UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1280

Population Genetic Methods and Applications to Human Genomes

LUCIE GATTEPAILLE

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2015

ISSN 1651-6214
ISBN 978-91-554-9319-6
urn:nbn:se:uu:diva-260998

Dissertation presented at Uppsala University to be publicly examined in Lindahlsalen, Norbyvägen 18A, Uppsala, Thursday, 22 October 2015 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Associate Professor Kevin Thornton (Department of Ecology and Evolutionary Biology, University of California, Irvine).

Abstract
Gattepaille, L. 2015. Population Genetic Methods and Applications to Human Genomes. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1280. 63 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9319-6.

Population Genetics has led to countless numbers of fruitful studies of evolution, due to its abilities for prediction and description of the most important evolutionary processes such as mutation, genetic drift and selection. The field is still growing today, with new methods and models being developed to answer questions of evolutionary relevance and to lift the veil on the past of all life forms. In this thesis, I present a modest contribution to the growth of population genetics. I investigate different questions related to the dynamics of populations, with particular focus on studying human evolution. I derive an upper bound and a lower bound for F_{ST}, a classical measure of population differentiation, as functions of the homozygosity in each of the two studied populations, and apply the result to discuss observed differentiation levels between human populations. I introduce a new criterion, the *Gain of Informativeness for Assignment*, to help us decide whether two genetic markers should be combined into a haplotype marker and improve the assignment of individuals to a panel of reference populations. Applying the method on SNP data for French, German and Swiss individuals, I show how haplotypes can lead to better assignment results when they are supervised by GIA. I also derive the population size over time as a function of the densities of cumulative coalescent times, show the robustness of this result to the number of loci as well as the sample size, and together with a simple algorithm of gene-genealogy inference, apply the method on low recombining regions of the human genome for four worldwide populations. I recover previously observed population size shapes, as well as uncover an early divergence of the Yoruba population from the non-African populations, suggesting ancient population structure on the African continent prior to the Out-of-Africa event. Finally, I present a case study of human adaptation to an arsenic-rich environment.

Keywords: Population genetics, Human evolution, Genetic diversity, Genetic differentiation, Adaptation, Population structure, Effective population size

Lucie Gattepaille, Department of Ecology and Genetics, Evolutionary Biology, Norbyvägen 18D, Uppsala University, SE-75236 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-9319-6
urn:nbn:se:uu:diva-260998 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-260998)

"I know one thing: that I know nothing."
Socrates

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Gattepaille, L. M.**, Jakobsson, M., Rosenberg, N. A. (–) Homozygosity constraints on the range of Nei's F_{ST}. *Manuscript*

II **Gattepaille, L. M.**, Jakobsson, M. (2012) Combining markers into haplotypes can improve population structure inference. *Genetics*, 190(1):159-174

III **Gattepaille, L. M.**, Jakobsson, M. (–) Popsicle: a method for inferring past effective population size from distributions of coalescent times. *Manuscript*

IV Schlebusch, C. M.*, **Gattepaille, L. M.***, Engström, K., Vahter, M., Jakobsson, M., Broberg, K. (2015) Human Adaptation to Arsenic-Rich Environments. *Molecular biology and evolution*, 32(6):1544-1555.

*These authors contributed equally to the study.

I am also co-author in the following articles that were published during my graduate studies.

Schlebusch, C. M.*, Skoglund, P.*, Sjödin, P., **Gattepaille, L. M.**, Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A, Blum, M. G. B., Soodyall, H., Jakobsson, M. (2012) Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science*, 338(6105):374-379

Gattepaille, L. M., Jakobsson, M., Blum, M. G. B. (2013) Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity*, 110(5):409-419

Shafer, A., **Gattepaille, L. M.**, Stewart, R. E. A., Wolf, J. B. W. (2015) Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: in silico evaluation of power, biases and proof of concept in Atlantic walrus. *Molecular ecology* 24(2):328-345

Duforet-Frebourg, N.*, **Gattepaille, L. M.***, Blum, M. G. B., Jakobsson M. (2015) HaploPOP: a software that improves population assignment by combining markers into haplotypes. *BMC Bioinformatics*, 16(1):242.

*These authors contributed equally to the study.

Contents

1. Introduction

Evolutionary biology is a scientific discipline aimed at studying species in the light of the ancestral relationships existing among them, at characterizing the forces leading to the divergence of subspecies into clear distinct species and at understanding life that we observe in all its diversity. Within the discipline, the field of population genetics provides mathematical tools to study how genetic variation evolves over time within a population and to quantify the effects of different evolutionary forces on the frequency of mutations within species.

Population genetics was born in the 1920s as a result of the successful attempt to reconcile the apparently separate two schools of thought regarding inheritance (Provine, 2001). On one side, the biometricians, focusing on continuous traits and relying heavily on statistical modelling, were viewing inheritance as the mixing of the parental traits into the offspring. On the other side, the Mendelians, influenced by the rediscovery of Mendel's work, considered the transmission from parent to offspring to be done via discrete characters, segregating with equal probability. While both sides could appreciate the arguments in support of each theory, they could not overcome the respective counter-arguments. Indeed, if characters are transmitted in a discrete fashion, with a chance of a half from parent to offspring, why do we observe apparently continuous traits such as height, and offspring being taller or smaller than their parents? If offspring are merely a blend of their parents, how would one explain the existence of discrete qualitative traits, such as the color of peas in Mendel's famous experiment?

By demonstrating how multiple genes of small quantitative effects could segregate according to Mendel's laws of inheritance but still create seemingly continuous traits (Fisher, 1919), Fisher added the first and strongest nail in the coffin of the long-standing debate on the means of evolution and heredity, leading later to the Modern Evolutionary Synthesis (Olby, 1989). The *gene* could finally be accepted as the unit of parent-offspring transmission, and phenotypes explained by the effects of one or multiple genes. It took, however, several decades after Fisher's work before the actual biological makeup of genes was found and before genetic variation could be investigated at the molecular level. The first assessment of genetic variation at a large number of loci (Lewontin and Hubby, 1966) revealed much more genetic variation within populations than previously anticipated. This surprising result challenged the view that natural selection was the main driving force of evolution. Under

such a view, high genetic homogeneity among individuals of the same species is expected, as most mutations that eventually prevail in a species would be adaptive while the other mutations would be purged out. In a seminal paper (Kimura et al., 1968), Kimura showed that the large amount of genetic variation found within populations could only be explained by the abundance of neutral or nearly neutral mutations and set later the mathematical foundation of the *neutral theory of evolution*. Evolution was then re-defined in the light of the mutation process creating variation and the fate of different alleles in populations, subject to both selective and neutral processes. Population genetics thus arrived at the center of evolutionary biology by providing means to study the evolution of allele frequencies over time with mathematical models. Thanks to its abilities for prediction and description of the most important evolutionary processes such as mutation, genetic drift and selection, population genetics has led to countless numbers of fruitful studies of evolution. The field is still growing today, with new methods and models being developed to answer questions of evolutionary relevance and to lift the veil on the past of all life forms. In this thesis, I present a modest contribution to the growth of population genetics, with a particular focus on human evolution.

1.1 Brief introduction to coalescent theory

The theory of the coalescent provides a simple mathematical description for the ancestral relationship between gene-copies sampled from a homogenous population. Derived by Kingman in 1982 (Kingman, 1982), it provides a useful alternative to the more complicated models at the time, such as diffusion theory where allele frequencies are modelled according to a brownian motion and followed forward in time in a large population (Watterson et al., 1962). The coalescent is a structure that follows ancestral relationships backward in time, hence removing the need for following all lineages, which is the downside of all forward in time approaches. Since its derivation and the subsequent large body of work fostering its development (*e.g.* Tavaré, 1984; Kaplan et al., 1988; Hudson and Kaplan, 1988; Griffiths and Tavare, 1994, among others), it has been widely used in population genetic studies. I give here a brief overview describing what a coalescent is, how it arises from the study of a sample and why it is an interesting structure for studying genetic data. Beforehand however, we need to introduce the Wright-Fisher model, a simple model describing a population generation after generation. The coalescent emerges naturally from the Wright-Fisher model as the population size grows large.

1.1.1 Wright-Fisher Model

The Wright-Fisher model is a model for describing the transmission of haploid gene-copies from a pool of parental gametes to the next generation. It assumes non-overlapping generations of constant size. Every new generation is formed by randomly sampling gene-copies in the parental generation (figure 1.1). It is, classically, the most used model of reproduction and leads to simple equations for describing ancestral relationships. In particular, if we define N as the population size, the probability of two particular gene-copies coming from the same parental gamete in the previous generation is $1/N$. More generally, the probability for two particular gene-copies to share their first common ancestor at exactly k generations back in time is $(1-1/N)^{k-1} \times 1/N$. We recognize the geometric probability distribution, with success probability $1/N$. For a sample of n gene-copies, the probability that *none* of the gene-copies come from a common parent in the previous generation is the probability α:

$$\alpha = (1-\frac{1}{N})(1-\frac{2}{N})\ldots(1-\frac{n-1}{N}), \tag{1.1}$$

so the probability that no common ancestor is shared within a sample of n gene-copies for exactly $k-1$ generations in the past and that at least a pair of gene-copies from the sample have a common ancestor at generation k is $\alpha^{k-1} \times (1-\alpha)$. This, once again, describes a geometric process, this time with a probability of success $1-\alpha$ (at least one common ancestor is found). Note that because the size of the population is finite and offspring are chosen at random from the parental generation, some parental gametes do not contribute to the next generation, while others contribute multiple times. The loss of some parental gene-copies at every generation due to random sampling is called *genetic drift*. Consequently, after a certain number of generations, all gene-copies in the population are descendants of a single ancestral gene-copy.

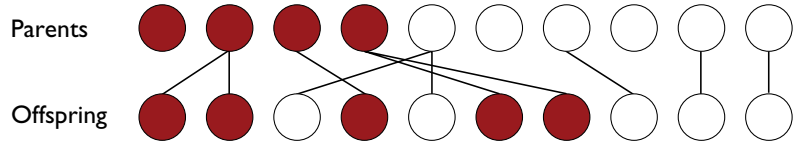

Figure 1.1. **Example of parent and offspring generations**. Offspring are produced by randomly sampling gene-copies from the parental generation, with the parent chosen indicated by a linking solid line. Three of the parent gene-copies do not contribute to the next generation. By chance, the number of red allele-copies is increased by one in the offspring generation.

Thanks to the simplicity of the reproduction process in the Wright-Fisher model, we can easily simulate a population over time and observe the fate of different alleles in the population. For example, if there are white and red alleles segregating in a population of N haploid gene-copies, as in figure 1.1, the

probability of observing a given number of white and red alleles depends only on the number of white and red alleles in the preceding generation. To be more precise, if there are k red and $N-k$ white alleles in the parental generation, the probability of observing exactly j red alleles in the offspring generation is $\binom{N}{j}\left(\frac{k}{N}\right)^{j}\left(\frac{N-k}{N}\right)^{N-j}$, which represents a binomial distribution. By chance, the number of red and white alleles varies from one generation to the next. Genetic drift thus represents a fundamental concept in population genetics, as it can greatly affect the frequency of alleles over time. Another important consequence of genetic drift is that, in the absence of additional mutations at the locus, one allele will eventually make up all gene-copies while the other allele is permanently removed from the population. This event is called *fixation* and when it occurs, the remaining allele is characterized as *fixed*. Different populations from the same species, if they are sufficiently isolated from one another, might accumulate fixed differences over time. Such differences can lead to reproductive isolation and eventually evolution of the two populations into separate species.

The Wright-Fisher model can be used beyond the realm of haploid individuals into polyploids and it can accommodate for sex as well. When generalized to diploid and sexually reproducing individuals, the random sampling step can be achieved by what is called *random mating*: parents make pairs at random and each parent contributes to one gamete selected randomly from their two gametes to form the offspring. This treatment is similar to considering the pool of gametes the individuals are harboring instead of the individuals themselves and using the standard haploid Wright-Fisher as described above on the pool of gametes.

1.1.2 From Wright-Fisher to the coalescent

In population genetic studies, we rarely (if ever) have access to genetic data from the entire population. Instead, we sample a number of individuals and by studying their genetic data, we hope to understand the processes that have shaped the entire population. It is therefore useful to mathematically model the information that can be extracted from a sample. Under the Wright-Fisher model, we computed the probability that the gene-copies from a sample of n gene-copies would have exactly n parent gene-copies in the previous generation (α in equation 1.1). If we consider N being large and n being small relative to N, then α is approximately equal to $1-\binom{n}{2}/N$. This approximation implies that we neglect the probability that more than one pair of gene-copies can share a common parental gene-copy in the previous generation, as such an event occurs with a probability in the order of $1/N^2$. In addition, the probability β that a common ancestor is found for the first time at generation k back in

time becomes

$$\beta = (1 - \binom{n}{2}/N)^{k-1} \times \binom{n}{2}/N. \tag{1.2}$$

When the ancestral lineages of two gene-copies meet in a common ancestor, we say that they *coalesce* and such an event is called a *coalescence* (figure 1.2). Because N is considered large, the probability of the first coalescence in the sample is approximately exponentially distributed with mean $\binom{n}{2}/N$. So if the time is rescaled in units of N generations, the probability of coalescence becomes independent of the population size and only dependent on the sample size, according to an exponential distribution with mean $\binom{n}{2}$.

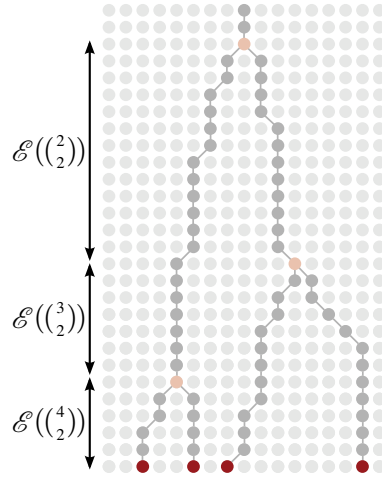

Figure 1.2. **Realization of a coalescent, for a sample size of 4**. Individuals from the sample are shown as dark red dots. Light grey dots represent individuals from the population that are not ancestors to the sample. Dark grey dots represent individuals that are ancestors to the sample and in beige we highlight ancestors where lineages coalesce. The waiting time for the first (second and third respectively) coalescence to occur follows an exponential distribution with mean $\binom{4}{2}$ ($\binom{3}{2}$ and $\binom{2}{2}$ resp.) when time is rescaled in units of N generations.

It might seem anecdotal at first but the consequence of the independence of the process from population size is important: all samples taken from populations following the assumptions of the Wright-Fisher model have the same underlying mathematical structure to describe their ancestral relationships, regardless of the population size (provided that the size is large). Thus, there is only one process to study: the coalescent. By rescaling the time appropriately, it can be used to study populations that can be quite different in size. The second advantage of the coalescent is its simplicity: waiting times to coalescence are modelled by exponential variables that only depend on the number

of lineages considered. The third advantage of the coalescent is its inherent property of following the lineages backward in time. This allows for fast simulation of samples because we do not have to keep track of the transmission patterns of the entire population, only the lineages leading to the sample at present are simulated.

1.1.3 Demography and effective population size

If the coalescent would be limited to populations following the assumptions of the Wright-Fisher model exactly, its use would also be rather limited, as such populations are likely to be rare. Populations usually experience a number of violations of those assumptions: generations can be overlapping, the size might not be constant over generations, and individuals might not reproduce at random, instead within smaller groups of proximity. Those demographic factors complicate the modelling of the population's evolution. However, in many cases, the coalescent still emerges from the model, when rescaling by an appropriate factor that integrates the violation of the Wright-Fisher model's assumptions and re-establishes the exponential distributions of the waiting times to coalescence. This factor is called *effective population size*.

The term *effective population size* has multiple definitions in population genetics. A population not following the ideal assumptions of the Wright-Fisher model will not behave according to the expectations of the model. We can often recover some expectations of the model by considering that the population has a different size (Sjödin et al., 2005). If the property that we want to fit the model's expectations with is the change in probability of identity by descent, we talk about inbreeding population size (Crow and Kimura, 1970). If the property is the change in variance of allele frequencies, we talk about variance effective size (Hartl and Clark, 1997). If the property is the waiting time for coalescence between individuals, we talk about coalescent effective size (Nordborg, 2001), as we do in the previous paragraph. If the property is the rate of loss of heterozygosity, we talk about eigenvalue effective size, in relation to the leading non-unit eigenvalue of the transition matrix in allele frequencies which equates to the loss of heterozygosity in one generation (Ewens, 1982). Note that we can rarely recover expectations of the Wright-Fisher model for *all* properties at once, so we have to be careful as to what type of effective size we are considering, *i.e.* which property has been chosen to fit the model's expectation. From this point on, effective population size will refer specifically to the coalescent effective size.

1.2 Genetic variation

Most of what makes us *us* is encoded in our DNA. We are different from our neighbours, our siblings, our parents, because our DNA is different from theirs. Those differences arise thanks to mutations. When DNA is replicated during meiosis (when reproductive cells are made), the copying process is not without error. Despite repair mechanisms, some errors make it into the gametes we transmit to our offspring, creating genetic variation. Over generations, mutations are passed on, lost because of genetic drift, get fixed by chance or with help of selective pressures, new mutations are constantly being produced, and pairing of particular alleles are being shuffled via recombination. All of these processes create a landscape of genetic diversity that is the testimony of the population's evolution. One aim of population genetics is to harness the information carried by the genetic data and unveil the demographic and/or selective processes that have given rise to the patterns observed.

1.2.1 Types of genetic data

Changes in DNA can occur at multiple levels, from large changes (e.g. genome duplications, copies or inversions of large portions of DNA), to small changes of a single base pair. Small changes are more common as they are less likely to cause serious damage to the offspring. Among all existing mutations, we only review here Single Nucleotide Polymorphisms (SNPs). A SNP mutation occurs when a single nucleotide is replaced by another. Because of the complementary structure of DNA, each strand can be used as a template for replication by pairing each nucleotide with its complement (A and T are complementary, as are C and G). Sometimes however, the DNA-polymerase performing the pairing makes an error and the template nucleotide does not get paired with its complement. This error is later detected by other enzymes which repair the mistake, either by replacing the incorrect complement nucleotide by the right complement nucleotide and thus restoring the original state (no mutation occurs then), or by replacing the template nucleotide (thereby creating a mutation). When comparing the DNA sequences of multiple individuals, we can observe these differences and use them to answer various questions, from finding potentially harmful variants, to reconstructing the demographic history of the population the individuals are from.

Thanks to the International HapMap Project, an audacious joint effort from the early 2000s between 6 countries (United Kingdom, Canada, Japan, China, Nigeria and United States) to map the genetic variants of humans using individuals sampled from Nigeria, China, Japan and U.S.A., many SNPs have been discovered (Gibbs et al., 2003). Those SNPs have since been used to develop SNP arrays: DNA microarrays that are designed to target the particular positions of the genome where SNPs have been observed. SNP arrays have

grown large over the years, from 1494 targeted positions in 1998 (LaFramboise, 2009; Wang et al., 1998), to more than 5 million positions on today's human SNP arrays. Over the years, the cost per SNP on arrays has decreased greatly, facilitating access to the technology for numerous research groups worldwide. These advances have resulted in a large body of fruitful genetic studies, and will continue to make this type of genetic data accessible to many. Nevertheless, SNP arrays are not without disadvantages. The main source of complications in genetic studies based on SNP arrays is *ascertainment bias* (Clark et al., 2005). SNPs are discovered on a sample of individuals, so if those individuals are not taken at random in the entire population or species, the SNP array will not give a representative picture of the genetic diversity in non-sampled groups. In particular, variants that are private to the non-sampled groups are going to be missed entirely and this may bias genetic studies of these groups. In humans, most SNP arrays contains many SNPs that have been ascertained in samples of European, Asian or West African ancestry.

In recent years, the cost of genome sequencing has gone down tremendously, from about 100 million dollars per human genome in 2001 to around 5,000 dollars in 2014 (Wetterstrand, 2014). Unlike SNP arrays, sequences do not suffer from ascertainment bias, as they capture all genetic variation present in the sampled individuals. Sequence data represent the ultimate form of genetic data, as it encompasses all DNA variation. However, there is still a rather high rate of sequencing error and some genomic regions are still very difficult to sequence (highly repetitive regions for example).

Haplotypes are another type of genetic data used in genetic studies. A haplotype is a rather generic term, but in all cases it represents a combination of alleles physically linked to each other on the DNA strand. When the zygote is formed at conception, the nucleus of the egg and the nucleus of the sperm fuse together into a diploid nucleus. The new genome is then composed of pairs of chromosomes: one set of chromosomes from the mother and one homologous set from the father. This implies that the alleles coming from a given parental chromosome are physically sitting on the same DNA strand. When genotyping using SNP arrays or when sequencing with the usual techniques, we only know whether the individual is homozygous or heterozygous at a given position but not how two heterozygous positions are physically linked to each other. Separating the maternal alleles from the paternal alleles is called *phasing*. For many studies of genetic data it is necessary to know the *phase* of the individual. Algorithms have been developed to perform statistical phasing (their principles are described in the method section) and recently, efforts have been made to obtain the phase molecularly, by sequencing longer stretches of DNA so as to capture the pairing of heterozygous positions (see *e.g.* Kitzman et al., 2011). Haplotype data can represent the combination of alleles in sequences of a given physical or genetic length, or a given number of variant

positions. The variant positions can be SNPs on a SNP array, or can be variants in sequences. What matters is that the alleles are phased, so that alleles are paired according to the gametes they are coming from, paternal or maternal.

1.2.2 Recombination and linkage disequilibrium

During meiosis, homologous chromosomes exchange genetic segments due to chromosomal crossovers and, therefore, the four gametes produced at the end of the meiosis of one reproductive cell contain haploid genomes that are mosaics of the paternal and maternal haploid genomes of the individual. This process of exchange of genetic material between homologous chromosomes is called *recombination*. Recombination breaks the association between alleles from the same parent and allows the creation of new combinations. The further two genes are on a chromosome, the higher the probability that a recombination event will occur between them in the formation of gametes. By looking at the transmission of allele combinations in studies of trios or in pedigrees, we can build a map indicating how likely recombination is to occur in a certain region at each meiosis. Such a map is called *genetic map* or *recombination map*.

Instead of investigating the transmission patterns in trios and pedigrees, we can also look at the statistical association between alleles at different positions in unrelated individuals sampled from a population. The dependent association of alleles at different positions in the genome is referred to as *linkage disequilibrium*. Comparably to genetic maps, linkage disequilibrium can give an idea of how strong the recombination probability is at a given position. However, it also depends on the local genetic ancestry of the samples at the particular position. Thus, the relationship to the probability of recombination is not a simple one. When looking at two physically close regions in the genome, we often observe a non-random pairing of alleles between those regions because they tend to be transmitted together without recombination. The further apart two genes are, the less likely it is for their alleles to be correlated because recombination breaks the association. There are different ways to measure linkage disequilibrium, but one popular statistic is r^2 (r-squared). It measures the statistical squared correlation between the alleles at the two positions. We generally observe a decay in r^2 values with physical distance. However, the strength of the decay varies among species and populations. It is affected both by the local recombination probabilities and by the demographic history of the sampled population, making it an interesting statistic to study recombination rate and demography.

Sometimes correlation between alleles can cause problems for certain analyses, as this violates the assumption of independence between sites that some population genetic methods require. In such cases, the problem can be addressed by extracting a subset of sites for which the correlation of alleles from one site to the next is below a chosen threshold. This procedure is called *pruning*. It usually results in a significant reduction in number of variable sites (see *e.g.* Novembre et al., 2008), but the remaining sites can be treated as independent and methods requiring independence can be applied to them. In contrast, some population genetic methods actually benefit from linkage disequilibrium. In association studies for instance, where genome-wide genetic data is scanned for association to a given phenotype, often a disease phenotype, correlation between alleles at neighbouring sites can lead to a local genetic signal of association, even when the causal variant is not genotyped in the sample, as can be the case when using a SNP array. Numerous genome-wide association studies have been performed using thousands of individuals thanks to the relatively low cost of SNP arrays and multiple genetic associations with diseases have been found (see *e.g.* Harold et al., 2009; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium and others, 2011).

1.2.3 Genetic variation in Humans

In humans, the genome-wide rate of single nucleotide mutations has been estimated to around 2.5×10^{-8} per base pair per generation when calibrated by the divergence time from chimpanzee, and to around 1.2×10^{-8} per base pair per generation when studying trios and pedigrees. On average, a human is heterozygous at a site approximately every 1000 base pairs (Prado-Martinez et al., 2013) and differs from a chimpanzee at around 1.2% of all sites (The Chimpanzee Sequencing and Analysis Consortium, 2005). It has often been said that 85% of the genetic variation in humans is accounted for by differences between individuals and the remaining 15% by differences among populations (Lewontin, 1972). While this is correct for the variable sites taken separately, there is information about population membership and shared ancestry among groups in the correlation between sites (Edwards, 2003) so that defining broad human groups has meaning for population genetic studies. In fact, surveys of genetic variation in individuals sampled worldwide have revealed a clear genetic structure among human populations, at different geographic scales (see *e.g.* Rosenberg et al., 2002; Jakobsson et al., 2008; Wang et al., 2007; Schlebusch et al., 2012). A study of mitochondrial DNA haplotypes in a worldwide sample of individuals revealed that mitochondrial genetic variation outside of Africa is a subset of the variation within Africa (Cann et al., 1987), suggesting an African origin of today's human populations. Support for the Out-of-Africa model of human demographic history has been further strengthened by the study of the *Y* chromosome (Hammer et al., 2001) and au-

tosomes (Goldstein et al., 1995), with patterns of genetic diversity that are consistent with serial founder events (DeGiorgio et al., 2009). Both genome-wide homozygosity and linkage disequilibrium increase with the distance from Africa (Jakobsson et al., 2008), consistent with an African origin of all modern human populations today. However, the origin of anatomically modern humans within Africa is still under debate, and some have argued that there may be no single geographic origin to begin with (Schlebusch et al., 2012).

1.3 Mapping genotype and phenotype

We are all the product of our genes, of our environment and the interaction between them. Disentangling the different factors contributing to observed phenotypic diversity is one of the main goals of genetic studies and has major applications notably in the field of human health (Visscher et al., 2008). Twin studies and other pedigree-based studies can give estimates of the heritability of a given phenotype, namely the proportion of phenotypic variance in a population that can be attributed to additive genetic effects (Lynch et al., 1998). Some traits, such as height, are found to be highly heritable, thus influenced greatly by genetic factors (Macgregor et al., 2006; Yang et al., 2010), while other traits are mainly influenced by the environment (Price and Schluter, 1991), leaving genes only a small role to play in their variance. Once it has been established that genetic factors contribute significantly to the phenotype of interest (often a disease status or a health-related quantitative trait, such as blood pressure or lipid levels), it is of great interest to identify the particular genes that influence the phenotype. A number of study designs have been developed to address this problem, such as candidate-gene association studies, linkage mapping, admixture mapping and Genome-Wide Association Studies (GWAS) (Hirschhorn and Daly, 2005). I briefly provide in the methods section an overview of the main concepts behind GWAS, as it was the approach chosen for identifying genetic contributors to arsenic metabolism in the study presented in paper IV of this thesis.

1.4 Selection

The immense diversity of life on Earth is undeniable. If one takes notice of this fact, one may wonder about the forces that have shaped the multitude of species observed today and in fossil records, and why do we see so many differences among species and yet so many similarities as well. In his seminal book *On the Origin of Species* (Darwin, 1859), Charles Darwin provided an answer: all life forms are related to one another via ancestral species (explain-

ing the similarities) and every species has taken its own evolutionary path under the constraints of natural selection (explaining the differences). This revolutionary idea, very controversial at the time of publication and sadly still challenged despite the large body of evidence supporting it, had profound implications in the study of life. Due to mutations, individuals carry different genetic make-up, which in turn leads to a variety of different phenotypes. If the phenotype of an individual grants it a reproductive advantage, the underlying genetic factor causing the phenotype is passed onto the next generation with higher probability, thus increasing the frequency of the advantageous phenotype in time. The process by which the succession of generations in a population is made up of individuals that are increasingly fit to live in their environment is called *adaptation*. Adaptation is one of the driving forces of evolution and species diversification.

When a given genetic variant grants a reproductive advantage to its carriers, the variant is said to be under *positive selection*. In humans, a handful of examples of positive selection have been identified. Some are examples of adaptation to the environment (Yi et al., 2010; Ruff, 1994; Perry and Dominy, 2009; Norton et al., 2007; Hamblin and Di Rienzo, 2000, see also paper IV), some are driven by dietary practices (Enattah et al., 2008), and some even by cultural practices (Asante et al., 2015). Nevertheless, it is still unclear how much positive selection has participated in shaping the evolution of mankind. Genetic drift is likely to play an important role as well, especially in small populations. Disentangling evolution due to genetic drift from adaptive evolution is a great challenge in evolutionary biology (Stajich and Hahn, 2005). The field is marked by a long-lasting debate between the neutralists who believe that species and populations mostly evolve due to genetic drift randomly bringing alleles to fixation or eliminating them (Kimura et al., 1968), and selectionists who believe that, on the contrary, evolution occurs mostly under selective constraints (Gillespie, 2010). Everyone nowadays agrees that both processes of genetic drift and natural selection are acting, but the degree to which they participate in the evolution of species is still hotly debated. Currently however, most methods aimed at detecting or measuring the amount of selection assume that most of the variation in the genome is neutral or nearly so, and that sites under positive selection are the exception. Hence, by looking at properties of variants, the genome-wide distributions of those properties represent the neutral expectation and outlier regions might be the result of selective processes (Nielsen, 2005). In the methods section, I provide examples of genome-scans to detect regions under positive selection.

New mutations can also be disadvantageous. The life of an organism is built on a complex and delicate mechanism and there are many places where it can fail when altered. Mutations in coding regions for example might alter a protein's conformation and hence, disrupt its function. Some mutations are lethal

and are immediately purged from the population. Some mutations are deleterious and lead to a survival or reproductive disadvantage for the individual. The frequency of such an allele is thus likely to decrease over time because of *purifying selection*, potentially also eliminating other variants that are physically linked to it. Purifying selection is likely to play an important role in creating the patterns of genetic diversity we observe, especially in genetic regions of central importance, but I do not address its effects in this thesis.

1.5 Population structure

As I mentioned previously, individuals in populations rarely reproduce at random. Instead, the pairing of individuals can depend on various factors, such as geographical proximity, sexual selection or cultural practices. The departure from random mating is referred to as *population structure*. There can be different reasons as to why random mating is hindered. Geography is an important contributor to population structure in humans for example, as people tend to pair with individuals that live in their vicinity. In time, this creates a particular pattern of genetic diversity, where the genetic similarity of individuals is correlated with the geographical distance that separates them. This model, called *isolation by distance*, seems to be holding well in Europe for example, where genetic data has been shown to mirror geography surprisingly well (Novembre et al., 2008). Sexual selection is another factor that can cause population structure, when individuals choose their mates according to the amount of similarity (assortative mating) or dissimilarity (disassortative mating) they have with them. In humans, the choice of a mate can be influenced by cultural factors, such as education level, religious views (Hur, 2003) or cooperativeness (Tognetti et al., 2014). Preferences for mates with different HLA alleles have also been shown (Wedekind et al., 1995).

Population structure complicates genetic analyses as it violates the important assumption of random mating of most population genetic models. In association studies for instance, if the cases are more related to each other due to cryptic population structure, variants that correlate with the disease status are more likely to be the result of shared ancestry than to be causing the phenotype, creating large amounts of false positive associations (Cardon and Palmer, 2003). To account for population structure, some corrections can be applied (*e.g.* Price et al., 2006). It is however difficult to characterize the extent of population structure in a population as, in general, many factors influence the choice of a mate. Sometimes, the difficulties caused by population structure in genetic analyses can be alleviated by applying some genomic control to the data, in the example of genome-wide association studies (Clayton et al., 2005). Population structure can also sometimes lead to valuable information.

In the case of population structure based on geography, the correlation of genetic data to spatial coordinates can help shed light on the movement of people or animals in time. It can also help in determining the geographical origin of a DNA sample of unknown origin, a useful piece of information in forensic science for instance.

2. Methods

Now that I have presented some key concepts of population genetics, I will present in the following section the key methods I have used to investigate questions of population genetic relevance.

2.1 Measuring genetic diversity and differentiation

Mutation and recombination processes create variation among individuals or among populations, but there are various ways to quantify the actual amount. We review here three statistics of interest that are used in the studies presented in this thesis: heterozygosity, r^2 and F_{ST}, which are measures of genetic diversity, linkage disequilibrium and differentiation respectively.

2.1.1 Heterozygosity

When looking at a single position in the genome where a variant is known to exist in a population, individuals can either be *homozygous* (carrying the same allele at both the paternally and maternally inherited chromosomes) or *heterozygous* (paternal and maternal alleles are different). *Observed heterozygosity* is defined as the proportion of heterozygous individuals in the population. Since genetic information can almost never be collected for all individuals, heterozygosity has to be estimated using a sample from the population. Under the assumption of random mating, with no other evolutionary forces at play, heterozygosity can also be computed using allele frequencies. More precisely, for a bi-allelic locus with allele frequencies p and $q = 1 - p$, the expected proportion of heterozygous individuals in a randomly mating population is $2pq$, which represents the probability of randomly sampling two different alleles from the population. When the observed heterozygosity is equal to the expected heterozygosity, the population is said to be under *Hardy-Weinberg Equilibrium*, named after Godfrey Harold Hardy and Wilhelm Weinberg who independently worked out the frequencies for each genotype under the equilibrium. Evolutionary processes such as mutation, selection, drift and population structure can lead to a discrepancy between the expected and observed genotype frequencies. Tests for deviation from the Hardy-Weinberg equilibrium have been built and they are often used in the context of revealing the presence of population structure or detecting selection.

Considering now an entire sequence of DNA, if the mutation rate per site per generation is small enough, most sites are monomorphic and polymorphic sites are likely to segregate for only two alleles. The average heterozygosity H over the sites of the sequence is expected to be:

$$H = 4N_e\mu, \tag{2.1}$$

with N_e the diploid effective population size and μ the mutation rate per site per generation. Thus, if the mutation rate is known, the effective population size can be computed from estimates of heterozygosity. In human populations, estimates of effective population size based on heterozygosity are in the order of 10,000 (Yu et al., 2004), with African population sizes larger than non-African sizes due to the founder effects of the Out-Of-Africa event and subsequent colonization of the entire world. This type of computation provides a single estimate for the effective population size, which then represents an average of the effective population size over the entire history of the population. Recently, methods have been developed to harness information contained in the rates of coalescence within samples and provide estimates of the effective population size over time (see for example paper III and Li and Durbin, 2011; Sheehan et al., 2013), which gives a finer insight into the population's history than the simplistic N_e estimate from heterozygosity.

2.1.2 r^2

As a measure of the non-random association of alleles at different sites, r^2 can be a measure of haplotypic diversity. A genomic region where all sites are heavily correlated contains less haplotype-alleles than a genomic region where sites are independent, as many more combinations of alleles can be observed in the latter case. For two biallelic genetic markers of minor allele frequency p and q, having a frequency x for the haplotype-allele formed by minor alleles at each marker, r^2 can be computed as:

$$r^2 = \frac{(x-pq)^2}{p(1-p)q(1-q)}. \tag{2.2}$$

The term $x-pq$ in the numerator of r^2 represents another statistic for measuring linkage disequilibrium, called D. D measures the deviation between the observed haplotype frequency x and the expected haplotype frequency if the two loci are independent, which is the product of the frequencies of the alleles at each locus, namely pq.

2.1.3 F_{ST}

Characterizing the amount of differentiation between populations is important in population genetics, as it carries information about how long ago the populations shared common ancestors and when they started to diverge. Variation

in the amount of genetic differentiation at the genome level can also be informative on levels of admixture since divergence, or identify particular regions of accelerated evolution that represent putative evidence for positive selection acting on the region. One of the most used statistics to characterize differentiation is F_{ST}, one of the three fixation indices introduced by Sewall Wright. F_{ST} aims at measuring the correlation of alleles of two homologous gene-copies randomly sampled from a sub-population relative to alleles randomly sampled from the entire population (Excoffier, 2008). There are different estimators of F_{ST} (Hudson et al., 1992; Nei, 1986; Weir and Cockerham, 1984), but one very often used is Nei's F_{ST} (Nei, 1973) which is computed as follows:

$$F_{ST} = \frac{H_T - H_S}{H_T}, \tag{2.3}$$

where H_T represents the heterozygosity of the total population and H_S the heterozygosity averaged across subpopulations, with each subpopulation given an equal weight in the summation.

2.2 Inferring the phase

As diploid individuals, humans receive half of their autosomal genomes from their mother and the other homologous half from their father. Most sequencing technologies today produce sequencing reads that are a couple of hundred base pairs long, hence variant positions are likely to sit on different reads and information on the joint origin (paternal or maternal) of alleles at different sites is lost. However, constructing the maternal and paternal haplotypes within one individual may be necessary for some genetic analyses. This procedure is called *phasing*. Accurate phase estimation in samples is becoming more and more important because phase information improves applications to disease association studies (Tewhey et al., 2011), imputation of untyped genetic variation (Marchini et al., 2007), inference of demographic history (Harris and Nielsen, 2013), identification of recombination breakpoints (Kong et al., 2008) or detection of regions under positive selection (Sabeti et al., 2002). Phase can be obtained either empirically, by sequencing long stretches of DNA, or statistically, by building methods that pair alleles at consecutive heterozygous sites using sample information. Only a few full genomes have been phased with molecular methods (*e.g.* Kitzman et al., 2011; Suk et al., 2011) as the cost of such techniques is two- to five-fold higher than the regular sequencing methods that produce unphased genetic data (Browning and Browning, 2011). In contrast, statistical phasing is rather inexpensive but can be computationally costly if the sample and the number of variant positions are large (Browning and Browning, 2011).

There are primarily two large classes of statistical phasing methods. Identity-by-descent methods are most often used on known pedigrees. They aim at detecting the long stretches of DNA that are shared via a very recent common ancestor - typically first to third degree relationships (Kong et al., 2008). In a parent-offspring comparison for example, ignoring all the *de novo* mutations in the offspring, both individuals share at least one allele identical-by-descent at every site. When the second parent is included, the phase within each of the three individuals can be inferred at all positions, except sites where all are heterozygous. To resolve such sites, one can turn to the second class of statistical phasing methods: the haplotype-frequency based methods. Such methods rely on computing the frequency of the haplotypes observed in the sample or a reference panel, using the frequencies to determine the likelihood of a given haplotypic configuration within an individual, and choosing the final configuration either by the help of a rule (like choosing the most likely configuration in parsimonious methods) (*e.g.* Wang and Xu, 2003; Gusfield, 2003) or according to a stochastic model (*e.g.* Scheet and Stephens, 2006; Browning and Browning, 2007; Williams et al., 2012). The latter way of choosing is the most commonly observed in phasing algorithms today. In many cases, it uses the posterior distribution of haplotypes given the genotypes, with the haplotypes being the hidden states of an underlying Hidden Markov Model that models the approximate coalescent with recombination. Haplotype-frequency based phasing algorithms can be used on any dataset of individuals, even when the dataset contains cryptic relatedness between individuals - such relatedness has been shown to actually improve the accuracy of the result (Browning and Browning, 2011) - but performs best in large samples, when computing time is not prohibitive. However, as new advances in sequencing technologies emerge, we may have to rely less and less on statistical phasing, obtaining haplotype information directly from molecular data. Empirical phasing remains the only method to address the issue of phasing *de novo* mutations and really rare variants, a problem that seems important for disease association studies (Bansal et al., 2010).

2.3 Visualizing and inferring population structure

2.3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a convenient statistical tool to observe multi-dimensional data in a space of fewer dimensions (usually in 2D) and at the same time preserving most of the features of the data. For genetic data, individuals are represented by points in the orthogonal space defined by each variable site. For example, let us consider 10 individuals genotyped on a 5 million SNP array. For each SNP, a reference allele can be defined so that the genotype of an individual is encoded as 0 if it is homozygous for the reference allele, 1 if heterozygous or 2 if homozygous for the other allele. The

encoded genotype represents the individual's coordinate on the axis defined by the SNP. The data representing all 10 individuals is thus a cloud in a space of 5 million dimensions. Despite the reduced number of individuals, it is difficult to picture the data as is because spaces of dimension higher than 3 are hard to visualize. PCA performs a rotation of the axes so that most of the variation in the data is captured on the first rotated axes which are called *principal components* (hence the name of the method). To be more precise, the first principal component represents the direction of the space where the data has the most variance. Then, in the space defined orthogonally to the first component, the second principal component is the direction that captures the most of the remaining variance. By sequentially projecting the data onto the orthogonal spaces of previously defined principal components and identifying the direction of highest variance in the data, PCA produces a new rotated set of axes, of which the first axes are most informative about the data. When sampling individuals from a population containing population structure (which can be from spatial constraints or other factors), a PCA can visually reveal the structure (figure 2.1).

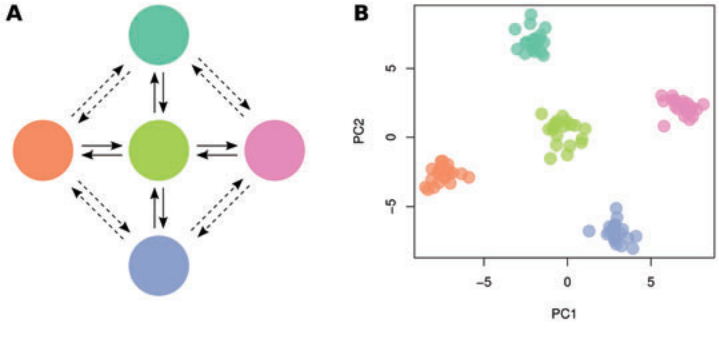

Figure 2.1. **Example of a PCA on genetic data**. We simulate a five island model using ms (Hudson, 2002), with 20 haploid individuals sampled from each island and 100,000 independent sites segregating among the 100 individuals. A) The island model. Solid arrows indicate a scaled mutation parameter of 20 and dashed arrows a scaled mutation parameter of 4. B) Results of the PCA applied to the 100 haploid individuals, for the 2 first principal components. The colors indicate the origin of each individual according to the model shown in A).

PCA is neither a data transformation technique nor a statistical test; it is merely a visualization tool that can be helpful to generate hypotheses about the data. Proper hypothesis testing is required to confirm or reject the hypotheses that were derived from looking at PCA results. The computation of the principal components can be sensitive to outliers and to the sampling scheme, when investigating spatial genetic correlation for example. When sampling from two diverged populations, a difference in the coordinates of in-

dividuals on the first principal component has been shown to be related to the average coalescence time between the individuals (McVean, 2009).

2.3.2 Bayesian inference of population structure

In the previous section, we presented PCA in the context of visualizing genetic data. As an exploratory tool, PCA may reveal structure in the studied sample but by no means does it model that structure formally, in a way that would make the structure quantifiable. In contrast, methods that model population structure explicitly have been developed (Pritchard et al., 2000; Alexander et al., 2009). One of the most cited programs that implement such methods is STRUCTURE (Pritchard et al., 2000; Falush et al., 2003; Hubisz et al., 2009). In the original version, Pritchard et al. (2000) use a Bayesian framework to estimate the membership of individuals in a given number of clusters using genetic data at unlinked loci. In general terms, STRUCTURE attempts to account for Hardy-Weinberg and linkage disequilibria by introducing a structure formed by clusters in which genotype frequencies and linkage are close to equilibrium. More formally, using the genotypes of the individuals as observed data, it estimates the allele frequencies within each cluster and the admixture proportions of each individual (the relative membership of individuals to each cluster) by computing their posterior distribution given the data via a MCMC Gibbs sampler. They use uninformative priors and assume Hardy-Weinberg equilibrium within each cluster. STRUCTURE has been widely used to study population structure in humans (*e.g.* Rosenberg et al., 2002) and many other organisms (*e.g.* Harter et al., 2004; Rosenberg et al., 2001). Extensions have been developed to include new aspects in the model such as linkage between loci (Falush et al., 2003), dominance and null alleles (Falush et al., 2007) and sample information (Hubisz et al., 2009). Another program called ADMIXTURE (Alexander et al., 2009) uses the same principles as STRUCTURE but improves computational speed greatly by the use of a quasi-Newton convergence acceleration method (Dennis and Moré, 1977).

2.4 Genome-Wide Association Studies

The principle behind Genome-Wide Association Studies (GWAS) is to survey the genome for association between the observed genotypes and a phenotype measured on the individuals in a study sample (Hirschhorn and Daly, 2005). The phenotype can be discrete (a disease status) or continuous (blood pressure for instance). The variants that are causing the phenotype might not be represented in the data (they may have not been genotyped). However, because of linkage disequilibrium, neighboring variants for which individuals have been genotyped might present an association to the phenotype due to the correlation of their alleles to the unobserved causal site. Thus, GWAS have

benefited greatly from the technological advances in high throughput genotyping (McCarthy et al., 2008). As the number of SNPs on arrays has increased over the years, better resolution has been achieved at the local level, providing hopes for identifying particular genes contributing to the phenotype. Population structure or cryptic relatedness in the sample can lead to false positives (Hirschhorn and Daly, 2005) and needs to be accounted for, either by direct modelling or by statistical correction of the effects, with genomic control for example (Price et al., 2010). Also, as the number of genotyped sites increases thanks to ever growing size of SNP arrays, more stringent significance thresholds for association need to be used to keep the number of false positives low.

Since 2005, more than 2,000 regions have been robustly associated with complex diseases and traits (Manolio, 2013). Nonetheless, the heritability of many common complex diseases or traits remains poorly explained by the variants found in association studies. One potential explanation is the relatively small effect of each variant on the phenotype (Hirschhorn and Daly, 2005). Complex traits and diseases might involve a large amount of genomic regions, each of them contributing only a small amount to the total variance in phenotypic values. Another explanation lies in the potential contribution of rare variants, which are very likely to be absent from SNP arrays or may be poorly correlated with neighboring genotyped sites (Bansal et al., 2010).

2.5 Detecting selection

When a new mutant allele is introduced in a population and is highly beneficial, it tends to increase rapidly in frequency, dragging along the particular haplotype it appeared in, so that the alleles forming the haplotype also increase rapidly in frequency (Nielsen, 2005). This effect is called *genetic hitchhiking*. The high frequency of those neutral alleles is mainly due to their proximity with a beneficial allele, and not due to an inherent positive effect on fitness. The phenomenon of an entire haplotype increasing in frequency due to positive selection on a *de novo* mutation and eventually reaching fixation is referred to as a *hard selective sweep*. The genetic variation gets swept away around the beneficial allele. When an allele that is already present in a population becomes beneficial (perhaps due to a change in environment), all the different haplotypes that the allele sits in tend to increase in frequency, creating a *soft selective sweep*. Soft sweeps are usually harder to detect as they resemble more the expected patterns of diversity under neutrality. In humans, only a handful of hard selective sweeps have been identified and it is believed that most selective events act on standing variation, thus causing soft selective

sweeps (Pritchard and Di Rienzo, 2010).

I describe here three statistics that can be used to detect signals of positive selection. The three statistics can be computed for every variable site in the genomes of a given sample. The main assumption of these types of genomic-scans are that most sites are evolving neutrally. Regions with high values compared to the genome average background level suggest potential candidate regions for positive selection. These outlier approaches are not well defined statistical tests *per se*, unless a formal computation or simulation of the distribution of background values under the neutral null model is performed. They can be useful however for generating hypotheses and can provide strong additional evidence for selection when a particular region is identified by other analyses prior to the scan (see paper IV for example).

2.5.1 iHS

The *iHS* statistic (Voight et al., 2006) aims at detecting signals of strong recent positive selection on *de novo* variation, when the beneficial allele has not yet reached fixation. It relies on the contrast between decays of haplotype homozygosity around either the ancestral or the derived allele and standardizes this contrast at genome-wide level, within classes of derived allele frequency. Indeed haplotypes around older alleles are more likely to be diverse, as recombination has had time to break down and re-shuffle the haplotypic background around the derived allele. By standardizing the iHS values within classes of allele frequencies, we limit the potential effect of allele age. In particular, within a class of derived allele frequency p, iHS is computed as

$$\text{iHS} = \frac{\ln(\frac{\text{iHH}_A}{\text{iHH}_D}) - \mathbb{E}_p[\ln(\frac{\text{iHH}_A}{\text{iHH}_D})]}{\text{SD}_p[\ln(\frac{\text{iHH}_A}{\text{iHH}_D})]}, \tag{2.4}$$

with iHH_A (resp. iHH_D) the integrated value of the decay of homozygosity in both directions around the ancestral (resp. derived) allele, $\mathbb{E}_p[\ .\ ]$ and $\text{SD}_p[\ .\ ]$, the genome-wide average and standard deviation within the derived allele frequency class p. As the value of iHS rather than its sign is important, genome scans are usually performed using the absolute value of iHS. Any site with a value of $\ln(\text{iHH}_A/\text{iHH}_D)$ that largely exceeds or largely falls below the genome average will be considered, regardless of the direction.

2.5.2 F_{ST} and LSBL

I talked about F_{ST} earlier, in the context of measuring population differentiation. The effect of population divergence on the genetic differences between two populations is expected to be the same throughout the genome. However,

if one of the two populations have experienced a selective event, targeting a particular region of the genome, differences accumulate faster in the selective sweep than on another randomly selected region of the genome. So in a genomic computation of local F_{ST} values, selected regions should appear elevated from the background level of neutral divergence. To detect signals of selection using this method, the ideal situation occurs when using two populations that diverged somewhat recently, so that traces of the selective events since divergence do not get lost in the background of F_{ST} values. A related statistic also aimed at detecting signals of selection in the genome is the Locus Specific Branch Length statistic (LSBL) (Shriver et al., 2004). Based on the divergence between 3 populations, it uses F_{ST} as a proxy for the temporal distance between populations and tries to extract the length of the branch leading to a particular population (figure 2.2).

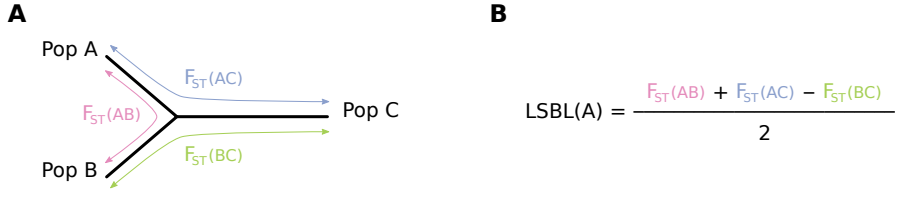

Figure 2.2. **LSBL**. A) Unrooted tree between three populations and the corresponding F_{ST}. B) Equation for LSBL as function of the three F_{ST} values.

Positive selection that is private to one population should result in more differences with the other two populations in the region targeted by selection, thus locally the tree representing the ancestry between the three populations should have a longer branch leading to the population under selection. Like in the F_{ST} scan, it is by comparing to the background of typical branch lengths that potential signals of selection can be found.

2.6 Inferring demographic parameters

Populations usually evolve in a complex manner. Their size can change over time, as a result of climate changes (Ruzzante et al., 2008) such as glacial cycles or of movements into new territories (DeGiorgio et al., 2011). They can split into smaller groups for ecological or geographical reasons (Shapiro et al., 2012). They might come into contact with other populations and exchange migrants (Wang et al., 2008). Understanding the demographic history of given populations can shed light on the impact of these extrinsic and intrinsic factors on their evolution (Gattepaille et al., 2013). Inferring demography is also important for deriving the neutral distribution of statistics of interest for which outlier values can then be interpreted as potential signals for selec-

tion (Nielsen, 2005).

Several methods have been developed to infer demographic parameters, such as population size, divergence times, migration rates, admixture times and proportions. Most methods assume a given parametric model, which can be quite complex, including split times, migrations, admixture events, bottlenecks and so on, and provide estimates for all parameters involved given the observed data. Some methods use the full-likelihood of the data (*e.g.* Kuhner et al., 2000; Beerli and Felsenstein, 2001) or the likelihood of summary statistics (*e.g.* Gutenkunst et al., 2009; Naduvilezhath et al., 2011), others use Approximate Bayesian Computation to estimate the parameters (*e.g.* Beaumont et al., 2002; Excoffier et al., 2005) or use a full Bayesian approach on the data (*e.g.* Li and Durbin, 2011; Sheehan et al., 2013; Steinrücken et al., 2013), coupled with the use of the Sequentially Markov Coalescent approximation (McVean and Cardin, 2005; Marjoram and Wall, 2006). I do not review here all the different methods and their specificities, however I will give a brief overview of PSMC (Li and Durbin, 2011), an approach to infer N_e over time, as I use it for comparison to the method of inferring N_e over time that I develop in paper III.

PSMC, which stands for Pairwise Sequentially Markovian Coalescent, is a method for inferring variable population size over time. The population size is modelled as a piecewise constant function whose breakpoints are defined by the user on a logarithmic scale. PSMC can be employed on the entire genome of one individual and utilizes the patterns of local heterozygosity to estimate local gene-genealogies. It models a Hidden Markov Model on the times to coalescence between the paternal and maternal DNA sequences of the individual, and uses the Sequentially Markovian Coalescent model (McVean and Cardin, 2005) to incorporate recombination into the method. The population sizes for every defined period are computed via Expectation Maximization during the course of the MCMC chain. The probabilities of the hidden states of times to coalescence obtained at the end of the run can be used to estimate the local gene-genealogies and the break-points between non-recombining segments. Since its publication in 2011, PSMC has been used a great number of times for various species, due to its simplicity, the little amount of parameters to specify and its computational speed. It has been shown to perform relatively well to estimate population size in ancient times, but performs poorly in the very recent past (Li and Durbin, 2011). Extensions to include more individuals have been developed (Sheehan et al., 2013; Schiffels and Durbin, 2014), which could potentially alleviate the problem of recent population size inference.

3. Research Aims

The main objective of this thesis was to develop new methods for providing answers to various population genetic questions, with application to the investigation of human evolution. More specifically, the aims were:

I Investigating the mathematical relationship between F_{ST} and homozygosity, by providing upper and lower bounds between the quantities, and survey F_{ST} estimates among human populations in the light of their homozygosities.

II Developing a criterion for deciding when to combine SNP markers into a haplotype in order to improve the assignment of individuals of unknown origin to populations represented in a reference panel, and applying the criterion on human SNP data to separate populations that cannot be distinguished by using the SNPs separately.

III Deriving an analytical correspondence between distributions of coalescence times and population size over time, verifying the robustness of the mathematical result on simulations and using it to infer past population sizes of different human populations.

IV Investigating human evolution in response to toxic environments, in particular identifying genetic regions involved in metabolizing arsenic using the genome-wide association framework and explore the potential presence of evidence for positive selection in the associated regions.

4. Summary of the papers

4.1 Paper I

Homozygosity Constraints on the Range of Nei's F_{ST}

F_{ST} is a widely used statistic to characterize the amount of differentiation between populations or species. Potentially ranging from 0 to 1, many studies have obtained estimates of F_{ST} that are typically lower than 0.2. It has been shown that F_{ST} estimators are sensitive to the levels of genetic diversity within populations which can explain why we rarely observe high values of F_{ST} in nature. We investigate this issue by looking at the mathematical properties of one estimator of F_{ST}, Nei's G_{ST}, between two populations harboring fixed levels of homozygosity, for loci carrying a given finite number of alleles. We derive upper and lower bounds for F_{ST} as functions of the homozygosities and the number of alleles. We illustrate our results on a human dataset of 131,834 haplotype-loci formed by combining 2,285,342 SNPs into windows of 20kb. Samples from 3 populations are used, 101 individuals sampled in Utah (United States) with European ancestry, 103 Yoruban individuals sampled in Ibadan (Nigeria) and 30 individuals sampled from two Northern San populations, the !Xun and Ju/'hoansi living close to the border of Namibia and Botswana.

The average homozygosity for these haplotypes was 0.31 for the European individuals, 0.20 for the Yoruban individuals and 0.26 for the Northern San individuals. The genome-wide averages of the haplotype F_{ST} were 0.092 between Europeans and Northern Sans, 0.061 between Europeans and Yorubans, 0.044 between Northern Sans and Yorubans, with average relative positions within the admissible range predicted by our theorem of 0.48, 0.34 and 0.27 respectively. We investigate the values of F_{ST} between Northern San and Europeans as well as between Northern San and Yoruba for three classes of haplotype-loci depending on the value of homozygosity of these haplotype-loci in Northern San: low homozygosity (0.05), intermediate homozygosity (0.2) and high homozygosity (0.6).

All values are within the range predicted from our theoretical result, however all F_{ST} values are close to the upper bound in the class of low homozygosity in Northern San, for both population comparisons. In addition, the F_{ST} values in the class of high homozygosity, haplotype-loci are close to the lower bound in the Northern San/Yoruba comparison while they are embracing almost the full admissible range in the Northern San/Europeans comparison

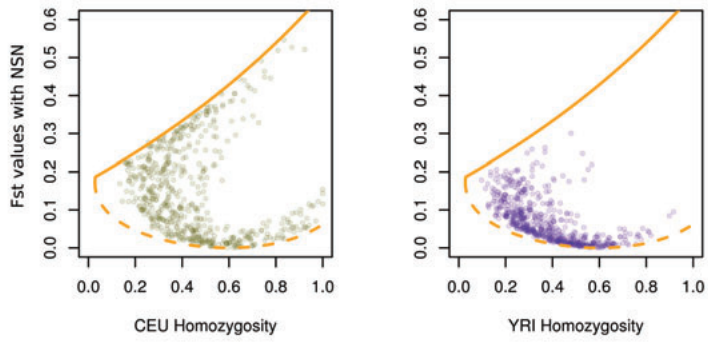

Figure 4.1. F_{ST} **values within their constrained range**. F_{ST} values for the Northern San vs. European comparison (left panel) and for the Northern San vs. Yoruba comparison (right panel). Only haplotype-loci having homozygosity between 0.6 and 0.62 in Northern San were included (281 loci). Upper and lower bounds on F_{ST} are indicated by orange solid and dashed lines.

(figure 4.1). We observe the least amount of differentiation between Yoruban and Northern San individuals despite the early divergence of the ancestors of Northern San populations from the ancestors of other extant humans found in earlier studies, which suggests a complex demographic history in Africa, possibly involving admixture events after the divergence of these two populations, confirming previous findings of admixture between !Xun and West African populations. Additional explanations for this differentiation pattern include potential effects of ascertainment bias on the SNP array for the African samples (as the array might exclude a number of African variants on which Yoruban and Northern San individuals could segregate), stronger differentiation in the European population as a result of the Out-of-Africa event as well as potential effects of archaic admixture in the European population. We believe that discussing F_{ST} values in the light of their constrained range will generate interesting hypotheses regarding the ancestral relationships between different populations, help better characterize levels of differentiation and further our understanding of the F_{ST} statistic and its dependency on genetic diversity.

4.2 Paper II

Combining markers into haplotypes can improve population structure inference

With the advent of high-throughput genotyping and sequencing technologies, the amount of genetic data available has tremendously increased. Today, SNP arrays can provide genotype information for individuals at more than 5 million positions in their genomes and the entire genomes themselves can be sequenced at a fairly low price compared to 10 years ago. This means that

when considering variable sites in a collection of genomes or genome-wide genotypes, many of the sites considered are in linkage disequilibrium and thus carry somewhat redundant information. While the correlation of alleles at different sites might not be a problem in population genetic inference methods that explicitly model this correlation, numerous methods actually do require independence of the sites considered. In particular, most population structure inference and genetic clustering methods are based on the genetic regions being independent and it is rather unclear how reliable the results can be if this assumption is violated. To alleviate this issue and obtain a collection of sites that are closed to independent, most studies perform pruning steps, where sites are discarded according to a certain threshold of linkage disequilibrium, resulting in a loss of data that can be quite considerable. In this paper, we investigate the feasibility of another strategy: combining the sites into larger segments (haplotypes) and use them as multi-allelic genetic markers. We evaluate the combining strategy for the problem of assigning individuals of unknown origin to a panel of already identified genetic groups.

Being able to correctly assign individuals to their group of origin can be of great importance in conservation genetics and forensic science. We introduce a new criterion, called the *Gain of Informativeness for Assignment* (GIA), that allows us to decide when two genetic markers can be combined to form a new haplotype marker and improve the assignment of individuals to groups. GIA is derived from the *Informativeness for Assignment* (IA), a statistic based on information theory which quantifies how helpful a given genetic marker is for assigning individuals to groups (Rosenberg et al., 2003). Let us consider a marker with multiple alleles: if the frequencies of the alleles are the same in all groups, then the marker carries no information for assignment and IA is zero; if all alleles are private to exactly one group, then the information carried by the marker is maximal. Our approach to the problem of combining markers is thus simple: we compute GIA as the difference between the informativeness of the haplotype and the sum of the informativenesses for each separate marker. If GIA is positive, meaning if IA for the haplotype is higher than the sum of the IA of the two markers to be combined, then the haplotype carries more information than the markers taken separately and combination is advised, otherwise both markers should be considered separately.

After performing a study of GIA as a function of allele frequencies and amount of LD, where we showed that there is no easy correspondence between the sign of GIA, the allele frequencies and the level of LD, but showed however that independent markers lead to a negative GIA, we test the usefulness of our criterion for two simulated scenarii. In the first scenario, we make a proof of concept for GIA where we generate 20 SNP pairs with given allele frequencies in two populations and given amount of LD. We sample 100 individuals from each population. Since each SNP pair has the same

characteristics in terms of LD and allele frequencies, they all have the same value of GIA. We perform the assignment of the individuals using the software STRUCTURE (Pritchard et al., 2000) under the unsupervised clustering setting. We can then compare the ability of the software to correctly assign the individuals to their group when using the 40 SNPs separately or when using the pairs of SNPs merged into 20 independent haplotypes, and look at the result in the light of the GIA value. Out of the 11 cases where assignment with haplotypes performed better, 9 cases had a positive GIA. For all of the 9 cases where the assignment was better using the SNPs separately, GIA was negative. This observation showed GIA's ability to indicate whether markers should be combined and improve the assignment.

In the second scenario, we generated genetic data for 200 individuals sampled in equal proportions from two populations evolving under a two-islands model with a proportion m of migrants per generation. The data consist in 1000 SNPs on a chromosome fragment where recombination can occur (two scaled recombination rates were used: $\rho = 150$ and $\rho = 1500$). We perform the assignment of individuals using STRUCTURE on different versions of the data set, either using all 1000 SNPs separately, or by removing SNPs in LD, or by combining SNPs into haplotypes either at random or with the help of GIA or F_{ST}. As we increase the migration rate, the accuracy of the assignment decreases as expected. Also, assignment was generally more difficult when recombination was small. We found that the most accurate assignments were obtained when SNPs were combined into haplotypes using GIA or F_{ST}, whereas randomly combining SNPs lead to worse results than the assignment based on the separate SNPs. This observation showed that the improvement obtained by combining SNPs did not originate from the sole process of building haplotypes, but in the guidance of *GIA* to choose which SNPs to combine.

Finally, we tested the effect of combining SNPs into haplotypes using GIA on human genetic data from POPRES. In particular, we chose to perform a cross validation study using genetic data from chromosomes 1 to 3 for 89 French and 70 German individuals, where half of the French individuals and half of the German individuals were considered of known origin and used as reference panel to compute all allele frequencies, representing thus a training set, and all remaining individuals considered of unknown origin and constituting a validation set. Assigning individuals in the validation set using STRUCTURE under supervised clustering with the training set did not perform well, regardless of the SNPs being used separately or combined, which might not be so surprising in the light of the low genetic differentiation between the French and German samples ($F_{ST} = 0.00068$). However assignment using the first principal component of a Principal Component Analysis (PCA) on either the SNP data or the GIA-combined data lead to interesting results. French and German individuals from the validation set could not be separated when PCA

was performed on separated SNPs (only 53.2% of individuals correctly assigned) but the assignment based on the GIA-combined dataset was markedly improved (87.3% of individuals correctly assigned) (figure 4.2).

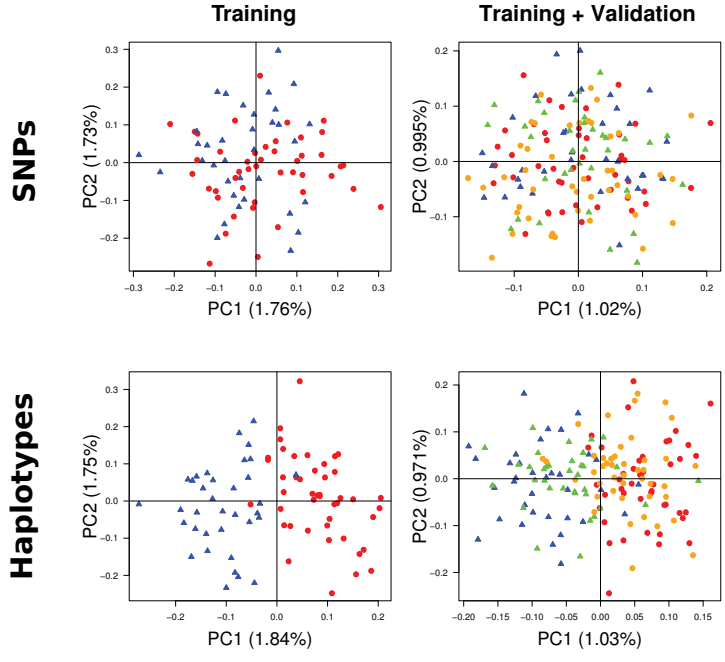

Figure 4.2. **Principal component analysis for French and German individuals from the POPRES data**. Each population sampled is divided into two samples of equal size: a training sample and a validation sample. The French and German training samples are used to build haplotypes using GIA. 105,341 SNPs on chromosomes 1, 2, and 3 were converted into 54,762 haplotype loci. Upper panels show the two first components of the PCA on the separate SNPs of the training set (panel A) and all individuals (panel B). Lower panels show the same individuals when plotted onto the two first principal component of the PCA based on the haplotypes. German individuals are represented by triangles (blue for training, green for validation) and French individuals by circles (red for training and orange for validation).

To take the application further, we used all French and German individuals as a two group reference panel (training set) to build the haplotypes using GIA, and tried to assign Swiss-French and Swiss-German individuals to the two clusters, to investigate the potential presence of assortative mating based on language. Using GIA-combined haplotypes only lead to a slight decrease (7%) in incorrect assignments suggesting that Swiss individuals do not separate on the base of French/German genetic differences. In a cross validation study of the Swiss samples alone, the decrease in incorrect assignments was higher (28.6% decrease) with GIA-combined haplotypes, but the overall

assignment remained difficult, probably due to the very low genetic differentiation between Swiss-French and Swiss-German individuals ($F_{ST} = 0.00012$).

With this study, we have demonstrated that haplotypes contain additional information about population structure and that using haplotypes instead of single SNPs can improve assignment of individuals to populations, even for difficult cases. The GIA statistic determines when it is possible to improve the assignment of individuals to populations by combining markers into haplotypes and it can be used as a tool for population structure inference methods to capitalize on dense sets of genetic markers.

4.3 Paper III

Popsicle: a method for inferring past effective population size from distributions of coalescent times

Natural populations rarely evolve under constant size. Instead, their size can vary as a consequence of climatic changes such as glacial periods, epidemics, founder effects when colonizing new territories, admixture when coming into contact with other populations, or just by chance due to the stochasticity in number of offspring produced and surviving. Investigating the effective size of a population over time has received considerable interest in recent years, with the introduction of new methods that utilize genomic data to reconstruct the profile of past population sizes. Most methods of population size inference are dealing internally with two steps (sometimes circularly), one step that uses genetic data to determine local ancestral relationships between genecopies sampled from the population, the other step that updates the population size profile by maximizing the likelihood of the computed genealogies under the model defined by the profile. In this paper, we facilitate the second step by providing the analytical relationship between the population size over time and the distributions of coalescence times. With this result, we can accurately approximate the population size over time when given a large number of independent gene-genealogies, that allows us to estimate the distributions of coalescence times with good accuracy.

Our main result derives from inverting a previously found result where distributions of coalescence times were expressed as linear combinations of a family of functions that depend on population size over time (Polanski et al., 2003). From the inverted result, we can have access to the population size over time using the distributions of coalescence times. Though this theoretical result is exact and valid for continuous functions, the fact the genome sizes are finite makes the distributions of coalescence times impossible to obtain in

their continuity and we can only have access to approximate distributions. We show on simulated data that our analytical result is stable, as using approximate distributions of coalescence times generated by collecting 1,000,000 gene-genealogies under the population size model, leads to estimates of the population size that are very close to the truth. In addition, we find that using much fewer loci (around 10,000) can still lead to good estimates of the population size. The accuracy of the inferred population size improves significantly when increasing the sample size from 2 to 10, but further increasing the sample size does not improve the accuracy much. We also simulated gene-copies with recombination and found that when recombination is ignored, our method can lead to biases in the estimated population sizes.

In reality, local gene-genealogies in the genome are unknown. The only visible testimonies of the underlying gene-genealogies are the mutations that arose onto the different ancestral lineages leading to the sample. In this paper, we do not address the challenging problem of inferring the succession of local gene-genealogies onto a recombining locus (inference of the ancestral recombination graph), instead we use a simple algorithm of gene-genealogy reconstruction for non-recombining loci, so as to evaluate the performance of our population size inference method under a range of different mutation rates and to apply the method to empirical data. Unsurprisingly, we found that with increasing mutation rate, the method performs better, as the gene-genealogies get inferred with better accuracy when the number of polymorphisms increases. For an empirical data application, we used phased sequences obtained from the Complete Genomics Trios data from the 1000 Genomes Project public data, and extracted 22,321 non-recombining regions according to the Decode genetic map. We applied our method and compared the results to the results of PSMC, a commonly used method for population size inference.

We found that even for a small number of loci and a simple algorithm for genealogy inference, our method was able to recover the general pattern of population size observed by PSMC (figure 4.3, upper panel). We also note an extra feature of an early divergence of the Yoruba population from the non-African populations (figure 4.3, lower panel), suggesting a long standing population structure on the African continent prior to the Out-of-Africa event. Our method is very fast compared to PSMC, as the computation of the population size from the distribution of coalescence times is virtually instantaneous, and the number of loci used is small. The main result presented here could potentially be integrated in algorithms such as PSMC in order to decrease the computing time and increase accuracy of the effective population size reconstruction.

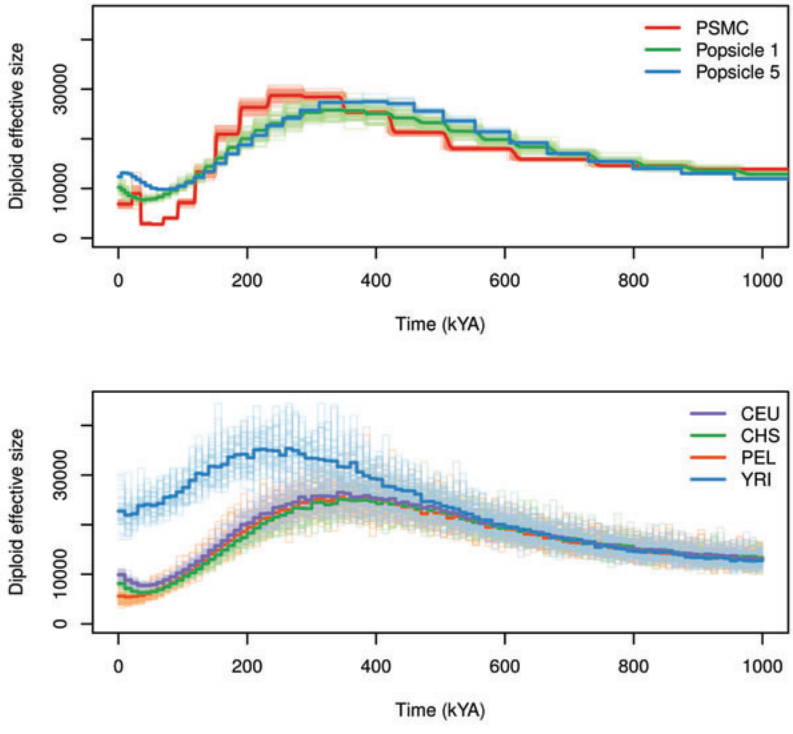

Figure 4.3. **Results of our effective population size inference method on human sequences from Complete Genomics**. Upper panel: comparison between effective population size profiles computed with PSMC and our method on single individuals (Popsicle 1) or on samples of 5 individuals (Popsicle 5), based on 64 European individuals. Lower panel: results of Popsicle 1 for four different populations, individuals of European ancestry (CEU), Southern Han Chinese individuals (CHS), Peruvian individuals (PEL) and Yoruban individuals (YRI). Averages are indicated by solid lines.

4.4 Paper IV

Human Adaptation to Arsenic-Rich Environments

In humans, the extent of evolution that has been shaped via selective pressure compared to neutral demographic processes remains unknown. In fact even today, there are only a handful of well identified cases of human adaptations. We know for example that some populations in Europe and Africa have evolved lactose tolerance to adapt to milk drinking diets, that adaptation to high altitude has been observed in Tibetans, that some people have developed a resistance to Malaria in regions of the globe where the disease is rampant, that lighter skin could have evolved from a need to produce more vitamin-D in regions of high latitude where the incidence of the sun's rays is reduced. The study presented in paper IV is an addition to this small list of known human adaptations to the environment, and offer a rather comprehensive view on the

process, from the phenotypes to the biological mechanism and to the genotypes.

The study deals with adaptation to arsenic-rich environments. In some villages in the Andes, it has been noticed previously that, despite a rather high concentration of arsenic in the drinking water, the inhabitants seemed not to suffer as much from the typical adverse effects of regular arsenic consumption. In one of those villages, San Antonio de Los Cobres (SAC) in Argentina, our collaborators collected blood and urine samples from 385 women living in the area and having a family history of at least 2 generations back in the region. Urine samples were analysed for inorganic arsenic and organic compounds containing arsenic: monomethylarsonic acid (MMA) and diethylarsinic (DMA). MMA has been shown to be the most toxic compound of the two and DMA to be more easily expelled from the body. Using the blood samples, we genotyped 124 of the women using a genome-wide dense SNP array of around 5 million SNPs.

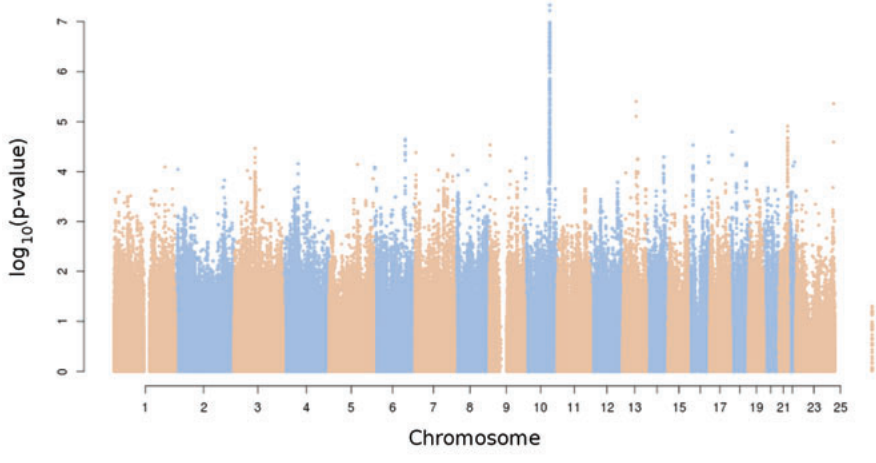

Figure 4.4. **GWAS result for the MMA concentration phenotype.**

We performed a Genome-Wide Association Study (GWAS) to identify regions of the genome involved in arsenic metabolism and found a strong and significant association on chromosome 10 for SNPs in a region upstream of the gene encoding for the enzyme arsenic (+3 oxidation state) methyltransferase (AS3MT) and both concentrations in MMA and DMA (figure 4.4). This result confirms previous findings regarding the central role of AS3MT in arsenic metabolism. Other significant regions were found, one on chromosome 21 which was significant in the GWAS based on MMA as well as in the GWAS based on DMA, and on chromosomes 2, 12 and 13, though significantly associated either with MMA only or DMA only. None of these regions have been

associated with arsenic metabolism previously and if the association signals are true positives, functional studies would be required to identify how these candidate genes influence the pathway of arsenic metabolism.

In addition to GWAS, we performed scans for selection using several types of statistics: F_{ST} between the SAC individuals and a closely related Peruvian population, the Locus-Specific Branch Length (LSBL), which evaluates the level of differentiation of a given site for a given population compared the same level in two comparative populations, and the integrated Haplotype Statistic (iHS) which uses patterns of homozygosity decay to identify putative regions under recent positive selection. We found elevated values of F_{ST} (figure 4.5), LSBL and mean |iHS| in the region of AS3MT, providing strong support for adaptation to a more efficient arsenic metabolism. In fact, out of the 100 SNPs with highest LSBL value genome-wide, 13 are located in the AS3MT region and the greatest 1Mb-window averaged |iHS| value in the region was in the 97 percentile of the genome distribution of 1Mb-window averaged |iHS| values.

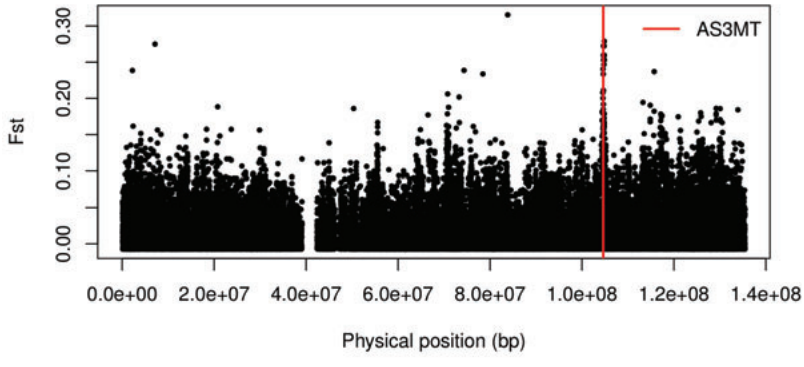

Figure 4.5. F_{ST} **values between the individuals from SAC and Peruvian individuals from the 1000 Genomes Project dataset.**

Zooming in the AS3MT region, we identify a combination of SNP alleles that is associated with higher level of DMA in urine (which is the more efficient phenotype) and represents a protective haplotype. Investigating the same haplotype-locus in other regions of the world reveals that the protective haplotype, or very similar haplotypes, is present in some Native American and East Asian samples. The sampled women of SAC were shown to have little hispanic ancestry and their native ancestors, the Atacameños, seem to have peopled the area as early as 11,000 years ago. This has left enough time for natural selection to act on the resistance phenotype and increase the frequency of the associated protective haplotype. The existence of the protective haplotype in other parts of the world suggests that selection acted on standing

variation, but because our genetic data consists in ascertained SNPs and not sequences, we cannot exclude the existence of a causal mutation explaining the phenotype. We evaluated the selection coefficient of the beneficial variant to be between 0.003 and 0.005. This study shows that combining GWAS to genome-scans for selection can yield interesting results and more power to support evidence for positive selection in a population.

5. Conclusions and future prospects

In this thesis, I have investigated different topics relative to human evolution, by developing population genetic methods and applying them to genetic data from human genomes. These topics include how differentiation between populations is related to homozygosity in the populations, how assignment of individuals of unknown origin can be improved by combining markers into haplotypes, how past population size can be computed using distributions of coalescent times and how some individuals from a region in the Andes responded genetically to the presence of arsenic in the drinking water. The breadth of the topics presented in this thesis illustrates the richness of tools population genetics has to offer for studying evolution. By bringing mathematical models into the study of patterns of genetic diversity, population genetics together with the advances in SNP genotyping and sequencing technologies have brought the field of evolutionary biology into a new level where many hypotheses of ecological or anthropological relevance can finally be tested.

In particular, with recent advances in sequencing techniques of ancient DNA material, we can start to build a more refined picture of the major demographic events of human history, using both present day and ancient samples. For instance, it has been shown that agriculture spread from the fertile crescent to northern Europe via a demic diffusion followed by subsequent admixture of the agriculturalists with the local hunter-gatherers (Skoglund et al., 2012). Also, ancient native American samples have been used to study the peopling of the Americas (Raghavan et al., 2015). Archaeology can now benefit from genetic analyses of discovered remains and I believe that incorporating multiple lines of evidence such as language phylogenies, movement of ancient crops and livestock, ancient bacterial metagenomes and retroviruses incorporated in the genomes, human history can be studied in great detail and exciting results are to be expected in the near future. Such composite data will require advances in modelling techniques to account for spatial and temporal structure, as well as the multiplicity of the nature of the data.

Advances in molecular techniques are also expected. In particular, progress in single cell genomics and molecular phasing will allow us to study the process of recombination in more detail and build more accurate recombination maps that do not rely on linkage disequilibrium for their computation. This will prove very useful in numerous population genetic analyses, such as association studies or the inference of the ancestral recombination graph from multiple sequences, which contains all the ancestral information a sequence can

harbor (Rasmussen et al., 2014). With increasing number of trios sequenced, we can also hope for accurate mutational maps in the future. Indeed, we know already that regions in the genome can have different mutation rates, as the conservation constraints can vary from region to region (Smith et al., 2002), and using a single mutation rate to calibrate times to coalescence (as we do in paper III for example) might bias the result of the analysis at hand. In addition, most human populations are growing super-exponentially since the industrial revolution, resulting in a large number of rare genetic variation that can only be observed by sequencing a sample, whose size becomes in the order of the effective population size or even larger (Keinan and Clark, 2012). Such large samples might not be suitable for analyses built on the coalescent model, which assumes a sample of negligible size compared to the effective size of the population. Models that account for the recent super-exponential growth of human populations might be needed. In addition, rare variants are believed to play an important role in health traits (Cirulli and Goldstein, 2010), and increasing efforts are likely to be made in the future to study rare variation. However, improvements in sequencing technologies are necessary to reliably capture rare variants. In particular, the rate of sequencing errors remains to-date too high for being able to call rare variants with good accuracy.

Data quality and quantity has drastically increased over the past decades, allowing for models that used to be purely theoretical to be tested and creating a need for new models as well. In this context, being a population geneticist is quite exciting. The use of mathematics has been proven invaluable for studying evolution, and there are surely numerous unforeseen fruitful outcomes that are going to emerge from the trans-disciplinary field that population genetics is. I am looking forward to see the future development of the field and eager to add my modest participation to the associated endeavor.

6. Svensk Sammanfattning

Populationsgenetik har tack vare sin förmåga att förutsäga och beskriva de viktigaste evolutionära processerna såsom mutation, genetisk drift och naturligt urval lett till ett stort antal viktiga studier av evolutionära processer. Nya populationsgenetiska metoder och modeller utvecklas ständigt för att besvara evolutionära frågor vilket ger oss nya pusselbitar om livsformers historia. I den här avhandlingen presenteras nya populationsgenetiska verktyg och resultat. Jag undersöker frågor angående populationers förändring över tid med särskilt fokus på människans evolution. Dessa frågor inkluderar hur differentiering mellan populationer är relaterad till homozygositet i populationer, hur populationsassignment av individer av okänt ursprung kan förbättras genom att kombinera markörer, hur en populationsstorlek över tid kan beräknas med hjälp av fördelningarna av koalescenstider och hur vissa människor från några regioner i Anderna med höga halter av arsenik i dricksvattnet nu bär på genvarianter anpassade till att delvis tåla arsenik.

Jag härleder en övre gräns och en undre gräns för F_{ST}, ett klassisk mått på populationsdifferentiering, som funktioner av homozygositet i subpopulationer. Jag tillämpar resultaten för att diskutera observerade populationsdifferentieringar. Jag inför en ny statistik, *Gain of Informativeness for Assignment* (GIA), som kan användas för att avgöra om två genetiska markörer bör kombineras i en haplotyp för att förbättra assignment av individer till populationer i en panel av referenspopulationer. Tillämpning av metoden på SNP-data för franska, tyska och schweiziska individer visar hur haplotyper konstruerade med hjälp av GIA kan leda till bättre assignment och en tydligare bild av kryptisk populationsstruktur. Jag härleder också matematiska formler för hur en populations historiska storlek över tiden är relaterad till fördelningarna av koalescenstider; visar hur robusta dessa formler är för antal loci och antal individer; och med hjälp av en enkel algoritm för att uppskatta en gens koalescenttider tillämpar jag min metod på regioner av det humana genomet med låg rekombinationstakt för fyra globalt spridda populationer. Jag visar att metoden kan fånga tidigare observerade populationsstorleksförändringar långt tillbaka i tiden och kan dessutom visa på nya populationsstorleksförändringar i mer modern tid genom att basera analyserna på flera individer samtidigt. Slutligen presenterar jag en studie av mäniskans anpassning till en arsenikrik miljö. Tillsammans med samarbetspartners från Karolinska Institutet och Lunds universitet genotypade vi 124 kvinnor från San Antonio de los Cobres (SAC) – en by i de argentinska Anderna där höga halter av arsenik i dricksvattnet har uppmätts men där de skadliga effekterna av arsenikkonsumtion är

mindre än i andra utsatta områden. Eftersom bosättningen i regionen tros vara åtminstone 10 000 år gammal tyder detta på att metabola mekanismer för att tåla höga halter av arsenik har utvecklats hos många individer i den här gruppen. Vi använde fenotypinformation i form av koncentrationer av oorganisk arsenik och organiska arsenikföreningar i urinen och utförde en associationsstudie för att upptäcka regioner i genomet associerade med låga koncentrationer av monomethylarsonic syra (MMA, en ganska giftig form av organisk arsenikförening) och höga koncentrationer av dimethylarsinic syra (DMA, en mildare form av organisk arsenikförening som är relativt lätt för kroppen att göra sig av med). Vi hittade framför allt en association i en region på kromosom 10, uppströms genen *AS3MT*, en gen som kodar för enzymet arsenik (+3 oxidationstillståndet) metyltransferas som tidigare påvisats en nyckelroll vid arsenik metylering. Dessutom fann vi förhöjda värden av iHS och LSBL (statistika mått som är känsliga för selektion) i denna region, vilket ger en stark indikation på att denna region har förändrats som ett led i anpassningen till en arsenik-rik miljö. Den här studien representerar det idag enda kända exemplet av mänsklig anpassning till giftiga miljöer.

Bredden av de ämnen som presenteras i min avhandling illustrerar en mångfald av verktyg som populationsgenetiken har att erbjuda för att studera evolution. Genom att applicera matematiska modeller på mönster av genetisk variation, har populationsgenetiken tillsammans med framstegen inom SNP genotypning och sekvenseringsteknologi möjliggjort bättre och mer avancerade evolutionsstudier där många hypoteser av ekologisk eller antropologisk betydelse slutligen kan testas.

7. Résumé en Français

Grâce à ses capacités de prédiction et de description des processus les plus importants de l'Évolution, tels la mutation, la dérive génétique et la sélection, la génétique des populations a généré un grand nombre d'études fructueuses sur l'Évolution. Ce domaine de recherche est en expansion encore aujourd'hui, avec de nouvelles méthodes et modèles en développement pour répondre à des questions de grand intérêt au regard de l'Évolution et pour lever le voile sur le passé des différentes formes de vie. Dans cette thèse, je présente ma modeste contribution au développement de la génétique des populations. J'examine différentes questions relatives à l'évolution de populations, en particulier de populations humaines. Ces questions inclusent le lien entre la differentiation génétique et l'homozygotie, comment l'assignement d'individus d'origine inconnue peut être améliorée en combinant des marqueurs génétiques en haplotypes, comment la taille des populations au cours du temps peut être calculée à partir de la distribution des temps de coalescence et comment une population des Andes a répondu génétiquement à la pression de sélection imposée par des milliers d'années de haute teneur en arsenic dans l'eau courante.

Dans une première étude, en particulier, je calcule une borne supérieure et une borne inférieure pour F_{ST}, une mesure classique de différentiation génétique entre populations, comme fonctions de l'homozygotie dans chacune des deux populations étudiées. J'applique ce résultat à des données humaines afin de discuter les niveaux de différentiation observés entre différentes populations. Dans une seconde étude, j'introduis un nouveau critère, intitulé le *Gain d'Information pour l'Assignement* (GIA), pour guider la décision de combiner deux marqueurs génétiques en un haplotype et améliorer l'assignement d'individus d'origine inconnue à un panel de référence de populations. En appliquant cette méthode à des données de SNP sur un échantillon de Français, d'Allemands et de Suisses, je montre comment l'utilisation des haplotypes peut conduire à un assignement plus exacte lorsque les haplotypes sont construits à l'aide de GIA. Dans une troisième étude, je calcule analytiquement la taille de population au cours du temps en utilisant les fonctions de densité des temps de coalescence. Je montre la robustesse de ce résultat mathématique lorsque les densités ne sont plus connues dans leur continuité mais estimées à partir d'un nombre fini de généalogies génétiques et lorsque l'on augmente le nombre d'individus de l'échantillon. À l'aide d'un algorithme d'inférence de généalogies à partir de données de séquences phasées, j'applique la méthode sur des régions non ou peu recombinantes du génome humain pour quatre populations de continents différents. Je retrouve dans le passé ancien des

profils de taille de population similaires à ce qui à été observé à l'aide de méthodes précédemment publiées, mais découvre de nouvelles charactéristiques dans le passé récent, en particulier une séparation ancienne de la population Africaine échantillonnée dans l'étude (Yoruba) des populations non-Africaines, ce qui pourrait suggérer une structure de population ancienne sur le continent Africain. Enfin, dans une quatrième étude, je présente un exemple d'adaptation humaine à un environnement riche en arsenic. Avec nos collaborateurs de l'Institut Karolinska et de l'Université de Lund, nous avons génotypé 124 femmes de San Antonio de los Cobres, un village dans les Andes argentines où des niveaux d'arsenic conséquents ont été mesurés dans l'eau courante alors que les effets sanitaires négatifs typiquements observés pour de tels niveaux ne sont pas aussi prévalents qu'attendu. Ce fait est suggestif de l'évolution d'un mécanisme métabolique pour surmonter les hauts niveaux d'arsenic depuis l'établissement de la population dans la région, il y a plus de 10,000 ans. Nous avons collecté des informations phénotypiques, sous la forme de concentrations en composés organiques et inorganique d'arsenic dans l'urine et nous avons réalisé une Étude d'Association Pangénomique pour détecter les régions du génome associées avec de faibles concentrations en acide monométhylarsonique (une forme organique assez toxique de l'arsenic) et avec de fortes concentrations en acide diméthylarsinique (une autre forme organique de l'arsenic qui est beaucoup moins toxique et plus facile à éliminer de l'organisme). Nous mettons en évidence une claire association avec une region du chromosome 10, en amont du gène *AS3MT*, qui encode pour l'enzyme arsenic (+3 oxidation state) methyltransferase dont le role dans la méthylation de l'arsenic a précédemment été prouvé. En outre, nous avons trouvé des valeurs élevées pour iHS et LSBL dans cette même région, qui sont deux statistiques développées pour détecter la sélection positive. La correspondance claire entre l'association de la region du gène *AS3MT* au phenotype de resistance à l'arsenic et des valeurs élevées des deux statistiques au même endroit du génome tend fortement vers l'hypothèse d'une évolution génétique sur des milliers d'années en réponse à la présence d'arsenic dans l'eau courante. Cette étude présente le seul exemple à nos jours d'adaptation humaine à un environment toxique.

La variété des sujets traités dans les études présentées dans cette thèse témoignent de la richesse des outils que la génétique des populations peut offrir pour étudier l'évolution. En apportant des modèles mathématiques dans l'étude de la diversité génétique, la génétique des populations, avec l'aide du développement téchnologique rapide en matière de génotypage et de séquençage, a propulsé la biologie de l'Évolution dans une nouvelle ère, où de nombreuses hypothèses à caractère écologique ou anthropologique peuvent enfin être testées.

8. Acknowledgements

I have a lot of people to be thankful for, as you can see from the length of this section, but the first person I would like to thank is my main supervisor Mattias. Thank you for believing in me and my work, for your patience against my stubbornness, my precipitancy and my over-emotional self at times, and for teaching me so much about proper scientific writing, I really needed it. It's been hard, but it's been fun too! I add my thanks to my second supervisors Michael and Martin, for good discussions and inspirations. To Noah and Hideki who supervised part of my work as well, thank you for giving me the opportunity to travel and work with you, discover other working environments and scientific focuses, meet numerous interesting people and enjoy life at Stanford and Sokendai.

When I started doing research with Mattias, it was in 2009, I was a master student, doing a research internship for my engineering degree back in Grenoble. The Jakobsson Lab consisted in 3 people back then: Mattias, Sen and I. Boy, has it grown since! I would like to thank all the members of the Huge Jakobsson Lab, past and present, for all these great times and fruitful conversations. I would like to thank in particular Sen and Pontus, as senior PhD students before me, you were great role models and gave me useful tips and tricks. I will not forget those. Agnès, going to work has always been more fun when you were around. Thank you for being your awesome self, so funny, so open and carefree, for being a great friend and confidant. Never forget that I am always so proud of you! Good luck with your endeavors in the last sprint. And keep on the great drawing work, you are truly talented! I should buy one of you custom printed T-Shirts! Carina, it has been really nice to sit in the same office as you, I have been touched by your kindness, impressed by your knowledge at the many places where my own was (and still is) seriously lacking. I am really happy to count you as a friend. Per, as an office mate, you speak way too much on the phone, but you make it up nicely with a great combination of scientific help and entertainment. We had really great times together and I hope we will have many more. Also, thanks for the big help on the thesis! I would like to add thanks to Torsten and Federico, for great career advices and fun conversations, and last but not least TJ for reviewing the kappa with his great language skills.

Sitting in front of my computer all day, one could think that I am always in my own world. That could not be less true! I've been blessed by great company in my offices during all these years! Some were good at bringing whiskey

or rum or beer and spark nice conversations, others lit up the party mood with a bit of Lady Gaga and other pop songs, some brought farting frogs, others boring plants, some shared silly pictures or even sillier videos, others brought gossip and juicy conversations to my ears. I am not gonna say who did what, but they will recognize themselves for sure. Thanks Emma, Nicklas, Carina M., Rebecca, Carina S., Reto, Per, Agnès, Hanna, Joaquin. Thanks for all the fun, and for being quiet most of the time nevertheless...

It is amazing how many great encounters you make during the course of PhD studies! The Department of Evolutionary Biology at EBC is really full of interesting people. I would like to name them all here and thank them for making work so much lively and stimulating, but considering the size of the department (past and present), it would make a paragraph the size of this thesis. However, I would like to specially thank Urban Friberg for his inspiring energy as a teacher and course coordinator. It's been a real pleasure to be teaching assistant under your coordination. Aaron, you have no idea how proud I was that you valued my input and knowledge on ABC techniques and the coalescent. It's been really interesting to be a sort of consultant on your project and gave me a taste for collaboration. I thank you for bringing me in this. Jelmer, thank you for the finishing tips and the helpful information for organizing the end of my PhD. Writing these acknowledgements in advance, I am currently following one of those helpful advices. Nina, Torsten, Willian and Sergio, I am looking forward for more board games evening with you, it's just so much fun! Thanks for letting me win from time to time.

Work is not the only place two colleagues can bond, and babies are especially useful to open to new social circles. Thank you Marta, Cosima, Reto and Frederik for putting up with the concerns, complaints associated with parenting but also sharing the joys and cuteness attacks. It was so nice to hang out around the babies. Special thanks to Maël's best friends Annika and Jakob, you guys are the best! I am looking forward to see you grow.

One cannot survive a PhD in sanity without home support. Johan, I am not sure how I can express the love and the gratefulness I feel when I think about you, but I'll try somehow. You have been here for me (well, not all but still a great deal of my PhD time), putting up with (and I suspect, secretly enjoying) my craziness, stubbornness, pickiness, laziness, pain-in-the-ass-ness, being a great partner, a fantastic father, helping me, supporting me, surprising me. So many things I owe you... I thank you, from the bottom of my heart. I am looking forward with excitement to spend many years with you in the future, living all the great experiences life has in store us. Maël, my son, my favorite person in this entire world. People could say that you have been mostly a hinderance to this entire thesis writing endeavor, by jamming my poor sleep-deprived post-pregnancy-breastfeeding-forgetful brain with pictures of your

irresistible cuteness, but that would not be *entirely* true. The perspective of seeing you earlier everyday after work, if I could finish my daily task early, really boosted my productivity. And who knows, weren't it be for you, I might have been the regular procrastinator I am and finish all this very close to the deadline under a huge amount of stress. But look, your mom is all grown up now, thanks to you! I love you more than I can ever express.

There are also many friends I would like to thank. Flora, I thank you for the great discussions we had, on all possible levels, for the soul searching and fun sharing. Congratulations on your newly obtained position at the CNRS! It is really impressive. Claire and Ola, you are truly amazing friends! I have never met anyone as kind and generous. You've been here for me through good and difficult times. Now I can only repay you by showering the future little Hjerpenetti with lots of love and kisses. I will. Definitely. Colin, it was such a fun surprise that a childhood friend could end up in Uppsala! Thanks for the fun role playing nights and other cool hangouts (in all senses of the term). I thank the Girls-Night-Out crew for making me feel young and fun again: Bong, Agnès, Ling, Valeria. Let's book a another night soon! Thanks to the people of the Swedish crew (that's how I like to call them): Sara, Peter, David & Anna, Karolina, Jonas, Ellinor & Leo, Karl & Emelie, Johan & Malin, it's great fun to hang out with you guys! Sara, thanks for cheering me up when I was down, for a great time together as flatmates and I cannot say enough thanks for introducing me to Johan! Peter, you owe me a dinner! :P Thanks for allowing this crazy French into your life and your circle of friends. I enjoy every time we hang out, you are a lot of fun! David and Anna, you came to visit me when I was in most need of company, you spoke English when I was in most need of speaking English, you organized game afternoons when I was in most need of playing. I am grateful for your kindness and how you welcome me often at your home.

Being together with someone often opens to whole new social perspectives and I made great friends thanks to you Johan. Especially, I would like to thank Anders and Perra, for being around, helping us, entertain us. You should babysit also sometimes. Good practice! Also, Johan's colleagues Kaspars, Victor, Wei, Ling, Eduard, Valeria and Johan, Francisco, you are all so much fun!

Not just friends, I also got a new family here. Carin and Lars-Olov, thank you so much for welcoming me in your family, for giving me a sense of parental home when I cannot get my own and feel the childish need of being spoiled. Your help with Maël has been extremely invaluable. And thanks to you, I have improved my Swedish (though not enough so that I feel comfortable to write this entire paragraph in Swedish... Seems like I should hang out with you some more). Stort tack för allt ni har gjort och kommer att göra

för mig i framtiden.

Mes remerciements partent vers la France maintenant. Merci Arthur et Julien, mes colocs de choc, meilleurs amis de Grenoble... et d'ailleurs, surtout d'ailleurs maintenant. On s'est bien amusés toutes ses années! Ça fait trop longtemps qu'on n'a pas codé ensemble, il faudrait qu'on s'organise un petit TP algo à distance un de ces quatres. On se tient au courant! Agnès (la Niaudet, pas la Sjöstrand), merci pour cette amitié fantastique, la plus longue de mon existence, pour toutes ces après-midis perles ou Tomb Raider, pour le partage de nos joies, de nos difficultés d'enfants et d'ados. Un large pan de ma vie n'aurait vraiment pas été le même sans toi. Je garde précieusement tous ces beaux souvenirs. Julie, la voyageuse, business woman, quand je pense à toi, je voyage entre hotels luxueux, restaurants gastronomiques, plages de sable fin. C'est pas toujours comme ca, je sais (t'as quand même vécu à Nyköping, faut pas l'oublier...), mais j'aime bien rêver. Avec toi, j'ai vécu une aventure incroyable dans la mer Baltique et des vacances fabuleuses en Andalousie. Dis moi quand sont tes vacances, je fais mes sacs et on repart à l'aventure! Merci de me garder comme amie, même si je n'appelle pas souvent. Sami, ça m'a fait drolement plaisir de recevoir la visite de ton auguste personne sur mes terres uppsalaiennes. J'espère pouvoir un jour venir te rendre la pareille à NYC! Merci de me faire profiter de tes voyages via Facebook, ça donne envie... Chloé et Al, merci aussi de votre visite! J'espère que la vie à quatre se passe aussi bien sinon mieux que la vie à trois.

Du côté de ma famille, il y a plein de mercis et de câlins à donner. En particulier, je voudrais faire un petit clin d'oeil à mes soeurs et mon frère. Marie, Amélie, Anaïs, Natacha et Christophe. Je suis contente d'avoir une famille aussi grande, même si toujours compliquée à expliquer au nouveau venu. Tellement de beaux souvenirs nous relient, et malheureusement si peu d'occasions d'en créer de nouveaux... Je vous aime et vous me manquez. Y'en a quelques uns qu'il faudrait quand même qu'ils se mettent à Skype un de ces jours! Ils (Elles ?) se reconnaîtront. Jean-Jacques, sans toi je ne suis pas sûre que j'aurais connu l'ordinateur avant mes 18 ans! Merci donc pour ca, et pour plein d'autres choses encore: ta patience, ta générosité, ta chaleur, ton énergie et ton humour. Tu es une personne que j'estime énormément, humainement, et je te remercie d'avoir été un réel parent pour moi pendant toutes ses années. Et de m'avoir servi de supers apéros. Ma Titi, ma tata Titi, le M. de mon nom sur chacun de mes articles publiés, que te dire sinon que je t'aime, que j'ai chéri et chéris encore toutes ces conversations que nous avons tenues de vive-voix ou par le truchement de quelqu' appareil digital, que tu es une personne formidable, d'intelligence et sensibilité rare, et que j'ai beaucoup appris de toi. Merci pour tout ce que tu me donnes, j'espère pouvoir un jour t'en donner tout autant.

Papa, Maman, être devenue moi-même maman m'a donné beaucoup plus de perspective quant à notre relation et j'ai l'impression de vous avoir redécouverts, sous un jour nouveau. Ça vous va bien d'être Mormor et Morfar! Merci pour les millions de bisous que vous m'avez probablement donné, pour vous être levés la nuit afin de calmer mes angoisses, pour m'avoir laissée grandir à mon rythme, découvrir mon petit monde, pour m'avoir fait confiance dans ma voie mais aussi montré parfois un autre chemin, pour m'avoir toujours aimée sans condition, telle que je suis. Je vous dois, en grande partie, d'être qui je suis, et normalement si tout va bien je suis docteur, ce qui n'est pas si mal (même si c'est le genre de docteur qui ne sert à rien). Je vous aime énormément et je serai toujours profondément reconnaissante pour tout ce que vous avez fait pour moi, dans des circonstances qui n'ont pas toujours été faciles. Beau boulot, votre Lulu est heureuse, trrrllltt trrrllltt ! Papa, je te souhaite le grand succès que tu mérites dans ta carrière de chanteur. Et j'espère que la Suède cette fois-ci t'auras remercié de ta visite avec autre chose que du guano! Hihi! Maman, profite bien de la vie trépidante Toulousaine ainsi que de ta retraite anticipée. On viendra vous voir avec plaisir.

Mon Pilou, tu n'aurais sans doute pas compris ce bouquin, surtout étant donné qu'il est en Anglais, mais tu aurais apprécié son existence. Je pense souvent à toi et je te remercie pour tous les bons souvenirs et les choses merveilleuses que tu m'as apprises. Tu me manques.

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664.

Asante, E. A., Smidak, M., Grimshaw, A., Houghton, R., Tomlinson, A., Jeelani, A., Jakubcova, T., Hamdan, S., Richard-Londt, A., Linehan, J. M., Brandner, S., Alpers, M., Whitfield, J., Mead, S., Wadsworth, J. D. F., and Collinge, J. (2015). A naturally occurring variant of the human prion protein completely prevents prion disease. *Nature*, page 10.1038/nature14510.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773–785.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98(8):4563–4568.

Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.

Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714.

Cann, R., Stoneking, M., and Wilson, A. (1987). Mitochondrial dna and human evolution. *Nature*, 325:31–36.

Cardon, L. R. and Palmer, L. J. (2003). Population stratification and spurious allelic association. *The Lancet*, 361(9357):598–604.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425.

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome research*, 15(11):1496–1502.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature genetics*, 37(11):1243–1246.

Crow, J. F. and Kimura, M. (1970). *An introduction to population genetics theory.* New York, Evanston and London: Harper & Row, Publishers.

Darwin, C. (1859). *On the origin of species by means of natural selection: or the preservation of favoured races in the struggle for life*. John Murray, Albemarle Street.

DeGiorgio, M., Degnan, J. H., and Rosenberg, N. A. (2011). Coalescence-time distributions in a serial founder model of human evolutionary history. *Genetics*, 189(2):579–593.

DeGiorgio, M., Jakobsson, M., and Rosenberg, N. A. (2009). Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from africa. *Proceedings of the National Academy of Sciences*, 106(38):16057–16062.

Dennis, Jr, J. E. and Moré, J. J. (1977). Quasi-newton methods, motivation and theory. *SIAM review*, 19(1):46–89.

Edwards, A. W. (2003). Human genetic diversity: Lewontin's fallacy. *BioEssays*, 25(8):798–801.

Enattah, N. S., Jensen, T. G., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., El-Shanti, H., Seo, J. K., Alifrangis, M., Khalil, I. F., et al. (2008). Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *The American Journal of Human Genetics*, 82(1):57–72.

Ewens, W. (1982). On the concept of the effective population size. *Theoretical Population Biology*, 21(3):373–378.

Excoffier, L. (2008). Analysis of population subdivision. In Balding, D., Bishop, M., and C., C., editors, *Handbook of Statistical Genetics*, pages 980–1020. John Wiley & Sons, Ltd.

Excoffier, L., Estoup, A., and Cornuet, J.-M. (2005). Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, 169(3):1727–1738.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular ecology notes*, 7(4):574–578.

Fisher, R. A. (1919). Xv. the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52:399–433.

Gattepaille, L. M., Jakobsson, M., and Blum, M. G. (2013). Inferring population size changes with sequence and snp data: lessons from human bottlenecks. *Heredity*, 110(5):409–419.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The international hapmap project. *Nature*, 426(6968):789–796.

Gillespie, J. H. (2010). *Population genetics: a concise guide*. JHU Press.

Goldstein, D., Linares, A. R., Cavalli-Sforza, L. L., and Feldman, M. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences*, 92(15):6723–6727.

Griffiths, R. C. and Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 344(1310):403–410.

Gusfield, D. (2003). Haplotype inference by pure parsimony. In *Combinatorial Pattern Matching*, pages 144–155. Springer.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet*, 5(10):e1000695.

Hamblin, M. T. and Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the duffy blood group locus. *The American Journal of Human Genetics*, 66(5):1669–1679.

Hammer, M. F., Karafet, T. M., Redd, A. J., Jarjanazi, H., Santachiara-Benerecetti, S., Soodyall, H., and Zegura, S. L. (2001). Hierarchical patterns of global human y-chromosome diversity. *Molecular Biology and Evolution*, 18(7):1189–1203.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., Pahwa, J. S., Moskvina, V., Dowzell, K., Williams, A., et al. (2009). Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. *Nature genetics*, 41(10):1088–1093.

Harris, K. and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, 9:e1003521.

Harter, A. V., Gardner, K. A., Falush, D., Lentz, D. L., Bye, R. A., and Rieseberg, L. H. (2004). Origin of extant domesticated sunflowers in eastern north america. *Nature*, 430(6996):201–205.

Hartl, D. L. and Clark, A. G. (1997). *Principles of population genetics*. Sinauer associates Sunderland, 4th edition.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108.

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular ecology resources*, 9(5):1322–1332.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338.

Hudson, R. R. and Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics*, 120(3):831–840.

Hudson, R. R., Slatkin, M., and Maddison, W. (1992). Estimation of levels of gene flow from dna sequence data. *Genetics*, 132(2):583–589.

Hur, Y.-M. (2003). Assortative mating for personality traits, educational level, religious affiliation, height, weight, and body mass index in parents of a korean twin sample. *Twin Research*, 6(06):467–470.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003.

Kaplan, N. L., Darden, T., and Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics*, 120(3):819–829.

Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *science*, 336(6082):740–743.

Kimura, M. et al. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626.

Kingman, J. (1982). On the genealogy of large populations. *Journal of Applied Probability*, 19A(1):27–43.

Kitzman, J. O., MacKenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., Ng, S. B., Alkan, C., Qiu, R., Eichler, E. E., et al. (2011). Haplotype-resolved genome sequencing of a gujarati indian individual. *Nature biotechnology*, 29(1):59–63.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068–1075.

Kuhner, M. K., Beerli, P., Yamato, J., and Felsenstein, J. (2000). Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, 156(1):439–447.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, page gkp552.

Lewontin, R. (1972). The apportionment of human diversity. In Dobzhansky, T., Hecht, M., and Steere, W., editors, *Evolutionary Biology*, pages 381–398. Springer US.

Lewontin, R. C. and Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. ii. amount of variation and degree of heterozygosity in natural populations of drosophila pseudoobscura. *Genetics*, 54(2):595.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.

Lynch, M., Walsh, B., et al. (1998). *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Macgregor, S., Cornes, B. K., Martin, N. G., and Visscher, P. M. (2006). Bias, precision and heritability of self-reported and clinically measured height in australian twins. *Human genetics*, 120(4):571–580.

Manolio, T. A. (2013). Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics*, 14(8):549–558.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913.

Marjoram, P. and Wall, J. D. (2006). Fast "coalescent" simulation. *BMC genetics*, 7(1):16.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369.

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet*, 5(10):e1000686.

McVean, G. A. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393.

Naduvilezhath, L., Rose, L. E., and Metzler, D. (2011). Jaatha: a fast composite-likelihood approach to estimate demographic parameters. *Molecular Ecology*, 20(13):2709–2723.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12):3321–3323.

Nei, M. (1986). Definition and estimation of fixation indices. *Evolution*, pages 643–645.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.*, 39:197–218.

Nordborg, M. (2001). Coalescent theory. In *Handbook of statistical genetics*. Wiley Online Library.

Norton, H. L., Kittles, R. A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V. A., Bradley, D. G., McEvoy, B., and Shriver, M. D. (2007). Genetic evidence for the convergent evolution of light skin in europeans and east asians. *Molecular biology and evolution*, 24(3):710–722.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., et al. (2008). Genes mirror geography within europe. *Nature*, 456(7218):98–101.

Olby, R. (1989). The dimensions of scientific controversy: the biometric-mendelian debate. *The British Journal for the History of Science*, 22(03):299–320.

Perry, G. H. and Dominy, N. J. (2009). Evolution of the human pygmy phenotype. *Trends in Ecology & Evolution*, 24(4):218–225.

Polanski, A., Bobrowski, A., and Kimmel, M. (2003). A note on distributions of times to coalescence, under time-dependent population size. *Theoretical population biology*, 63(1):33–40.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., et al. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459):471–475.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463.

Price, T. and Schluter, D. (1991). On the low heritability of life-history traits. *Evolution*, pages 853–861.

Pritchard, J. K. and Di Rienzo, A. (2010). Adaptation–not by sweeps alone. *Nature Reviews Genetics*, 11(10):665–667.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Provine, W. B. (2001). *The origins of theoretical population genetics: With a new afterword*. University of Chicago Press.

Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M. C., Malaspinas, A.-S., et al. (2015). Genomic evidence for the pleistocene and recent population history of native americans. *Science*, 349(6250):aab3884.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PloS Genetics*, 10(5):e1004342.

Rosenberg, N. A., Burke, T., Elo, K., Feldman, M. W., Freidlin, P. J., Groenen, M. A., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., et al. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics*, 159(2):699–713.

Rosenberg, N. A., Li, L. M., Ward, R., and Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics*, 73(6):1402–1422.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *science*, 298(5602):2381–2385.

Ruff, C. B. (1994). Morphological adaptation to climate in modern and fossil hominids. *American Journal of Physical Anthropology*, 37(S19):65–107.

Ruzzante, D. E., Walde, S. J., Gosse, J. C., Cussac, V. E., Habit, E., Zemlak, T. S., and Adams, E. D. (2008). Climate control on ancestral population dynamics: insight from patagonian fish phylogeography. *Molecular Ecology*, 17(9):2234–2244.

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(6909):832–837.

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. *The American Journal of Human Genetics*, 78(4):629–644.

Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature genetics*.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium and others (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature genetics*, 43(10):969–976.

Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G., et al. (2012). Genomic variation in seven khoe-san groups reveals adaptation and complex african history. *Science*, 338(6105):374–379.

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F., and Alm, E. J. (2012). Population genomics of early events in the ecological differentiation of bacteria. *science*, 336(6077):48–51.

Sheehan, S., Harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics*, 194(3):647–662.

Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., Akey, J. M., and Jones, K. W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal snps. *Human Genomics*, 1(4):274.

Sjödin, P., Kaj, I., Krone, S., Lascoux, M., and Nordborg, M. (2005). On the meaning and existence of an effective population size. *Genetics*, 169(2):1061–1070.

Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., Gilbert, M. T. P., Götherström, A., and Jakobsson, M. (2012). Origins and genetic legacy of neolithic farmers and hunter-gatherers in europe. *Science*, 336(6080):466–469.

Smith, N. G., Webster, M. T., and Ellegren, H. (2002). Deterministic mutation rate variation in the human genome. *Genome Research*, 12(9):1350–1356.

Stajich, J. E. and Hahn, M. W. (2005). Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*, 22(1):63–73.

Steinrücken, M., Paul, J. S., and Song, Y. S. (2013). A sequentially markov conditional sampling distribution for structured populations with migration and recombination. *Theoretical population biology*, 87:51–61.

Suk, E.-K., McEwen, G. K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D. T., McLaughlin, S., Peckham, H., et al. (2011). A comprehensively molecular haplotype-resolved genome of a european individual. *Genome research*, 21(10):1672–1685.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical population biology*, 26(2):119–164.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., and Schork, N. J. (2011). The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223.

The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

Tognetti, A., Berticat, C., Raymond, M., and Faurie, C. (2014). Assortative mating based on cooperativeness and generosity. *Journal of evolutionary biology*, 27(5):975–981.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era – concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS biology*, 4(3):446.

Wang, D. G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082.

Wang, L. and Xu, Y. (2003). Haplotype inference by maximum parsimony. *Bioinformatics*, 19(14):1773–1780.

Wang, S., Lewis Jr, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M. V., Molina, J. A., Gallo, C., et al. (2007). Genetic variation and population structure in native americans. *PLoS Genetics*, 3:2049–2067.

Wang, S., Ray, N., Rojas, W., Parra, M. V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., Hurtado, A. M., et al. (2008). Geographic patterns of genome admixture in latin american mestizos. *PLoS Genet*, 4(3):e1000037.

Watterson, G. et al. (1962). Some theoretical aspects of diffusion theory in population genetics. *The Annals of Mathematical Statistics*, 33(3):939–957.

Wedekind, C., Seebeck, T., Bettens, F., and Paepke, A. J. (1995). Mhc-dependent mate preferences in humans. *Proceedings of the Royal Society of London B: Biological Sciences*, 260(1359):245–249.

Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *evolution*, pages 1358–1370.

Wetterstrand, K. A. (2014). Dna sequencing costs: Data from the nhgri genome sequencing program (gsp). `www.genome.gov/sequencingcosts`. Accessed: 2015-06-10.

Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H., and Reich, D. (2012). Phasing of many thousands of genotyped samples. *The American Journal of Human Genetics*, 91(2):238–251.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329(5987):75–78.

Yu, N., Jensen-Seaman, M. I., Chemnick, L., Ryder, O., and Li, W.-H. (2004). Nucleotide diversity in gorillas. *Genetics*, 166(3):1375–1383.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1280

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-260998

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2015