Università degli Studi di Napoli "Federico II"

Dottorato in Biologia Computazionale e Bioinformatica

XXIV ciclo

Complex diseases: a genome-wide assessment of the role of selective pressure on the human genome

Tutor: Prof. Gennaro Miele
Co-tutor: Prof. Sergio Cocozza

Examinee:
Roberto Amato

November 2011

Contents

List of Figures

List of Tables

Abstract

Genetic based diseases are commonly thought as an "error", i.e. as the result of one or few rare mutations in the DNA. While this explanation can work fine for Mendelian, high penetrance, diseases, it is less plausible for complex diseases (CD). Indeed, this important class of diseases is characterized by the fact of being caused by hundreds of variants, each of them "common" in the population and with a small effect. As a possible explanation for such an apparent paradox, it has been proposed that variants associated with CD are the result of direct or indirect evolutionary pressures in ancient times. According to this hypothesis, those variants (or variants close to them) were selected in our ancestors for being advantageous and that they became dangerous only recently because of the totally different environment we live in. Then, the very recent changes in the environment, together with the late onset characterizing these diseases, provided no time for natural selection to act against them.

As a first step toward addressing this hypothesis, I analyzed the genomic distribution of a specific marker of selective pressure, namely F_{ST} . I examined, in particular, its relationship with genes associated to human CD finding indeed suggestions of positive selection occurred on them.

To better understand the role of natural selection on genes associated with CD I then focused on two different cases of study corresponding to two different scenarios. In the first one, I found hints for schizophrenia to be, at least partially, a maladaptive

by-product of natural selection which acted on vitamin D related genes when first humans moved to higher latitudes.

In the second example, I focused on the case in which variants increasing the risk for autoimmune diseases are most likely to be the actual and direct target of selection. I found that a plausible hypothesis is that the diseases themselves are the results of an environmental mismatch with respect to that our ancestors, when these variants were selected.

Chapter 1

Complex diseases in an evolutionary perspective

In the past years, the main contributions of genetic to medicine were directed towards the finding of visible chromosomal defects and mutations in genes that interfere with the specific function of a single gene and thereby cause "Mendelian" or "monogenic" diseases. The result of this big effort is such that, to date, more than 4,000 disorders are known and, for the majority of them, the molecular mechanism is known as well (`www.omim.org`). Many reasons account for this tendency, not last the technologies not available up to few years ago and the relative simplicity (compared to other phenotypes) of the study of these disorders. Mendelian diseases, indeed, although rare are characterized by simple and clear-cut pattern of transmission and are usually due to one or few highly penetrant and strongly deleterious alleles that segregate in families.

However, the vast majority of human diseases shows incomplete pattern of penetrance and they are usually caused by many genetic and environmental factors and by (non-linear) interactions among them. To this class of disorders, known as "complex diseases" (CDs), belong those pathologies with the highest prevalence in the

human being, such as cancer, cardiovascular diseases, asthma, diabetes mellitus, etc [1].

At least until few years ago, the search for genetic mechanisms underlying CDs was not successful as it was for Mendelian ones [2]. One of the most important reason is that, because of their intrinsic multi-factoriality, each allele is expected to influence the disease risk only in a small amount and this can explain why traditional methods for mapping (e.g. linkage analysis) work poorly. But, with the introduction of high-density single nucleotide polymorphisms genotyping platforms, the scenario became different. Genome-wide association studies (GWAS) allowed to unveil many, well replicated, risk loci associated to common diseases [3]. Moreover, the decrease of their cost and the possibility of studying even more individuals, also allowed to identify very subtle signals in a relatively cheap and quick way [4].

A deeper look at these results, however, unveil a subtle evolutionary paradox. Power of GWAS strongly depends on the frequency of the variant in the sample and in the population [5]. Indeed, variants which are rare in the population could not even be present on the array while those that have low frequency in the sample are usually unable to reach a genome-wide significance threshold. The high consistency of many of these results, and consequently the likelihood of them being true, supports the evidence that susceptibility variants are "common" (frequency $> 5\%$) in the population. It is worth stressing that this doesn't means that all the variants affecting the risk for common diseases needs to be common. Unfortunately, because of the limits of current technologies, right now this is the range of frequencies we can more comfortably focus on [6].

Even under the conservative hypothesis that they represent just a piece, more or less big, of the genetic puzzle explaining CDs, the contribution of those variants is non negligible. In one of the first large scale GWAS, for example, Zeggini and colleagues found about 10 variants associated with type 2 diabetes (T2D) mellitus which frequencies range from 15% up to 45%[7] in a UK samples, and in general

few variants associated with T2D show an allelic frequency $< 10\%$. Five variants strongly associated with age-related macular degeneration risk ($\sim 50\%$ of explained variance) shows frequencies above 20% [8].

Overall, the average minor allele frequency of CD-risk variants is about 30% [9]. If we instead look at genetic variation responsible for Mendelian diseases they are essentially rare ($\ll 1\%$). From an evolutionary perspective, indeed, it is reasonable that variants providing strongly deleterious phenotypes, as many Mendelian disorders are, are likely to be removed by purifying selection. But, the fact that a similar reasoning doesn't work for CDs, led to the question which represents the aim of this thesis: *why mutations that increase the risk for a disease should be present at such an high frequency in the human population?*.

Several explanations can be hypothesized. First of all, it cannot be excluded that common variants are just proxies for a set of surrounding rare mutations. Moreover it should be kept in mind that CDs can be sometimes characterized by milder phenotypes, with late onset and a modest or null impact on reproductive fitness. However, in some cases both this assumptions have been showed to be false and some more advanced and general explanations are needed. One of the most intriguing is the idea that natural selection played, in the past, an important role. Indeed, many mechanisms underlying CDs become clearer and more understandable if looked from an evolutionary perspective. Aim of this work will be to clarify how selective pressure could have played a role in the evolution of CDs, to assess on a genome-wide scale its role, and to exhibit some concrete examples relative to different models.

As an example of how studying and understanding evolution has an even more important impact in medicine and, eventually, on health, it is worth to remark that the National Academy of Sciences of USA held on April 2009 in Washington DC one of its famous Arthur M. Sackler colloquia on the topic "Evolution in Health and Medicine" [10].

1.1 Evolutionary models for CDs

The most basic mechanism one can imagine acting on diseases is purifying or negative pressure. As discussed for Mendelian diseases, indeed, one should expect that disease alleles are removed from the population due to their deleterious effect. At the same time, the mutational process re-introduce new disease alleles. At a certain point, these two forces will reach a balance (*mutation–selection balance*) between the input of new mutations and purifying selection that removes them [11] and susceptibility allele (if rare) will reach an equilibrium frequency approximately equal to the mutation rate divided by the selection against the allele in heterozygotes. Under this model, traits are expected to feature an high levels of allelic heterogeneity and a low prevalence, as it is the case for Mendelian diseases [12].

There are some cases in which also Mendelian disorders are caused by mutations with a frequency ($> 1\%$). This is the famous case, for example, of β-globin defects that lead to sickle cell anaemia in homozygotes individuals but protect heterozygotes against malaria. In this case is reasonable to hypothesize that the allele is kept in the population because of a *balancing selection* acting on the heterozygotes.

In 2001, Pritchard investigated the effect of the mutation–selection balance in the framework of CDs [13]. In this work, he simulated under this model two sets of neutral and slightly deleterious susceptibility alleles (e.g. with a late and an early onset effect) contributing susceptibility to a disease and looked at their overall frequency. Interestingly, while in the neutral case susceptibility alleles tend to be either rare or close to fixation, in the selected case the probability of having alleles at an intermediate frequencies was considerably higher. This result seems to suggest that CDs are unlikely to be due to selectively neutral alleles.

More interesting for their usefulness in shedding light on the biological mechanisms leading to a disease, and central focus of this thesis, are the cases in which susceptibility alleles for CDs underwent a positive selection. This apparently paradoxical

cases, that as we will see in **Chapter 2** are not unusual, can be explained by different general models.

The first and more simple scenario is the case in which one gene, one allele or linked alleles, have an antagonistic pleiotropic effect and one of the phenotypes has a beneficial effect in terms of fitness which overcomes (or reaches a trade-off) the detrimental effects of another trait. For example, the gene *p53* helps, in human, to prevent cancer by preventing cells with DNA damages from dividing, but it can also suppress the division of stem cells, which allows the body to renew and replace deteriorating tissues during aging [14]. In **Chapter 3** I will present one of this cases of "indirect" positive selection in which genes increasing the risk for schizophrenia also have a beneficial effect of vitamin D metabolism. Indeed, while vitamin D deficiency can severely compromise the capability of individuals to reproduce (causing for example rickets and pelvic deformities), schizophrenia has a late onset and a scarce influence on reproductive fitness.

A similar case is the one in which different adaptations conflict, which requires a compromise between them to ensure an optimal cost-benefit trade-off. This is, for example, the very well known case of skin pigmentation, which needs to guarantee protection from UV (favored in dark skins) and at the same time allow the skin synthesis of vitamin D (favored in light skin).

All the models discussed so far assume that the selective force is constant and doesn't change in the time. However, this assumption is in many case far from being true both because environment changes over time and because our ancestors moved in several moments of our history to different places finding different living conditions.

A first scenario, of "spatially varying selection", is the one in which peoples adapted to environmental conditions specific for one place and, after moving to different places and different living conditions, the variants that so far provided an advantage became deleterious. One of the best known example is the "sodium retention"

hypothesis, proposed to try to explain the inter-ethnic differences in the the prevalence of hypertension. The idea is that the ancestral populations living in equatorial Africa adapted to the hot and humid climate thanks to an increase of the rate of sodium retention. When then peoples moved to cooler and drier climate, those variants lost their advantageous and became even detrimental making individuals more prone to hypertension.

Conceptually similar is the case of "timing varying selection", according to which populations adapted to specific environment and life styles and, when these changed, variants conferring adaptation became dangerous. Stated in a different way, this hypothesis asserts that CDs, or at least a subset of them, could be the result of an environmental mismatch between our ancestors' and ours life styles. The impact of the selective shifts resulting from these transitions has been formalized, for type 2 diabetes and obesity, by Neal in 1962 and is known as "thrifty genotype". During the hunter-gatherer period, populations underwent continuous cycles of feast and famine. In such environment, it is easy to figure out that genetic variants allowing a better and more efficient fats and carbohydrates storage were extremely advantageous, increasing the chances for individuals to survive and reproduce. As the transition from that life style occurred to more reliable food sources and different dietary patterns, that "thriftiness" became detrimental.

In both the previous scenarios of varying selection, there is a relatively recent change in the selective pressures acting on biological processes responsible for maintaining the correct balance between the organism and its environment. The recent environmental change disrupts this balance leading, in turn, to new detrimental phenotypes since it was too rapid to allow the gene pool to re-adapt accordingly. According to this model it is thus plausible that the ancestral version of susceptibility alleles to CDs should reflect adaptations in early populations, when these alleles were maintained by purifying selection [15].

I will discuss in **Chapter 4** the case of autoimmune diseases. It has been hypothe-

sized, indeed, that variants increasing the risk for autoimmune diseases were actually advantageous in environments with high pathogens load and variability. Today, the higher sanitary conditions typical of industrial societies "transformed" that ancestral advantage in an increased risk for autoimmune diseases. It will be also clear that being able to estimate the timing when selection occurred/started helps in testing this hypothesis and in understanding how those big changes in our ancestors living styles had an impact today on CDs risk.

In the entire work I will assume that selection acted on a single gene at a time. However, this is a reductive hypothesis which will discard all the cases in which the change in the fitness is due to many genes, each of them with a moderate to small effect. The search for signatures of "polygenic adaptation" is still at the very beginning and in **Chapter 5** I will present one of the first case of study.

Results presented in chapters 2, 3 and 5 are published on scientific journals and the respective references are reported at the beginning of each chapter. Results discussed in chapter 4, instead, are in part preliminary and the manuscript is in preparation. For this last project I want to acknowledge the principal investigator Anna Di Rienzo from the University of Chicago and all the peoples in her group. I want to specifically acknowledge Gorka Alkorta-Aranburu, which was responsible for all the biological and experimental aspects of the project, and David Witonsky, which worked on all the preliminary aspects, for their constant support and the interesting discussions. I also want to thank Dick Hudson and Molly Przeworski from the University of Chicago for their illuminating insights.

Chapter 2

A genome-wide assessment of selective pressure on human genome and its relationship with diseases

Genetic differences are present in humans at both individual and population level. Human genetic variations are studied for their evolutionary relevance and for their potential medical applications. This studies can help scientists in understanding ancient human population migrations as well as how selective forces act on the human being [16, 17].

According to the theory of neutral variation, most of the genetic variability within species are caused by random drift of selectively neutral polymorphic alleles [18]. Genetic drift should affect all loci across the genome in a similar manner. Therefore, when a locus shows extraordinary high or low levels of variability this may be interpreted as evidence for natural selection [19]. High levels of population differentiation can suggest the acting of a positive selection of advantageous alleles in one or more populations. On the contrary, lower levels of population differentiation can be considered as the effect of balancing selection that tends to maintain a constant proportion of alleles across all populations [20].

Population differentiation is sensitive to a variety of demographic factors (including the rate of drift within populations and the extent of gene flow among them), making it difficult to rule out demographic scenarios that could account for the observed variations. Another class of tests is aimed to detect signature of natural selection by comparing data from different species. These tests explore the fact that mutations can be synonymous and non synonymous, and that non-synonymous mutations are more likely to have an effect on individual fitness. This method is also known as d_N/d_S. Results obtained by this comparative approach are rarely interpreted in terms of population genetics theory [21].

The human population is also not homogeneous in terms of disease susceptibility. Risks of common diseases are substantially different among ethnic groups [22]. The understanding of population genetic differentiation, especially in genes associated with diseases, can help to explain the observed variations in the prevalence of diseases. It is not difficult to forecast that, in the future, genetic structure of populations can be used in public health management [23]. Moreover, natural selection on genes that underlie human disease susceptibility has been invoked. In this framework, ancestral alleles reflect ancient adaptation. With the shift in the environment, these alleles increase the risk for common diseases [15].

Different strategies to quantify the population genetic differentiation have been elaborated [24–30]. One of the most used is a measure devised by Wright and known as fixation index, or F_{ST} [31, 32], which is the amount of genetic variation among groups relative to a panmictic state. As a test of selection, observed F_{ST} values are compared to those expected under neutrality. The main difficulty of this approach is to determine the distribution of F_{ST} values under neutrality [24]. Recently, however, the abundance of genetic data available allows the creation of an empirical genome-wide distribution to be used for the comparisons. Rather than statistically testing specific loci, we can use their position relative to this distribution to gain insights about their selective histories. In addition, the abundance of information about variability of many genes makes it possible to analyze not only single genes,

but also sets of functionally related genes. International HapMap Project [33] by supplying data of a large number of Single Nucleotide Polymorphisms (SNPs) across many human populations, is providing an exceptional tool for studying the genetic structure of human populations.

In the present work we report the results of a genome-wide estimation of F_{ST} on $3,917,301$ SNPs from the latest release of HapMap data. Our results show a heterogeneous distribution of F_{ST} values among genomic regions. Furthermore, we studied the relationship between F_{ST} and an evolutionary measure obtained by a comparative interspecific approach. We applied a gene set approach, widely used for microarray data, to detect biochemical pathways under selection. Finally, we detected a signature of selection within genes associated with complex diseases. [1]

2.1 Results

Using F_{ST} , we estimated populations differentiation for $3,917,301$ SNPs in population samples from the International HapMap Project data (Public release 27, merged II + III). To retain the largest number of SNPs broadly reflecting a continental subdivision, we used data from Yoruba (Africa), Japanese (Asia), Han Chinese (Asia) and CEPH (European descendant) individuals. Combining data from these populations we were able to compare the largest set of genotyped SNPs up to now available. We pooled Japanese and Han Chinese samples due to their geographical closeness. Furthermore, this pooling allowed us to compare our data with previous studies [25, 34]. F_{ST} was estimated according to Weir and Cockerham [32, 35].

After exclusion for Minor Allele Frequency (MAF), we obtained a final SNP sample

[1] The results presented in this chapter are published in: Amato R, Pinelli M, Monticelli A, Marino D, Miele G and Cocozza, S. (2009) Genome-wide scan for signatures of human population differentiation and their relationship with natural selection, functional pathways and diseases. PLoS ONE 4(11): e7927.

of $2,125,440$ SNPs. The mean F_{ST} was 0.122 (SE $= 5 \times 10^{-5}$, median = 0.091, interquartile range = 0.131; see Table 2.1 for more detailed statistics). Figure 2.1 shows distribution of F_{ST} values for each chromosome. The median F_{ST} values of SNPs on the autosomal and sexual chromosomes were statistically different (Kruskal-Wallis test, p-value $< 10^{-16}$). The median F_{ST} values for X and Y chromosomes were 0.129 (mean = 0.174) and 0.676 (mean = 0.606) respectively and were notably higher than those of autosomal chromosomes. Also medians between autosomal chromosomes showed significant differences, but in a very small range of values (median range = 0.084 to 0.098).

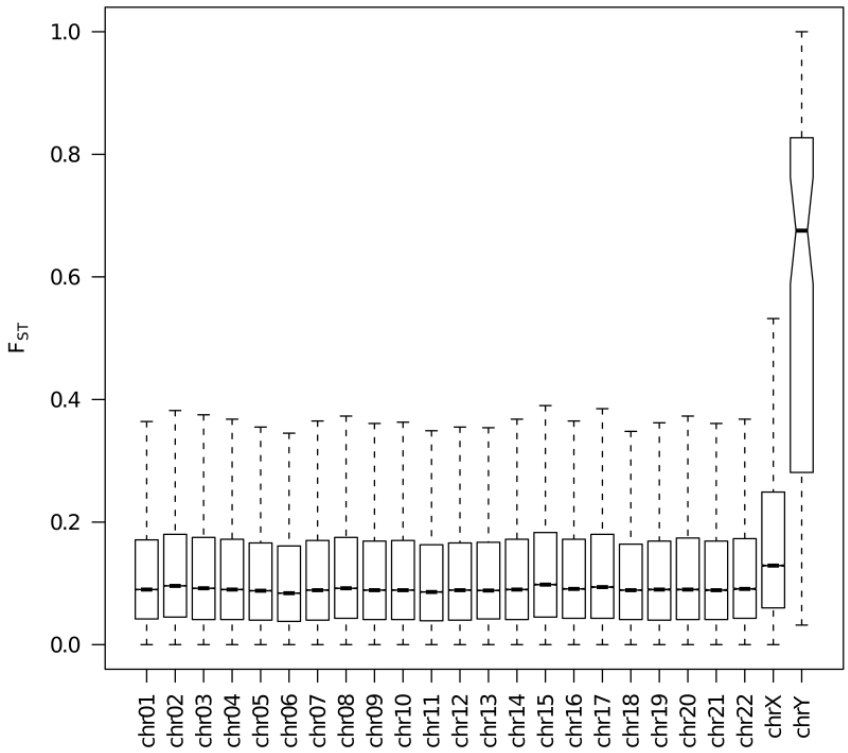

Figure 2.1: **Distribution of F_{ST} values across chromosomes.** For each chromosome, the box length is the interquartile range while the horizontal line inside it is the value of the median. The whiskers extend to the most extreme data point < 1.5 times the interquartile range from the box. Extremes of the notches represents 95% confidence interval of the median.

Chr	Number of SNPs	Mean	SD	Median	IQR
1	160631	0.121	0.110	0.090	0.129
2	182756	0.127	0.114	0.096	0.135
3	141273	0.123	0.111	0.092	0.134
4	131470	0.122	0.112	0.090	0.131
5	138353	0.118	0.107	0.088	0.126
6	142779	0.114	0.104	0.084	0.123
7	114767	0.120	0.110	0.089	0.130
8	119383	0.123	0.110	0.092	0.132
9	97952	0.120	0.108	0.089	0.128
10	110485	0.120	0.110	0.089	0.129
11	106880	0.115	0.105	0.086	0.124
12	96493	0.119	0.108	0.089	0.126
13	83084	0.117	0.104	0.089	0.125
14	67477	0.120	0.107	0.090	0.131
15	59040	0.130	0.116	0.098	0.138
16	59604	0.121	0.108	0.091	0.129
17	50163	0.127	0.114	0.094	0.137
18	62453	0.116	0.103	0.089	0.123
19	32325	0.119	0.109	0.090	0.129
20	51481	0.123	0.113	0.090	0.133
21	28532	0.120	0.110	0.089	0.128
22	26757	0.122	0.109	0.091	0.130
X	61204	0.174	0.153	0.129	0.189
Y	98	0.606	0.282	0.676	0.528
Overall	2125440	0.122	0.111	0.091	0.131

Table 2.1: **Per-chromosome statistics of $\mathbf{F_{ST}}$ values**

For each chromosome, we computed the correlations of all pairs of F_{ST} values for neighbouring SNPs separated by a fixed number of SNPs (1 to 30). This method is commonly used to assess whether F_{ST} values are non randomly distributed across chromosomes [19, 36]. As expected, we found that correlation plots are different from those expected from a noisy signal (Figure 2.2). Moreover, scrambling F_{ST} values across each chromosome produced vanishing correlation values demonstrating that the distribution of data is non-random (data not shown). This result was also supported by a test for non-randomness of data (Ljung-Box test, p-value $< 10^{-16}$). Figure 2.2 shows a clear difference between correlation plots of autosomal and X-linked SNPs, the latter showing higher autocorrelation values. Chromosome Y was excluded from this analysis because of the small number of SNPs sampled.

To attribute F_{ST} value to genes we followed the approach by Akey et al. and Pikrell et al. [19, 30, 36], considering F_{ST} of a gene the maximum F_{ST} value in the gene region (see 2.3). It is worth stressing that this approach is very conservative for genes with low F_{ST} values.

Selection affects both interspecific (between-species) and intraspecific (within-species) variability. F_{ST} is a measure of intraspecific variability. Estimation of genic d_N/d_S is an interspecific measure of variability [21]. We compared the gene F_{ST} values that we obtained with previously reported data from a genome-wide estimation of genic d_N/d_S [37]. In that article the authors divided genes into subgroups with strong, weak and no evidence of positive selection. We compared F_{ST} values of genes belonging to these subgroups. Genes with both weak and strong evidence of positive selection showed lower F_{ST} values than genes with no evidence of positive selection (ANOVA, p-value < 0.001; Bonferroni post-hoc, no evidence vs. weak evidence p-value < 0.02, no evidence vs. strong evidence p-value < 0.005, weak evidence vs. strong evidence = N.S.; Figure 2.3).

To identify functions potentially under selective pressure, we used an innovative approach, focusing on gene pathways instead of outliers. We performed this "gene

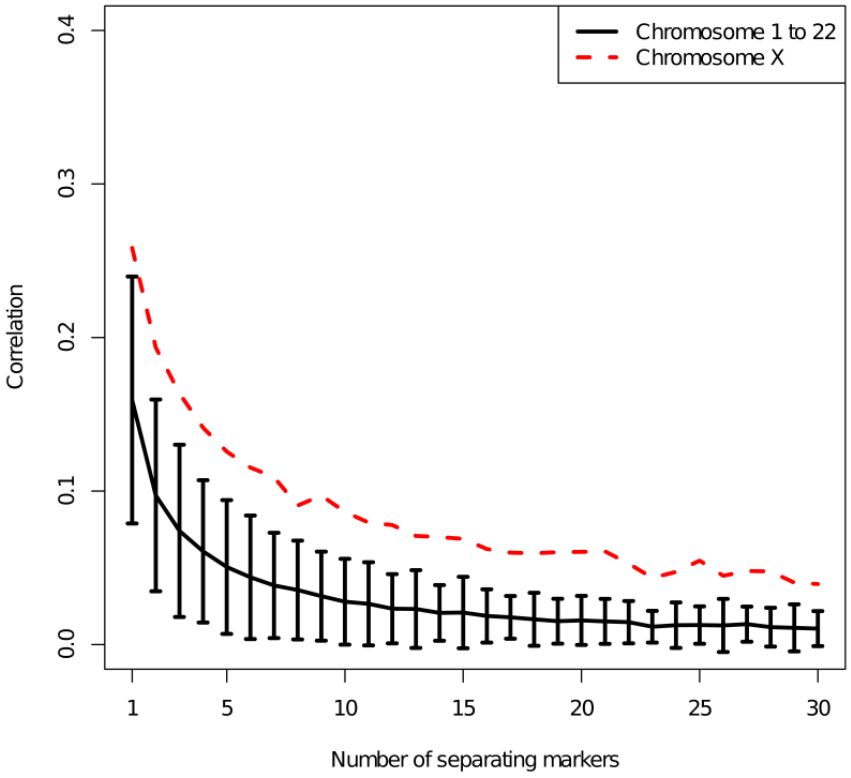

Figure 2.2: **Correlation between F_{ST} values.** The correlation is calculated, for each chromosome, for all pairs of SNPs separated by a fixed number of intervening SNPs. Black line shows mean value and 2σ error bars of the correlation of SNPs belonging to autosomal chromosomes. Red line shows correlation among X-linked SNPs.

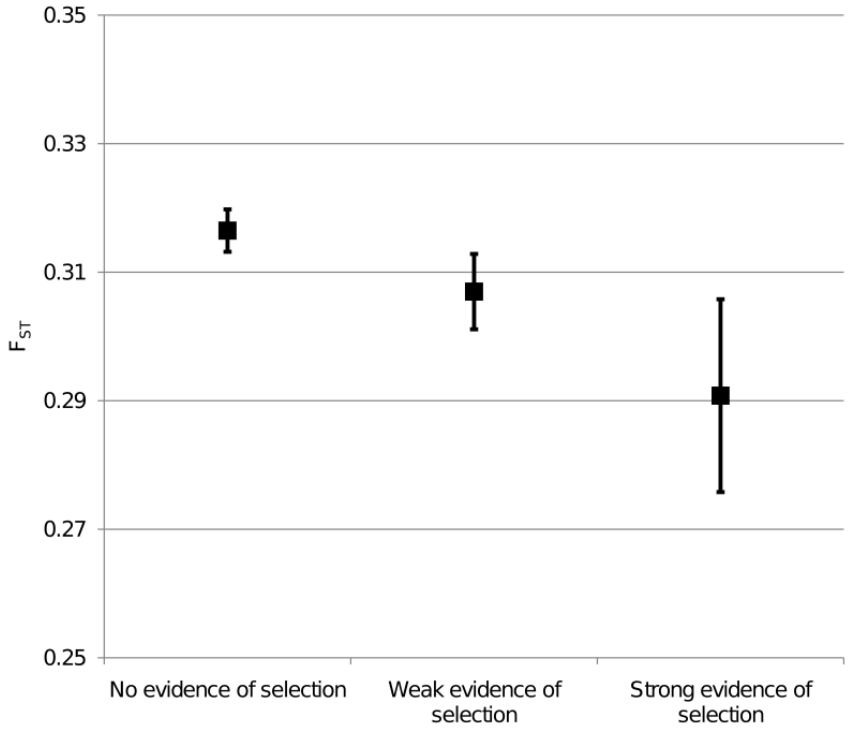

Figure 2.3: **Mean F_{ST} value of genes with and without interspecific evidence of positive selection.** Genes were grouped according to the strength of evidence of their positive selection across six species [37]. Vertical bars represent 95% confidence interval.

Pathway	Name	KEGG ID	Size	FDR
Enriched by high F_{ST} genes				
	Axon guidance	HS04360	126	< 0.001
	Focal adhesion	HS04510	194	0.008
	ECM receptor interaction	HS04512	85	0.009
	Regulation of actin cytoskeleton	HS04810	199	0.010
	Adherens junction	HS04520	75	0.010
	Calcium signaling pathway	HS04020	168	0.010
Enriched by low F_{ST} genes				
	Antigen proc. and presentation	HS04612	70	0.001

Table 2.2: **Leading edge genes of the high F_{ST} enriched KEGG pathways identified by GSEA.** For each pathway is showed the name, the KEGG ID, the number of genes included in the pathway and the p-value after the False Discovery Rate (FDR) correction.

set" analysis using the Gene Set Enrichment Analysis (GSEA) algorithm [38, 39]. GSEA is oriented to identify sets of functionally related genes and is currently used in the analysis of microarray data. Screening the KEGG pathway database by GSEA, we identified 6 KEGG pathways enriched by genes with high values of F_{ST} and one pathway enriched by genes with low values of F_{ST} (Table 2.2). In this method, the enrichment of a pathway is mainly driven by a group of genes that are called "leading edge genes" (see 2.3). Figure 2.4 shows the leading edge genes for the six pathways with high F_{ST} values. A partial overlap of genes among pathways is present.

We then studied populations differentiation of genes associated with complex diseases. We used the Genetic Association Database (GAD) to select genes annotated

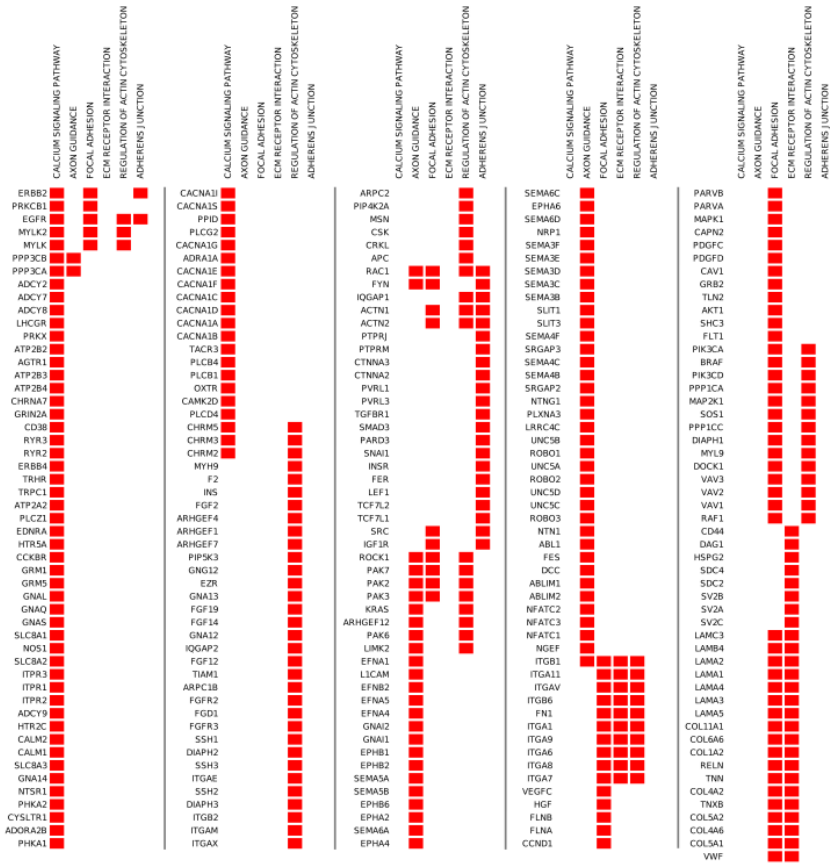

Figure 2.4: **Leading edge genes of the high F_{ST} enriched KEGG pathways identified by GSEA.** Genes are indicated by gene symbols. Red box marks the presence of that gene, as leading edge gene, in that pathway.

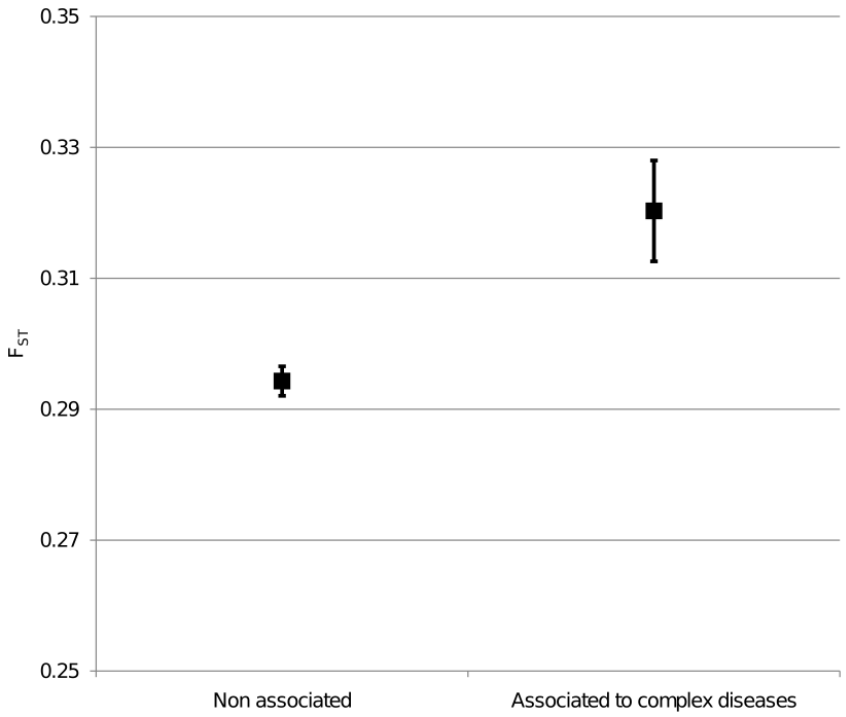

Figure 2.5: **Mean F_{ST} value of genes associated to complex diseases.** Genes found positively associated with complex diseases according to the Genetic Association Database are compared with the remaining ones. Vertical bars represent 95% confidence interval.

as having positive association with complex diseases. We compared F_{ST} values of these genes with those where no association had been positively found. Genes associated with complex diseases showed a significant higher mean value of F_{ST} (t-test, p-value < 0.001; Moving Block Bootstrap, empirical p-value $= 0.0005$; Figure 2.5). Then, we divided diseases in subgroups according to the GAD classification of diseases. Figure 2.6 shows that large differences of F_{ST} values exist among disease classes, while mean F_{ST} values are usually higher than those of non associated genes.

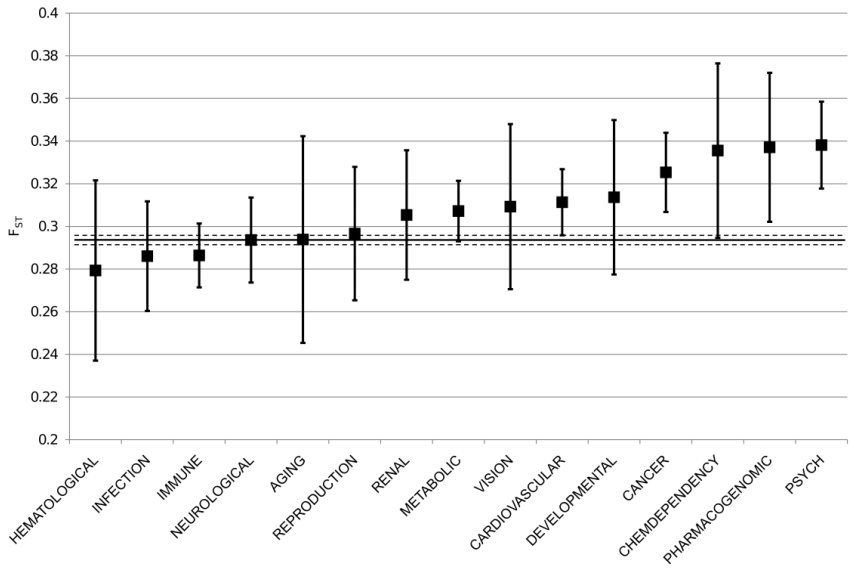

Figure 2.6: **Mean F_{ST} values of genes in different disease classes.** Genes were grouped according to the diseases classification of Genetic Association Database. Vertical bars represent 95% confidence interval. Horizontal solid and dashed lines represent mean value and 95% confidence interval of the set of non associated genes.

2.2 Discussion

The study of the evolutionary forces acting in diseases and physiological traits is an exciting field that may drive further researches and, in the future, public health policies. The study of population genetic differentiation could help the understanding of human evolution, demographic history and disease susceptibility [40]. To study population differentiation we performed a genome-wide F_{ST} calculation using the latest available data release from the HapMap. Using this release we were able to increase both the number of SNPs and the number of individuals analyzed in comparison to recent analogous studies [29]. We focused on samples from three different continents (Africa, Asia, Europe) to obtain a broad but sound measure of populations differentiation.

We found an overall mean F_{ST} value (0.122) broadly consistent with previous estimations [19, 29, 36]. The slightly higher value that we obtained could be explained by the exclusion of SNPs with MAF < 0.05 and the inclusion of heterochromosomes in the calculation. Indeed, as expected [19], we observed a significantly higher median F_{ST} value of X-linked SNPs with respect to the autosomal ones. Furthermore, we found median F_{ST} value of Y-linked SNPs to be significantly higher than both the autosomal and the X-linked ones. Previous data from smaller datasets suggested a similar phenomenon [41], but, in our knowledge, this is the first observation made on Y chromosome F_{ST} in a more robust framework. The higher population differentiation for X and Y chromosomes can be due to various causes: their smaller effective population size (three-quarter and one-quarter of autosomes, respectively), the lower mutation and recombination rates and the different selective pressure between genders have been invoked [19, 21, 42].

Keinan et al. showed that there was a period of accelerated genetic drift on chromosome X associated with the human dispersal out of Africa. In particular, they estimated the autosome-to-X genetic drift ratio between North Europeans and East Asians is consistent with the expected 3/4 while it is significantly reduced between

North Europeans and West Africans, and between East Asians and West Africans [43]. As possible explanations they suggested that a gender-biased process reduced the female effective population size, or that an episode of natural selection affecting chromosome X was associated with the founding of non-African populations. Our results are consistent with these finding. We computed population pair-wise F_{ST} and we found that the autosome-to-X genetic drift ratios (Q), estimated as in [43], are compatible with those reported in [43] (Asia-Europe Q = 0.72; Asia-Africa Q = 0.66; Europe-Africa Q = 0.65).

The weak but significant correlation that we found among F_{ST} values of neighbouring markers demonstrated that they are non-randomly distributed along chromosomes. This result confirms previous observations made on smaller datasets [19, 36]. We extended for the first time this observation to the X chromosome and we found that correlation was slightly stronger than that of autosomes. It has been observed that correlation between SNPs is proportional to Linkage Disequilibrium (LD) [36]. Therefore, the higher value of autocorrelation that we found can be explained by the higher value of LD in X chromosome [36].

Population genetics approach has been largely used for studying natural selection. Other approaches include the comparative one, in which data from different species are used. The most commonly used method is to compare the ratio of nonsynonymous mutations per nonsynonymous site to the number of synonymous mutations per nonsynonymous site (d_N/d_S). Data from comparative studies and from population genetics are poorly connected. We found that genes with a high d_N/d_S ratio, indicating positive selection, showed a significantly lower F_{ST} mean value. In our knowledge this represents the first attempt to connect human population genetic data and comparative data at a genome-wide level. Our finding does not conflict with previous studies performed on a restricted number of genes [44]. It is well established that comparative data provides the most unambiguous evidence for selection, but relatively vague assertion on the type of selection and if the selection is currently acting in a population [21]. For such reasons the connection with pop-

ulation genetic data is needed. Further studies, mainly focused on this topic, are required to confirm and understand the relationship that we found.

We used a gene set approach to identify pathways with extraordinary levels of population genetic differentiation. The traditional approach used to perform this analysis is based on the identification of those loci outliers in a given statistic. This approach has been recently reviewed and its limits explored [24, 45–47]. Interestingly, similar criticisms are arising on analogous methods used in transcriptomic data analysis. In this field, alternative approaches, as the "gene set" ones, are gaining increasing interest. Among the tools implementing this approach, Gene Set Enrichment Analysis [38, 39] is one of the most used [48, 49]. The key idea underlying GSEA is to focus on gene sets, which are defined as groups of genes sharing common features (e.g. biological pathways, chromosomal position, etc.). In microarray data analysis, GSEA aims to determine whether a gene set shows statistically significant, concordant differences between two biological states or phenotypes. This method has been tailored for microarray data, however its use is being explored also in different fields [50, 51]. To the best of our knowledge, the present report is the first attempt to functionally analyse genes under selective pressure by a gene set statistical approach.

Using very conservative statistics, the GSEA analysis found differential F_{ST} values on seven KEGG pathways, one enriched by low F_{ST} genes and six enriched by high F_{ST} genes. However, it is important to note that the discrepancy between the number of low and high F_{ST} pathways is a consequence of the way by which we attributed F_{ST} values to genes rather than underlying evolutionary forces. The only pathway with decreased degree of differentiation among populations was the "antigen processing and presentation" pathway (Figure 2.7). Included in this pathway are genes involved in the antigen-presenting machinery as (i) the expression of major histocompatibility complex (MHC) molecules, (ii) the mechanism of cross-presentation, and (iii) the interaction of antigen-presenting cells. Opposing views exist concerning the evolutionary forces that shaped the innate immune system. In particular, the relative impact of purifying and balancing selection is under discussion [52, 53]. Bar-

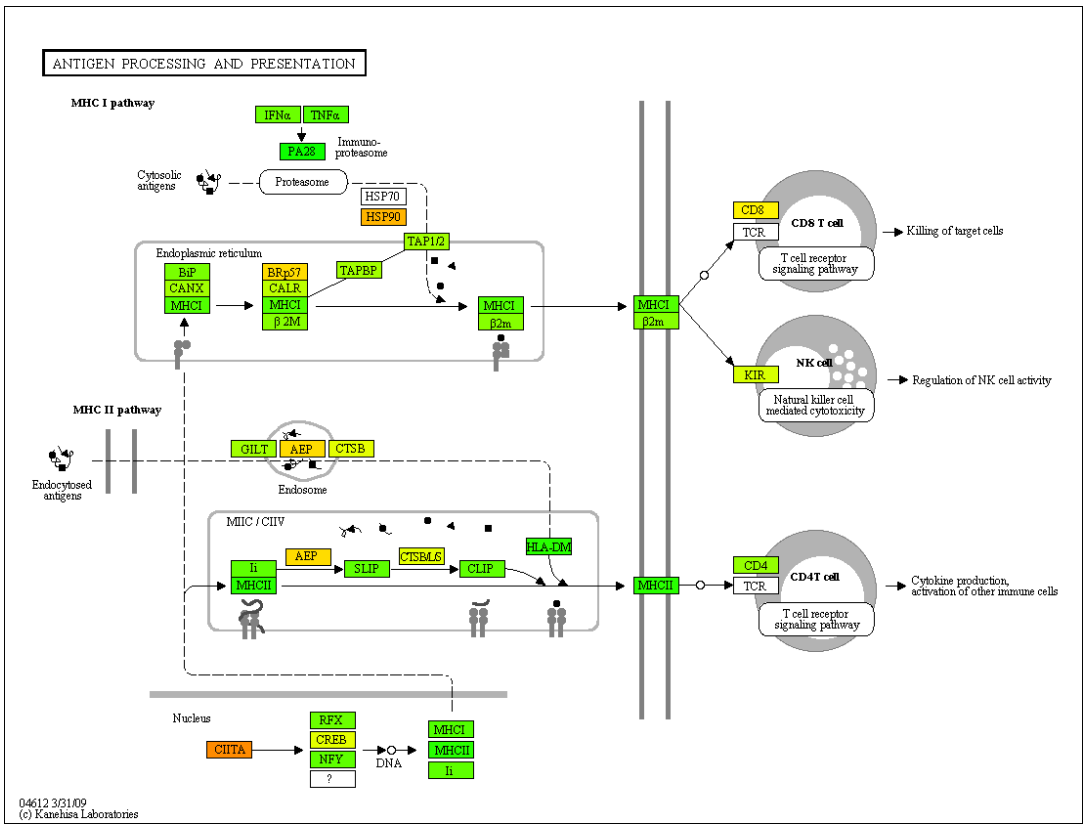

Figure 2.7: **Antigen processing and presentation.** Genes are colored according to their F_{ST} value.

reiro et al demonstrated that several SNPs of genes related to the immune response to pathogens showed very high F_{ST} values [29]. On the other hand, Akey et al. reported a four times increase of proteins that perform a defense/immunity function in the group of the low F_{ST} genes [19]. Moreover, low levels of population differentiation have been previously detected at loci that are involved with host-pathogen responses (HLA class I and class II genes, beta-globin, *G6PD*, glycophorin A, interleukin 4 receptor-alpha and *CCR5*) [20]. Further evidence arises from the group of genes that we studied and that were previously described to be under positive selection. This group of genes, which we found with low F_{ST} values, was described to be enriched for several functions related to immunity and defense [37].

Among the six gene sets enriched by high F_{ST} genes, we found the "calcium signal-

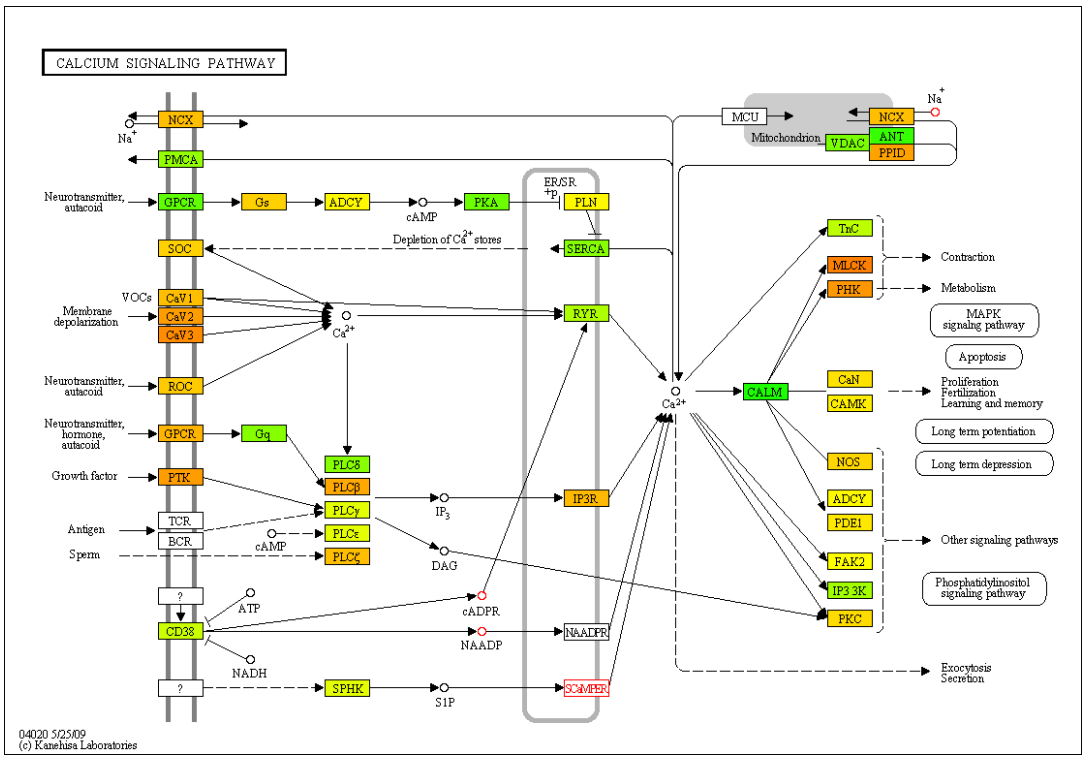

Figure 2.8: **Calcium signaling pathway.** Genes are colored according to their F_{ST} value.

ing" pathway. Calcium is the most abundant mineral in the body. It is also a highly versatile intracellular signal that regulates many cellular processes in response to different external stimuli, as growth factors [54]. We found very high F_{ST} values in three genes belonging to the growth factor stimulated calcium signaling pathway, namely *EGFR*, *ERBB2*, and *ERBB4*. It is interesting to note that a previous study from Pickrell et al. found that *ERBB4* showed extreme signs of haplotype selective sweep in non-African populations [30]. The authors suggested that this gene could affect an unidentified phenotype that experienced a strong recent selection in non-African population. Our gene set approach seems to confirm this finding and expands this observation to other members of the *ERBB* gene family.

The other five high F_{ST} pathways are involved in the control of cell shape and mobility. Among them, four interconnected pathways ("focal adhesion", "regulation of the

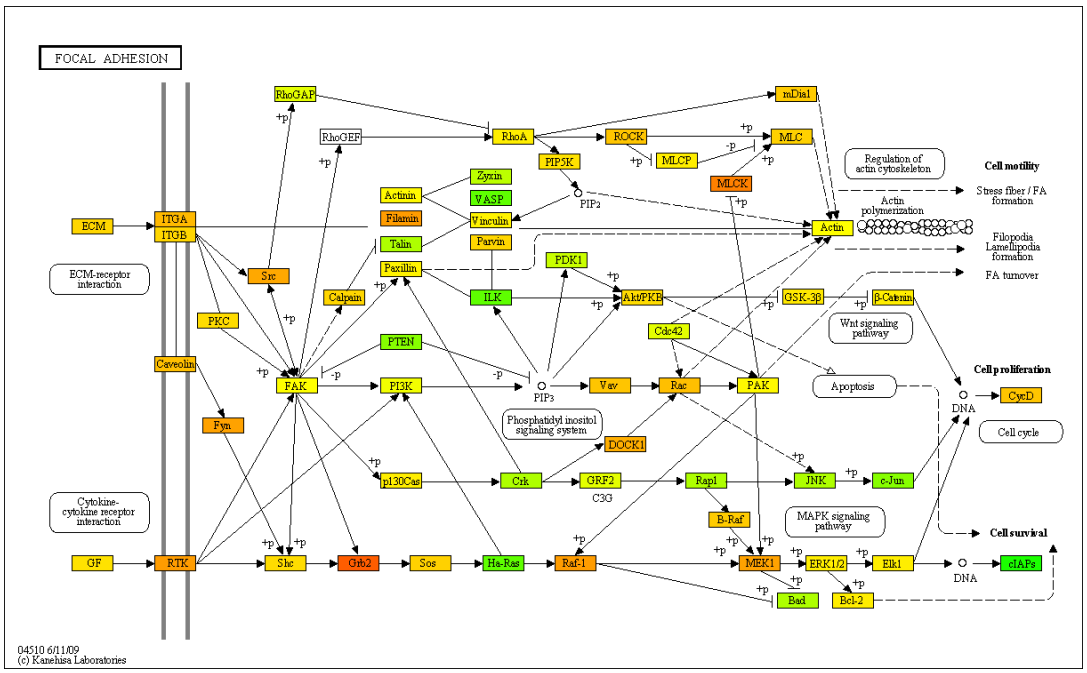

Figure 2.9: **Focal adhesion.** Genes are colored according to their F_{ST} value.

actin cytoskeleton", "adherens junction" and "extra cellular matrix receptor interaction"; Figures 2.9, 2.10, 2.11 and 2.12, respectively) govern growth-related processes and morphogenesis. Morphological traits have been demonstrated to show strong signature of positive selection [29]. These pathways were found also to be altered in a mouse model of fetal alcohol syndrome, associated with a low birth-weight phenotype [55]. Indeed, human body shape and size varies among populations showing a correlation with geographic and climate variables [56]. In addition, in the "adherens junction" pathway, one of the strongest F_{ST} values was showed by *TCF7L2*, the gene with largest type 2 diabetes effect size found to date [57]. This last finding is consistent with previous observations [30, 57]. Since it has been demonstrated that *TCF7L2* variants also substantially influence normal birth-weight variations [58], a complex interplay between pathways that govern growth-related processes and susceptibility to type 2 diabetes could be hypothesized.

The last high F_{ST} pathway, the "axon guidance" (Figure 2.13), is involved in brain wiring during fetal development and repair throughout life. Axon guidance proteins

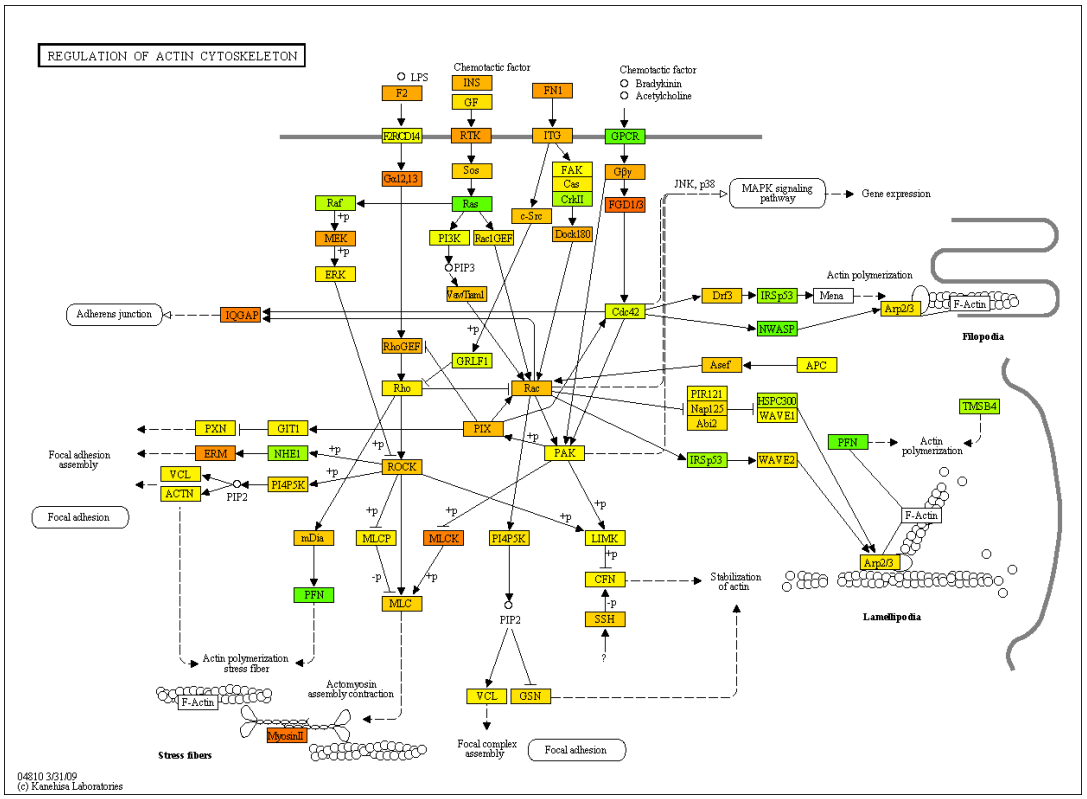

Figure 2.10: **Regulation of actin cytoskeleton.** Genes are colored according to their F_{ST} value.

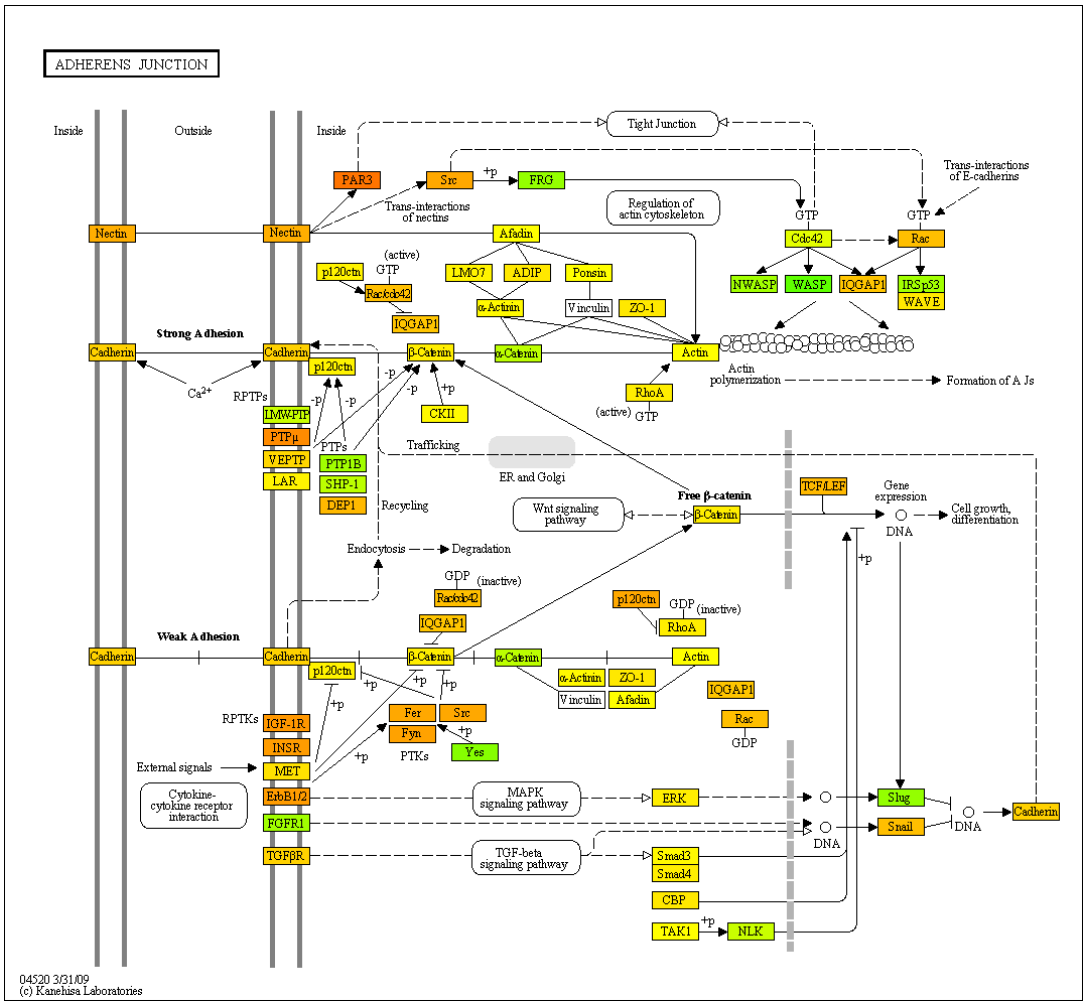

Figure 2.11: **Adherens junction.** Genes are colored according to their F_{ST} value.

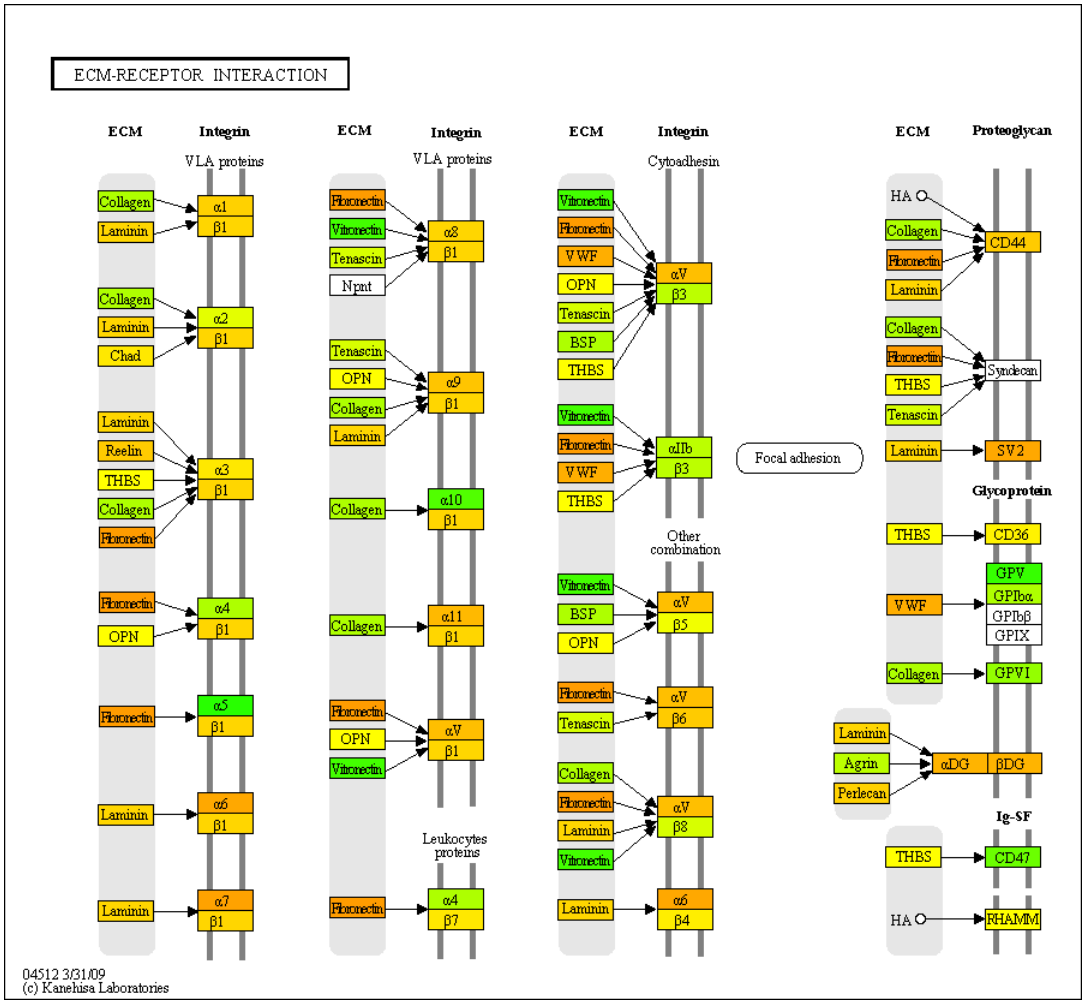

Figure 2.12: **ECM receptor interaction.** Genes are colored according to their F_{ST} value.

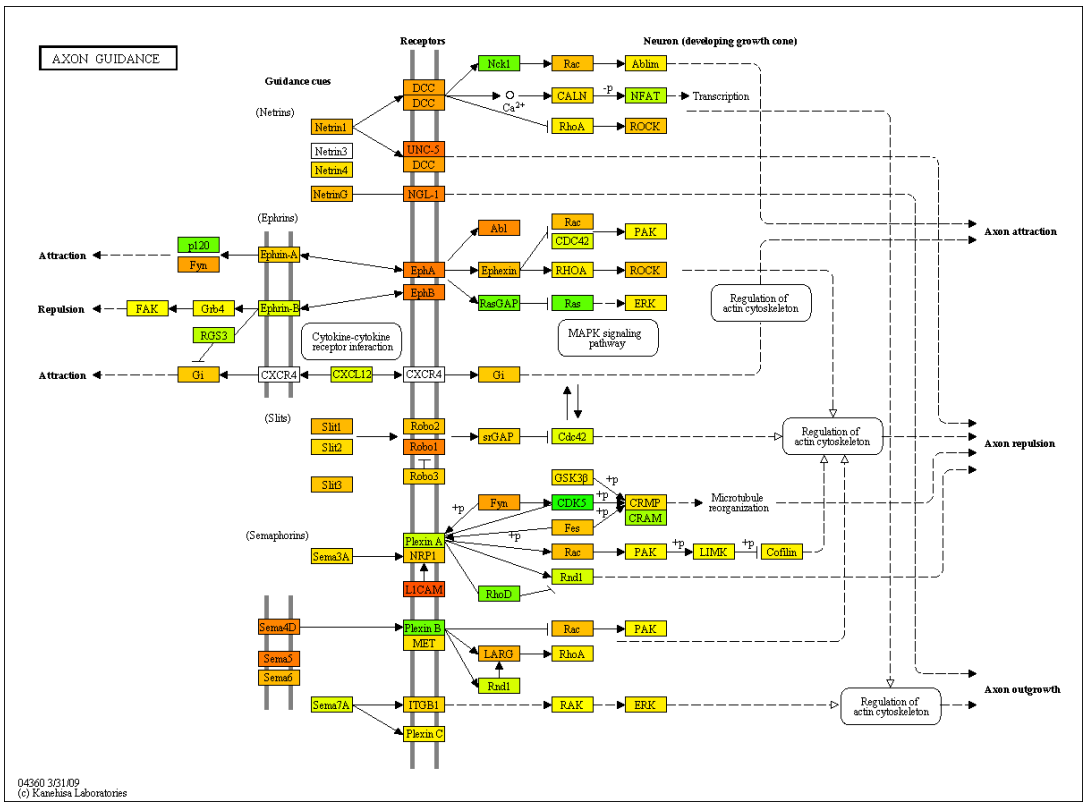

Figure 2.13: **Axon guidance.** Genes are colored according to their F_{ST} value.

and their relative binding partners have also an emerging role in the pathogenesis of several neurodegenerative and psychiatric diseases such as schizophrenia [59, 60]. Signature of recent positive selection inferred by identification of selective sweeps in specific populations was found in genes involved in schizophrenia [61]. Moreover, population dependent results were obtained when gene-association studies were performed using several high F_{ST} genes present in this gene set [62, 63].

It has been suggested that alleles involved in common disease could be targets of selection [15, 56, 64, 65]. The common disease/common variant (CD/CV) hypothesis proposes that common diseases are usually caused by one or a few common disease susceptibility alleles. These genetic variants represent ancestral alleles, presumably under selective pressure, that have become disadvantageous after changes in environment and of lifestyle [64, 66, 67]. We found that genes associated with complex

diseases showed a significant higher mean value of F_{ST} , supporting the CD/CV hypothesis. However, several previous studies of SNPs associated with complex diseases did not find significant evidence of population differentiation [68, 69]. On the other hand, further studies observed that the distribution of maximum F_{ST} was shifted upward in regions associated with type 2 diabetes mellitus [30]. Moreover SNPs known to protect against obesity and diabetes showed very high F_{ST} values [29]. Simulation studies also provided support for the CD/CV hypothesis [70].

According to the GAD classification of diseases, we divided the overall group of the genes associated with complex diseases. Clear differences in F_{ST} means among the various classes were present. In particular, several disease classes, namely "hematological", "infection", and "immune", showed an F_{ST} mean value slightly lower than the mean value of non-associated genes. Nevertheless, the majority of the classes showed F_{ST} mean values to be higher than the non-associated one. Highest F_{ST} values were detected in "pharmacogenomics" and "psychiatric" classes. GAD classifies in "pharmacogenomics" those diseases related to drug effects. It is well established that drugs effects are ethnic specific [71]. The GAD "psychiatric" class includes mental disorders. Why genes that confer susceptibility to mental diseases are still maintained by natural selection, is an old question which, up to now, is still unanswered. The compensatory advantage for genes associated to intermediate phenotypes has been invoked as explanation for this phenomenon, also called "psychiatric paradox" [72]. Further studies should be performed to determine if the high level of population differentiation that we found for this disease class could be related to the psychiatric paradox.

2.3 Materials and Methods

2.3.1 Data

All analysis are based on the HapMap Public Release #27 (merged II+III) datafiles (`http://www.hapmap.org`). We analyzed the data from the CEPH (Utah residents with ancestry from northern and western Europe; CEU, $n = 165$), Yoruba in Ibadan, Nigeria (YRI, $n = 167$), Han Chinese in Beijing, China (CHB, $n = 84$) and Japanese in Tokyo, Japan (JPT, $n = 86$) samples. We pooled the CHB and JPT samples to form a single sample. Additional SNP information about physical positions and SNP-gene association were obtained from dbSNP build 129 (`http://www.ncbi.nlm.nih.gov/projects/SNP`). In particular, according to dbSNP classification, we considered all SNPs within 2 kb of a gene (locus region) as associated to that gene. Data from the International HapMap Project and dbSNP were merged in a local MySQL database by a set of script from Amigo et al. [73]. When we consider the whole Hap map dataset (autosomes and heterochromosomes) we analyzed a total of $3,917,301$ SNPs.

We excluded by this analysis SNPs that were non sampled or non polymorphic in all the three samples. We excluded also SNPs with a minor allele frequency $< 5\%$ in any of the 3 samples, getting a final SNP sample of $2,125,440$ SNPs.

2.3.2 Estimation of F_{ST}

Fixation index (F_{ST}) was calculated using the unbiased estimator proposed by Weir and Cockerham [32, 35]. We implemented this calculation in a Perl script available upon request.

All analyses presented in this work were also performed by using the original F_{ST} estimator proposed by Wright [31] and results are almost identical to that obtained

by the Weir and Cockerham method. This result is not surprising considering previous reports [19, 74] and the strong correlation that we found between these two measures (Spearman's $\rho = 0.97$, p$< 10^{-16}$; see Figure 2.14).

The maximum F_{ST} values among those of the SNPs associated to the gene according to dbSNP (see 2.3.1) was used to assign a F_{ST} value to each gene. This approach is consistent with previously described ones [19, 30]. We studied the correlation between F_{ST} value and gene length and we found that the former have a quite marginal effect on the latter ($r^2 = 0.2$).

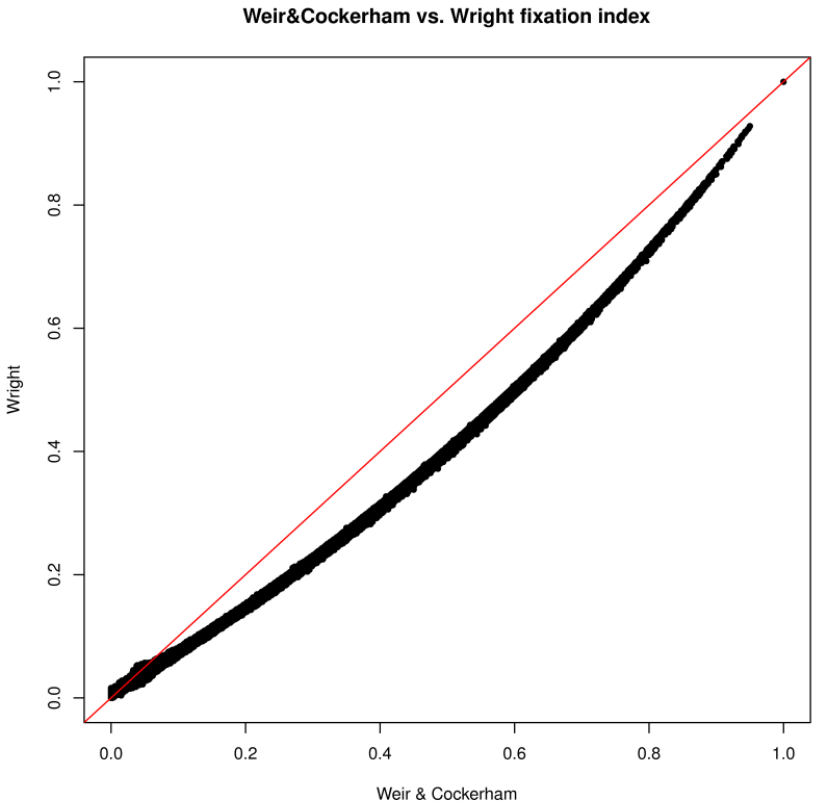

Figure 2.14: **Correlation between Wright's and Weir and Cockerham's estimators for F_{ST} .** Red line represent the diagonal.

2.3.3 Statistical Analysis

SNPs F_{ST} values are not normally distributed across chromosomes. Thus to detect differences among medians F_{ST} values of chromosomes we used the non-parametric Kruskal-Wallis test. Conversely, F_{ST} values of genes are normally distributed (Kolmogorov-Smirnov/Lilliefor test, $p< 0.001$) thus comparison among these values were performed by using parametric tests (ANOVA and t-test).

All statistical analyses were performed with R ver. 2.9 (R Foundation for Statistical Computing, Vienna, Austria; `http://www.r-project.org/`). Non-randomness of data was assessed by using a Ljung-Box test (R function "Box.test"). We calculated the autocorrelation of each chromosome which can be seen as the mean correlation of all pairs of F_{ST} values separated by a fixed number of values (R function "acf").

"Positively Selected Genes" (database "hg18", table "mammalPsg") in UCSD Genome browser (`http://genome.ucsc.edu`). This list was produced by a genome wide scan in six mammalian genomes performed by Kosiol et al. [37]. In particular they identified (i) 400 genes with strong evidence of positive selection across species, (ii) 144 genes with strong evidence of positive selection in one or more branches, (iii) 3705 genes with weak evidence of positive selection on one or more branches, and (iv) 12280 (orthologs) genes with no significant evidence of positive selection. We pooled first and second group into a single "strong evidence of positive selection" group. Differences among groups were evaluated by ANOVA with Bonferroni post-hoc calculation.

Genes associated with complex diseases were obtained from the Genetic Association Database (GAD; October 1 2007 update; `http://geneticassociationdb.nih.gov`). We only kept genes with positive evidence of association, for a total of 1789 genes. According to GAD, these genes are divided into 15 classes of diseases. We excluded from the analysis four diseases classes (Other, Unknown, Mitochon-

drial and Normal variations) because they were not informative. Differences among groups were evaluated by a t-test and a resampling approach. In particular, we used a Moving Block Bootstrap (MBB) strategy [75]. Briefly, (i) we resampled 10000 times 1789 set of adjacent SNPs $\{n_i\}_j$ with $i = 1, \ldots, 1789$ and $j = 1, \ldots, 10000$ and with each set n_i having the same number of SNPs as the i-th GAD associated gene; (ii) for each resample, we computed the F_{ST} of each set n_i according to our method (the maximum F_{ST} values among those of the SNPs in the set); then, (iii) we computed the mean F_{ST} value of each resample j obtaining a distribution to which compare the mean F_{ST} value of the GAD associated genes.

2.3.4 Functional Analysis

We used Gene Set Enrichment Analysis (GSEA) 2.0 [76] to detect KEGG pathways enriched by genes with low or high values of F_{ST} . We provided GSEA, by its "Preranked" feature, with a list L of genes ranked according to their F_{ST} value. Given an a priori defined set of genes S representing a pathway (e.g., genes encoding products in a metabolic pathway), the goal of GSEA is to find out whether the members of S are randomly distributed throughout L or mainly found at the top or bottom (i.e. being "enriched"). Since GSEA preferably expect the values to rank for (in our case F_{ST}) to vary from negative to positive values, we linear shifted these values to get vanishing mean.

We explored the enrichment of KEGG pathways included in the software. For each pathway a False Discovery Rate (FDR) is computed, representing the statistical significance of the enrichment. For experimental conditions similar to the ours, GSEA user's guide suggests a threshold of significance FDR ≤ 0.05. Because of the exploratory nature of this study, we used a more conservative threshold of significance (FDR ≤ 0.01). Overlap among pathways was examined by the "Leading edge analysis" feature of GSEA.

Chapter 3

Adaptation to latitude: a possible model for schizophrenia

Among the environmental factors that strongly influenced our evolutionary history, geographical latitude deserves particular attention. Latitude, indeed, severely affects many natural phenomena such as climate, flora and fauna, light-dark cycle, and all of them, in turn, have an impact on many aspects of our life. For sake of brevity hereafter we refer to all these phenomena simply as "latitude".

Genetic traits following a latitudinal gradient have been observed for several polymorphisms in humans as well as in natural populations of model organisms like *Drosophila* and *Arabidopsis thaliana* [77–80]. The best known example of this kind of spatial variation in *Homo Sapiens* is skin pigmentation. The clinal gradation of skin colouration is correlated with UV radiation levels and represents a compromising solution to the conflicting physiological requirements of photoprotection and vitamin D UV-dependent synthesis [81]. The latter is very important since vitamin D is involved in many health outcomes (e.g. cardiovascular diseases, rickets, pelvic deformities, infections, etc.).

A possible influence of latitude on the circadian phenotype has also been suggested

[82, 83]. Circadian rhythm is a ubiquitous feature of living systems. Daylight hours vary with latitude and seasons therefore adaptability of circadian clocks is of fundamental importance for the adaptation of organisms to the alternating light/dark cycles.

Another example of spatial variation is human body size and shape. These phenotypes show a correlation with climate (that in turn has a strong relationship with latitude) suggesting also for humans the adaptation to the classical ecological rules that individuals living in colder regions are bulkier and have shorter limb lengths [84, 85].

Nevertheless, it is also possible that an allelic variant increasing the fitness of individuals at particular latitudes will no longer be advantageous or even increase the risk for some pathologies at different latitudes. In particular, several diseases show latitudinal clinals. A well known example regards sodium homeostasis. It have been postulated that in hot climate regions, genetic variants inducing enhanced sodium retention were positively selected. This adaptive process would allow a proper vascular tone and salt storage in conditions of excessive sweating. These same variants, when carried by individuals migrated in colder climates (i.e. African American), would increase the risk for sodium retention-related hypertension [86]. Supporting this hypothesis, several studies reported a strong correlation between latitude and the frequencies of hypertension susceptibility variants [86, 87].

In addition, there is growing evidence supporting the idea that these diseases are frequently due to a negative by-product of adaptive changes during human evolution [61]. This is the case when natural selection favours a vital phenotype at the price of predisposing to some other pathology that, for example, do not directly affect the reproduction or are characterized by a late onset-age. Indeed, contrasting forces often affect the outcome of natural selection. For instance, depigmentation is crucial for vitamin D synthesis at higher latitudes but it also exposes a higher risk for skin cancer. As most individuals do not develop cancer until they are past their

reproductive age, from an evolutionary point of view, skin cancer represents a less powerful selective force than vitamin D availability in serum [88].

Many other common diseases like different types of cancer, dismetabolic conditions, schizophrenia, Parkinson's disease, etc. have an incidence following a latitudinal gradient [56, 89–91]. However, the relative importance of variation of environmental exposures or genetic predisposition is not yet fully defined. In some cases, a genetic adaptation has been suggested [56, 61], even if the direct target of this process is often unclear.

In this work, we investigated the latitude-driven adaptation phenomena, for the first time, on a wide genomic scale. In particular, we selected a set of SNPs and genes showing signs of latitude-dependent population differentiation, and by a biological characterization of the genes, we found enrichment for neural-related processes. In light of this result, we investigated whether genes associated to pathological phenotypes, namely psychiatric and neurological diseases, were enriched for Latitude-Related Genes (LRGs). Remarkably, we found a strong enrichment of LRGs in the set of genes associated with schizophrenia. In an attempt to try to explain this possible link between latitude and schizophrenia, we investigated their association with vitamin D, which had been previously associated, separately, to both of them. Our findings suggest a molecular link among latitude, schizophrenia and vitamin D. [1]

3.1 Results

A set of SNPs showing high levels of latitude-dependent population differentiation was selected by using a two-step approach. Our starting point consists of geno-

[1] The results presented in this chapter are published in: Amato R, Pinelli M, Monticelli A, Miele G and Cocozza S. (2010) Schizophrenia and Vitamin D Related Genes Could Have Been Subject to Latitude-driven Adaptation. BMC evolutionary biology 10: 351

type data concerning about 660,000 SNP loci of 938 unrelated individuals from 51 populations of the Human Genome Diversity Panel [42].

The first step was the estimation of the population differentiation level of each SNP. After the exclusion for minor allele frequency and for SNPs falling in intergenic regions, we obtained a set of 224,501 SNPs. For all of them we calculated F_{ST} according to the Weir and Cockerham estimator [35], and to select SNPs with high levels of population differentiation, we extracted those at the top of the empirical distribution of F_{ST} values. Because of the differences existing between the distribution of FST values for autosomic and X-linked SNPs 2.1, these two sets were handled separately. In particular, we selected 22,132 autosomic and 459 X-linked SNPs falling in the top 10% of their own distributions (namely with F_{ST} value greater than 0.153 and 0.262, respectively).

In the second step, we computed for each SNP the absolute value of correlation between the frequency of the ancestral allele in the 51 populations and the absolute value of geographical latitude of the population location. We again handled the distributions of correlation values of autosomic and X-linked SNPs separately because of their differences. Moreover, since population sizes were very different, we computed the correlation by taking into account the number of individuals in each population. We selected the autosomic and X-linked SNPs with absolute value of correlation greater than 0.567 and 0.575 respectively, corresponding to the highest 10% of their respective distributions. We finally obtained two sets of 2193 autosomic and 46 X-linked SNPs corresponding to 1307 and 29 unique genes. Hereafter, we denote by LRGs (Latitude-Related Genes) this set of genes whose SNPs showed both high F_{ST} (mean = 0.230) and high latitude correlation (mean = 0.616) values.

High F_{ST} values can be produced both by selection and by demography. The effect of population histories is potentially the major confounding factor in the interpretation of genetic differences among populations. We tried to estimate the extent of population differentiation that could be ascribed to selection following the approach

proposed by Barreiro et al. [29]. To achieve this goal, in their paper Barreiro and colleagues used a genome-wide approach analyzing both genic and non-genic SNPs. The main difference between ours and BarreiroÕs approach is that, in our study, we considered intragenic SNPs only, since we were focused on functional analysis at a gene level. In Figure 3.1 we report the proportion in the set of LRGs of SNPs according to their classes (dbSNP classification). The intronic set does not show any significant variation with respect to the expected 1% (red line). This value represents the expected ratio between the number of LRGs SNPs and the number of all the intragenic SNPs (10% of 10%), under neutral hypothesis. Conversely, non-synonymous SNP class shows an enrichment of about 30% ($p = 0.016$; Fisher's exact test). This kind of enrichment has been interpreted by Barreiro and colleagues as a signature of selection.

As a further test for the procedure, we chose a well-known latitude dependent phenomenon such as skin pigmentation and we checked whether LRGs were statistically overrepresented in the set of genes associated to this phenotype. As expected a significant enrichment was found (7 out of 24 genes in common, Fisher's exact test p-value 0.0003).

To characterize the set of LRGs we used two methods, tissue localization and functional characterization. Firstly, we explored the tissue localization of proteins encoded by these genes. To achieve this, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID) focusing on the "Uniprot Tissue" list, a curated list of localization based on literature mining. We found that genes expressed in brain and brain-related tissues were significantly enriched in LRGs, accounting for more than a half of them (Table 3.1).

LRGs were also functionally characterized by looking for overrepresentation of Gene Ontology (GO) annotation terms [92]. Because of the high redundancy of GO, we used the Model-based Gene-Set Analysis (MGSA) method included in Ontologizer 2.0 [93, 94]. This promising and novel approach analyses all categories together by

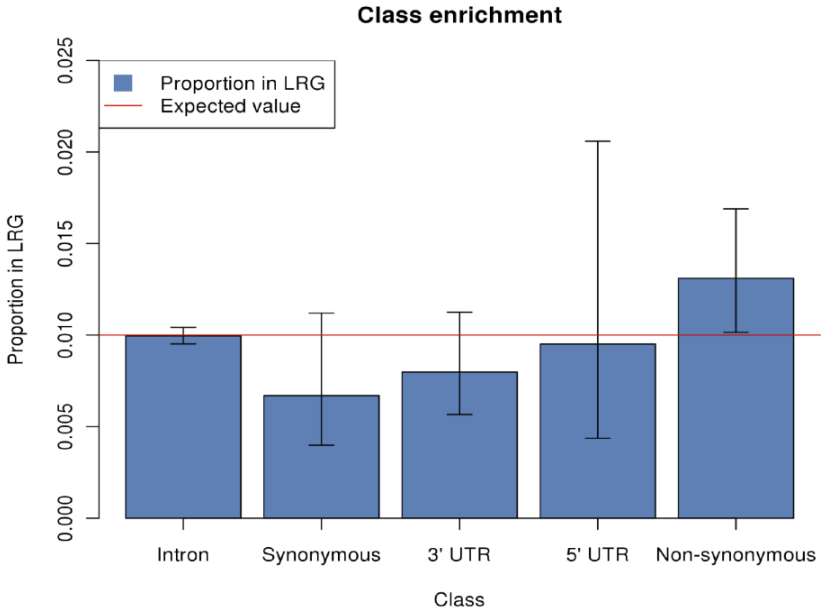

Figure 3.1: **Proportion, in the set of LRGs, of SNPs according to their dbSNP classification.** Red line represents the expected ratio between the number of LRGs SNPs and the number of all the intragenic SNPs (1% = 10% of 10%), under neutral hypothesis. Error bars represent the 95% confidence level of the observed proportion obtained in the analytical way by using the Wilson score interval method.

Tissue	LRGs count (%)	p-value
Brain	683 (56.6%)	3×10^{-18}
Amygdale	112 (9.3%)	8×10^{-6}
Thalamus	76 (6.3%)	1.4×10^{-4}

Table 3.1: **Enrichment in tissue for LRGs computed using the DAVID's Uniprot Tissue category.** For each tissue is reported the number and the percentage of LRGs expressed in the tissue and the significance (Fisher's exact test, Bonferroni adjusted) of that enrichment.

Name	**Sub Ontology**	**Marginal mean (Min-Max)**	**LRGs count**	**Total Count**
Synapse (GO:0045202)	CC	0.998 (0.980 - 1)	60	351
Neuropetide signaling pathway (GO:0007218)	BP	0.793 (0.764 - 0.828)	13	86
Cell morphogenesis (GO:0000902)	BP	0.734 (0.680 - 0.789)	58	420

Table 3.2: **Enrichment for GO terms by LRGs.** For each term is reported the sub ontology to whom it belongs (CC: cellular component; BP: biological process), the mean, minimum and maximum marginal posterior probability of being involved (among 20 runs) and the number of genes annotated in the list of LRGs and in total.

embedding them in a Bayesian network. Differently from other methods, it provides for each term a marginal posterior probability that reflects a measure of certainty in its involvement in the process. Following the authors' recommendation, we repeated the analysis 20 times in order to see whether the reported marginal probabilities of the top terms fluctuated. We found that two terms out of the three showing a posterior probability above 0.5 consistently among the 20 runs were related to neural processes (Table 3.2).

All these results suggested a further investigation about a possible relationship between LRGs and genes involved in neuropsychiatric diseases related with latitude. Indeed, several neurological diseases were previously described to have a latitude-shaped incidence and/or prevalence [89, 91, 95]. To perform this task we compared the list of LRGs with publicly available collections of genes involved in schizophre-

Disease	Overlap with LRGs	Total count	p-value	Adjusted p-value
Schizophrenia	85	885	4×10^{-6}	1.6×10^{-5}
Parkinson's disease	40	490	0.021	0.084
Multiple sclerosis	16	178	0.058	0.232
Alzheimer's disease	45	618	0.075	0.3

Table 3.3: **Enrichment of neuropsychiatric diseases lists by LRGs.** For each disease genes list, it is reported the number of LRGs present, the total size of the list and the significance of the overlap (Fisher's exact test and Bonferroni's correction.

nia, multiple sclerosis, Parkinson's and Alzheimer's disease [60, 96–98]. While there is a weak or non significant overlap with genes related to multiple sclerosis, Parkinson's and Alzheimer's disease, we found a significant enrichment of 85 LRGs in genes related to schizophrenia (Fisher's exact test, Bonferroni adjusted p-value 1.6×10^{-5}; Table 3.3 and Figure 3.2).

We explored if the enrichment found for schizophrenia with the used list was also present in pruned sub-lists. Four different lists were used and all fairly support the presence of a relationship between latitude and schizophrenia.

"Association": It was obtained from the SchiZophreniaGene (SZGene) database, cleaning data according to a risk-allele evaluation pipeline developed by Sun et al. [99]. According to this meta-analysis, 278 protein-coding genes were selected having significant p-values using a combined OR method or at least one positive association result in publication. In this gene set we found an enrichment of 27 LRGs ($p = 0.0068$).

"Core": It contains genes that have been manually collected to include those that have been commonly considered as candidate genes in expert review or had signif-

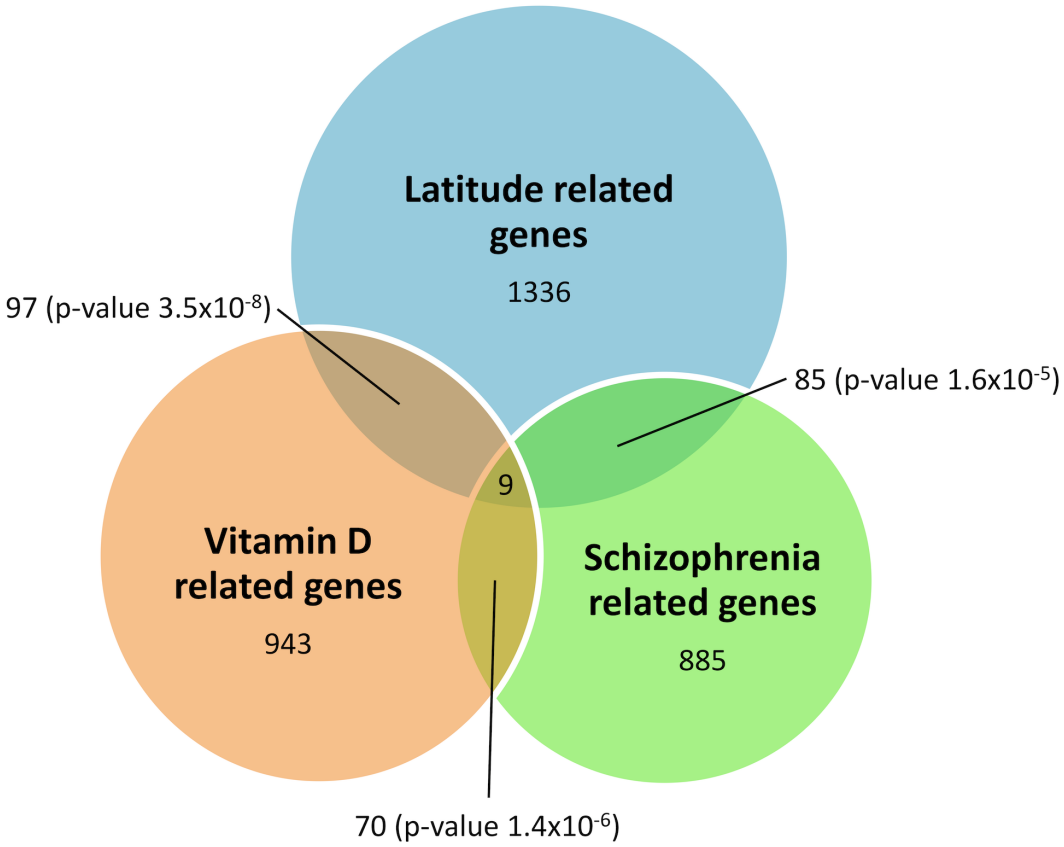

Figure 3.2: **Overlaps among latitude, vitamin D and schizophrenia related genes.** In each circle is reported the number of genes present in the list. For the two-way intersections is reported the size and the significance of the overlap (Fisher's exact test).

icant results in the meta-analysis of association studies. Ross et al. [100] reviewed the evidence in four domains (association with schizophrenia, linkage to gene locus, biological plausibility, and altered expression in schizophrenia) and suggested 19 genes being candidates. They also included 27 genes with significant meta-analysis results performed by the SchizophreniaGene team. The genes were selected by having a nominally significant summary OR in all ethnic groups or Caucasian samples. After removing redundancy, the core gene set contains 38 genes. In this set we found an enrichment of 7 LRGs ($p = 0.0059$).

"75 genes by COR": This list contains 75 genes that were prioritized by ranking about 500 genes from more than 2000 association studies [99]. This list shows an enrichment of 11 LRGs ($p = 0.0041$).

"173 by Ng et al.": This list contains 173 genes and it is based on genetic studies for schizophrenia from four major categories: association studies, linkage analyses, gene expression, and literature search. Genes in these data sets are initially scored by category-specific scoring methods. Then, an optimal weight matrix is searched by a two-step procedure (core genes and unbiased P values in independent genome-wide association studies). Finally, genes are prioritized by their combined scores using the optimal weight matrix. This set shows an enrichment of 19 LRGs ($p = 0.0061$).

We then investigated for possible latitude-dependent biological mechanisms linking latitude to neural development. An important factor hypothesized to be both latitude dependent and neural development-related is vitamin D. We checked for an enrichment of LRGs in Vitamin D Related (VDR) genes by manually creating a list of 943 genes that broadly comprised the most important processes in which it is involved, since a comprehensive list is not yet present in the literature. To achieve this, we merged: a list of 6 genes implied in the metabolism of vitamin D obtained from Reactome, a list of 26 genes in the pathway of the control of the expression by vitamin D receptor from Biocarta and a list of 911 genes from a large-scale identification of $1,25(OH)_2D_3$ target genes by Wang and colleagues [101].

We computed the overlap between LRGs and VDR genes and found a significant overlap of 97 genes (p-value 3.5×10^{-8}, Fisher's exact test; Figure 3.2). This result suggests that VDR genes show signature of latitude-dependent population differentiation. Also the overlap of 70 genes between VDR and schizophrenia related genes was significant (p-value 1.4×10^{-6}, Fisher's exact test; Figure 3.2) confirming the role of vitamin D in schizophrenia pathogenesis.

Finally, we found 9 genes (*SMARCA2*, *MITF*, *DLGAP1*, *MAGI1*, *IL4R*, *NTRK3*, *RUNX1*, *PPP3CA* and *INPP4B*) in common among those ones related to latitude, vitamin D and schizophrenia (Figure 3.2). We checked whether or not in these 9 genes belonged SNPs, selected by our procedure, that were previously studied in relationship to either schizophrenia or vitamin D related phenotypes. One SNP resulted from the analysis, rs3793490 ($F_{ST} = 0.202$, correlation = 0.626), an intronic SNP of the *SMARCA2* gene. In Figure 3.3 is reported its alleles geographic distribution (A) and the values of cross-Population Extended Haplotype Homozygosity (XP-EHH) (B) of the genomic region. This test detects alleles that have risen to high frequency rapidly, enough that long-range association with nearby polymorphisms (the long-range haplotype) have not been eroded by recombination. The analysis showed a strong sign of recent selective pressure.

3.2 Discussion

Many natural phenomena are directly or indirectly related to latitude. Living at different latitudes has consequences in being generally exposed to different climates, diets, light/dark cycles, etc. Therefore, it is reasonable to presume that exposure of individuals to different latitudes could have shaped genetic background as a result of the adaptation process. Indeed, relationships between allelic frequencies of specific genes and latitude have been identified in plants [102–104] and animals [105, 106].

Previous studies in humans, at specific loci, have found evidence of correlation be-

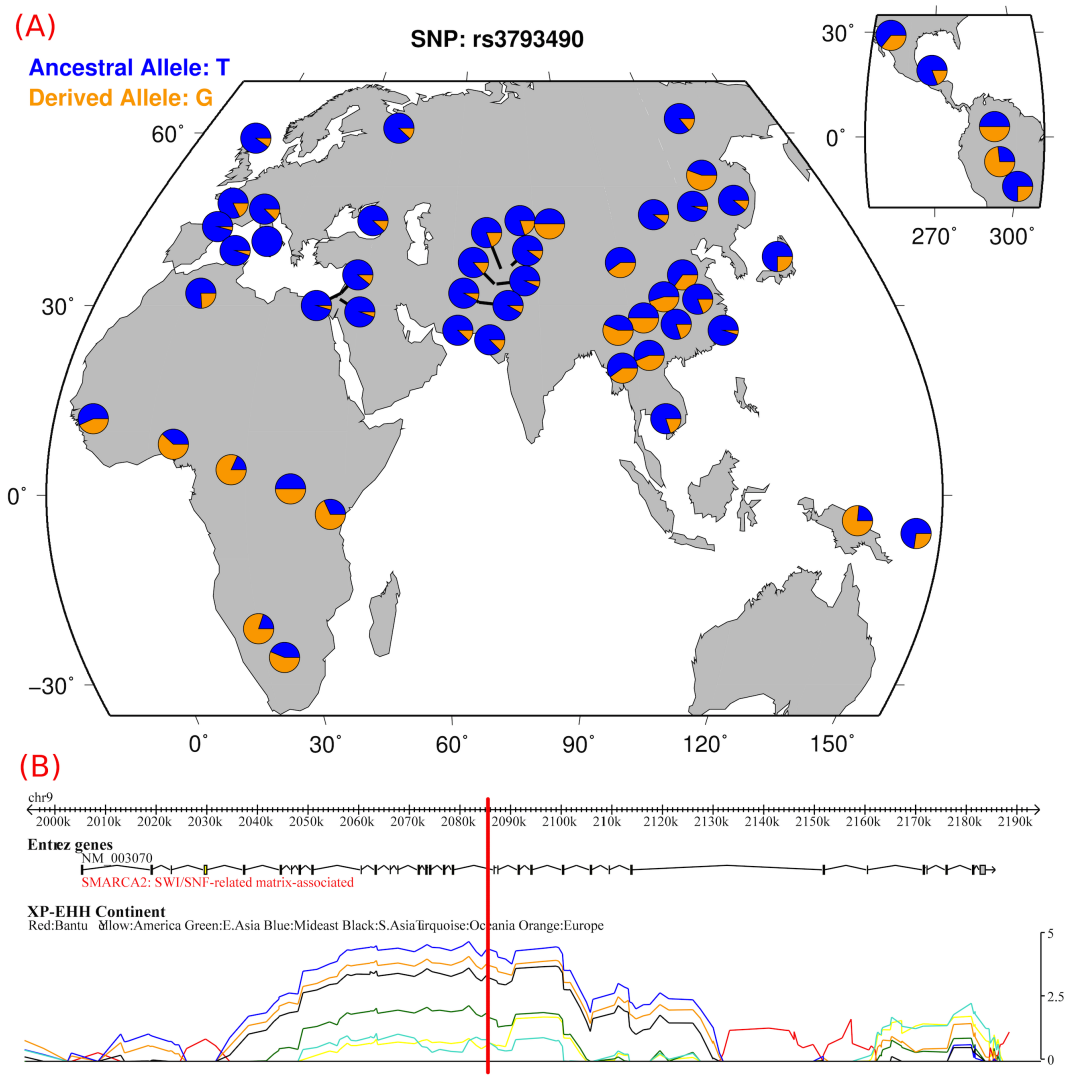

Figure 3.3: **Alleles geographic distribution (A) and cross-population extended haplotype homozygosity (B) for the SNP rs3793490.** Images were readapted from the "HGDP Selection Browser". The red line marks the SNP's position inside the *SMARCA2* gene region.

tween allelic frequencies and latitude of sampled populations. For example allelic frequencies of a genetic polymorphism in the prion gene *PRNP* showed a clear correlation with the latitude within Europe [107] and one in the *ACP1* worldwide [108].

To explore possible genetic adaptations to latitude, within this work we defined a set of latitude-related genes (LRGs) following a two step approach. Firstly, we identified SNPs with a high level of population differentiation (F_{ST}) with the aim to enrich for variants under selective pressure. From these we then extracted those SNPs showing high values of correlation of allelic frequencies with the geographical latitude. To the best of our knowledge this is the first search at a wide genomic level for loci showing latitude-dependent populations differentiation.

Both functional characterization and expression localization of LRGs resulted in a strong enrichment of neural-related processes (Tables 3.1 and 3.2). The relationship between neural development and latitude is partially known. In particular, there exists evidence of a latitude correlation of some physiological neural-related phenotypes. It was reported for humans that cranial capacity is different between populations, probably, as results of adaptation for brain cooling and that the craniofacial diversity results from the tissues of neural crest origin [85]. It is worth stressing that although population differences are observed in these phenotypes, no relationship exists with mental functioning. Also, for several pathological neural phenotypes there has been described previously a relationship with latitude. For multiple sclerosis and schizophrenia there was a latitudinal variation in incidence and prevalence [91, 95]. An association between mortality related to Parkinsonism and birthplace geographical latitude was also found [109].

When we compared the list of LRGs with those containing genes associated to neurological and psychiatric diseases, we found a vanishing enrichment of LRGs in genes related to multiple sclerosis, Alzheimer's and Parkinson's disease (Table 3.3). This is not surprising. Concerning Alzheimer's and Parkinson's diseases, it

should be noted that the relationships with latitude reported in the literature are not largely confirmed as in the case of the other two diseases. Multiple sclerosis has a strong immunity component in the etiology. Since genes involved in immunity related processes usually exhibit low levels of F_{ST} [19] (see also 2.1), they are likely to be excluded by our procedure, in this way weakening the enrichment.

In contrast, we found a strong significant enrichment of LRGs in genes associated with schizophrenia (Table 3.3 and Figure 3.2). The correlation of schizophrenia prevalence with latitude is both large and robust [110]. In previous surveys it was noted that there is a strong tendency for schizophrenia prevalence to increase with increasing latitude [91, 111]. The present result also agrees with that obtained in one of our previous works, based on different data and different statistical approaches. In this previous study, we explored for signs of natural selection in genes associated to complex diseases, and the strongest level of population differentiation was observed in genes associated to psychiatric diseases (see 2.1). In addition, testing for recent selective sweeps in human populations, Crespi and colleagues found significant evidence for adaptive evolution of several genes underlying schizophrenia [61]. Our results may suggest that a latitude-related adaptation occurred for some schizophrenia associated genes, but to which extent this phenomenon is related to the latitude-shaped prevalence of schizophrenia is of course still an open question.

Searching for a mechanism that could connect latitude and schizophrenia at a molecular level, we reasoned that a well-known phenomenon separately linked to both of them is vitamin D. Indeed, vitamin D is essential for normal growth, calcium absorption and skeletal development. The cutaneous synthesis of vitamin D is a function of skin pigmentation and of the solar zenith angle which, in turn, depends on latitude, season, and time of day [112]. With the same dietary intake, the most important determinant for vitamin D levels is considered to be where individuals live, because of the dependence on geographical location of the availability of UV radiation for vitamin D synthesis. Ancestral populations who migrated out of Africa, moving from south to north were exposed to less incident sunlight. It is commonly accepted

that, in these conditions, depigmentation was favored [81]. This adaptation process was compulsory, since pelvic deformities due to vitamin D deficiency could prevent normal childbirth, but it is reasonable to suggest that also other molecular mechanisms apart from depigmentation could have been subject to selective pressure in order to increase the levels of vitamin D in the body.

To test this link at a genomic level, we created a hand-curated list of genes that, at different levels, are related to vitamin D (see 3.1). We found among vitamin D related genes a significant enrichment for genes latitude-related (Figure 3.2). This result may suggest that vitamin D related genes could have been subject, at least in part, to a latitude-dependent adaptation, supporting the hypothesis that selection also acted on other mechanisms different from depigmentation.

On the other hand, the relationship between vitamin D and schizophrenia is well established. Vitamin D receptors were found in most tissues other than those classically involved in the vitamin D action (bone, gut, kidney, etc.). In particular, receptors for vitamin D are widely distributed in the nervous system and vitamin D has been recently implicated in brain function [113]. There is growing evidence that low vitamin D levels adversely impact on brain development [114]. In mice, it has been suggested that changes in brain development induced by prenatal vitamin D deficiency lead to specific functional alterations in hippocampal synaptic plasticity [115]. Also in humans, maternal vitamin D insufficiency has been associated with childhood rickets and longer term problems including schizophrenia [116]. In addition, epidemiological data suggest that vitamin D deficiency may be associated with increased risks of mental health disorders such as schizophrenia [117]. Despite the consistence of this relationship, to our knowledge studies connecting schizophrenia and vitamin D at a molecular level are not yet available.

We found a significant overlap of 70 genes between schizophrenia and vitamin D related genes (Figure 3.2). Our analysis provides the first hint, at a genomic level, of the existence of a relationship between them. The largest part of vitamin D related

genes in our list is made by genes differentially expressed in epithelial cells after treatment with the biological active form of vitamin D [101]. It is therefore possible that the same genes could be regulated by vitamin D levels also in neurons during brain development that, in turn, have been associated to schizophrenia [110].

According to our hypothesis that vitamin D could be the link between latitude and schizophrenia, we focused on the 9 LRGs present in both the lists of vitamin D and schizophrenia related genes (Figure 3.2). Among these genes, that were related to schizophrenic phenotypes, several of them were also previously described in association with bone development. However, no evidence of selective adaptation was present in the literature until now.

For example, one of these genes is the neurotrophic tyrosine kinase receptor type 3 (*NTRK3*) which encodes a member of the NTRK family. These neurotrophins (NTs) receptors are best known for their role in the differentiation and survival of various types of neurons [118]. Gene expression of *NTRK3* has been reported to be reduced in patients with schizophrenia [119, 120]. Furthermore, it has been suggested that the *NTRK3* gene influences hippocampal function and may modify the risk of schizophrenia [121]. Nevertheless, NTs and their receptors are also produced by a growing list of non-neuronal cells [122] including osteoblastic cell lines [123]. NTs receptors were observed in the bone forming area during fracture healing. *NTRK3* was observed in osteoblast like cells and hypertrophic chondrocytes [124].

Another example is *PPP3CA* (protein phosphatase 3 catalytic subunit alpha isoform). This gene, also known as calcineurin A alpha, acts as a calcium-sensor and regulator of calcium homeostasis. For this reason, it shapes calcium and cyclic AMP dependent processes like synaptic activity, receptor desensitization, cell survival and neuroplasticity. Expression of *PPP3CA*, previously described by microarrays analysis from cortical [125] and rat hippocampus tissues [126], was found down regulated in schizophrenic anterior temporal lobe [127]. On the other hand, *PPP3CA* is also expressed in osteoclasts, playing a role in the regulation of bone resorption and its

deletion results in osteoporosis [128, 129].

At least in one case, we also found independent evidence of an adaptive process, reasonably suggesting a molecular mechanism. This is the case of the *SMARCA2* (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2) gene. The protein encoded by *SMARCA2* (also known as *BRM*) is part of the large ATP-dependent chromatin remodelling complex SNF/SWI, which is required for transcriptional activation of genes normally repressed by chromatin. Mammalian SWI/SNF complex actually consists of a small series of compositionally distinct assemblies distinguished by the presence of alternative subunits. The complexes contain either one of two closely related alternative ATPases: Brahma (*BRM* i.e. *SMARCA2*) or Brahma-related gene 1 (*BRG1*). The combinatorial assembly of these complexes could account for the specificity of their functions in different tissues and development phases. SWI/SNF components and DNA replication-related factors form, in turn, a human multiprotein complex (WINAC) that directly interacts with vitamin D receptor [130]. WINAC and vitamin D receptor are targeted to vitamin D responsive promoters in the absence of ligand to both positively and negatively regulated genes. WINAC may rearrange the nucleosome array around the positive and negative vitamin D responsive elements (VDREs), thereby facilitating the coregulatory complexes access for further transcription control. Subsequent binding of coregulators requires ligand binding. Several studies have revealed that one family of the SWI/SNF complexes based on the *BRG1* and *BRM* ATPases has particular critical dosage-dependent roles in the development of the nervous system [131, 132].

One of the five SNPs with the highest levels of both population differentiation and correlation with latitude that we identified at this locus was rs3793490. In a previous study this SNP was associated with *SMARCA2* expression levels in the human brain [133]. In particular, the T variant, whose frequency increases in a latitude dependent way, is associated with low *SMARCA2* expression levels. Much evidence suggests that low levels of *SMARCA2* may play a role in the pathophysiology of schizophre-

nia. *SMARCA2* knockout mice showed impaired social interaction and prepulse inhibition. In the mouse brain, psychotogenic drugs lowered *SMARCA2* expression while antipsychotic drugs increased it [133]. On the other hand, in MC3T3-E1 cell line, deficiency of *SMARCA2* results in an accelerated rate of mineralization with higher levels of expression of osteogenic markers [134]. In addition, the most prominent phenotype of *SMARCA2* null mice is a larger (about 14% more than normal) body size with a disproportionately increased bone and muscle mass [135]. Putting together all this data seems to suggest that *SMARCA2* deficiency is associated with different effects on bones and neurons. It is possible to speculate that a selective pressure could have favoured a haplotype containing the T allele in an environment with low vitamin D availability, for its positive effect on the bone phenotype. In turn, the low *SMARCA2* expression levels linked to this variant could be responsible for the increased risk of schizophrenia. The XP-EHH analysis that we performed (Figure 3.3) confirms the presence of a strong recent positive selection in the *SMARCA2* locus. Interestingly, the homozygosity geographical pattern detected by XP-EHH recognises the effect that we found with correlation analysis. These results agree with a previously described hypothesis of schizophrenia being, at least partly, a maladaptive by-product of adaptive changes during human evolution [61].

In Figure 3.2 besides the described overlaps among the lists used, large non-intersecting areas exist. This can be partially due to technical limitations of our work. Anyway, these areas of non-intersection were, at least in part, expected. As stated, latitude acts directly or indirectly, on a wide variety of phenomena, therefore it is expected that its effects are not limited just to vitamin D and neural development. For similar reasons, we expected that also the link between latitude and schizophrenia cannot be completely explained by vitamin D. For example, in the intersection between LRGs and genes related with schizophrenia there is a gene involved in the circadian rhythms, *TIMELESS*.

Circadian rhythm consists of light and dark phases which coincide with the phases

of the solar day and that is, obviously, correlated with the different photoperiods existing at different latitudes. *TIMELESS* is required for normal progression of S-phase and is involved in the circadian rhythm autoregulatory loop. Associations between circadian gene polymorphisms, including *TIMELESS*, and some mental disorders have been found, including schizophrenia [136]. In addition, expression of *TIMELESS* was investigated in the pitcher-plant mosquito, *Wyeomyia smithii*, and was found to vary with latitude of origin. This suggests that other mechanisms should be taken into account [137].

It is worth stressing that this work represents just an initial exploration of this complex problem and can suffer from some limitations. The first aspect to take into account is the arbitrary choice of the thresholds used as inclusion criteria in the list of LRGs. However, it should be underlined that our aim was just to obtain a list enriched for genes showing both high levels of population differentiation and correlation with latitude. Another important aspect is that the largest part of our analysis is based on hand-curated lists. Nevertheless, it should be noted that these lists are widely used [138, 139]. The only exception is the list of vitamin D related genes that we had to build *ex novo* since no others were present in literature. In addition, a cause-effect relationship can never be conclusively established based only on a reciprocal enrichment between sets of genes. Moreover, alternative explanations of the observed enrichments can be imagined. For example, we cannot exclude the presence of an unknown common factor, different from vitamin D, linking together schizophrenia and latitude. A further limit of our approach concerns the gene level analysis. In fact, focusing on genes rather than on SNPs, does not allow to conclusively assess whether variants related to latitude coincide with those involved in the studied phenotypes. In other words, even if a gene is associated to both latitude adaptation and schizophrenia, we cannot exclude that this is due to functionally independent variants. For this reason our conclusions should be corroborated by further and more detailed studies. Finally, we are aware that population differentiation is influenced by demographic history and thus it cannot be straightforwardly

interpreted as a sign of natural selection. Some approaches have been proposed to evaluate the impact of natural selection on population differentiation. Under the assumption of neutrality, any set of SNPs, even if classified according to their physical location and functional impact, should show the same degree of population differentiation. According to Barreiro and colleagues, any deviation from this expectation should be attributable to selection [29]. In particular, the authors observed that variants leading to amino-acid changes (non-synonymous mutations) are overrepresented among SNPs showing high level of F_{ST} . They interpreted this excess as the result from the action of natural selection. We used a similar approach to analyse our data. We found an excess of non-synonymous polymorphisms in LRGs. This result seems to suggest the appreciable presence of genes under selective pressure in LRGs.

Recently, an international team of researchers presented the first detailed analysis of the draft sequence of the Neanderthals' genome [140]. In particular, they showed the presence of interbreeding between Neanderthals and *Homo Sapiens* occurred after the Out-of-Africa migration. Affecting mainly the non-African populations, this genetic flow could mimic a latitudinal effect on the Africa-Europe axis and therefore potentially influences our conclusions. First of all, LRGs are selected according to a worldwide latitude correlation, which is only partially due to the Africa-Europe axis. This is the case, for example, of the rs3793490 SNP (Figure 3.3) where it is also clear a latitudinal effect out of this axis. In addition, Europeans and Asians share only 1% to 4% of their nuclear DNA with Neanderthals suggesting a limited impact, if any, on our conclusions. Finally, we observed that approximately 10% of LRGs are present in the list of genes that Green et al. showed to have evolved recently in our lineage after we split from Neanderthals, and thus, at least in these cases, we can confidently exclude the influence of interbreeding.

3.3 Materials and Methods

3.3.1 Data

The whole analysis is based on genotypes data of 660,918 single-nucleotide polymorphisms (SNPs) in samples from the Human Genome Diversity Panel (HGDP-CEPH), which represents 938 unrelated individuals from 51 populations from sub-Saharan Africa, North Africa, Europe, the Middle East, South/Central Asia, East Asia, Oceania, and the Americas. We removed the two small heterogeneous Southern Bantu populations as in [141]. Genotypic data was retrieved from `http://hagsc.org/hgdp/` [42] while geographical information was obtained from `http://www.cephb.fr/en/hgdp/`.

For all the SNPs, we computed the allelic frequencies within each population by using the R package "genetics" version 1.3.4. Additional SNP information about physical positions and SNP-gene mapping were obtained from dbSNP build 129 `http://www.ncbi.nlm.nih.gov/projects/SNP`. Data from the HGDP-CEPH project and dbSNP were merged in a local MySQL database.

After excluding SNPs with minor allele frequencies less than 5% in all of the populations, we obtained a set of 655,810 SNPs. Since we were interested in performing this study at a gene level, we retained only intragenic SNPs obtaining a final set of 224,501 SNPs. In particular, we considered all SNPs within 2 kb of a gene, according to dbSNP classification. We also excluded the Y-linked SNPs because of scarce numbers.

Genes underlying skin pigmentation were obtained from a study by Myles and colleagues [142]. Genes related to neuropsychiatric diseases were obtained from a set of publicly available comprehensive, uniform and regularly updated database of genes considered involved in schizophrenia, multiple sclerosis, Parkinson's and Alzheimer's disease [60, 96–98]. These databases collect genes related to these phenotypes re-

sulting from different approaches (association, genome-wide, candidate, etc.). We used the January 30th 2010 update of these databases, containing 892, 197, 511 and 622 genes, respectively. The list of vitamin D related genes was manually created by merging three different lists of genes related in different ways to vitamin D. The first list is the "Vitamin D (calciferol) metabolism" pathway by Reactome (REACT_13523.2). The second one was extracted from the Biocarta's pathway "Control of the expression by vitamin D receptor" (h_vdrPathway) as present on the Cancer Genome Anatomy Project `http://cgap.nci.nih.gov/`. Finally, candidate transcriptional target genes of vitamin D were obtained from a set of genes differentially expressed in SCC25 cells treated with $1,25(OH)_2D_3$ by a genome-wide microarray analysis [101].

A common problem in using lists of genes is that different types of gene identifiers are used. The effect is that usually only part of the genes is recognized. Therefore, each tool that we used lost a variable number of genes during the analyses. Nevertheless, to obtain the best performances from each tool we decided to provide them with the whole lists. In particular, DAVID recognized 1207 identifiers and MGSA recognized 1101 ids. In addition, overlaps between lists of genes were computed by using MatchMiner web tool, which also takes into account gene aliases and/or different types of gene identifiers [143]. This tool recognized 885 schizophrenia, 178 multiple sclerosis, 490 Parkinson's disease, 618 Alzheimer's disease and 1254 latitude related genes.

3.3.2 Statistical analysis

For all SNPs we calculated the fixation index (F_{ST}) according to the Weir and Cockerham estimator [35] using the previously developed Perl script (see 3.3). We then computed the absolute value of the correlation between the frequency in each population of the ancestral allele and the absolute value of the latitude of the population weighted by its sample size. This was to avoid the overweighting of allelic frequen-

cies inside small populations and was implemented simply by using the weighted mean, standard deviation and covariance.

Statistical significance of overlaps was estimated by using Fisher's exact test, considering 21463 genes (number of distinct gene symbols present in the "refGene" track of UCSC Genome Browser) as background population. All statistical analyses were carried out with R ver. 2.10.1 [144]. The whole study was conducted considering a p-value of 0.001 as statistical significance threshold.

To check for a potential bias toward larger genes (since they can contain more genotyped SNPs), we applied the non -parametric Mann-Whitney test to compare the difference among the lengths of LRGs with respect to the remaining genes in the "refGene" track of UCSC Genome Browser. There is no statistical evidence for a difference ($p = 0.07$), with median length of LRGs slightly smaller than median of the others.

3.3.3 Biological characterization

The analysis of gene expression localization was performed by using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 and the "UP_TISSUE" category [145, 146]. The "Uniprot tissue" (UP_TISSUE) list is based on literature mining and reports for each gene in which tissues it has been found to be expressed, by using a curated vocabulary.

Overrepresentation of GO terms was assessed by using Ontologizer 2.0 and the Model-based Gene Set Analysis (MGSA) method. The idea behind this approach is to estimate the marginal posterior probability of a term being enriched by using a Bayesian Network. The greater probability for a GO term being near to 1, the higher is the certainty of its involvement in the process. All parameters were left to "auto", allowing the system to automatically estimate the optimal a priori probabilities and the false positive and false negative rates. Probabilities are estimated

by MGSA using 10^7 Markov-Chain Monte Carlo (MCMC) steps. Since MCMC is not guaranteed to converge in any a priori defined number of steps, we repeated the analysis 20 times and retained only terms having a marginal posterior probability greater than 0.5 in every run, as recommended by authors.

Chapter 4

Evolutionary forces shaping autoimmunity

In human history, infections have been a powerful selective pressure [147]. After the human species originated in sub-Saharan Africa more than 100 kya, it spread much of the globe. In this journey, humans were exposed to new pathogens that challenged their immune system; as a result, genetic variants conferring resistance to infections would have been preferentially selected. However, several genetic variants that protect against infections may also predispose to autoimmune diseases (ADs) in contemporary populations (e.g., alleles of solute carrier family 11 member 1, *SLC11A1*) gene that protect against tuberculosis predispose to rheumatoid arthritis (RA) and type 1 diabetes (T1D) [148, 149]. Consequently, genetic perturbations of the immune system that occurred as a result of adaptations to pathogen exposure may involve variation that influences risk to ADs, and might also be responsible for population differences in AD prevalence. [1]

[1]The manuscript based on the results presented in this chapter is currently in preparation. This project was developed at the University of Chicago under the supervision of Anna Di Rienzo. Gorka Alkorta-Aranburu performed all the experiments and was responsible for all the biological contents. David Witonsky actively contributed in the preliminary phases and during the whole project.

Common disorders of the immune response and AD in particular. The purpose of the body's immune system is to fight infectious agents, such as viruses or bacteria. However, several things can go wrong with the immune system. Disorders of the immune system can be broken down into four main categories: immunodeficiency disorders (when a part of the immune system is not present or is not working properly), cancers of the immune system, allergic disorders (in which the immune system overreacts in response to an antigen, for example, asthma, eczema and allergies), and ADs (in which the body's own immune system attacks its own tissue). Normally, AD is prevented by a self-tolerance mechanism, which depends on a succession of control layers that operate at different sites and stages of development. Interestingly, genetic variation in tolerance control layer genes has been associated with autoimmunity: (1) genes that affect autoantigen availability and clearance (e.g., component C3, C1q, C2, and C4 of the complement pathway, Mannan-binding lectin, autoimmune regulator *AIRE*); (2) genes that affect apoptosis (e.g., Fas and FasL mutations in autoimmune lymphoproliferative syndrome *ALPS*); (3) genes that affect signaling threshold (for example decreasing receptor sensitivity in the thymus can fail to negatively select autoreactive cells, and increasing sensitivity in the periphery can cause an exaggerated immune response causing autoimmunity; and (4) genes affecting co-stimulatory molecules, for instance, citotoxic T lymphocyte-associated antigen 4 (*CTLA4*). CTL4 is a member of the *CD28* family that binds to *CD80* and *CD86* and transduces an inhibitory signal to T cells. When the gene for *CTLA4* is knocked out in mice, it results in a severe and lethal autoimmune syndrome [150]. Even in humans, *CTLA4* variation, together with protein tyrosine phosphatase (*PTPN22*) and TNF-α variation, has been associated with different ADs indicating that autoimmune phenotypes could represent pleiotropic outcomes of non-specific diseases' genes that underline similar immunogenetic mechanisms [151].

Still, the strongest influence on susceptibility is the major histocompatibility complex (MHC), in particular human leukocyte antigens (HLA) [151]. In addition to genetics, environmental factors also influence AD risk. Certain drugs elicit autoim-

munity as a side-effect (e.g., procainamide induces lupus-like autoantibodies [152]) and toxins can also cause autoimmunity (e.g., heavy metals such us gold or mercury [153]). Infections are also associated with the onset of AD. Infections destruct tissues, so inflammatory mediators (e.g., cytokines and chemokines) are released, which can activate self-reactive lymphocytes. Many pathogens can secret their own cytokines or prevent lymphocyte apoptosis [154, 155]. Even in healthy individuals isolated breakdowns of one or more layers can be detected. However, ADs only develop when enough safe guards are overcome to lead to a sustained and persistent immune response against self, including the generation of effectors of tissue destruction. This is because each tolerance layer is partly effective in preventing anti-self responses, and all of them together provide efficient protection against autoimmunity without inhibiting the immune system's ability to fight against pathogens. Consequently, some degree of autoimmunity can be thought of as the evolutionary price of being able to make effective responses against pathogens.

The identification of genes that promote autoimmunity might help in AD diagnosis, prognosis, identifying risk individuals who could benefit from medical intervention before disease and suggesting new therapeutic strategies. Recent advances, such as the International HapMap project [34, 156] and high throughput genotyping platforms (e.g., Illumina, and Affymetrix) offer new opportunities to rapidly screen common genetic variation across the genome, making GWASs powerful and approachable. As a result of these advances, several GWASs have identified a large number of common AD risk variants (ADRVs) without prior knowledge of position or function. At the same time, methods to scan the human genome for selection signals have been developed and applied to large-scale data sets such as those from the HapMap project [25, 27, 157]. Interestingly, several of these studies found signals of selection in genes involved in the immune response, further supporting the notion that pathogens have exerted strong selective pressures on the human genome. For example, using SNP data, Voight et al. [25] detected a significant excess of MCH-1 mediated immunity genes with evidence for partial sweeps in the CEU (CEPH

European population) (p-value < 0.0001), but not in the YRI (Yoruba population, Ibadan, Nigeria) (p-value $= 0.002$) or the ASN (the Han Chinese and Japanese populations) populations. In addition, Sabeti et al. [27] also reported that like-glycosyltransferase (*LARGE*) and dystrophin (*DMD*) genes, which are both related to infection by the Lassa virus, are suspected to be under positive natural selection in the YRI.

4.0.4 Major transitions in pathogen exposure during human history.

Many significant human diseases are thought to have arisen with the introduction of agriculture. Before the Neolithic period, the main form of human subsistence was hunting and gathering (HG), which was characterized by small and scattered populations. Approximately 10,000 years ago, however, with the introduction of agriculture and animal husbandry, humans transitioned to a sedentary agricultural (AG) subsistence, population size/density increased, and humans became attractive pathogen hosts because large populations in small areas maximize the chance of transmission between longer-lived barriers. Therefore, the human population growth during the Neolithic created the conditions that favored the emergence of pathogens that specialize in human hosts [158]. In addition, many pathogens became endemic. As most individuals reaching reproductive age were exposed to pathogens in early life and acquired resistance, variants conferring resistance were preferentially transmitted to the next generation and, therefore, were selectively advantageous. Nutritional deficiency reinforced this selection. First, since AGs rely on 1 or 3 crops, not having back-up crops, starvation, mortality and disruption of fertility cycles are expected [159]. Second, in order for mothers to quickly return to reproductive readiness, AG infants were weaned earlier; in addition, the low-protein and high-carbohydrate diet typical of the agricultural subsistence made these infants more prone to diarrhea and infections. This high AG infant mortality was balanced by reducing birth spacing,

which worsens infant and mother's nutrition and health [159].

In addition to subsistence, other features of the physical environment are known to be important determinants of pathogen load and diversity. It is well established that plant and animal species diversity is strongly correlated with distance from the equator. It was recently shown that this correlation holds for pathogens, and that it is mainly due to climatic factors [160]. This raises the question of whether the main shift in the selective pressures acting on ADRVs occurred (1) when humans changed the environment with the introduction of agriculture, (2) when climate changed, for instance, during the last glaciations and humans moved to higher latitudes, or (3) when humans became exposed to different climates and environments as a result of the exodus from Africa. The first two scenarios are not distinguishable in terms of the timing of selection as both transitions occurred within a relatively recent window of time (the last 14 kya). The third scenario, however, is expected to result in older selection signals compared to the other two. Therefore, the end of the Ice Age (coupled with the advanced agriculture) and the dispersal out of Africa represent two major moments when genetic variants affecting autoimmunity could have been selected.

To understand the evolutionary forces shaping autoimmunity, a population genetics approach is proposed. When a genetic variant is advantageous, chromosomes carrying it quickly increase in frequency while chromosomes that do not carry the selected variant are lost. This process creates a signature in the patterns of neutral variation tightly linked to the selected site, which may affect multiple aspects of genetic variation (i.e., nucleotide diversity, variation frequency spectrum, and patterns of linkage disequilibrium). In order to characterize these aspects of variation, a full re-sequencing approach is necessary. Full re-sequencing data are also necessary to uncover the variant that underlies a disease association signal and to narrow down the location of an advantageous variant. Next-generation sequencing technologies offer new opportunities for population genetics beyond the conventional and low throughput capillary-based sequencing.

Finally, under a high pathogen load, genetic variants resulting in over-responsive immunity (e.g., ADRVs) can be advantageous. This raises the possibility that some of the ADRVs were evolutionarily advantageous at some point during human history. In addition, if pathogen loads vary across populations, the frequency of over-responsive immune variants may differ substantially as a result of local adaptations. Indeed, a preliminary analysis of ADRVs identified in genome wide association studies (GWAS) showed that several of these variants carry signals of positive natural selection. In addition, consistent with the idea that these variants were the targets of local adaptations, their geographic distribution is often restricted to closely related populations. For example, some ARDVs seem to have been locally selected in Europeans, perhaps during the Neolithic period. Moreover, autoimmune disease prevalence differs among ethnic groups even when they live in similar environments, e.g. ethnically diverse populations living in the United States.

In conclusion, reconstructing the chronology of local adaptations will yield important insights into the origin of ADs. Answering where and when selective factors shaped genetic variation has the potential to understand the evolutionary history of the autoimmunity phenotype, which in turn will be used to guide the future search to define the key populations and the causative ADRVs to be functionally studied.

4.1 Methods

4.1.1 Selection of the AD-risk regions.

Two general AD loci with signatures of European local adaptation were selected: 4q27 and 12q24. Interestingly, the protective/derived allele in 4q27 and the risk/derived allele in 12q24 are associated with signals of positive selection in Europe (i.e., restricted to Europe and significant iHS signal).

Chromosome 4q27 region

It represents a general AD risk locus. WTCCC found a risk haplotype highly associated with T1D [161], which was later replicated with rs17388568 for T1D and associated with Grave's disease (GD) [162]. To be noted, the intergenic rs6822844*C (ancestral) allele was a perfect proxy for the risk haplotype highly associated with T1D [163, 164], RA [163–167], CD [164, 168, 169], ulcerative colitis (UC) [168, 169] , juvenile idiopathic arthritis (JIA) [170], psoriatic arthritis (PSA) [171], psoriasis (PS) [171], early onset psoriasis [172] and SLE [163]. In addition, in a CeD GWAS, besides the HLA region, rs13119723 was the most significant finding, but a follow-up in two cohorts highlighted rs6822844*C with even stronger risk-association p-value (meta-analysis p-value = 1.3×10^{-14}) [173, 174] and replicated in a Scandinavian CeD cohort [175].

All those polymorphisms were present in a strong LD block containing four genes: *KIAA110*, ATP-dependent DNA helicase (*ADAD*), interleukin (IL)-2, and *IL-21*. *IL-2* and *IL-21* are strong AD candidate genes. For example, *IL-2* is an important cytokine in T and B lymphocyte proliferation; about half of *IL-2* deficient mice die of autoimmune haemolytic anaemia before 2 months of age, and the survivors develop inflammatory bowel disease [176, 177], which suggests an essential role of *IL-2* in the immune response to antigenic stimuli. In addition, the receptor of the type I cytokine *IL-21* is expressed on T, B, and NK cells; and BXSB-Yaa mice, which develop a SLE-like disease, have greatly elevated *IL-21*, suggesting a role for *IL-21* in the development of AD [178]. Overall, the ancestral/risk alleles/haplotype of T1D, GD, RA, CD, UC, JIA, PSA, PS, early onset PS, SLE and CeD appear to be the same, and of similar frequency, which are significantly correlated with climate variables, and latitude. Interestingly, the derived/protective rs6822844*T allele shows evidence of recent selection in the HapMap CEU - European population ($\text{iHS}_{CEU} = -1.958$). Furthermore, this protective/derived/selected allele is only present in Europe, Middle East, West Asia, and the Siberian Yakut population,

while the risk allele is fixed in the remaining HGDP populations (Figure 4.1).

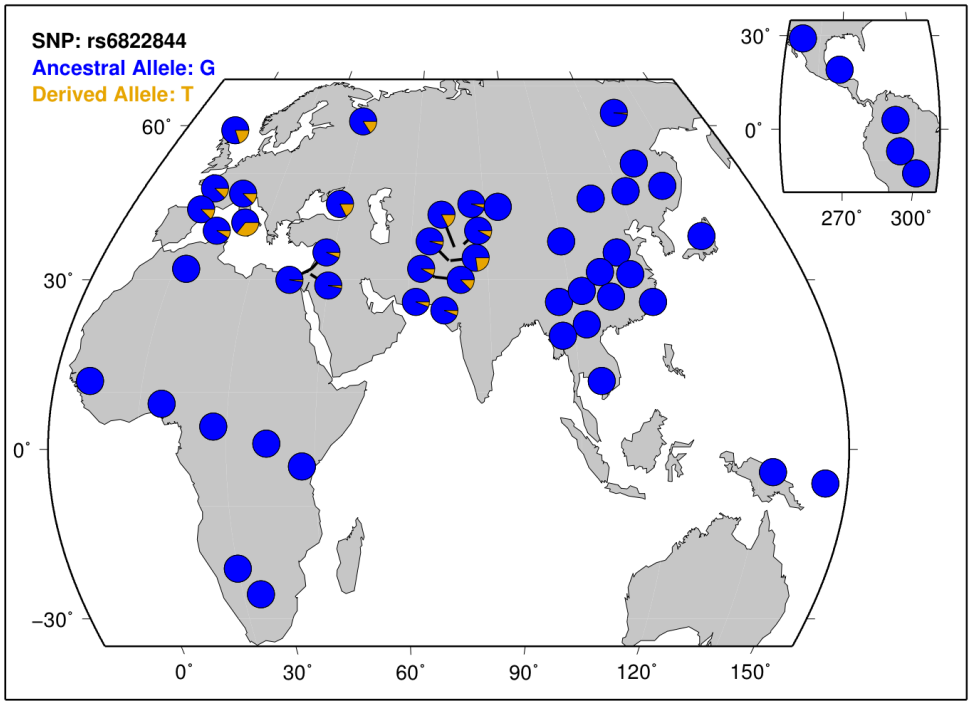

Figure 4.1: **Pie chart of the allele frequencies worldwide for the SNP rs6822844.** Pies report the allele frequency distribution in each one of the 53 HGDP population (derived allele is in orange). From the HGDP Selection Browser.

Chromosome 12q24 region

The chromosome 12q24 also represents a general AD risk locus. 12q24 harbors a large number of SNPs significantly associated with T1D [161]. The ancestral/risk allele of rs17696736 - GW significant p-value - showed evidence of recent selection in the European population ($\text{iHS}_{CEU} = -3.21$) and its frequency is correlated with latitude. Interestingly, in the follow-up replication study [162], four SNPs for which the LD r^2 values with rs17696736 ranged from 0.59 to 0.82 were tested and the nsSNP in exon 3 of *SH2B3* encoding a pleckstrin homology domain (R262W- rs3184504) had the highest association (p-value $= 1.73 \times 10^{-21}$; odds ratio (OR) $= 1.33$, 95%

c.i. $= 1.26 - 1.42$). This single nsSNP was sufficient to model the association of the entire region [162]. Not only the derived rs3184504*T allele increased risk for T1D [161], but also for CeD [173], multiple sclerosis (MS) [179], myocardial infarction [173] and is correlated with the Soluble ICAM-1 (*sICAM-1*) which is and endothelium-derived inflammatory marker that has been associated with diverse conditions such as myocardial infarction, diabetes, stroke, and malaria [180].The derived/risk rs3184504*T allele is also present in the longer haplotype suggesting that it was recently advantageous.

Interestingly, (1) multiple SNPs in this region have $|\text{iHS}_{CEU}| > 2$, suggesting strong recent selection on variation in this region in the European population; (2) this region is significantly enriched for high F_{ST} (CEU vs. ASN and CEU vs. YRI) values; and (3) the the worldwide distribution of rs3184504 is similar to other AD-risk alleles in this candidate region: Europe, Middle East, West Asia, and the Siberian Yakut population (Figure 4.2). *SH2B3* is strongly expressed in monocytes and dendritic cells, as well as to a lesser extent in resting B, T and natural killer (NK) cells (Genomic Novartis Foundation SymAtlas [181]). The observed higher small intestine *SH2B3* expression in inflamed celiac biopsies may reflect leukocyte recruitment and activation [173].

SH2B3 regulates T-cell receptor, growth factor and cytokine receptor-mediated signaling implicated in leukocyte and myeloid cell homeostasis [182]. The R262W amino acid change in the pleckstrin homology domain may be important in plasma membrane targeting [173]. $SH2B3^{-/-}$ mice have increased responses to multiple cytokines [183].

OCA2

In addition to the previous regions, we selected a third genomic region in chromosome 15 which is not associated with AD but blue eye-color in Europeans. It is another variant that even if it is associated with a completely different phenotype,

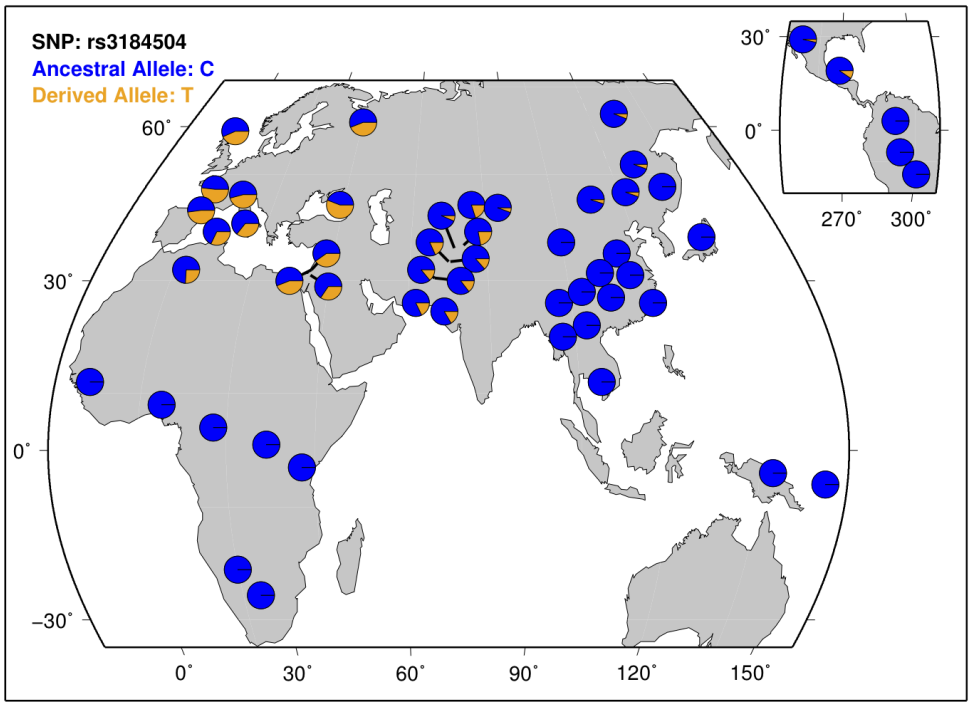

Figure 4.2: **Pie chart of the allele frequencies worldwide for the SNP rs3184504.** Pies report the allele frequency distribution in each one of the 53 HGDP population (derived allele is in orange). From the HGDP Selection Browser.

shares similar features with the above mentioned AD loci. Variation in rs12913832 is relatively common in Caucasians though rare among other ethnic groups [184] and correlates with skin, eye, and hair color variation [185] (Figure 4.3). For example, the G/G genotype is associated with blue eye color [184, 186]. As a note, rs12913832 is part of the "h-1" haplotype, spanning 166kB and found in homozygous state in 97% of individuals with blue eye color [184]: rs4778241(C), rs1129038(A), rs12593929(A), rs12913832(G), rs7183877(C), rs3935591(G), rs7170852(A), rs2238289(T), rs3940272(C), rs8028689(T), rs2240203(A), rs11631797(G), and rs916977(G). rs12913832 is near the OCA2 gene and may be functionally linked to eye color due to a lowering of promoter activity of the OCA2 gene.

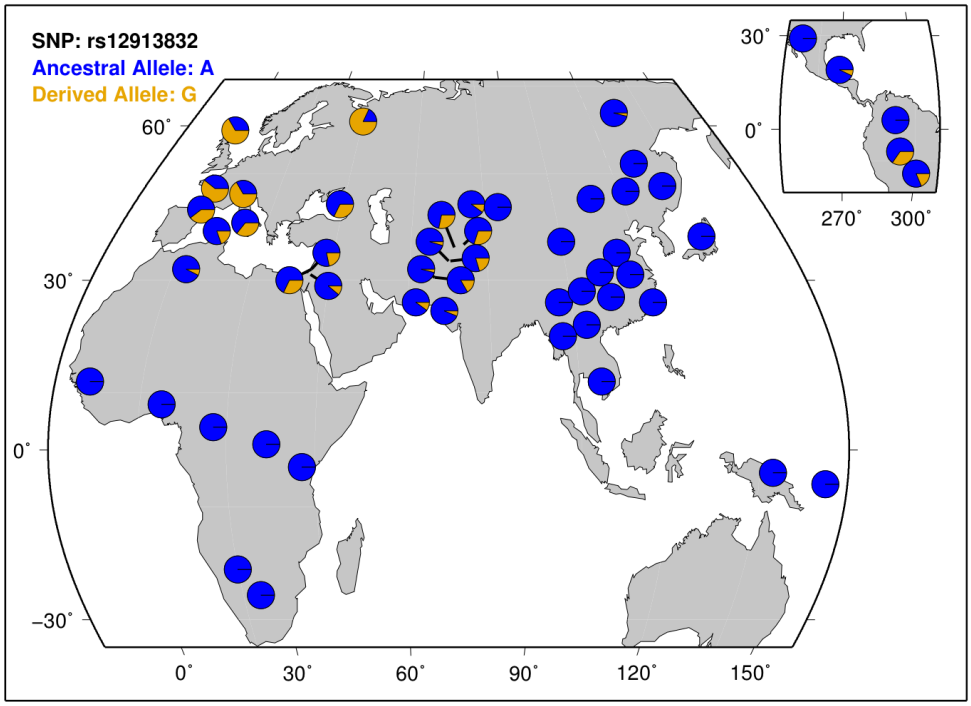

Figure 4.3: **Pie chart of the allele frequencies worldwide for the SNP rs12913832.** Pies report the allele frequency distribution in each one of the 53 HGDP population (derived allele is in orange). From the HGDP Selection Browser.

We included this region for two reason. First of all, it is known that this gene

underwent selective pressure after the out of Africa migration. For this reason, this region will represent a sort of benchmark for the age estimation method. On the other hand, skin pigmentation has a direct effect on vitamin D metabolism that, in turn, play a central role on the immune system. Hence, it could be important to understand the timing of the selection process of this gene with respect to those of the previous regions.

4.1.2 Selection of samples

DNA of 14 European individual genomes from the Centre dÕEtude du Polymorphism Humain (CEPH) was purchased from the Coriell Cell Repository (`http://ccr.coriell.org`) with the corresponding genotypes for the targeted-SNPs (i.e., chr4:rs6822844; chr12:rs3184504 and chr15:rs12913832). As a note, allele T in chr4 and chr12, and allele G in chr15 show evidence of recent selection. Individuals were selected in order to enrich the sample for the selected alleles. This led to the following composition:

rs6822844 3× T/T homozygotes and 11× G/T heterozygotes

rs3184504 5× T/T homozygotes and 9× C/T heterozygotes

rs12913832 11× G/G homozygotes and 3× A/G heterozygotes

4.1.3 Design of the capturing arrays and sequencing

First, to define the coordinates of the region to be captured/sequenced, per each region, the targeted-SNPs (i.e., chr4:rs6822844; chr12:rs3184504 and chr15:rs12913832) were used as the reference. Per targeted-SNPs, all SNPs with $r^2_{CEU} > 0.2$ and within 2 Mb were selected. The SNP with the lower and the SNP with the higher coordinate defined the region to be captured. This resulted in a 2.26Mb region for chr4; 1.9 Mb for chr12; and 1.47 Mb for chr15. However, the SNPs whose coordinates

limited the length of the chr4 and chr15 region resulted to be outliers among the rest of the LD-SNPs. Once removed, the coordinates of the regions to be captured and sequenced were defined as follows (hg_18): chr4: 122729299-123782528 (1.053 Mb); chr12: 109769404-111681303 (1.911 Mb), and chr15: 25542017-26213429 (0.671 Mb). Second, NimbleGen designed the capturing arrays (OID20617) by first masking the repeats of the requested regions and designed unique probes (with an average of $\sim$ 85bp) as determined by the SSAHA algorithm. These unique probes allowed for up to 5 insertions, deletions or mismatches. Only unique probes were designed. Finally, 72.3% of the targeted regions were directly covered by probes, and 86.5% of the regions were either directly covered or within 100bp of a probe.

Capturing was performed on Roche 385K NimbleGen Capturing Arrays. The capturing protocol was modified so that Illumina next-generation sequencing method could be used instead of 454 sequencing.

Two different sequencing runs were performed in the Illumina Genome Analyzer II using paired-end 76 bp reads. The number of reads obtained per sample is summarized in Table 4.1.

4.1.4 Raw data preprocessing

Sequencing data alignment

First, the human reference genome version b36 was obtained from GenomeAnalysisToolKit (GATK) website as part of the GATK resource bundle. Second, the obtained reads were aligned to the reference genome using BWA 0.5.9rc1 [187]. The search for suffix alignment coordinates (bwa aln) and the alignment (bwa sampe) were both performed with the default parameters (see `http://bio-bwa.sourceforge.net/bwa.shtml`); but within the alignment step (bwa sampe), we changed the following parameter values from the default: (i) -a (maximum insert size for a read pair to be considered being mapped properly) 600 was used

Sample	Total number of reads	% of unique reads	% of unique reads aligned on targeted regions	Chr4		Chr12		Chr15	
				% of reads	median coverage (expected)	% of reads	median coverage (expected)	% of reads	median coverage (expected)
C1	64,011,962	85%	20%	26%	181 (171)	54%	199 (212)	20%	200 (187)
C2	66,612,726	34%	10%	22%	31 (30)	57%	45 (47)	21%	44 (41)
C3	71,681,848	71%	39%	23%	297 (275)	57%	416 (400)	20%	408 (349)
C5	67,897,518	53%	17%	20%	57 (72)	55%	93 (120)	26%	113 (135)
C6	71,136,622	58%	54%	23%	350 (311)	58%	487 (463)	18%	423 (354)
C7	65,299,828	58%	35%	21%	171 (172)	56%	251 (266)	23%	269 (262)
C8	73,460,858	56%	61%	25%	448 (386)	54%	500 (489)	20%	500 (448)
C10	71,389,744	76%	36%	26%	352 (306)	54%	400 (379)	20%	401 (336)
C11	73,994,098	62%	30%	23%	201 (196)	56%	264 (276)	21%	280 (252)
C12	70,014,962	46%	23%	15%	42 (70)	58%	114 (157)	26%	134 (171)
C13	71,423,260	42%	26%	17%	51 (79)	58%	118 (163)	25%	141 (172)
C14	67,098,136	32%	33%	15%	46 (67)	60%	126 (156)	24%	134 (154)
C15	73,636,302	51%	54%	23%	292 (279)	56%	414 (409)	21%	443 (376)
C16	72,597,604	49%	48%	22%	242 (231)	56%	351 (344)	21%	373 (316)

Table 4.1: **Reads and coverage statistics.** For each sample is reported the total number of reads obtained, the percentage of unique reads and of unique reads mapping into one of the targeted region. For each region is then reported the percentage of unique reads mapping in that region, the median and the expected coverage.

instead of the default value of 500; and (ii) -r @RG\tID:<*unique id*>\tSM:<*sample id*>\tPU:<*lane*>\tPL:ILLUMINA was used instead of the default 'none' to specify the read group for each sample. The latter was needed by GATK since samples were going to be pooled before the SNP calling step. Third, following GATK's "Best practice for variant detection v. 2", realignment around indels was performed. We performed sample-level realignment with known and novel indels using GATK 1.0.4905 and default parameters. Known indels were from obtained from dbSNP 130. Novel potential indels were obtained with GATK RealignerTargetCreator. The number of aligned reads per sample is summarized in Table 4.1.

Using Picard-tools 1.38 and default parameters (`http://picard.sourceforge.net/`), a different BAM file is created per sample with unique, sorted and indexed reads. Also the duplicated reads (i.e. read pairs with the same orientation and alignment position) were removed using Picard-tools.

Base quality score recalibration

The base quality scores were recalibrated to get them closer to the actual probability of a read mismatching the reference genome using GATK 1.0.4905 with default parameters and QualityScoreCovariate, CycleCovariate and DinucCovariate as covariates (`http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration`). ReadGroupCovariate was not used because for each sample all reads belong to the same group. GATK walks over all of the reads and tabulates data about the selected covariates: assigned quality score $\times$ machine cycle producing this base $\times$ current base + previous base (dinucleotide). For each of such bin, it counts the number of bases within the bin and how often such bases mismatch the reference base, excluding loci known to vary in the population (according to dbSNP). The new quality score is the sum of the global difference between reported quality scores and the empirical quality, plus the quality bin specific shift, plus the quality-per-cycle and quality-per-dinucleotide effect.

Indels and genotyping call

Indels were called, per sample, using GATK and default parameters according to GATK's "Best practice for variant detection v. 2".

Genotypes were called using the default parameters and the "pooled" method. Only variants in the targeted regions or 400bp up- or down-stream of a targeted region were called.

A set of filters was then used to discard low quality genotype calls. The "called/accepted" variants satisfied the following conditions: (1) $QUAL \geqslant 50$ (PHRED-scaled quality of the variant); (2) $QD \geqslant 5$ - Quality by Depth; (3) $SB \geqslant -0.10$ - Strand bias (variant allele is supported by reads that map to both strands?); (4) $HRun \geqslant 5$ - largest contiguous homopolymer run of variant allele in either direction; (5) $MQ0 \geqslant 4$ or $MQ0 \geqslant 0.1 \times DP$ - number of covering reads with mapping quality score zero; (6) Not in cluster, defined as 3 SNPs in 10bp and (7) not inside an indel.

Comparing our data with the HapMap genotype calls, our mean concordance is 99.97%; our false positive (i.e. HapMap homozygote called heterozygote) was 0.01%; our false negative rate (i.e. HapMap heterozygote called homozygote) was 0.02%; our non-reference sensitivity (i.e. fraction of variant sites in HapMap that we call variant) was 96.91%; and finally, our non-reference discrepancy rate (i.e. discrepancy of genotype call excluding concordant reference homozygotes calls) was 0.08%.

Haplotypic phase determination

IMPUTE2 was used for imputation and phasing using the whole HapMap panel as reference panel. The default values were used, except (1) the number of iterations was increased from 30 to 50; and (2) the maximum number of copying states to use for diploid phasing updates was also incremented from 80 to 100. To increase phasing performances, also variants sampled in the HapMap panel and falling in

non-targeted regions were considered. Those variants were then discarded. Also, we discarded all variants were at least one of the genotype was imputed.

Identification of the subregions

For the age estimation purpose, we simulated only a smaller part of the regions. This allowed us to decrease the computational cost of the method and also to improve the estimation. Indeed, in this way summary statistics are mainly affected by the haplotype, without dilution of the information due to larger, neutral regions. To select the subregions, we took into account two aspects, namely the position of the haplotype and the presence of strong recombination hotspots. First, selected haplotypes were roughly identified by visual inspection for each region. Then, the "selected subregion" was obtained as the subregion including the haplotype and enclosed between either the edge of the targeted regions and/or a strong ($> 20\times$ the background recombination rate) recombination hotspot.

For Chr4, haplotype spans about 500 kbp and selected subregion goes from (hg_18) 123,100,000 to 123,800,000 for a total of 700 kbp; for Chr12, haplotype spans about 700 kbp and selected subregion goes from 109,900,000 to 111,500,000 for a total of 1.6 Mbp; for Chr15, haplotype spans about 300 kbp and selected subregion goes from 25,850,000 to 26,100,000 for a total of 350 kbp (Figure 4.4).

4.1.5 Allele age estimation

One of the most interesting question in population genetics, today, is to understand "why" (the functional relevance) and "when" (the timing) loci underwent positive selection. The two questions are, of course, strictly related and answering to one can help in the understanding of the other one.

Regarding the timing, what researchers are primarily interested into is the age of the onset of the selective pressure. This quantity, indeed, can potentially provide

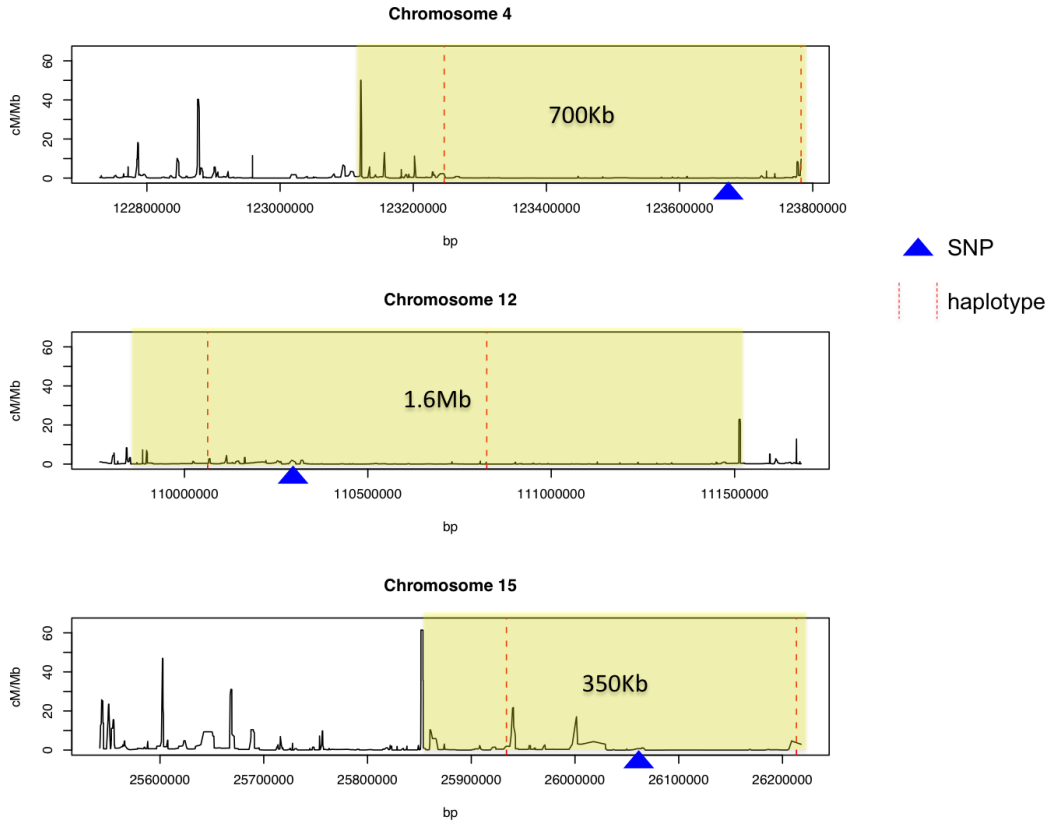

Figure 4.4: **Genetic map of the captured regions.** On the y-axis is reported the recombination rate in cM/Mb. Yellow boxes represent the "selected subregions" while vertical red dashed lines mark the haplotypes. With a blue triangle is marked the position of the selected SNP.

important hints on why the allele was selected and, eventually, its biological role. Unfortunately, this quantity is hard to estimate also because of its vague definition. For this reason, two proxies commonly used are the time since the most recent common ancestor (tMRCA) and the age of the allele (that is the time of the mutation, assuming that the selection occurred on a new allele). Relative relationship among these quantities are shown in Figure 4.5.

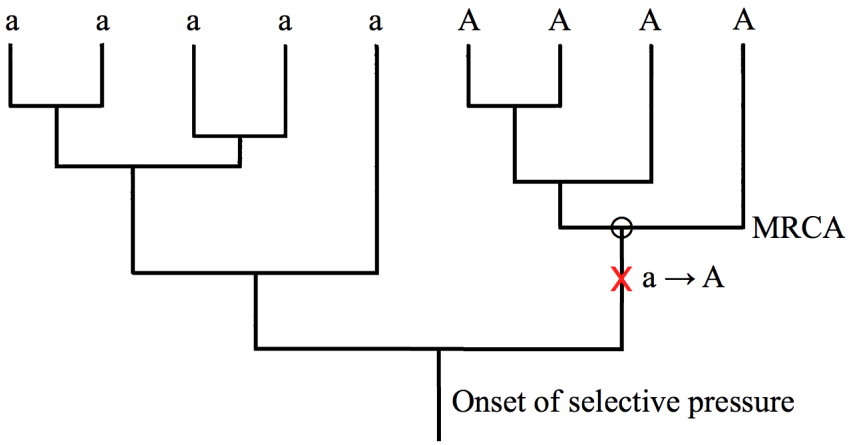

Figure 4.5: **Cartoon of selection acting on a new mutation.** Coalescent tree showing the relative relationship among onset of selection, age of the mutation (red cross) and time since most recent common ancestor (black circle).

Several analytical methods have been proposed to estimate the age of specific classes of alleles, namely rare, near to fixation or fixed [188–190]. However, it is possible to show that given the estimated human effective population size and realistic selection coefficients, sweeps occurred after the out of Africa are, on average, incomplete [191] and thus selected alleles are likely to be at intermediate frequencies.

Among the most used allele age estimation methods, particular attention deserves the *haplotype decay* method. A relatively simple and widely used implementation of this method was proposed by Voight and colleagues [25]. The main idea comes from the fact that the probability of observing two identical chromosomes at a given

recombination distance, decay exponentially with the time. In particular, given r the recombination distance at which the two chromosomes are still identical and T the tMRCA expressed in number of generations,

$$P\left(\text{Homo}\right) = e^{-rT} \ .$$

This method makes some simplifying assumptions about the demographic model which make it unsuitable in many real application. In particular, they assume a population that is panmitic and constant in size, which is clearly unrealistic for human.

Another interesting approach, known as *counting method*, comes from Thomson and colleagues [190]. This method starts building a tree for the region of interest and counting the number of mutations. Since neutral mutations accumulate at a constant rate, the total number results to be proportional to the tMRCA. Given x_i the number of mutations on branch i (from the root to the tip, and thus the number of mutations carried by the i-th chromosome, n the number of chromosomes and μ the mutation rate,

$$T = \frac{\sum_{i=1}^{n} x_i}{n\mu} \ .$$

Contrarily from the haplotype decay, this method makes no assumptions on the demographic model. However, it only consider one tree per region and thus its interpretation in presence of recombination is not straightforward.

Moreover, both these methods as well as many others, usually just provide a punctual estimation of the age and no general rules exist to assess the confidence interval of these estimations.

Many methods for parameter estimation that are also able to provide confidence/-credible intervals are based on the concept of likelihood, that it the probability of the data given the parameter. Given an unknown (set of) parameter θ and the data D, for the Bayes' rule

$$P\left(\theta|D\right) \propto P\left(D|\theta\right) P_\theta = \mathcal{L}\left(\theta\right) P_\theta \tag{4.1}$$

where $\mathcal{L}(\theta)$ is the likelihood of θ and P_θ is the *a priori* distribution of θ.

Usually, the likelihood function is derived from theory. In the case of age estimation, this can include population genetics and coalescent theory. The problem with this approach, however, is the difficulty of deriving such a likelihood function, and this restrict their applicability to very simple scenarios.

A different class of methods tries to overcome this problem bypassing the exact likelihood calculation by means of simulations and a set of quantities (hereafter called "summary statistics" or SSs) that summarize the data. One of these approaches is known as Approximate Bayesian Computation (ABC) [192]. The main intuition is very simple. Suppose you can calculate from the data a minimal set of quantities able to represent all the information contained in it and let's call $S(D)$ this set of summary statistics calculated on the data set D. Moreover, suppose an appropriate model M for the data is available. In other words, suppose it is possible to generate simulated data sets where the underlying model is the same as the real data. The naïve implementation, using a simple rejection schema, is as follows:

1. sample a θ^* value from P_θ,
2. generate a new simulation X^* under the model M with parameter θ^*,
3. if $S(X^*) = S(D)$ then accept θ^*.

The idea behind ABC is that it is possible to approximate the *a posteriori* distribution of the parameter θ using the *a posteriori* distribution of the accepted parameter θ^*. That is

$$P(\theta|D) \approx P(\theta^*|X^*) \tag{4.2}$$

In the practice, the probability that a simulation X yields to a $S(X)$ that is exactly equal to $S(D)$ is almost vanishing. For this reason, instead of considering simulations that are "exactly" as expected, one can consider all simulations that are "similar" or "close enough" to the expected value. Formally, given a distance metric $\Delta(\cdot, \cdot)$

(e.g. Euclidean distance), it is possible to reformulate the third step as: *3. if* $\Delta\left(S(X^*, S(D)\right) \leq \epsilon$ *then accept* θ^*, where ϵ is an arbitrary similitude threshold.

A statistic S is "sufficient" if $P\left(D|S(D), \theta\right)$ does not depend on θ. In other words, if S provides as much information to estimate θ as the whole data set. It is possible to show that if S is sufficient, for vanishing ϵ the relationship in Eq. 4.2 tends to an equality.

A clear advantage of the ABC approach is that it can deal with models on any complexity, given the fact that simulations of the data under the model still remain feasible. On the other hand, one of the limitations is given by threshold ϵ. Indeed, it is worth to stress that ϵ has a double role: on one hand one wants to keep this value as low as possible to increase the precision of the approximation, but this also decrease the acceptance rate yielding a prohibitive computational cost. On the other hand, increasing the value will increase the acceptance rate as well but can also distort the approximation because all the retained simulations are treated equally, independently from the actual distance $\Delta\left(S(X^*, S(D)\right)$.

One of the first strategies to overcome this limitation was proposed by Beaumont in 2002 [192]. Without going into details, I just want to recall the main idea that is to improve the approximation by weighting the θ^* according to the actual distance $\Delta\left(S(X^*, S(D)\right)$ and adjusting their values using local-linear regression to weaken the effect of the discrepancy between $S(D)$ and $S(X^*)$.

4.1.6 Algorithm and implementation

The method for age estimation of selected allele under positive selection via ABC can be summarized as follows:

1. generate a new simulations X^* where the mutation of the allele under positive selection occurred t^* generations ago;

2. emulate all the characteristics of the re-sequencing data;

3. calculate $S(X^*)$;

4. if $\Delta\left(S(X^*, S(D)\right) \leq \epsilon$ then accept t^*;

5. finally, calculate the *a posteriori* $P\left(t|D\right)$ using Eq. 4.2

I'm now going to discuss more in details each one of the previous steps. It is worth to notice that in the implementation of the method I'm going to estimate the logarithm of the age of the allele (t) instead of t. This is because we are more interested in differences in orders of magnitude of the age than on the exact value. Moreover, many SSs have an almost linear linear relationship with $\log(t)$, and this allows to improve the estimation since there are steps where a linear relation is assumed (e.g. local-linear regression).

Step 1 – Generation of simulations

The first, crucial step for using ABC is the capability of producing simulations with the same underlying model as the real data. To this aim, we used *mssel*, a modified version of *ms* [193] kindly provided by Richard Hudson. This simulator uses a coalescent approach to generate neutrally-evolving regions linked to a site at which an allele is under selection, allowing recombination, gene-conversion and a variety of demographic model. Some parameters are required in order to resemble the specific region.

First of all, one needs to specify the sample composition desired in the output in terms of the number of haplotypes carrying the ancestral and the derived allele at the selected site. We used, for each region, the same composition as in the real data. As stated before, samples were enriched for the derived allele, hence in this way we kept the same bias also in the simulations, allowing us to directly compare real and simulated summary statistics.

The second set of parameters are the population mutation (Θ) and recombination (ρ) rates. For both of them, we assumed $N_e = 12000$. ρ was estimated, for each region, from the HapMap data. The mutation rate, instead, was calculated based on the divergence from other species and using 1000 Genomes Project data (`www.1000genomes.org`). The ancestral sequence for each base were defined by the 4-way (human, chimp, orangutan, rhesus) EPO alignments (ftp://ftp.ensembl.org/pub/release-54/emf/ensembl-compara/epo_4_catarrhini/). Only sites where the mutation occurred in the human lineage (uppercase in the 1000 Genomes Project notation: chimp, orangutan and rhesus sequence agree, human disagree) were considered. We then counted the number of derived alleles fixed in the whole 1000 Genomes Project sample (June 2011 release) and thus estimated the mutation rate assuming 5My split time and 25 years per generation. We obtained a consistent estimation of 1.2 mutation/generation/bp for all of the three regions. Interestingly, the region in chromosome 12 showed a decrease of nucleotide diversity in all the population (also in YRI, where there should be no selection). Using π estimator, we obtained a mutation rate of 0.6 for this region, roughly consistent in CEU, YRI and ASN.

Demographic model was obtained from [194]. Briefly, a 10-fold reduction of the population (bottleneck from 12,000 to 1,200) occurred between 27,500 and 40,000 years ago. A 10-fold expansion (from 12,000 to 120,000) then started 12,500 years ago.

Last input provided is the trajectory for the selected allele. While the previous parameters are kept, a new trajectory is generated for each simulation, varying the time when the mutation occurred but such that all of them ends with the actual frequency of the allele in the real population (CEU population, HapMap data). This device will ensure that all simulations will match the real data for the final frequency of the allele. First, a generation t^* is randomly drawn from a log-uniform (distribution uniform in the log space) *a priori* distribution between log(50) and log(3000). As stated before, indeed, we are going to estimate $\log(t)$ instead of t and the use of a flat *a priori* will not affect the estimation (maximum uncertainty).

After that, a selection coefficient s is drawn from a uniform distribution between 0 and 0.1. Given the starting point t^*, a trajectory is simulated forward time using s and a co-dominant diffusion process. If the final frequency of the allele matches the desired one (accounting for a sampling error), then the trajectory is saved and a coalescent simulation is generated. Otherwise, a new value for s is drawn and the process is repeated. Notice that in this process any time t has equal probability despite of different population size across time. We explicitly take into account this aspect in the step 5.

Step 2 – Introducing potential sources of error

Re-sequencing data are affected by some biases that could, in principle, influence the estimation. The strategy adopted was to try to identify these potential sources of errors and try to include them in the simulations as far as possible. In this way, even though these factors have an effect, they are took into account and simulations and real data are still comparable.

Among the features that more likely could have an effect, the most relevant are (i) the small sample size (14 individuals), (ii) the presence of region un-targeted and, most likely, un-sequenced sub-regions (gaps), (iii) phasing uncertainty for novel and rare variants and (iv) the complex demographic history of sequenced individuals (Europeans). In the following I'm going to present how each feature is introduced in the simulations and the comparison, in terms of relative error, with a vanilla model in which the feature is absent.

Even if the cost per base is decreasing every day, re-sequencing is still an expensive technology. For this reason, the number of individuals that can be analyzed is strongly limited by the available budget. Hence, the first analysis was aimed to assess the effect of a small sample size. In Figure 4.6 is reported the relative error of the ABC point estimates (posterior median) for different sample sizes (number of chromosmes). It is calculated as $\log_2\left(\frac{\hat{t}}{t}\right)$, where, for each simulations, t is the age of

the allele in that particular realization and $\hat{t}$ represent its ABC estimation. A value of 1.0 or -1.0 corresponds to a two-fold over- or under-estimation, respectively. Is it possible to see that, even though a bigger sample size actually decreases both the bias and the variance of the estimation, the improvement is not justified by the increasing sequencing cost.

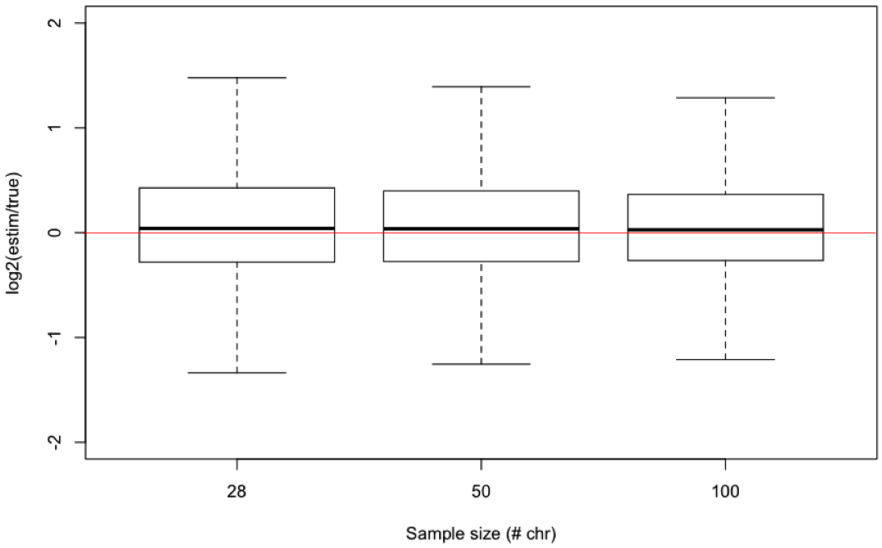

Figure 4.6: **Boxplot of $\log_2$ ratio of the estimate to the true age of the allele using different sample sizes.** $10,000$ simulations are produced as described in step 1, but with a constant population size ($N = 10,000$), uniform recombination rate, additive dominance and with the selected allele at a frequency of 50% in the population and in the sample.

As discussed in Section 4.1.3, the probes on the capturing array were designed masking repetitive elements. This means that subregions for which no probes were inserted on the chip are likely to remain un-sequenced. Actually, this is not totally true because a single fragment of DNA (and thus the two reads associated) can straddle a targeted and an un-targeted subregion. Hence, considering the typical length of the DNA fragments and the length of the reads, it is plausible that variants lying in the un-targeted subregions but within 200bp from a the edge (i.e. a targeted

subregion) still have enough coverage to be confidently called. On the other hand, variants to far from a targeted subregion will be missed and this can result in an overall decrease in the number of segregating sites in the regions. We emulated this potential bias in the simulations with a very simple approach. We removed from the simulated haplotypes all the segregating sites distant more than 200bp from a targeted subregions, according to their positions in the real regions[2]. Again, the presence of gaps influences in a negligible way the quality of the estimation (Figure 4.7).

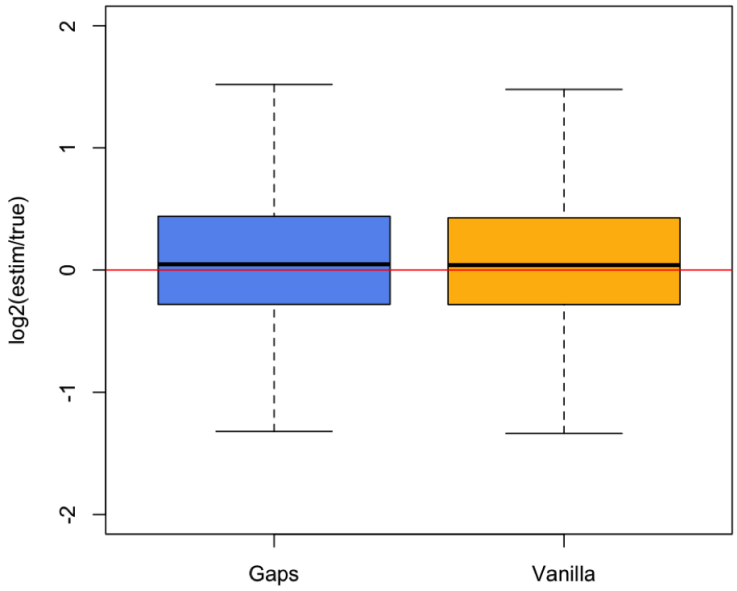

Figure 4.7: **Boxplot of $\log_2$ ratio of the estimate to the true age of the allele in simulations with (blue) and without (yellow) gaps.** Simulations parameters are the same as those reported in the caption of Figure 4.6

Next generation sequencing is one of the best way to identify novel and rare variants. Although they are crucial for population genetics in general and for the aim of this work in particular (2 out of the 3 summary statistics used relies on them), those

[2]NimbleGen provided a BED file with the positions, for each of the three regions, of the regions for which probes were designed

variants are, unfortunately, hard to correctly phase since no other examples are presents either in reference panels (novelty) or in the sequenced population (rarity). Of course, this problem is less severe with the increase of the allele count (intermediate or high frequency variants). On the other hand, all variants outputted by the simulator are, by construction, correctly phased. We introduced this uncertainty in the phasing in the simulations as well, but under some simplistic assumptions: (i) all variants with a count greater than 2 are correctly phased and (ii) all variants with a count equal or less than 2 (singletons and doubletons) have a 50% chance of being placed on the wrong chromosome. It is worth noticing that the latter is a very conservative assumption since it is assumed that singletons and doubletons are placed completely randomly on the chromosomes. The implementation of these two constraints is straightforward. As a first step, we coupled chromosomes in order to emulate individuals, respecting the actual composition of homozygotes and heterozygotes of the selected SNP. Then, for each variant with a count equal or less than 2 we randomly place, with a probability of 50%, the allele on the other chromosome of the individual. Once again, the effect of this uncertainty is almost absent (Figure 4.8).

Finally, we compared the performances of the estimation in simulations with constant population size versus a complex demographic history (as reported in step 1). Results are reported in Figure 4.9 divided per demographic phase (pre-bottleneck, bottleneck, recovery, expansion). It can be seen that demography has an impact on the estimation in the sense that both bias and variance changes over time.

Step 3 – Summary statistics

Another very important aspect of ABC is the choice of the right set of SSs. As discussed, we wish this set to be sufficient in order to obtain an accurate estimation of the *a posteriori* distribution. Moreover, because of the local-linear regression step and of the used distance metric Δ, this set needs to be non-redundant. Several

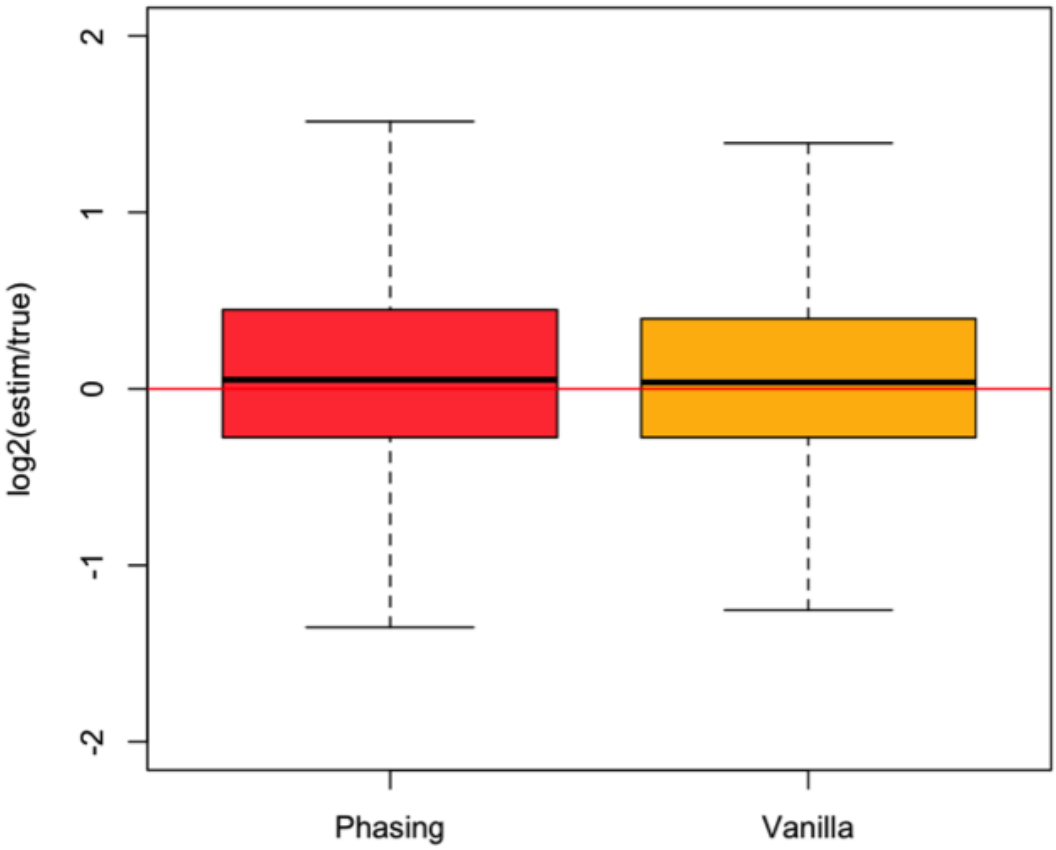

Figure 4.8: **Boxplot of $\log_2$ ratio of the estimate to the true age of the allele in simulations with (red) and without (yellow) phasing uncertainty.** Simulations parameters are the same as those reported in the caption of Figure 4.6

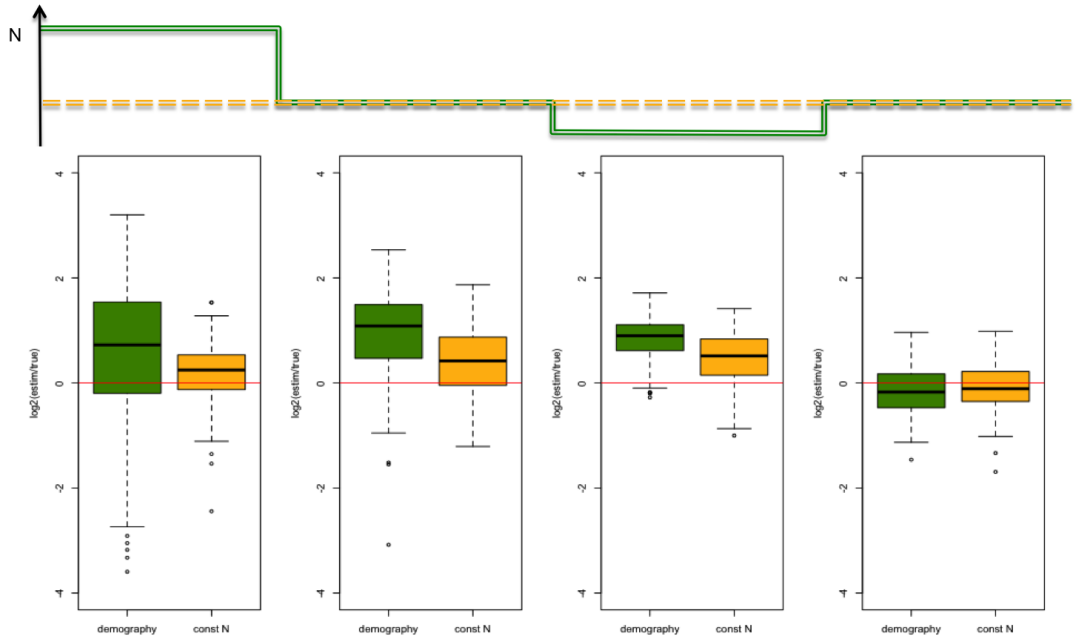

Figure 4.9: **Boxplot of $\log_2$ ratio of the estimate to the true age of the allele in simulations with constant population size (yellow) and with complex demography (green).** Four panels represents the four demographic phases: (from right to left) before the bottleneck, during the bottleneck, after the recovery and during the expansion. On the top is reported a cartoon of the demographic model (green: complex demography, yellow: constant population size) for convenience (left: present, right: past). Simulations parameters, but demograpy, are the same as those reported in the caption of Figure 4.6

strategies have been proposed to face this problem [195, 196], including PCA reduction and feature selection techniques, but up to now there are no general rules and the optimal choice still depends on the specific problem.

For the aims of our problem, we used two different sets of SSs. A first set (S_C) is used to be sure that simulated regions are actually representative of the real data and thus that the model is an appropriate one. This set is composed by 4 SSs, namely the number of segregating sites in the region (σ), nucleotide diversity of the region (π), Tajima's D (D_T) and Fay and Wu's H (H). These summary statistics are calculated considering all the 28 haplotypes (i.e. derived and ancestral haplotypes together). It is worth stressing that this set is not directly used for the estimation of the age, but it only has check purposes. For this reasons, eventual correlations among these 4 SSs are non influential. For each region, we checked that $S_C(D)$ was in the bulk of the distribution of simulated values.

The second set (S) was used for the age estimation. We choose three uncorrelated quantities known to be informative about the age of the selected allele. In particular, we considered (i) the length in genetic distance of the non-recombining haplotype surrounding the selected allele (L_H), (ii) the number of mutations accumulated in this haplotype (M_H) and (iii) the number of singleton variants divided by the total number of segregating sites in the haplotype (R_H). This set of SSs is calculated only on the set of chromosomes carrying the derived (selected) allele.

The haplotype was defined using the same approach proposed by Voight and colleague to define the region to integrate in the calculation of their iHS statistic [25]. Namely,the boundaries are defined as the positions away from the selected allele (core) where the extended haplotype homozygosity reaches 0.05. The first SS is the length of this haplotype. However, since the actual genetic map for the real data is non-uniform (as assumed in *mssel*), L_H is normalized, separately in the simulated and in the real data, by the total recombination length of the region. Interestingly, this SS is strictly related to the Voight's implementation of the haplotype decay

	L_H	M_H	R_H	age
L_H	1	0.09	0.156	0.90
M_H	-	1	0.155	0.44
R_H	-	-	1	0.56

Table 4.2: **Correlation among summary statistics.** Spearman's correlation coefficients, corrected for age.

method.

The second statistic, M_H, is instead related to the counting method described above. Together, these first two SSs are aimed to capture and put together the advantages of these two well established methods.

The last statistics, R_H, comes from the observation that allele frequency spectrum, and in particular the number of singleton variants, can be influenced by both demography and selection. Because all our simulations assume the same demographic scenario (and, hopefully, the real data too), the main influence on this number is more likely to be due to selection. But, since the mode of selection, again, is the same in all the simulations except for the age/strength of the selection, this number eventually result to be informative on the age.

In order to quantify the influence on the estimation and to exclude collinearity, we calculated the correlation among these three SSs as well as the correlation between each of them and the age (Table 4.2). Correlations among SSs is calculated correcting for age in order to exclude spurious correlations.

Step 4 – Acceptance criterion

We chose Euclidean distance as $\Delta(\cdot, \cdot)$ function, so that acceptance regions are spheres [192]. Regarding the value of ϵ, we again followed the approach originally

proposed by Beaumont and we set ϵ to be a quantile, P_ϵ, of the empirical distribution function $\Delta\left(S(X^*, S(D)\right)$. In the specific, we used $P_\epsilon = 0.001$ meaning that the 0.1% of simulated X^* that are closest to D are retained.

Step 5 – A posteriori density estimation

Once obtained the set of $\{X^*\}$ as defined in the previous step, the *a posteriori* density is estimated using a Gaussian kernel with a bandwidth chose following Silverman's "rule of thumb" (R default option). As pointed out in step 1, the size of the population at the particular t^* is not considered. But, since the probability for a mutation to occur is proportional to N,the density estimation process explicitly take into account this fact by weighting each t^* by $N\left(t^*\right)$.

4.2 Results and Discussion

In the previous section I described how data for the three regions were obtained and how ABC works. As stated, one of the step is to check that the model used to generate simulations is a good model for the data. In general, this is an hard problem and no definitive solutions exist yet. To assess, at least qualitatively, this point we checked whether the the simulator was able to produce haplotypes "compatible" with the real ones. In particular, we verified that the all the summary statistics in the "control set" (S_C, see 4.1.6) were in the bulk of the respective distributions of simulated values. Results are reported in Figure 4.10, 4.11 and 4.12.

It is possible to see that the simulations of regions on chromosome 4 and 15 do fit the real data. Unfortunately, this is not true for the region on chromosome 12. As it can be seen, the region exhibits a lower nucleotide diversity than what the simulator is able to reproduce. It is worth remarking that these four SSs are calculated considering both the derived and the ancestral (which are supposed to be neutral)

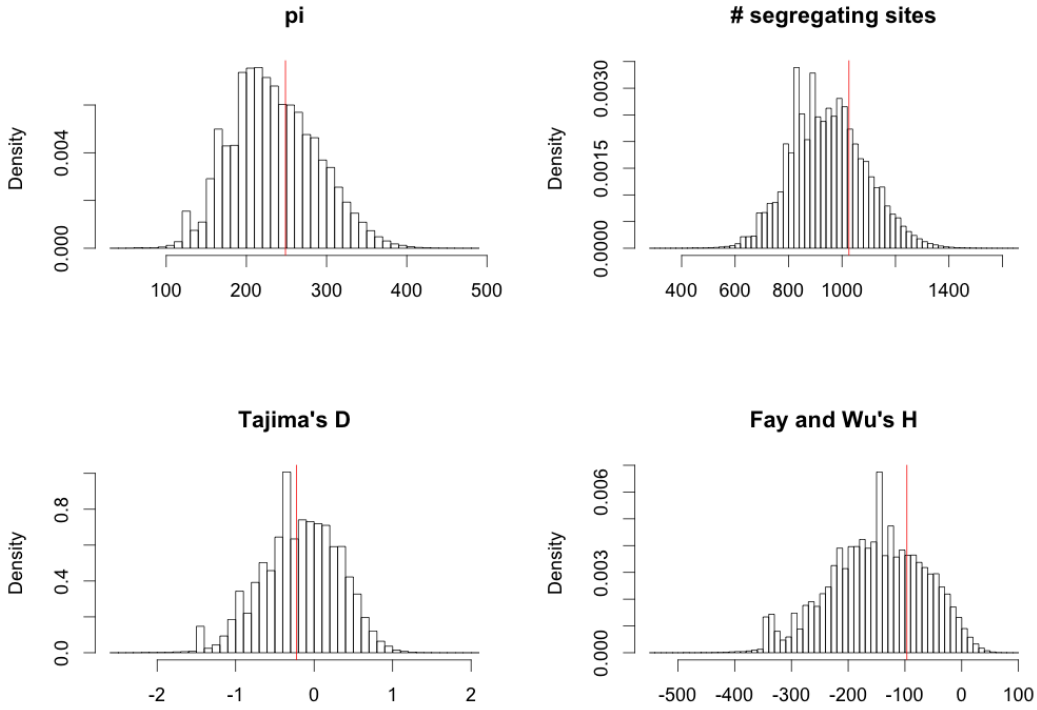

Figure 4.10: **Comparison between real and simulated "control" summary statistics (S_C) for region on chromosome 4.** Distribution of the four summary statistics in the control set (S_C) for the simulated region. Red line indicates the true value of the SS for the real region.

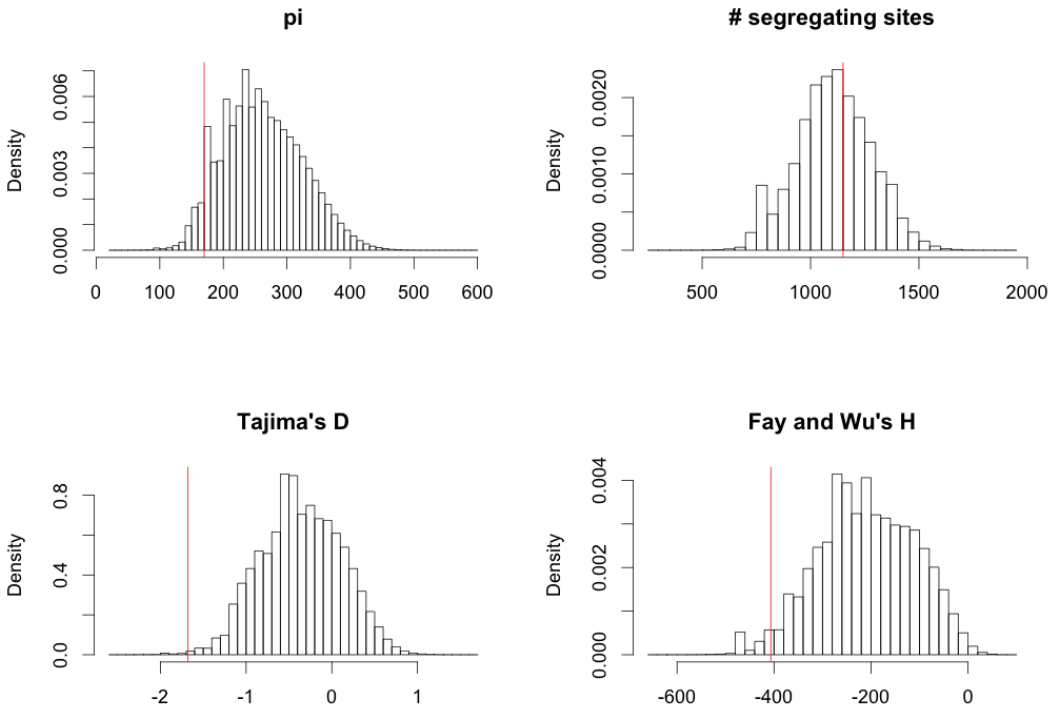

Figure 4.11: **Comparison between real and simulated "control" summary statistics (S_C) for region on chromosome 12.** Distribution of the four summary statistics in the control set (S_C) for the simulated region. Red line indicates the true value of the SS for the real region.

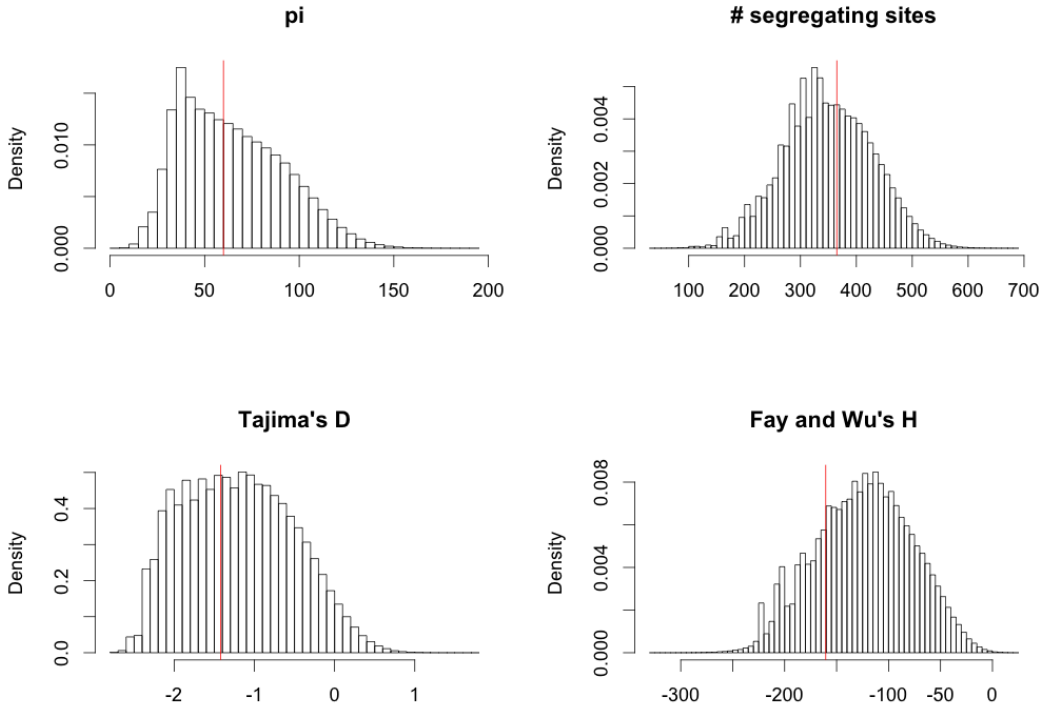

Figure 4.12: **Comparison between real and simulated "control" summary statistics (S_C) for region on chromosome 15.** Distribution of the four summary statistics in the control set (S_C) for the simulated region. Red line indicates the true value of the SS for the real region.

haplotypes and, given the nature of the statistics and the composition of the sample, they should be dominated by these latter. If we look at the same statistics, but calculated only considering the derived haplotypes, the situation is different (Figure 4.13). This time, simulations seem to fit the real data. Putting together, these results suggest the presence of some uncontrolled phenomenon affecting the "neutral" background (background selection, higher gene content, the presence of a previous sweep). However, the statistics calculated only considering derived haplotypes suggest that the model is good enough to represent, at least, the selection phenomenon. For this reason, and given the fact the the set of SSs used for the estimation only consider derived haplotypes, we are confident in proceeding with the age estimation also for the allele in this region, even though further investigations are needed.

We estimated the *a posteriori* distribution of the age for the three alleles using $1,000,000$ simulations for each region (Figure 4.14).

According to this estimation, the allelic variant associated with blue eyes and skin pigmentation (on chromosome 15) arose, as expected, most probably after the out-of-Africa when people moved to higher latitude. In particular, even tough the 95% credible interval is wide ([7,700 − 60,000]), the most reasonable range of dates for this allele is between the out-of-Africa migration ($\sim$ 30,000 years ago) and the end of the last glaciation/beginning of the agricultural era ($\sim 10 - 20{,}000$ years ago). As mentioned, we included this variant essentially for checking purposes since it is known that it arose after the migration out of Africa, and our estimation reflects the expectation.

More interesting are the estimations for the two variants associated with AD. The first thing that leaps out is that the two mutations occurred in two different moments. Credible intervals for variant on chromosome 4 and 12 are [23,000 − 75,000] and [3,300 − 18,000], respectively. Hence, the variant on chromosome 4 arose in a period of time compatible with the out-of-Africa migration, while the variant on chromosome 12 is likely to have been selected with the introduction of agriculture.

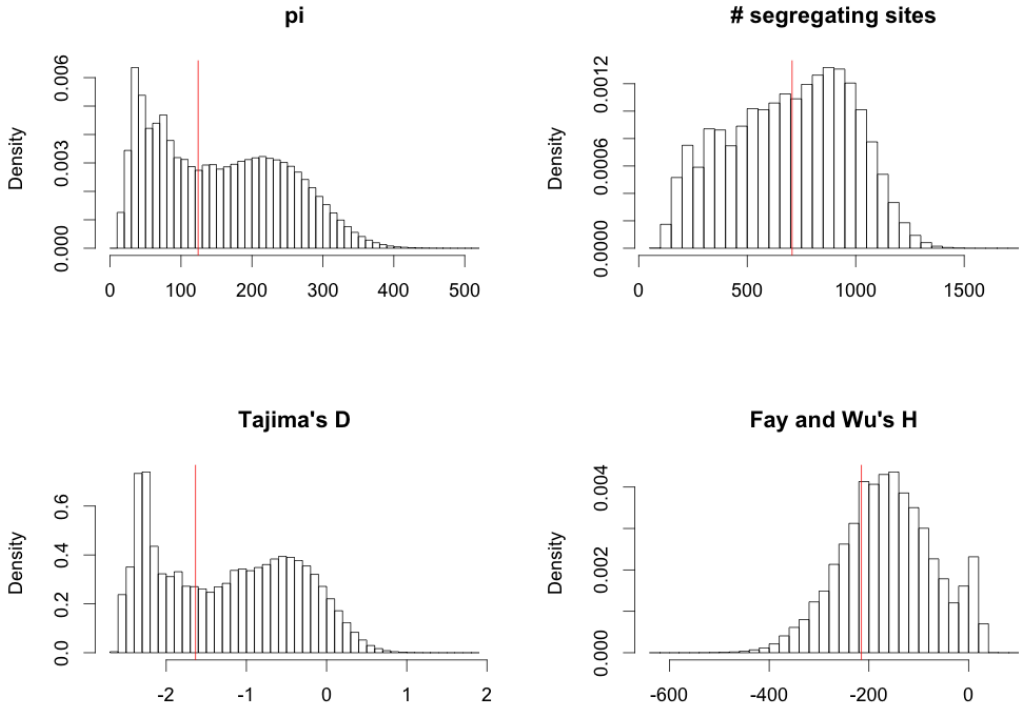

Figure 4.13: **Comparison between real and simulated "control" summary statistics (S_C) for region on chromosome 12, calculated considering only derived haplotypes.** Distribution of the four summary statistics in the control set (S_C) for the simulated region. Red line indicates the true value of the SS for the real region.

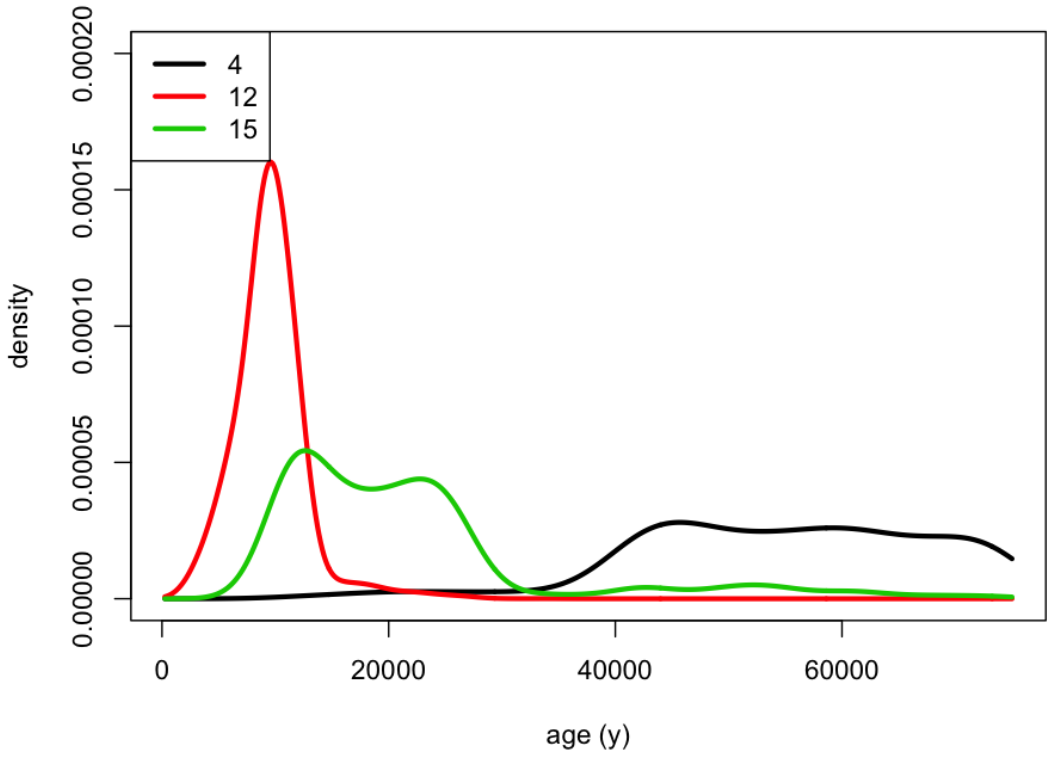

Figure 4.14: ***A posteriori* density estimation of the age of the three alleles.** Each line represents the *a posteriori* distribution of the age of the allele on chromosome 4, 12 and 15, respectively.

Ideally, variants associated with ADs and under positive selection should be protective. This is the case of the variant on chromosome 4, where the derived allele shows signature of positive selection and confers protection against ADs, too. Hence, the most plausible explanation for the protective allele being beneficial is because of the protection that gives by reducing ADs risk. It is worth noting that some ADs, like T1D, have an early onset, thus they can definitively affect the fitness.

The variant on chromosome 12, on the other hand, seems to be instead compatible with the scenario described in the introduction. With the introduction of agriculture and animal husbandry, humans transitioned to a sedentary agricultural subsistence, population size/density increased, and humans became attractive pathogen hosts because large populations in small areas maximize the chance of transmission be-

tween longer-lived barriers. Therefore, the human population growth during the Neolithic created the conditions that favored the emergence of pathogens that specialize in human hosts. In this scenario, as described more in details in Section 4.0.4, variants conferring a higher responsiveness to pathogens can be selected, even tough this can lead as a side effect to an over-responsiveness of the immune system and thus an increased risk for ADs. In a recent study, Corona and colleagues found that alleles under selection and increasing the risk of T1D are more frequent than alleles under selection and decreasing the risk [197]. Indeed, among the 80 SNPs most associated with T1D and showing strong signs of positive selection, 58 alleles associated with disease susceptibility show signs of positive selection, while only 22 associated with disease protection show signs of positive selection.

Although we examined just two variants and no general conclusions can be drawn, it seems that the balance between protection against pathogens and protection against ADs was broken towards two different directions in two different moments of our history.

Chapter 5

Future directions: polygenic adaptation and the case of study of height

Decades of work provided evidence of selective pressure in *Homo Sapiens* at the level of individual genes or loci [46, 65, 198]. Among the others, methods based on population differentiation were widely used to unveil their signature [199]. However, in most cases, variation in phenotype among individuals is the result of a polygenic effect, involving multiple genetic variations at multiple unlinked loci [200]. Up to now, only few studies investigated the presence of selective pressure on polygenic traits in humans. A possible explanation for this lack of evidence is that polygenic adaptation might be largely undetected by conventional methods able to look for selection [201, 202]. Furthermore, the identification of signatures of polygenic evolution could require genome-scale data sources, with a well-defined set of genetic variants involved in the polygenic effect.

Stature is one of most studied polygenic traits, because measuring height is easy and replicable and its inheritance well recognized. Recently, genome-wide association studies contributed in the identification of genes involved in this trait. Lango

Allen and colleagues, in particular, demonstrated that hundreds of genetic variants, in at least 180 loci, influence adult height [203]. Although this result explains approximately only 10% of the phenotypic variation in height, this study provides up to now the most detailed description of a polygenic effect in humans at molecular level.

From an evolutionary point of view, a complex interaction of different forces acts on stature and the complete dynamics is still not completely clear. In particular, several studies suggested a stabilizing selection on human height because of an increased number of health problems in very short and very tall individuals [204, 205]. At the same time, other studies invoked a directional sexual selection on male human height in that taller men often have more reproductive chances [206]. Anyway, a worldwide distributed sexual selection seems to represent an overall reasonable scenario, even though local adaptation phenomena that could favor particular heights and body shapes in particular environmental conditions cannot be excluded.

In this work we explored, in a simple simulated model, the behavior of F_{ST} on a generic polygenic trait. We then investigated whether a similar behavior was observable in a real case, namely in the set of loci related to height. [1]

5.1 Results and Discussion

To analyze the behavior of F_{ST} on a polygenic trait, we started by simulating the action of a polygenic stabilizing selection pressure in a very simple model. Such a model relies on the following simplifying assumptions: (i) each contributing allele has small and relatively equal additive effects, without neither environmental influences nor non-linear effects (dominance, epistasis, etc), (ii) individual fitness is given by

[1] The results presented in this chapter are published in: Amato R, Miele G, Monticelli A and Cocozza S. (2011) Signs of selective pressure on genetic variants affecting human height. PLoS ONE 6(11): e27588

a bell-shaped curve where the maximum is achieved by individual owning only a part of the advantageous alleles, (iii) no migration or demographic events affect the populations.

Under these simple assumptions, we observed that the set of alleles involved in the phenomenon has a higher mean F_{ST} value than those subject to genetic drift only. We thus checked whether the same behavior was observable in the 180 loci influencing adult height. Indeed, more than 80% of its variation within a given population is estimated to be attributable to additive genetic factors of small effects. Moreover, the authors found no evidence that non-additive effects including gene-gene interaction would increase the proportion of the phenotypic variance explained [203]. In the light of this, our simulations seem to reasonable model this scenario and, for this reason, we explicitly searched for an increase in the mean F_{ST} value of the 180 variants associated to height.

We found in the F_{ST} distribution for these variants an overrepresentation of higher values with respect to the genomic background (median=0.1 vs. 0.086; p=0.0356, one-tailed Mann-Whitney test; Figure 5.1). We investigated for potential confounding factors, first of all whether the increase was just due to the presence of outliers. We thus excluded from the initial set of height related variants those falling in the top or in the bottom 5% tail according to the genomic distribution of F_{ST} , obtaining a set of 161 SNPs (hereafter denoted as "core set"). The number of outliers excluded is compatible with the expected 5% (4% and 5% of the SNPs falling in the top and in the bottom tail, respectively). Moreover, the "core set" still exhibits a significant overrepresentation of high values F_{ST} value (median=0.1 vs. 0.086; p=0.0232, one-tailed Mann-Whitney test).

In a recent paper, Pritchard and Di Rienzo argued that a signature of selection on a polygenic trait should reasonably be small and spread across the loci [207]. Our result is, to some extension, in agreement with this hypothesis. Indeed, the increased F_{ST} value is not dominated by few outlier loci, but small and distributed.

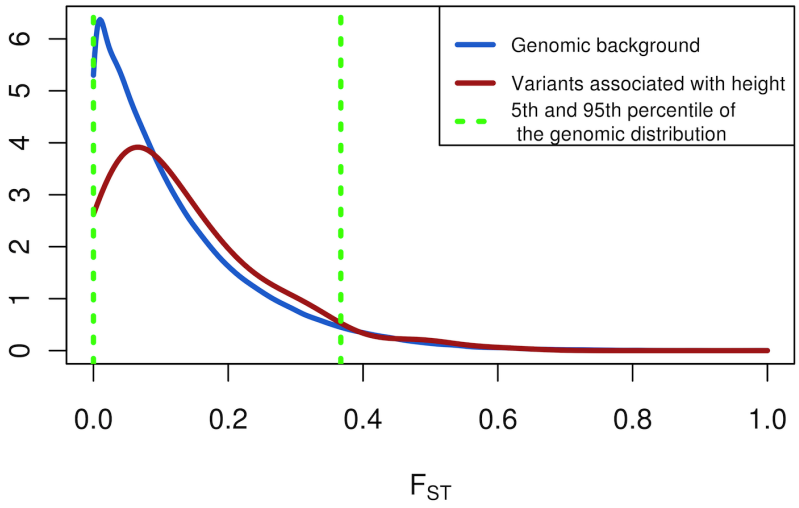

Figure 5.1: **F_{ST} density distribution for the genomic background and the height associated variants.** Green lines mark the 5th and 95th percentiles of the genomic distribution.

From this point of view, also the low value of statistical significance that we found could be expected.

On the other hand, if we consider height either under a stabilizing selective pressure or a worldwide distributed sexual selection, the higher mean F_{ST} value could seem unexpected. Under these hypotheses, indeed, we expected a result that is opposite to that obtained, even in presence of marginal phenomena of local adaptation. In these conditions, one should expect the majority of the genes having a vanishing F_{ST} and some outliers with very high values of F_{ST} . But looking at the simulated alleles trajectories over time, a possible explanation could be suggested. In simulations, one can observe that, in different populations, different sets of alleles become prevalent (Figure 5.2). In particular, the dynamics favors in each population the prevalence of a specific subset of alleles among the large amount of different subset choices all capable to maximize the phenotype under selective pressure. In absence of extraneous forces, the choice of the allele subset for each population is just randomly driven and, hence, it is highly probable to be different for each population.

This results in an increase of the mean diversity among the populations.

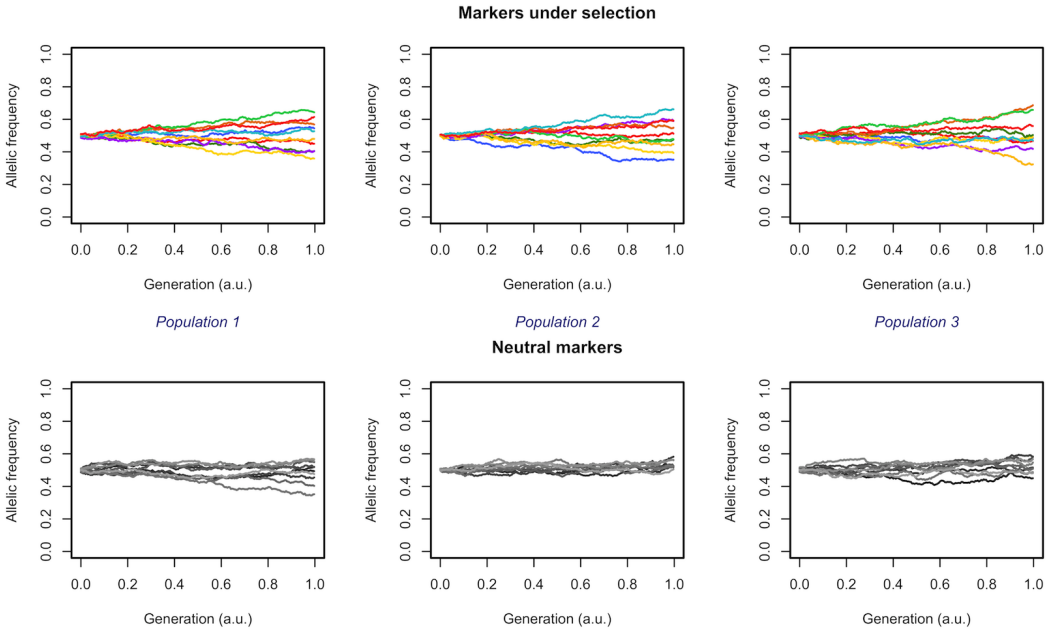

Figure 5.2: **Trajectories of the allelic frequencies for markers under polygenic selection in three simulated populations.** Each column, i.e. top and bottom panel together, represents a different population. Top panels show trajectories over time for the set of 10 alleles under polygenic selection; bottom panels show trajectories for set of 10 neutral alleles. Different colors mark different alleles, consistently across populations.

This very simple model resembles the behavior of a well-known statistical mechanics phenomenon denoted as Spontaneous Symmetry Breaking (SSB). This mechanism generally plays a relevant role in system self-organization, and it is common in many fields of Natural Sciences where a system described in a theoretically symmetrical way ends up in a non-symmetrical state. The physics of condensed matter probably provides the most striking examples of SSB phenomena. In a ferromagnet cooled below its critical temperature, as the thermal fluctuations slow down, will become energetically favorable the appearance of domains where all elementary magnets point in the same direction, randomly chosen, and hence breaking the original rota-

tional symmetry.

SSB has been widely observed in biological systems. Population genetics also provides examples of SSB even though in this case to lead the breaking are the initial conditions, stochastically modified by events like drift, bottlenecks, etc, and other stochastic events like the born of new mutations. Among the others, we can quote the role of symmetry breaking and coarsening in spatially distributed evolutionary processes relevant for genetic diversity and species formation [208], and the relevance of symmetry breaking in the long-term evolution of multilocus traits [209].

The symmetry-breaking scenario could represent a simple yet reasonable model of the selection acting on height. But, even though height is basically under stabilizing selection, local adaptive phenomena cannot be excluded. To explicitly explore the presence of local adaptive phenomena, we analyzed how iHS, another marker of selective pressure, is distributed in genetic variants involved in height. iHS is a score specifically oriented to detect recent adaptive phenomena with higher geographical resolution. We found that, in each population, the number of SNPs associated with height having a value of iHS falling in the highest 5% of the genomic distribution is compatible with the expectation (varying from 4-7% across populations; Figure 5.3). This finding seems to indicate that, even if recent local adaptation phenomena cannot be excluded, their role seems to be marginal.

Another hypothesis that we explored was that loci responsible of local adaptive phenomena on different phenotypes, in linkage disequilibrium with height related variants, were responsible of the selective signatures that we found. Under this hypothesis, the increased mean F_{ST} value that we attributed to the height associated genes could be the by-product of selective pressures acting on other traits. To test this hypothesis, we extracted height related SNPs that were in linkage disequilibrium ($r_2 > 0.8$ in at least one population) with alleles associated to other traits in genome-wide studies. We found 11 SNPs, 9 of which belonging to the "core set" (Table 5.1). As far as we know, there are no clearly identified signals of selec-

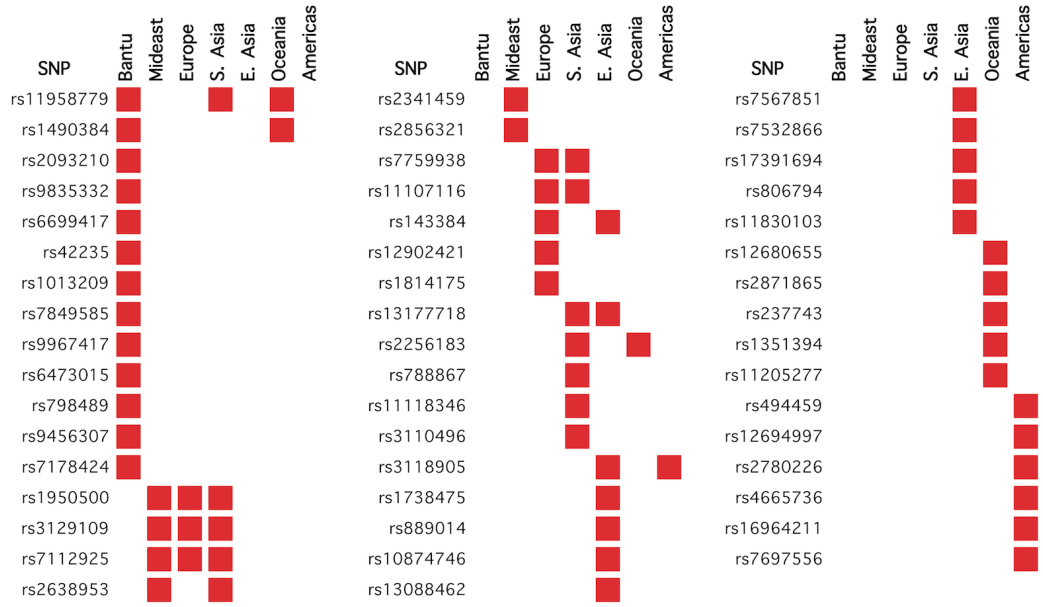

Figure 5.3: **Variants associated to height with high iHS score.** Red squares indicate mark the population in which the iHS score for the variant falls in the top 5th percentile of the respective distribution.

tive pressure on the phenotypes in linkage disequilibrium with the height associated SNPs. Moreover, also removing these SNPs from the analysis, the "core set" still show a significant higher F_{ST} (median= 0.099 vs. 0.086; p= 0.03452, one-tailed Mann-Whitney test).

We are aware that the increase in the mean F_{ST} value could be, at least in part, due to an eventual difference across the mean heights of the three populations considered. In other words, the increase in the mean F_{ST} value can be divided in two distinct components, where the first one accounts for the differences in the mean height while the second one accounts for the pure genetic differences. Unfortunately, at the best of our knowledge, data regarding the mean height of each population are not present in literature so far. For this reason, we were unable to estimate the relative weight of the two components or exclude the effect of the first one. Moreover, it is worth stressing that the 180 associated SNPs found by Lango Allen and colleagues only explain about 10% of the variance in adult height [203]. For this reason, wherever present, phenotypic differences among populations should have a marginal effect.

5.2 Materials and Methods

5.2.1 Data and statistical analysis

Analysis is based on the HapMap Public Release #27 (merged II+III) datafiles. We analyzed data from the CEPH (Utah residents with ancestry from northern and western Europe), Yoruba in Ibadan, Nigeria (YRI), Han Chinese in Beijing, China (CHB) and Japanese in Tokyo, Japan (JPT) samples. We pooled the CHB and JPT samples to form a single sample. Additional SNP information about physical positions and SNP-gene association were obtained from dbSNP build 129. We excluded by this analysis SNPs that were either non sampled or non polymorphic in all the three samples. We also excluded SNPs with a minor allele frequency $< 5\%$ in all of

Height variant	F_{ST}	GWAS variant	Disease/Trait	PubMed ID	r^2_{CEU}	r^2_{ASN}	r^2_{YRI}	Core set?	Note
rs10037512	0.218	rs1366594	Bone mineral density (hip)	19801982	0.967	0.931	1.000	Y	
rs10874746	0.313	rs12745968	Biplolar disorder and schizophrenia	20889312	0.928	1.000		Y	
rs17782313	0.007	rs10871777	Obesity (extreme)	20421936	1.000	1.000	0.930	Y	in LD with rs12970134 (r^2=0.813 CEU) and rs571312 (r^2=1 CEU; 0.963 ASN)
		rs12970134	Waist circumference and related phenotypes	18454146	0.813			Y	in LD with rs571312 (r^2=0.813 CEU) and rs10871777 (r^2=0.813 CEU)
		rs571312	Body mass index	20935630	1.000	0.963		Y	in LD with rs10871777 (r^2=1 CEU; r^2=0.963 ASN) and rs12970134 (r^2=0.813 CEU)
rs4665736	0.268	rs713586	Body mass index	20935630	0.811			Y	
rs2145272	0.122	rs2145270	Body mass index	19079261	0.898	1.000		Y	
rs3129109	0.295	rs4947339	Platelet aggregation	20526338	0.898			Y	
rs7759938	0.130	rs314276	Menarche (age at onset)	19448623	0.965	0.971		Y	in LD with rs314280 (r^2=1 ASN)
		rs314280	Menarche (age at onset)	19448622		0.971		Y	in LD with rs314276 (r^2=1 ASN)
rs6457620	0.010	rs6457617	Rheumatoid arthritis	18668548	1.000	1.000	0.967	Y	
rs2066807	0.044	rs2066808	Psoriasis	19169254	1.000			Y	
rs1490384	0.604	rs9388489	Type 1 diabetes	19430480	0.842	1.000	0.938	N	
rs2093210	0.377	rs10483727	Optic disc size (rim)	20395239	0.802		1.000	N	

Table 5.1: **Variants associated to height in linkage disequilibrium ($r^2 > 0.8$ in at least one population) with alleles associated with other traits in genome-wide association studies.** For each height variant is reported the GWAS variant in linkage, the trait and the PubMed ID for the study itself.

the 3 samples. Per-population linkage disequilibrium data (r^2) were obtained from HapMap Public Release #27 (merged II+III) as well. F_{ST} was calculated using the unbiased estimator proposed by Weir and Cockerham as discussed in 2.3. All data was merged in a local MySQL database. As "genomic background" we refer to all the SNPs in this database, for a total of $3,294,557$ SNPs.

Variants associated with height were collected from [203]. Of the 180 provided SNPs, 176 were present in our database and hence considered in our analysis.

The normalized iHS scores were obtained from UCSC Genome Browser "HGDP iHS" track. They were calculated using SNPs genotyped in 1043 individual coming from 53 populations worldwide by the Human Genome Diversity Project in collaboration with the Centre d'Etude du Polymorphisme Humain (HGDP-CEPH). The 53 populations were divided into seven continental groups: Africa (Bantu populations only), Middle East, Europe, South Asia, East Asia, Oceania and the Americas. Per-SNP iHS scores were smoothed in windows of 31 SNPs, centered on each SNP.

Data on the association of SNPs with diseases was obtained from a catalog of genome wide association studies available at `http://www.genome.gov/gwastudies`, (accessed 12/13/10).

All statistical analyses were performed with R ver. 2.10 (R Foundation for Statistical Computing, Vienna, Austria; `http://www.r-project.org/`) considering 0.05 as significance threshold.

5.2.2 Simulations

We simulated three populations of diploid organisms of fixed sample size evolving independently each other. Individuals are represented by 20 markers, where half of them are assumed to be neutral, and the remaining ones contribute additively and uniformly to the phenotype in a codominant way. Basically, markers evolve under a Wright-Fisher model with recombination. Polygenic selection is then simulated

through viability selection. Denoting with m the number of beneficial alleles carried by an individual, each contributing with x to the phenotype, the fitness is parameterized as $\exp\left(-(m-\mu)^2 \frac{x^2}{2s}\right)$. In the previous expression the quantity s measures the selection strength and μ is the number of beneficial alleles that maximizes the fitness. In figure 5.2 it is shown a particular case where the sample size is $N = 10000$, $s = 10$, $\mu = 10$ and $x = 5$. Furthermore, the initial allele frequency is set to 0.5 for all markers. Lower values of N increase the effect of genetic drift, while different values of μ change the number of alleles rising in frequency in each population. By tuning the value of x^2/s one can change the strength of selection hence affecting the time required to observe relevant variation in allele frequencies. Nevertheless, these changes do not qualitatively affect the results.

Bibliography

1. Van Heyningen, V. & Yeyati, P. L. Mechanisms of non-Mendelian inheritance in genetic disease. *Human molecular genetics* **13 Spec No,** R225–33 (Oct. 2004).

2. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405,** 847–56 (June 2000).

3. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science (New York, N.Y.)* **322,** 881–8 (Nov. 2008).

4. McCarthy, M. I. & Hirschhorn, J. N. Genome-wide association studies: potential next steps on a genetic journey. *Human molecular genetics* **17,** R156–65 (Oct. 2008).

5. Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. en. *Nature reviews. Genetics* **6,** 109–18 (Feb. 2005).

6. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–53 (Oct. 2009).

7. Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science (New York, N.Y.)* **316,** 1336–41 (June 2007).

8. Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M. J. & Seddon, J. M. Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. en. *Nature genetics* **38,** 1055–9 (Sept. 2006).

9. Luca, F. & Di Rienzo, A. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, Chichester, UK, 2007). doi:10.1002/9780470015902.a0020758.

10. Stearns, S. C., Nesse, R. M., Govindaraju, D. R. & Ellison, P. T. Evolution in health and medicine Sackler colloquium: Evolutionary perspectives on health and medicine. *Proceedings of the National Academy of Sciences of the United States of America* **107 Suppl,** 1691–5 (Jan. 2010).

11. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* 4th (Sinauer Associates, Inc., Sunderland, MA, 2006).

12. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends in genetics : TIG* **17,** 502–10 (Sept. 2001).

13. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics* **69,** 124–37 (July 2001).

14. Rodier, F., Campisi, J. & Bhaumik, D. Two faces of p53: aging and tumor suppression. *Nucleic acids research* **35,** 7475–84 (Jan. 2007).

15. Di Rienzo, A. & Hudson, R. R. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends in Genetics: TIG* **21,** 596–601 (Nov. 2005).

16. Barbujani, G. & Goldstein, D. B. Africans and Asians abroad: genetic diversity in Europe. *Annual Review of Genomics and Human Genetics* **5,** 119–150 (2004).

17. Cavalli-Sforza, L. & Menozzi, P. The history and geography of human genes (Aug. 1994).

18. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Feb. 1985).

19. Akey, J., Zhang, G., Zhang, K., Jin, L. & Shriver, M. Interrogating a high-density SNP map for signatures of natural selection. *Genome research* **12,** 1805 (Dec. 2002).

20. Bamshad, M. & Wooding, S. Signatures of natural selection in the human genome. *Nature Reviews Genetics* **4,** 99–111 (Feb. 2003).

21. Nielsen, R. Molecular signatures of natural selection. *Annual Review of Genetics* **39,** 197–218 (2005).

22. Ioannidis, J. P. A., Ntzani, E. E. & Trikalinos, T. A. 'Racial' differences in genetic effects for complex diseases. *Nature Genetics* **36,** 1312–1318 (Dec. 2004).

23. Davey Smith, G., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J. & Burton, P. R. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* **366,** 1484–1498 (Oct. 2005).

24. Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W. & Akey, J. M. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research* **16,** 980–989 (Aug. 2006).

25. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biology* **4,** e72 (Mar. 2006).

26. Wang, E. T., Kodama, G., Baldi, P. & Moyzis, R. K. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 135–140 (Jan. 2006).

27. Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449,** 913–918 (Oct. 2007).

28. Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D. & Nielsen, R. Localizing recent adaptive evolution in the human genome. *PLoS Genetics* **3,** e90 (June 2007).

29. Barreiro, L., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. *Nature* **200,** 8 (Mar. 2008).

30. Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* **19,** 826–837 (May 2009).

31. Wright, S. The genetic structure of populations. *Annals of Eugenic* **15,** 323–354 (1951).

32. Weir, B. S. & Hill, W. G. Estimating F-statistics. *Annual Review of Genetics* **36,** 721–750 (2002).

33. International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (Dec. 2003).

34. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (Oct. 2005).

35. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38,** 1358–1370 (Nov. 1984).

36. Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M. & Hill, W. G. Measures of human population structure show heterogeneity among genomic regions. *Genome Research* **15,** 1468–1476 (Nov. 2005).

37. Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R. & Siepel, A. Patterns of positive selection in six Mammalian genomes. *PLoS Genetics* **4,** e1000144 (2008).

38. Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34,** 267–273 (July 2003).

39. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–15550 (Oct. 2005).

40. Cavalli-Sforza, L. The human genome diversity project: past, present and future. *Nat Rev Genet* **6,** 333–340 (Apr. 2005).

41. Seielstad, M. T., Minch, E & Cavalli-Sforza, L. L. Genetic evidence for a higher female migration rate in humans. *Nature Genetics* **20,** 278–280 (Nov. 1998).

42. Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, N.Y.)* **319,** 1100–1104 (Feb. 2008).

43. Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nature Genetics* **41,** 66–70 (Jan. 2009).

44. Tarazona-Santos, E., Bernig, T., Burdett, L., Magalhaes, W. C. S., Fabbri, C., Liao, J., Redondo, R. A. F., *et al.* CYBB, an NADPH-oxidase gene: restricted diversity in humans and evidence for differential long-term purifying selection on transmembrane and cytosolic domains. *Human Mutation* **29,** 623–632 (May 2008).

45. McVean, G. & Spencer, C. C. A. Scanning the human genome for signals of selection. *Current Opinion in Genetics & Development* **16,** 624–629 (Dec. 2006).

46. Sabeti, P. C. Positive Natural Selection in the Human Lineage. *Science* **312,** 1614–1620 (June 2006).

47. Teshima, K. M., Coop, G. & Przeworski, M. How reliable are empirical genomic scans for selective sweeps? *Genome Research* **16,** 702–712 (June 2006).

48. Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 11667–11672 (July 2009).

49. Sardiello, M., Palmieri, M., di Ronza, A., Medina, D. L., Valenza, M., Gennarino, V. A., Di Malta, C., *et al.* A gene network regulating lysosomal biogenesis and function. *Science (New York, N.Y.)* **325,** 473–477 (July 2009).

50. Holden, M., Deng, S. & Wojnowski, L. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* **24,** 2784–2785 (Dec. 2008).

51. Iorio, F., Tagliaferri, R. & di Bernardo, D. Identifying network of drug mode of action by gene expression profiling. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* **16,** 241–251 (Feb. 2009).

52. Ferrer-Admetlla, A., Bosch, E., Sikora, M., Marquès-Bonet, T., Ramírez-Soriano, A., Muntasell, A., Navarro, A., *et al.* Balancing selection is the main force shaping the evolution of innate immunity genes. *Journal of Immunology (Baltimore, Md.: 1950)* **181,** 1315–1322 (July 2008).

53. Mukherjee, S., Sarkar-Roy, N., Wagener, D. K. & Majumder, P. P. Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. *Proceedings of the National*

Academy of Sciences of the United States of America **106,** 7073–7078 (Apr. 2009).

54. Berridge, M. J., Lipp, P & Bootman, M. D. The versatility and universality of calcium signalling. *Nature Reviews. Molecular Cell Biology* **1,** 11–21 (Oct. 2000).

55. Green, M. L., Singh, A. V., Zhang, Y., Nemeth, K. A., Sulik, K. K. & Knudsen, T. B. Reprogramming of genetic networks during initiation of the Fetal Alcohol Syndrome. *Developmental Dynamics: An Official Publication of the American Association of Anatomists* **236,** 613–631 (Feb. 2007).

56. Hancock, A. M., Witonsky, D. B., Gordon, A. S., Eshel, G., Pritchard, J. K., Coop, G. & Di Rienzo, A. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics* **4,** e32 (Feb. 2008).

57. Helgason, A., Pálsson, S., Thorleifsson, G., Grant, S. F. A., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., *et al.* Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nature Genetics* **39,** 218–225 (Feb. 2007).

58. Freathy, R. M., Weedon, M. N., Bennett, A., Hypponen, E., Relton, C. L., Knight, B., Shields, B., *et al.* Type 2 diabetes TCF7L2 risk genotypes alter birth weight: a study of 24,053 individuals. *American Journal of Human Genetics* **80,** 1150–1161 (June 2007).

59. Lin, L., Lesnick, T. G., Maraganore, D. M. & Isacson, O. Axon guidance and synaptic maintenance: preclinical markers for neurodegenerative disease and therapeutics. *Trends in Neurosciences* **32,** 142–149 (Mar. 2009).

60. Allen, N., Bagade, S., McQueen, M., Ioannidis, J., Kavvoura, F., Khoury, M., Tanzi, R., *et al.* Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nature genetics* **40,** 827–834 (July 2008).

61. Crespi, B., Summers, K. & Dorus, S. Adaptive evolution of genes underlying schizophrenia. *Proceedings. Biological Sciences / The Royal Society* **274,** 2801–2810 (Nov. 2007).

62. Clarimon, J., Scholz, S., Fung, H.-C., Hardy, J., Eerola, J., Hellstrom, O., Chen, C.-M., *et al.* Conflicting results regarding the semaphorin gene (SEMA5A) and the risk for Parkinson disease. *American journal of human genetics* **78,** 1082–4; author reply 1092–4 (June 2006).

63. Fujii, T., Iijima, Y., Kondo, H., Shizuno, T., Hori, H., Nakabayashi, T., Arima, K., *et al.* Failure to confirm an association between the PLXNA2 gene and schizophrenia in a Japanese population. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* **31,** 873–877 (May 2007).

64. Neel, J. V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American Journal of Human Genetics* **14,** 353–362 (Dec. 1962).

65. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. G. Recent and ongoing selection in the human genome. *Nature reviews. Genetics* **8,** 857–68 (Nov. 2007).

66. Watanabe, S., Kang, D.-H., Feng, L., Nakagawa, T., Kanellis, J., Lan, H., Mazzali, M., *et al.* Uric acid, hominoid evolution, and the pathogenesis of salt-sensitivity. *Hypertension* **40,** 355–360 (Sept. 2002).

67. Diamond, J. The double puzzle of diabetes. *Nature* **423,** 599–602 (June 2003).

68. Lohmueller, K. E., Mauney, M. M., Reich, D. & Braverman, J. M. Variants associated with common disease are not unusually differentiated in frequency across populations. *American Journal of Human Genetics* **78,** 130–136 (Jan. 2006).

69. Myles, S., Davison, D., Barrett, J., Stoneking, M. & Timpson, N. Worldwide population differentiation at disease-associated SNPs. *BMC Medical Genomics* **1,** 22 (2008).

70. Peng, B. & Kimmel, M. Simulations provide support for the common disease-common variant hypothesis. *Genetics* **175,** 763–776 (Feb. 2007).

71. Evans, W. E. & Relling, M. V. Moving towards individualized medicine with pharmacogenomics. *Nature* **429,** 464–468 (May 2004).

72. Keller, M. C. & Miller, G. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *The Behavioral and Brain Sciences* **29,** 385–404; discussion 405–452 (Aug. 2006).

73. Amigo, J., Salas, A., Phillips, C. & Carracedo, A. SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics* **9,** 428 (2008).

74. Weicker, J. J., Brumfield, R. T. & Winker, K. Estimating the unbiased estimator theta for population genetic survey data. *Evolution; International Journal of Organic Evolution* **55,** 2601–2605 (Dec. 2001).

75. Lahiri, S. *Resampling Methods for Dependent Data* 1st ed. (Springer, Aug. 2003).

76. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics (Oxford, England)* **23,** 3251–3253 (Dec. 2007).

77. Sezgin, E., Duvernell, D. D., Matzkin, L. M., Duan, Y., Zhu, C.-T., Verrelli, B. C. & Eanes, W. F. Single-locus latitudinal clines and their relationship to temperate adaptation in metabolic genes and derived alleles in Drosophila melanogaster. *Genetics* **168,** 923–931 (Oct. 2004).

78. Verrelli, B. C. & Eanes, W. F. Clinal variation for amino acid polymorphisms at the Pgm locus in Drosophila melanogaster. *Genetics* **157,** 1649–1663 (Apr. 2001).

79. Balasubramanian, S., Sureshkumar, S., Agrawal, M., Michael, T., Wessinger, C., Maloof, J., Clark, R., *et al.* The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of Arabidopsis thaliana. *Nature genetics* **38,** 711 (June 2006).

80. Caicedo, A. L., Stinchcombe, J. R., Olsen, K. M., Schmitt, J. & Purugganan, M. D. Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 15670–5 (Nov. 2004).

81. Jablonski, N. G. & Chaplin, G. The evolution of human skin coloration. *Journal of Human Evolution* **39,** 57–106 (July 2000).

82. Pereira, D. S., Tufik, S., Louzada, F. M., Benedito-Silva, A. A., Lopez, A. R., Lemos, N. A., Korczak, A. L., *et al.* Association of the length polymorphism in the human Per3 gene with the delayed sleep-phase syndrome: does latitude have an influence upon it? *Sleep* **28,** 29–32 (Jan. 2005).

83. Cruciani, F., Trombetta, B., Labuda, D., Modiano, D., Torroni, A., Costa, R. & Scozzari, R. Genetic diversity patterns at the human clock gene period 2 are suggestive of population-specific positive selection. *European Journal of Human Genetics: EJHG* **16,** 1526–1534 (Dec. 2008).

84. D F Roberts. Body weight, race and climate. *American Journal of Physical Anthropology* **11,** 533–558 (Dec. 1953).

85. Irmak, M. K., Korkmaz, A & Erogul, O. Selective brain cooling seems to be a mechanism leading to human craniofacial diversity observed in different geographical regions. *Medical Hypotheses* **63,** 974–979 (2004).

86. Young, J. H., Chang, Y.-P. C., Kim, J. D.-O., Chretien, J.-P., Klag, M. J., Levine, M. A., Ruff, C. B., *et al.* Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genetics* **1,** e82 (Dec. 2005).

87. Thompson, E. E., Kuttab-Boulos, H, Witonsky, D, Yang, L, Roe, B. A. & Di Rienzo, A. CYP3A variation and the evolution of salt-sensitivity variants. *American Journal of Human Genetics* **75,** 1059–1069 (Dec. 2004).

88. Parra, E. J. Human pigmentation variation: evolution, genetic basis, and implications for public health. *American Journal of Physical Anthropology* **Suppl 45,** 85–105 (2007).

89. Eisen, A & Calne, D. Amyotrophic lateral sclerosis, Parkinson's disease and Alzheimer's disease: phylogenetic disorders of the human neocortex sharing many characteristics. *The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques* **19,** 117–123 (Feb. 1992).

90. Karami, S., Boffetta, P., Stewart, P., Rothman, N., Hunting, K. L., Dosemeci, M., Berndt, S. I., *et al.* Occupational sunlight exposure and risk of renal cell carcinoma. *Cancer*. doi:10.1002/cncr.24939 (Mar. 2010).

91. Kinney, D. K., Teixeira, P., Hsu, D., Napoleon, S. C., Crowley, D. J., Miller, A., Hyman, W., *et al.* Relation of schizophrenia prevalence to latitude, climate, fish consumption, infant mortality, and skin color: a role for prenatal vitamin d deficiency and infections? *Schizophrenia Bulletin* **35,** 582–595 (May 2009).

92. Ashburner, M, Ball, C., Blake, J. & Botstein, D. Gene ontology: tool for the unification of biology.{T} he {G} ene {O} ntology {C} onsortium. *Nature Genetics* **25,** 25–29 (May 2000).

93. Bauer, S., Grossmann, S., Vingron, M. & Robinson, P. Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24,** 1650 (July 2008).

94. Bauer, S., Gagneur, J. & Robinson, P. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic acids research* **38,** 3523 (Feb. 2010).

95. Taylor, B., Lucas, R., Dear, K. & Kilpatrick, T. Latitudinal variation in incidence and type of first central nervous system demyelinating events. *Multiple.* doi:10.1177/1352458509359724 (Feb. 2010).

96. Lill, C., McQueen, M. B., Roehr, J., Bagade, S., Schjeide, B., Zipp, F & Bertram, L. *The MSGene Database* Jan. 2010.

97. Lill, C., Bagade, S., McQueen, M. B., Roehr, J., Kavvoura, F. K., Schjeide, B., Allen, N. C., *et al. The PDGene Database* Jan. 2010.

98. Bertram, L., McQueen, M., Mullin, K., Blacker, D. & Tanzi, R. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature genetics* **39,** 17–24 (Jan. 2007).

99. Sun, J., Kuo, P.-H., Riley, B. P., Kendler, K. S. & Zhao, Z. Candidate genes for schizophrenia: a survey of association studies and gene ranking. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* **147B,** 1173–81 (Oct. 2008).

100. Ross, C. A., Margolis, R. L., Reading, S. A. J., Pletnikov, M. & Coyle, J. T. Neurobiology of schizophrenia. *Neuron* **52,** 139–53 (Oct. 2006).

101. Wang, T.-T., Tavera-Mendoza, L. E., Laperriere, D., Libby, E., MacLeod, N. B., Nagai, Y., Bourdeau, V., *et al.* Large-scale in silico and microarray-based identification of direct 1,25-dihydroxyvitamin D3 target genes. *Molecular Endocrinology (Baltimore, Md.)* **19,** 2685–2695 (Nov. 2005).

102. Cober, E. R. & Morrison, M. J. Regulation of seed yield and agronomic characters by photoperiod sensitivity and growth habit genes in soybean. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* **120,** 1005–1012 (Mar. 2010).

103. Zhang, D., Zhang, H., Wang, M., Sun, J., Qi, Y., Wang, F., Wei, X., *et al.* Genetic structure and differentiation of Oryza sativa L. in China revealed

by microsatellites. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* **119,** 1105–1117 (Oct. 2009).

104. Guo, Z., Song, Y., Zhou, R., Ren, Z. & Jia, J. Discovery, evaluation and distribution of haplotypes of the wheat Ppd-D1 gene. *The New Phytologist.* doi:10.1111/j.1469-8137.2009.03099.x (Dec. 2009).

105. Van 't Land, J, Van Putten, W. F., Villarroel, H, Kamping, A & Van Delden, W. Latitudinal variation for two enzyme loci and an inversion polymorphism in Drosophila melanogaster from Central and South America. *Evolution; International Journal of Organic Evolution* **54,** 201–209 (Feb. 2000).

106. Johnsen, A, Fidler, A. E., Kuhn, S, Carter, K. L., Hoffmann, A, Barr, I. R., Biard, C, *et al.* Avian Clock gene polymorphism: evidence for a latitudinal cline in allele frequencies. *Molecular Ecology* **16,** 4867–4880 (Nov. 2007).

107. Lucotte, G. & Mercier, G. The population distribution of the Met allele at the PRNP129 polymorphism (a high risk factor for Creutzfeldt-Jakob disease) in various regions of France and in West Europe. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* **5,** 141–144 (Mar. 2005).

108. Greene, L. S., Bottini, N., Borgiani, P. & Gloria-Bottini, F. Acid phosphatase locus 1 (ACP1): Possible relationship of allelic variation to body size and human population adaptation to thermal stress-A theoretical perspective. *American Journal of Human Biology: The Official Journal of the Human Biology Council* **12,** 688–701 (Sept. 2000).

109. Treves, T. A. & de Pedro-Cuesta, J. Parkinsonism mortality in the US, 1. Time and space distribution. *Acta Neurologica Scandinavica* **84,** 389–397 (Nov. 1991).

110. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiologic Reviews* **30,** 67–76 (2008).

111. Torrey, E. F. Prevalence studies in schizophrenia. *The British Journal of Psychiatry: The Journal of Mental Science* **150,** 598–608 (May 1987).

112. Lu, Z, Chen, T. C., Zhang, A, Persons, K. S., Kohn, N, Berkowitz, R, Martinello, S, *et al.* An evaluation of the vitamin D3 content in fish: Is the vitamin D content adequate to satisfy the dietary requirement for vitamin D? *The Journal of Steroid Biochemistry and Molecular Biology* **103,** 642–644 (Mar. 2007).

113. Kuningas, M., Mooijaart, S. P., Jolles, J., Slagboom, P. E., Westendorp, R. G. J. & van Heemst, D. VDR gene variants associate with cognitive function and depressive symptoms in old age. *Neurobiology of Aging* **30,** 466–473 (Mar. 2009).

114. Kesby, J. P., Cui, X., Ko, P., McGrath, J. J., Burne, T. H. J. & Eyles, D. W. Developmental vitamin D deficiency alters dopamine turnover in neonatal rat forebrain. *Neuroscience Letters* **461,** 155–158 (Sept. 2009).

115. Grecksch, G., Rüthrich, H., Höllt, V. & Becker, A. Transient prenatal vitamin D deficiency is associated with changes of synaptic plasticity in the dentate gyrus in adult rats. *Psychoneuroendocrinology* **34 Suppl 1,** S258–264 (Dec. 2009).

116. Holmes, V. A., Barnes, M. S., Alexander, H. D., McFaul, P. & Wallace, J. M. W. Vitamin D deficiency and insufficiency in pregnant women: a longitudinal study. *The British Journal of Nutrition* **102,** 876–881 (Sept. 2009).

117. Lucas, R. M., Repacholi, M. H. & McMichael, A. J. Is the current public health message on UV exposure correct? *Bulletin of the World Health Organization* **84,** 485–491 (June 2006).

118. Levi-Montalcini, R & Hamburger, V. Selective growth stimulating effects of mouse sarcoma on the sensory and sympathetic nervous system of the chick embryo. *The Journal of Experimental Zoology* **116,** 321–361 (Mar. 1951).

119. Weickert, C. S., Ligons, D. L., Romanczyk, T, Ungaro, G, Hyde, T. M., Herman, M. M., Weinberger, D. R., *et al.* Reductions in neurotrophin receptor mRNAs in the prefrontal cortex of patients with schizophrenia. *Molecular Psychiatry* **10,** 637–650 (July 2005).

120. Schramm, M, Falkai, P, Feldmann, N, Knable, M. B. & Bayer, T. A. Reduced tyrosine kinase receptor C mRNA levels in the frontal cortex of patients with schizophrenia. *Neuroscience Letters* **257,** 65–68 (Nov. 1998).

121. Otnaess, M. K., Djurovic, S., Rimol, L. M., Kulle, B., Kähler, A. K., Jönsson, E. G., Agartz, I., *et al.* Evidence for a possible association of neurotrophin receptor (NTRK-3) gene polymorphisms with hippocampal function and schizophrenia. *Neurobiology of Disease* **34,** 518–524 (June 2009).

122. Shibayama, E & Koizumi, H. Cellular localization of the Trk neurotrophin receptor family in human non-neuronal tissues. *The American Journal of Pathology* **148,** 1807–1818 (June 1996).

123. Nakanishi, T, Takahashi, K, Aoki, C, Nishikawa, K, Hattori, T & Taniguchi, S. Expression of nerve growth factor family neurotrophins in a mouse osteoblastic cell line. *Biochemical and Biophysical Research Communications* **198,** 891–897 (Feb. 1994).

124. Asaumi, K, Nakanishi, T, Asahara, H, Inoue, H & Takigawa, M. Expression of neurotrophins and their receptors (TRK) during fracture healing. *Bone* **26,** 625–33 (June 2000).

125. Hakak, Y, Walker, J. R., Li, C, Wong, W. H., Davis, K. L., Buxbaum, J. D., Haroutunian, V, *et al.* Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 4746–4751 (Apr. 2001).

126. Eastwood, S. L., Salih, T. & Harrison, P. J. Differential expression of calcineurin A subunit mRNA isoforms during rat hippocampal and cerebellar development. *The European Journal of Neuroscience* **22,** 3017–3024 (Dec. 2005).

127. Martins-de Souza, D., Gattaz, W. F., Schmitt, A., Rewerts, C., Marangoni, S., Novello, J. C., Maccarrone, G., *et al.* Alterations in oligodendrocyte proteins, calcium homeostasis and new potential markers in schizophrenia anterior temporal lobe are revealed by shotgun proteome analysis. *Journal of Neural Transmission (Vienna, Austria: 1996)* **116,** 275–289 (Mar. 2009).

128. Sun, L., Zhu, L.-L., Zaidi, N., Yang, G., Moonga, B. S., Abe, E., Iqbal, J., *et al.* Cellular and molecular consequences of calcineurin A alpha gene deletion. *Annals of the New York Academy of Sciences* **1116,** 216–226 (Nov. 2007).

129. Awumey, E. M., Moonga, B. S., Sodam, B. R., Koval, A. P., Adebanjo, O. A., Kumegawa, M, Zaidi, M, *et al.* Molecular and functional evidence for calcineurin-A alpha and beta isoforms in the osteoclast: novel insights into cyclosporin A action on bone resorption. *Biochemical and Biophysical Research Communications* **254,** 248–252 (Jan. 1999).

130. Kitagawa, H., Fujiki, R., Yoshimura, K., Mezaki, Y., Uematsu, Y., Matsui, D., Ogawa, S., *et al.* The chromatin-remodeling complex WINAC targets a nuclear receptor to promoters and is impaired in Williams syndrome. *Cell* **113,** 905–917 (June 2003).

131. Khavari, P. A., Peterson, C. L., Tamkun, J. W., Mendel, D. B. & Crabtree, G. R. BRG1 contains a conserved domain of the SWI2/SNF2 family necessary for normal mitotic growth and transcription. *Nature* **366,** 170–174 (Nov. 1993).

132. Lessard, J., Wu, J. I., Ranish, J. A., Wan, M., Winslow, M. M., Staahl, B. T., Wu, H., *et al.* An essential switch in subunit composition of a chromatin remodeling complex during neural development. *Neuron* **55,** 201–215 (July 2007).

133. Koga, M., Ishiguro, H., Yazaki, S., Horiuchi, Y., Arai, M., Niizato, K., Iritani, S., *et al.* Involvement of SMARCA2/BRM in the SWI/SNF chromatin-remodeling complex in schizophrenia. *Human Molecular Genetics* **18,** 2483–2494 (July 2009).

134. Flowers, S., Nagl, N. G., Beck, G. R. & Moran, E. Antagonistic roles for BRM and BRG1 SWI/SNF complexes in differentiation. *The Journal of Biological Chemistry* **284,** 10067–10075 (Apr. 2009).

135. Reyes, J. C., Barra, J, Muchardt, C, Camus, A, Babinet, C & Yaniv, M. Altered control of cellular proliferation in the absence of mammalian brahma (SNF2alpha). *The EMBO Journal* **17,** 6979–6991 (Dec. 1998).

136. Lamont, E. W., Legault-Coutu, D., Cermakian, N. & Boivin, D. B. The role of circadian clock genes in mental disorders. *Dialogues in Clinical Neuroscience* **9,** 333–342 (2007).

137. Mathias, D, Jacky, L, Bradshaw, W. E. & Holzapfel, C. M. Geographic and developmental variation in expression of the circadian rhythm gene, timeless, in the pitcher-plant mosquito, Wyeomyia smithii. *Journal of Insect Physiology* **51,** 661–667 (June 2005).

138. Palsdottir, A., Helgason, A., Palsson, S., Bjornsson, H. T., Bragason, B. T., Gretarsdottir, S., Thorsteinsdottir, U., *et al.* A Drastic Reduction in the Life Span of Cystatin C L68Q Carriers Due to Life-Style Changes during the Last Two Centuries. *PLoS Genetics* **4.** doi:10.1371/journal.pgen.1000099 (June 2008).

139. Huffaker, S. J., Chen, J., Nicodemus, K. K., Sambataro, F., Yang, F., Mattay, V., Lipska, B. K., *et al.* A primate-specific, brain isoform of KCNH2 affects cortical physiology, cognition, neuronal repolarization and risk of schizophrenia. *Nature Medicine* **15,** 509–518 (May 2009).

140. Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., *et al.* A draft sequence of the Neandertal genome. *Science (New York, N.Y.)* **328,** 710–722 (May 2010).

141. Balloux, F., Handley, L.-J. L., Jombart, T., Liu, H. & Manica, A. Climate shaped the worldwide distribution of human mitochondrial DNA sequence variation. *Proceedings. Biological Sciences / The Royal Society* **276,** 3447–3455 (Oct. 2009).

142. Myles, S., Somel, M., Tang, K., Kelso, J. & Stoneking, M. Identifying genes underlying skin pigmentation differences among human populations. *Human Genetics* **120,** 613–621 (Jan. 2007).

143. Bussey, K., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W., Zeeberg, B., *et al.* MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol* **4,** R27 (2003).

144. R Development Core Team. *R: A Language and Environment for Statistical Computing* (Vienna, Austria, 2009).

145. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).

146. Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. & Lempicki, R. A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* **4,** P3 (2003).

147. O'Brien, V. Viruses and apoptosis. *The Journal of general virology* **79 (Pt 8),** 1833–45 (Aug. 1998).

148. Nishino, M., Ikegami, H., Fujisawa, T., Kawaguchi, Y., Kawabata, Y., Shintani, M., Ono, M., *et al.* Functional polymorphism in Z-DNA-forming motif of promoter of SLC11A1 gene and type 1 diabetes in Japanese subjects: association study and meta-analysis. *Metabolism: clinical and experimental* **54,** 628–33 (May 2005).

149. Searle, S & Blackwell, J. M. Evidence for a functional repeat polymorphism in the promoter of the human NRAMP1 gene that correlates with autoimmune versus infectious disease susceptibility. *Journal of medical genetics* **36,** 295–9 (Apr. 1999).

150. Tivol, E. A., Borriello, F, Schweitzer, A. N., Lynch, W. P., Bluestone, J. A. & Sharpe, A. H. Loss of CTLA-4 leads to massive lymphoproliferation and fatal multiorgan tissue destruction, revealing a critical negative regulatory role of CTLA-4. *Immunity* **3,** 541–7 (Nov. 1995).

151. Serrano, N. C., Millan, P. & Páez, M.-C. Non-HLA associations with autoimmune diseases. *Autoimmunity reviews* **5,** 209–14 (Mar. 2006).

152. Yung, R. L., Quddus, J, Chrisp, C. E., Johnson, K. J. & Richardson, B. C. Mechanism of drug-induced lupus. I. Cloned Th2 cells modified with DNA methylation inhibitors in vitro cause autoimmunity in vivo. *Journal of immunology (Baltimore, Md. : 1950)* **154,** 3025–35 (Mar. 1995).

153. Bagenstose, L. M., Salgame, P & Monestier, M. Murine mercury-induced autoimmunity: a model of chemically related autoimmunity in humans. *Immunologic research* **20,** 67–78 (Jan. 1999).

154. Aichele, P, Bachmann, M. F., Hengartner, H & Zinkernagel, R. M. Immunopathology or organ-specific autoimmunity as a consequence of virus infection. *Immunological reviews* **152,** 21–45 (Aug. 1996).

155. Steinman, L & Conlon, P. Viral damage and the breakdown of self-tolerance. *Nature medicine* **3,** 1085–7 (Oct. 1997).

156. International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (Oct. 2007).

157. Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G. & Bustamante, C. Genomic scans for selective sweeps using SNP data. *Genome research* **15,** 1566–75 (Nov. 2005).

158. Mira, A., Pushker, R. & Rodríguez-Valera, F. The Neolithic revolution of bacterial genomes. *Trends in microbiology* **14,** 200–6 (May 2006).

159. Armelagos, G. J., Goodman, A. H. & Jacobs, K. H. The origins of agriculture: Population growth during a period of declining health. *Population and Environment* **13,** 9–22 (Sept. 1991).

160. Guernier, V., Hochberg, M. E. & Guégan, J.-F. Ecology drives the worldwide distribution of human diseases. *PLoS biology* **2,** e141 (June 2004).

161. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–78 (June 2007).

162. Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., Bailey, R., *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature genetics* **39,** 857–64 (July 2007).

163. Maiti, A. K., Kim-Howard, X., Viswanathan, P., Guillén, L., Rojas-Villarraga, A., Deshmukh, H., Direskeneli, H., *et al.* Confirmation of an association between rs6822844 at the Il2-Il21 region and multiple autoimmune diseases: evidence of a general susceptibility locus. *Arthritis and rheumatism* **62,** 323–9 (Feb. 2010).

164. Zhernakova, A., Alizadeh, B. Z., Bevova, M., van Leeuwen, M. A., Coenen, M. J. H., Franke, B., Franke, L., *et al.* Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. *American journal of human genetics* **81,** 1284–8 (Dec. 2007).

165. Hollis-Moffatt, J. E., Chen-Xu, M, Topless, R, Dalbeth, N, Gow, P. J., Harrison, A. A., Highton, J, *et al.* Only one independent genetic association with rheumatoid arthritis within the KIAA1109-TENR-IL2-IL21 locus in Caucasian sample sets: confirmation of association of rs6822844 with rheumatoid

arthritis at a genome-wide level of significance. *Arthritis Res. Ther.* **12,** R116 (2010).

166. Daha, N. A., Kurreeman, F. A., Marques, R. B., Stoeken-Rijsbergen, G, Verduijn, W, Huizinga, T. W. & Toes, R. E. Confirmation of STAT4, IL2/IL21, and CTLA4 polymorphisms in rheumatoid arthritis. *Arthritis Rheum.* **60,** 1255–1260 (2009).

167. Teixeira, V. H., Pierlot, C, Migliorini, P, Balsa, A, Westhovens, R, Barrera, P, Alves, H, *et al.* Testing for the association of the KIAA1109/Tenr/IL2/IL21 gene region with rheumatoid arthritis in a European family-based study. *Arthritis Res. Ther.* **11,** R45 (2009).

168. Festen, E. A., Goyette, P, Scott, R, Annese, V, Zhernakova, A, Lian, J, Lefebvre, C, *et al.* Genetic variants in the region harbouring IL2/IL21 associated with ulcerative colitis. *Gut* **58,** 799–804 (2009).

169. Marquez, A, Orozco, G, Martinez, A, Palomino-Morales, R, Fernandez-Arquero, M, Mendoza, J. L., Taxonera, C, *et al.* Novel association of the interleukin 2-interleukin 21 region with inflammatory bowel disease. *Am. J. Gastroenterol.* **104,** 1968–1975 (2009).

170. Albers, H. M., Kurreeman, F. A. S., Stoeken-Rijsbergen, G, Brinkman, D. M. C., Kamphuis, S. S. M., van Rossum, M. A. J., Girschick, H. J., *et al.* Association of the autoimmunity locus 4q27 with juvenile idiopathic arthritis. *Arthritis and rheumatism* **60,** 901–4 (Mar. 2009).

171. Liu, Y., Helms, C., Liao, W., Zaba, L. C., Duan, S., Gardner, J., Wise, C., *et al.* A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS genetics* **4,** e1000041 (Mar. 2008).

172. Warren, R. B., Smith, R. L., Flynn, E, Bowes, J, Eyre, S, Worthington, J, Barton, A, *et al.* A systematic investigation of confirmed autoimmune loci in early-onset psoriasis reveals an association with IL2/IL21. *The British journal of dermatology* **164,** 660–4 (Mar. 2011).

173. Hunt, K. A., Zhernakova, A, Turner, G, Heap, G. A., Franke, L, Bruinenberg, M, Romanos, J, *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* **40,** 395–402 (Apr. 2008).

174. Van Heel, D. A., Franke, L, Hunt, K. A., Gwilliam, R, Zhernakova, A, Inouye, M, Wapenaar, M. C., *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.* **39,** 827–829 (2007).

175. Adamovic, S, Amundsen, S. S., Lie, B. A., Gudjónsdóttir, A. H., Ascher, H, Ek, J, van Heel, D. A., *et al.* Association study of IL2/IL21 and FcgRIIa: significant association with the IL2/IL21 region in Scandinavian coeliac disease families. *Genes and immunity* **9,** 364–7 (June 2008).

176. O'Shea, J. J., Ma, A & Lipsky, P. Cytokines and autoimmunity. *Nat. Rev. Immunol.* **2,** 37–45 (2002).

177. Ozaki, K, Spolski, R, Ettinger, R, Kim, H. P., Wang, G, Qi, C. F., Hwu, P, *et al.* {R}egulation of {B} cell differentiation and plasma cell generation by {I}{L}-21, a novel inducer of {B}limp-1 and {B}cl-6. *J. Immunol.* **173,** 5361–5371 (Nov. 2004).

178. Ozaki, K., Spolski, R., Ettinger, R., Kim, H.-P. P., Wang, G., Qi, C.-F. F., Hwu, P., *et al.* Regulation of B cell differentiation and plasma cell generation by IL-21, a novel inducer of Blimp-1 and Bcl-6. *Journal of immunology (Baltimore, Md. : 1950)* **173,** 5361–71 (Nov. 2004).

179. Alcina, A, Vandenbroeck, K, Otaegui, D, Saiz, A, Gonzalez, J. R., Fernandez, O, Cavanillas, M. L., *et al.* The autoimmune disease-associated KIF5A, CD226 and SH2B3 gene variants confer susceptibility for multiple sclerosis. *Genes and immunity* **11,** 439–45 (July 2010).

180. Paré, G., Ridker, P. M., Rose, L., Barbalic, M., Dupuis, J., Dehghan, A., Bis, J. C., *et al.* Genome-wide association analysis of soluble ICAM-1 concentration

reveals novel associations at the NFKBIK, PNPLA3, RELA, and SH2B3 loci. *PLoS genetics* **7,** e1001374 (Apr. 2011).

181. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 6062–7 (Apr. 2004).

182. Li, Y, He, X, Schembri-King, J, Jakes, S & Hayashi, J. Cloning and characterization of human Lnk, an adaptor protein with pleckstrin homology and Src homology 2 domains that can inhibit T cell activation. *Journal of immunology (Baltimore, Md. : 1950)* **164,** 5199–206 (May 2000).

183. Velazquez, L., Cheng, A. M., Fleming, H. E., Furlonger, C., Vesely, S., Bernstein, A., Paige, C. J., *et al.* Cytokine signaling and hematopoietic homeostasis are disrupted in Lnk-deficient mice. *The Journal of experimental medicine* **195,** 1599–611 (June 2002).

184. Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K. W. & Hansen, L. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Human genetics* **123,** 177–87 (Mar. 2008).

185. Cook, A. L., Chen, W., Thurber, A. E., Smit, D. J., Smith, A. G., Bladen, T. G., Brown, D. L., *et al.* Analysis of cultured human melanocytes based on polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci. *The Journal of investigative dermatology* **129,** 392–405 (Feb. 2009).

186. Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P. N., Stark, M. S., Hayward, N. K., Martin, N. G., *et al.* A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *American journal of human genetics* **82,** 424–31 (Feb. 2008).

187. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25,** 1754–60 (July 2009).

188. Przeworski, M. Estimating the time since the fixation of a beneficial allele. *Genetics* **164,** 1667–76 (Aug. 2003).

189. Slatkin, M & Rannala, B. Estimating allele age. *Annual review of genomics and human genetics* **1,** 225–49 (Jan. 2000).

190. Thomson, R, Pritchard, J. K., Shen, P, Oefner, P. J. & Feldman, M. W. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 7360–5 (June 2000).

191. Stephan, W, Wiehe, T & Lenz, M. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theoretical Population Biology* **41,** 237–254 (Apr. 1992).

192. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162,** 2025–35 (Dec. 2002).

193. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)* **18,** 337–8 (Feb. 2002).

194. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–6 (July 2011).

195. Nunes, M. A. & Balding, D. J. On optimal selection of summary statistics for approximate Bayesian computation. *Statistical applications in genetics and molecular biology* **9,** Article34 (Jan. 2010).

196. Csilléry, K., Blum, M. G. B., Gaggiotti, O. E. & François, O. Approximate Bayesian Computation (ABC) in practice. *Trends in ecology & evolution* **25,** 410–8 (July 2010).

197. Corona, E., Dudley, J. T. & Butte, A. J. Extreme evolutionary disparities seen in positive selection across seven complex diseases. *PloS one* **5,** e12236 (Jan. 2010).

198. Stearns, S. C., Byars, S. G., Govindaraju, D. R. & Ewbank, D. Measuring selection in contemporary human populations. en. *Nature reviews. Genetics* **11,** 611–622 (Aug. 2010).

199. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature reviews. Genetics* **10,** 639–50 (Sept. 2009).

200. Roff, D. A. *Evolutionary Quantitative Genetics* 516 (Springer, 1997).

201. Hancock, A. M., Alkorta-Aranburu, G., Witonsky, D. B. & Di Rienzo, A. Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **365,** 2459–68 (Aug. 2010).

202. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology : CB* **20,** R208–15 (Feb. 2010).

203. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–8 (Sept. 2010).

204. Nettle, D. Height and reproductive success in a cohort of british men. *Human Nature* **13,** 473–491 (Dec. 2002).

205. Nettle, D. Women's height, reproductive success and the evolution of sexual dimorphism in modern humans. *Proceedings. Biological sciences / The Royal Society* **269,** 1919–23 (Sept. 2002).

206. Pawlowski, B, Dunbar, R. I. & Lipowicz, A. Tall men have more reproductive success. *Nature* **403,** 156 (Jan. 2000).

207. Pritchard, J. K. & Di Rienzo, A. Adaptation - not by sweeps alone. *Nature Reviews Genetics* **11,** 665–7 (Sept. 2010).

208. Sayama, H, Kaufman, L & Bar-Yam, Y. Symmetry breaking and coarsening in spatially distributed evolutionary processes including sexual reproduction and disruptive selection. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics* **62,** 7065–9 (Nov. 2000).

209. Van Doorn, G. S. & Dieckmann, U. The long-term evolution of multilocus traits under frequency-dependent disruptive selection. *Evolution; international journal of organic evolution* **60,** 2226–38 (Nov. 2006).