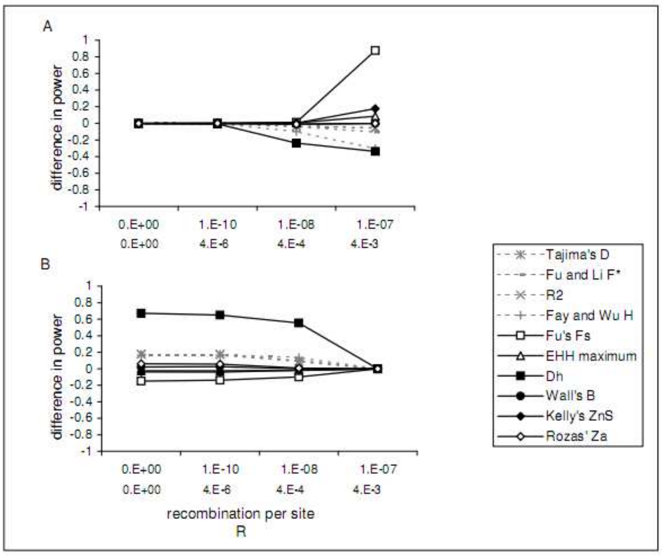
Supplemental Figure 4



Supplemental Figure 5



Supplemental Figure 6



2 CORRECTING ESTIMATORS OF THETA AND TAJIMA'S D FOR ASCERTAINMENT BIASES CAUSED BY THE SNP DISCOVERY PROCESS (Ramirez-Soriano and Nielsen. submitted): SUPPLEMENTAL DATA

Supplemental data from Ramirez-Soriano *et al.* (submitted) is included in the CD-ROM attached.

3 TAJIMA'S *D* CORRECTOR README FILE (Ramirez-Soriano and Nielsen. submitted)

TAJIMA'S D CORRECTOR

ANNA RAMÍREZ-SORIANO AND RASMUS NIELSEN

FEBRUARY 2008

Most SNP data suffers from an ascertainment bias caused by the process of SNP discovery followed by SNP genotyping. SNP genotyping data have an excess of common alleles compared to directly sequenced data, making standard genetic methods of analysis inapplicable to this type of data. We have derived corrected estimators of the fundamental population genetic parameter $\theta = 4N_e\mu$ (N_e = effective population size, μ = mutation rate) based on the average number of pairwise differences and based on the number of segregating sites. We also derive the variances and co-variances of these estimators, and provide a corrected version of Tajima's *D* statistic. Tajima's *D* Corrector implements the corrected estimators we have derived and provides the corrected Tajima's *D* value of a sample, working both over simulation and empirical data, and over constant and changing *d* size.

An example of what Tajima's *D* Corrector can be used for is to find traces of selection on a gene through a genotyping study in which the SNPs have been selected from a discovery sample. That is, the SNP discovery protocol should be a) to resequence a subsample belonging to the genotyping sample and b) to genotype the SNPs found in the subsample *d* in the whole sample. Once the gene is genotyped, the value of the corrected Tajima's *D* can be calculated using this program. Moreover, its significance can also be easily computed by means of simulations, performed using any program aimed at doing coalescent simulations such as ms (Hudson, R. R. 2002. Bioinformatics 18: 337-338). The output of ms can be directly introduced to Tajima's *D* Corrector, where it is possible to simulate the ascertainment scheme done over the sample and to calculate the corrected Tajima's *D* over each sample in te simulations. Then, the significance of the Tajima's *D* of the gene can be easily found by looking where in the distribution obtained by simulations falls.

The program has been developed using Java version 6.0 and compiled under Windows, but it can also be used in Linux with graphic environment. To run Tajima's *D* Corrector, just double-click the .jar file or, if working from the command line, write java –jar TajimaCorrection.jar.

References: Ramírez-Soriano A, Nielsen R. Correcting Estimators of θ and Tajima's *D* for ascertainment biases caused by the SNP discovery process

INPUT FILES

GENERAL INFORMATION

Input files can be entered to the program typing them directly by hand or selecting them using the Browse button.

The only infile format accepted is the ms output format, both for simulations and for empirical data. This format is characterised by having a header and, below, the samples separated by a space and a double bar (//).

The header has the name of the program, the number of chromosomes per sample, the number of runs, the parameter for running ms and the seed used:

```
ms <chromosomes> <runs> -s/-t <various parameters>
<seed>

ms 50 10 -s 111
111
```

Of the header, the only information which will be actually used by Tajima's *D* Corrector is the number of chromosomes per sample and the number of runs.

The samples start with the double bar and next have a line with the number of segregating sites and another with their positions. Finally, there are the chromosomes, one per line:

```
//
segsites: 10
positions:  0.0001 0.0193 0.0350 0.0442 0.0609 0.0864 0.0872 0.1004 0.1016 0.1071
1010000000
0010000101
0010001101
0010000010
0010000101
1010100000
0010000101
0010000000
0010000101
0101000000
0010000101
0010000101
0010000101
0010010101
0010000000
1010100000
0010001101
0010000101
0010000000
0010000000
```

SIMULATION DATA

Samples generated using the ms program can be introduced directly into the program.

If some other program has been used to generate the samples, it has to be transformed into the ms output format before being used in this program. Some scripts which transform the outputs of some programs (such as Cosi or SelSim) into ms outputs can be dowloaded from http://www.snpator.com/public/downloads/aRamirez/.

EMPIRICAL DATA

The samples obtained from empirical data must be transformed to an ms format. Thus, each position must be coded as 0 or 1; note that "0" does not necessarily denote ancestrality:

```
ms 50 10 -s 111
111

//
segsites: 10
positions:  0.0001 0.0193 0.0350 0.0442 0.0609 0.0864 0.0872 0.1004 0.1016 0.1071
1010000000
0010000101
0010001101
0010000010
0010000101
```

If multiple population samples are to be analysed, they should be in separate files, and the program should be run once for each population file.

If the *d* sample size is not constant, each population file has to be accompanied by a second file with the same name that the one containing the sample with the extension .asc. This file must have two columns: one with the position of the SNP and the other with the size of the discovery sample for this SNP:

```
position dsample
21452    18
21662    17
22106    16
22328    3
22925    2
23393    1
24224    5
24685    17
24808    3
25062    2
25690    5
26249    1
```

CALCULATE TAJIMA’S *D* CORRECTED

CONSTANT *D* SIZE

When all the SNPs in the samples where the corrected Tajima’s *D* needs to be calculated share the same *d* size, two parameters need to be specified: if the sample is already ascertained (which will usually be the case on empirical data but not on simulation data) and the discovery sample size.

This option does not allow data to have missings neither to work through windows along the region.

ASCERTAIN SAMPLES

“Ascertain simulations” specifies if the sample needs ascertainment or if it is already ascertained. If it is set to “No”, the corrected Tajima’s *D* will be calculated from the sample as it is introduced. This option should be used in empirical data and in simulation data if the ascertainment has been previously applied to the simulations.

If the sample needs to be ascertained, “ascertain simulations” should be set to “Yes”. In this case, for each position a subsample of the size specified will be randomly selected. The SNP will only be considered to calculate Tajima’s *D* if it is polymorphic in the subsample.

D SAMPLE SIZE

The discovery sample size should be specified here.

CHANGING *D* SIZE

This option should be set if the size of the *d* sample is not uniform over the SNPs. In this case the program accepts missings, which have to be coded as ‘?’, as well as a different discovery sample size for each position, which has to be specified in a separated file (see infiles section). The formulas used to treat missings and different discovery sample sizes are explained at Ramírez-Soriano A, Nielsen R. Correcting Estimators of θ and Tajima’s *D* for ascertainment biases caused by the SNP discovery process.

The corrected Tajima’s *D* applied to non-uniform *d* sample size works through windows. The size of the windows and the step size between windows need to be specified at “windows size” and “step size” fields respectively. Both sizes have to be expressed in kilobases (kb) and must be integer number, as the program does not accept decimals.

OUTPUT FILES

The corrected Tajima’s *D* will be given as an output file with the same name as the infile but with the extension .tcr. The output files are different depending if the input is simulation or empirical data.

CONSTANT *D* SIZE

Simulation data will provide an output file as follows:

```
Watterson's theta corrected       Tajima's theta corrected   Variance W_theta corrected       Variance
T_theta corrected          Covariance corrected       Tajima's D corrected
17.000000       15.642857       70.940928       109.347628      84.654057       -0.409558
2.000000        0.542857        2.396624        2.952241        2.497456        -2.449229
13.000000       11.300000       43.936709       66.437652       51.801591       -0.653306
5.000000        4.771429        8.966245        12.336695       9.974563        -0.196445
4.000000        1.814286        6.379747        8.547731        6.984738        -2.233110
0.000000        0.000000        0.000000        0.000000        0.000000        0.000000
1.000000        1.071429        1.000000        1.145714        1.000000        0.187122
5.000000        4.571429        8.966245        12.336695       9.974563        -0.368335
7.000000        6.142857        15.329114       21.897060       17.446582       -0.561171
1.000000        1.071429        1.000000        1.145714        1.000000        0.187122
```

This file has six columns, and the corrected Tajima’s *D* value is the last. The other columns provide the corrected estimators of Watterson’s and Tajima’s theta, their variances and the covariance among them.

Moreover, if “Ascertain simulations” has been set to yes, another file will be generated with the sample obtained after ascertainment, that is, the sample over which the program will work. This file starts with the same name than the infile, but will have written at the end “_ascertained_x”, where x will correspond to the size of the discovery sample. In that case, the output file with the corrected Tajima’s *D* will be named as the file with the final sample. For example, if the infile was named Test.out and the discovery sample has been set to 5, two new files will be generated: Test_ascertained_5.out and Test_ascertained_5.tcr. They will contain, respectively, the ascertained sample and the corrected Tajima’s *D.*

CHANGING *D* SIZE

Empirical data will provide an output file as follows:

```
Window start_pos        end_possnp_num         average_n      W_theta        T_theta
       average_Theta    Var_W_theta   VarT_theta      Cov     Tajima's D
1      0       49       49      139   23.432733       17.535322      23.432733
       79.241026        163.471632    106.042844      -0.192556
2      20      49       29      140   12.906103       8.235925       12.906103
       25.571425        51.268497     33.403851       -0.465518
3      28      49       21      139   8.681747        6.103672       8.681747
       12.306826        24.077158     15.766162       -0.531380
4      35      49       14      139   5.739914        4.164504       5.739914
       6.053543         11.337225     7.525796        -0.673489
5      41      49       8       139   2.749935        2.359691       2.749935
       1.745456         3.019952      2.061302        -0.607096
6      42      49       7       141   2.284971        1.744746       2.284971
       1.270244         2.165306      1.477794        -1.125558
7      44      49       5       141   1.670762        1.352576       1.670762
       0.826446         1.330367      0.931516        -1.083072
8      44      49       5       141   1.670762        1.352576       1.670762
       0.826446         1.330367      0.931516        -1.083072
9      46      49       3       142   1.028643        0.651729       1.028643
       0.439866         0.651925      0.473517        -2.603770
```

This file has 11 columns.

Column 1: number of window.

Columns 2 and 3: absolute position of the start and end SNP. That is, 0 represents the first SNP in the sample.

Column 4: average number of SNPs per window.

Column 5: average number of valid chromosomes per window. For each position in the windows, the number of chromosomes corresponds to the number of chromosomes without missings.

Columns 6 and 7: corrected estimators of Watterson's and Tajima's thetas.

Columns 8 to 10: variances and covariance.

Column 11: corrected Tajima's *D.*

4 FABSIM README FILE (Ramirez-Soriano and Calafell. submitted)

FABSIM

ANNA RAMÍREZ-SORIANO AND FRANCESC CALAFELL

JULY 2008

F_{ST} is widely used to find genes under local selection by comparing the F_{ST} value of a single locus against genome-wide, empirical values. However, empirical distributions suffer from ascertainment bias caused by the protocol used to select SNPs, which affects the shape of the distribution. An alternative is working with simulated distributions, but this procedure also produces unreliable distributions as F_{ST} is highly dependant on the demographic history of the samples, and simulations do not take into account ascertainment bias. Provided that there is an increasing amount of information on the demographic history of populations, we have developed a software that applies ascertainment bias on simulated sequences and calculates F_{ST} on them. Moreover, we also used our program to generate several simulated F_{ST} distributions with different ascertainment biases and have compared them against the F_{ST} values found in an empirical database.

The program has been developed using Java version 6.0 and compiled under Windows, but it can also be used in Linux with graphic environment. To run FABSIM, just double-click the .jar file or, if working from the command line, write java –jar FABSIM.jar.

References: Ramirez-Soriano, A., and F. Calafell, FABSIM: A software for generating fst distributions with various ascertainment biases. Submitted.

INFILE INFORMATION

GENERAL INFORMATION

Input files are entered to the program selecting them from their location using the Browse button. All infile names must have an extension separated from the name by a dot. However, FABSIM is not strict in the extension name.

FABSIM accepts three different infile formats, the ones produced by ms (http://home.uchicago.edu/~rhudson1/;{{76 Hudson,R.R. 2002; }}), cosi (http://www.broad.mit.edu/~sfs/cosi/;{{330 Schaffner,S.F. 2005; }}), and SelSim (http://www.stats.ox.ac.uk/~spencer/SelSim/Controlfile.html) {{346 Spencer,Chris C.A. 2004; }}. Depending on the input format, FABSIM requires additional information, as detailed below.

INFILE FORMATS

ms format

The ms format is characterised by having a header and, below, the samples separated by a space and a double bar (//).

The header has the name of the program, the number of chromosomes per sample, the number of runs, the parameter for running ms and the seed used:

```
ms <chromosomes> <runs> -s/-t <various parameters>
<seed>

ms 5 20 -s 10
111
```

Of the header, the only information that will be actually used by FABSIM is the number of chromosomes per sample and the number of runs.

The samples start with the double bar. After that, two lines indicate the number of segregating sites and their positions. Finally the chromosomes are listed, one per line:

```
//
segsites: 10
positions:  0.0001 0.0193 0.0350 0.0442 0.0609 0.0864 0.0872 0.1004 0.1016 0.1071
1010000000
0010000101
0010001101
0010000010
0010000101
```

As ms provides the relative position of each segregating site in a scale from 0 to 1, FABSIM requires the simulated sequence length from the user. Absolute positions are obtained multiplying the relative position by the length of the fragment and rounding to the nearest integer.

cosi format

cosi provides two files for each simulated population, named out.hap and out.pos, which contain the haplotypes and the information for each segregating site respectively. If run as provided, cosi only simulates one sample per file. However, FABSIM is able to process files containing multiple samples separated by a blank line (this must be implemented in both files). A multiple-sample files can be obtained using the script run_cosi, which can be dowloaded from http://www.snpator.com/public/downloads/aRamirez/FABSIM/. Information on run_cosi can be found at the end of this manual.

The cosi output haplotypes file only contains the sample or samples as follows:

```
0       1       2 2 1 2 2 2 2 1 2 2
1       1       1 2 2 1 2 2 2 2 2 1
2       1       2 2 2 2 2 2 2 2 2 1
3       1       2 1 1 2 1 2 2 1 2 2
4       1       2 2 2 2 2 1 2 1 2 2
5       1       2 2 2 2 2 2 2 2 2 1
```

The first column states the chromosome number, the second the population label and afterwards come the segregating sites, each position separated by a blank space.

The file containing the information per each segregating sites contains the SNP number, the population label, the position of the site, and the frequency of each allele:

```
SNP   CHROM  CHROM_POS     ALLELE1 FREQ1  ALLELE2 FREQ2
1       1      127.3788       1      0.1154   2      0.8846
2       1      215.9448       1      0.0000   2      1.0000
3       1      229.8352       1      0.0000   2      1.0000
4       1      623.0247       1      0.4231   2      0.5769
5       1      463.2629       1      0.1538   2      0.8462
```

When using infiles in a cosi output format, the user must introduce in the program the two files provided for each population included in the analysis. The name of the haplotype files must contain a label followed by a dash, a number indicating the population code and a point:

```
TC-1.testCosi.1
TC-2.testCosi.1
```

Information files must have the same label followed by a dot, the Word “pos”, a dash and the number indicating the population code:

```
TC.pos-1.testCosi.1
TC.pos-2.testCosi.1
```

FABSIM requires the number of samples (iterations) in the file from the user. The positions are rounded for analysis to the nearest integer.

SelSim format

Output files from SelSim start with a blank line followed by a header that contains the name of the control file, the seed, and the type of output used. FABSIM only accepts “sequences” files:

```
SelSimCON.txt  -1147959592  Sequences
```

The header is followed by the samples, separated by a space. Samples start with a line with a double bar (//) followed by the number of chromosomes, the number of segregating sites and the sequence length. Next the positions of each segregating site are specified and after a blank line the samples, with each locus separated by blank spaces. After another blank line the time of the marginal genealogy underlying each position is provided, and finally the total time in all marginal trees which has not mutated, as follows:

```
//5   11   2000
1  35  132  285  299  330  463  525  781  1528  1703

1  0  0  1  0  0  0  0  1  0  0
1  1  0  0  0  0  0  0  0  0  1
1  0  0  0  0  0  0  0  0  0  1
1  0  0  1  0  0  0  0  1  0  0
1  0  0  0  0  0  1  0  1  0  0
```

```
        3.33093    3.33093    3.33093    3.33093    3.33093    3.33093    3.33093    3.33093    3.33093
3.33093  3.33093  3.33093

        6621.9
```

FABSIM requires the number of samples (iterations) in the file from the user.

CALCULATIONS

ASCERTAINMENT BIAS

Seven different ascertainment biases (or no bias) can be applied to data. Each one of them requires information to be added by the user. The appropriate fields required turn white rather than grey after the bias to be applied is selected by the user.

More than one bias can be applied to data at the same time, except if "none" is selected. Other incompatibilities are listed when applicable. If more than one bias is selected, they are applied to the sample in the order FABSIM displays them (which is also the same in which they are explained here).

Except if the contrary is said, in case that more than one population is introduced for analysis the SNPs are selected over only one population (determined by the user), but the bias is applied over all populations. That is, the SNPs deleted from the chosen population are deleted from all populations.

Discovery sample per gene

Discovery sample per gene assumes that only some chromosomes (a subsample of *d* size) of a sample of size *n* have been resequenced, and the segregating sites found on them have been genotyped on the whole sample *n*.

When this bias is activated by the user, *d* sequences are randomly selected over the total number of sequences in the sample, and only those SNPs that are polymorphic in these *d* sequences are kept.

The information required to apply this bias is the population where the *d* sample is to be selected from and the *d* sample size. The latter must be an integer between 0 and *n*.

This bias cannot be applied together with *Discovery sample per SNP*.

Discovery sample per SNP

Discovery sample per SNP works similarly to *discovery sample per gene*, but a different *d* sample is chosen for each locus.

As above, the information required to apply this bias is the population where the *d* sample is to be selected from and the *d* sample size. The latter must be an integer between 0 and *n*.

This bias is incompatible with *Discovery sample per gene*.

SNPs polymorphic in a given population

SNPs polymorphic in a given population keeps only those SNPs that are polymorphic in the selected populations.

FABSIM requires from the user the population in which to select the polymorphic loci.

This bias cannot be applied together with *SNPs polymorphic in all populations*.

SNPs polymorphic in all populations

SNPs polymorphic in all populations keeps only those SNPs which are polymorphic in all the populations entered.

No parameters are needed from the user in this bias.

This bias cannot be applied together with *SNPs polymorphic in a given population*.

MAF > threshold

The *MAF > thresold* bias discards all the SNPs that have a minor allele frequency (MAF) under the threshold provided by the user.

The information required to apply this bias is the population to ascertain and the threshold. The last must be a positive number smaller than 0.5.

Fixed SNP spacing

The *Fixed SNP spacing* bias selects one SNP every given number of bases. To do so, FABSIM selects randomly one segregating site among the *x* first base pairs. From this first selected SNP, it counts the position that is found *x* base pairs further. If in this new position there is a segregating site FABSIM selects it; otherwise, it selects the nearest one, either upstream or downstream of the new position. FABSIM proceed as explained until the new position is found outside the simulated fragment.

The information required to apply this bias is the population from which the SNPs should be ascertained and *x*, the spacing in basepairs. The latter must be a positive integer.

This bias cannot be applied together with *Variable SNP spacing*.

Variable snp spacing

In this bias, the user can specify different segments in the simulated sequence in which different SNP spacings will be applied, as described above for *Fixed SNP spacing*.

The information required to apply this bias is the population from which the SNPs should be ascertained and a file containing the different SNP densities. This file must have two columns: the first indicates the position in bp and the second the spacing for SNPs in this fragment:

200	50
300	10
500	70

In this example we consider a simulated 500-bp region. For its first 200 bp, a SNP is chosen every 50 bp. From the SNPs located between position 200 and 300, one SNP is chosen every 10 bp. From position 300 to the end, the distance between SNPs is 70 bp.

This bias cannot be applied together with *Fixed SNP spacing*.

STATISTICS

FABSIM can calculate F_{st}, minor (MAF) and derived (DAF) allele frequencies, and a number of neutrality tests on simulation data. Several statistics can be computed together in the same execution of the program.

F_{ST}

F_{st} can be calculated according to several parameters. On one hand, it can be corrected or not by the different sample sizes between populations. On the other hand, FABSIM can output the F_{st} for each SNP in the sample, per gene, or both.

The number of populations to compare is not limited. However, all of them need to have the same number of segregating sites, located in the same positions.

MAF and DAF

FABSIM calculates minor (MAF) and derived (DAF) allele frequencies for all the SNPs in all the samples of the simulations.

To calculate DAF, FABSIM assumes as ancestral the locus coded as ‘0’ of ms and SelSim and the locus coded as ‘2’ for cosi, as stated in the documentation of the programs.

Neutrality statistics

The neutrality statistics included in FABSIM are the number of segregating sites, the number of pairwise differences, the number of singletons, Tajima’s *D* {{117 Tajima,F. 1989; }}; Fu and Li’s *D*, *F*, *D** and *F** {{66 Fu,Y.X. 1993; }}; Fay and Wu’s *H* {{63 Fay,J.C. 2000; }}, R_2 {{103 Ramos-Onsins,S.E. 2002; }}, Fu’s F_s {{68 Fu,Y.X. 1997; }}, *Dh* {{97 Nei,M. 1987; }}(equation 8.4 replacing 2n by n), Wall’s *B* and *Q* {{123 Wall,J.D. 1999; }}, Kelly's Z_{nS} {{81 Kelly,J.K. 1997; }}, Rozas’ Z_A and *ZZ* {{109 Rozas,J. 2001; }} and extended haplotype homozygosity *EHH* {{111 Sabeti,P.C. 2002;352 Ramirez-Soriano,A. 2008; }}.

OUTPUT FILES

OUTFILE FORMATS

FABSIM output file do not have an uniformly formatted content, given the diversity of the results FABSIM can produce. However, two general predefined formats are provided: information per sample and tabulated data.

Information per sample

Information per sample shows the information for all samples linearly, separating them by a blank line. Each sample starts with a line stating the sample number (e.g. SAMPLE 1), and is followed by a list with the desired statistic values. Examples of this format for each statistic are shown below.

F_{ST}

The F_{st} output file in the information per sample format has three different appearances depending on if the user wants the information per snp, per sample or both. In any case the first two lines, which are shared between this and the tabulated format, show the populations that are being compared (first) and the legend. Next come the F_{st} value for each SNP or the average, maximum and minimum F_{st} of the gene, depending on what it has been required. If both per gene and per locus F_{st} are requested, per locus F_{st} is given first, as shown in the example:

```
Fst comparison between: TC-1.testCosi.1 TC-2.testCosi.1
np = not polymorfic, fixed position

SAMPLE 1
Position 1          Fst value: 0.044446
Position 2          Fst value: 0.000000
Position 3          Fst value: 0.030770
Position 4          Fst value: 0.249997
Position 5          Fst value: 0.117649
Position 6          Fst value: 0.000000
Position 7          Fst value: 0.400003
Position 8          Fst value: np
Position 9          Fst value: 0.142855
Position 10         Fst value: 0.025975
Average Fst: 0.112411     Max Fst: 0.400003          Min Fst: 0.000000
```

MAF and DAF

If MAF and DAF are computed, for each sample information on every locus is displayed sequentially in three lines corresponding to the SNP number, the MAF, and the DAF:

```
SAMPLE 1
Snp 1
          Maf: 0.1111111111
          Daf: 0.1111111111
Snp 2
          Maf: 0.2222222222
          Daf: 0.2222222222
Snp 3
          Maf: 0.1111111111
          Daf: 0.1111111111
Snp 4
          Maf: 0E-10
          Daf: 0E-10
Snp 5
          Maf: 0.4444444444
```

```
Daf: 0.5555555556
```

Neutrality statistics

In the case of neutrality statistics, for each sample the outfile starts with a line containing some statistics describing the variability of the sequences. Next the neutrality statistics appear, classified according to whether they belong to Class I (based on the mutation spectrum of frequencies) or to Class II (based on haplotypes):

```
SAMPLE 1
Sequences: 9    Seg. sites: 4    Pi: 1.388889    Singletons: 2
Class I Statistics
Tajima's D: -0.228839
Fu and Li D*: -0.264179    Fu and Li F*: -0.284088
Fu and Li D: -0.467128     Fu and Li F: -0.483865
R2: 0.157288
Fay and Wu H: 0.777778
Class II Statistics
Fu's Fs: -1.686055
EHH average: 8.000000    EHH maximum: 8.000000
Dh: 0.805556
Wall's B: 0E-8    Wall's Q: 0E-8
ZnS: 0.116805     Za: 0.075890     ZZ: -0.040914
```

Tabulated data

The tabulated data format shows as many columns as satistics plus one first colum with the sample number, separated by tabulators. The first line is a header stating what each column is. Examples of this format for each statistic are shown below.

F_{ST}

The tabulated output file for F_{st}, as in the previous format, has three different appearances depending on the calculation chosen and contains the two lines showing the populations that are being compared (first) and the legend. If both SNP and gene information are displayed, the outfile will look as follows, with the average, maximum and minimum F_{st} added in three columns next to the last SNP in the sample:

```
Fst comparison between: TC-1.testCosi.1 TC-2.testCosi.1
np = not polymorfic, fixed position

sample  snp     fst         average_fst     max_fst min_fst
1       1       0.044446
1       2       0.000000
1       3       0.030770
1       4       0.249997
1       5       0.117649
1       6       0.000000
1       7       0.400003
1       8       np
1       9       0.142855
1       10      0.025975    0.112411        0.400003        0.000000
```

If information on SNPs is required exclusively only the first three columns are shown. Instead, if the information asked is F_{st} per gene, the “snp” and “fst” columns are not displayed.

MAF and DAF

MAF and DAF output tabulated format has four columns which, from left to right, correspond to the sample and locus number, MAF and DAF.

```
Sample  SNP     maf             daf
1       1       0.3333333333    0.3333333333
1       2       0.3333333333    0.3333333333
1       3       0.4444444444    0.4444444444
1       4       0.3333333333    0.3333333333
1       5       0.3333333333    0.6666666667
```

Neutrality statistics

The neutrality statistics outfile has 21 columns, the first for the sample and the next for the descriptors and the statistics, in the same order as in the information per sample format:

```
Sample  seq       segsites pi        singl   Ts_D      FL_D2   FL_F2    FL_D      FL_F      R2       FW_H
        Fs        EHH_a    EHH_m     Dh      W_B       W_Q     ZnS      Za        ZZ
1       9         4        1. 388889         2         -0. 228839       -0.264179           -0.284088
        -0.467128          -0. 483865        0. 157288         0. 777778          -1.686055
        8.000000           8.000000          0. 805556         0E-8     0E-8      0.116805
        0.075890           -0.040914
```

OUTFILE NAMES

With the exception of F_{st}, the name of the output is formed based on the name of the infile. The outfile name, then, is the infile without extension followed by a dash and an abbreviation for the calculation done. Its extension depends on the predefined outfile format selected.

In the case of F_{st}, the outfile name without extension must be provided by the user. FABSIM will use this name to code it as explained.

The codes for the calculations and the formats are:

CALCULATIONS	
_fst	F_{st}
_maf	MAF and DAF
_stats	neutrality statistics
FORMATS	
.smp	information per sample
.tab	tabulated data

Examples

The output file obtained from calculating neutrality statistics on a file named simulations.inp, specifying the information per sample format, would be named simulations_stats.smp.

If the user wants to calculate F_{st} and MAF and DAF on a set of simulations from two populations which are in the files population1.out and population2.out, and obtain the results in a tabulated format, it first must provide a name for the F_{st} output file. Let's say the given name is *populations*. FABSIM will then output three outfiles:

populations_fst.tab
population1_maf.tab
population2_maf.tab