1. Introduction

1.1. Natural selection

1.1.1. Types of natural selection

The theory of natural selection was introduced by Charles R. Darwin and Alfred R. Wallace in 1858, later described in detail in Darwin's book, 'On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life' (1). The process of natural selection is based on the assumption that individuals more suited to the environment are more likely to survive and reproduce, passing on their heritable traits to future generations (so-called survival of the fittest), so that the frequency of fitness-enhancing traits increases in the population over time (1, 2). Individuals with less favourable phenotypes are less likely to survive and reproduce, as all organisms are exposed to severe competition (2). According to the natural selection theory, populations change to adapt to their environments, which leads to the accumulation of variation over time and perhaps formation of a new species: a process of divergence that explains the diversity of living organisms (1). The key elements of the natural selection theory were inter-individual variation, inheritance with modification and the multiplication of new forms, although the hereditary mechanism and mutagenesis were unknown at that time (1). The advances in understanding the mechanism of inheritance made by Gregor J. Mendel and Thomas H. Morgan became the core of classical genetics and eventually enabled putting evolution to be understood in a molecular context (3). The integration of genetics with the theory of natural selection by Ronald A. Fisher, Sewall G. Wright and John B. S. Haldane formed the basis for population genetics and the modern evolutionary synthesis (4-7).

The concept of natural selection in genetics has evolved since then, and different types of selection were recognised, depending on whether an allele is advantageous or deleterious and on the fitness conferred by the genotype (8). In this thesis we will use terminology proposed by Akey and Nielsen to describe simplified modes of natural selection (8, 9). Imagine an initial (ancestral) single allele A_1 and a new (derived) allele A_2 introduced by mutation into the population. Their possible genotype combinations are A_1A_1, A_1A_2 and A_2A_2, each with its own genotype fitness. We can represent the fitness of the new genotypes relative to the fitness of the initial ancestral genotype (equal to 1) as $1 + hs$ and $1 + s$ for A_1A_2 and A_2A_2 respectively, where h is the heterozygote effect and s is the selection coefficient (8). If $s = 0$ there is no difference in fitness between genotypes and allele frequencies are assumed to evolve naturally. Directional selection occurs if their fitnesses are not all equal (8, 9). In the case of incomplete dominance ($0 < h < 1$) and $s > 0$, the new mutation is advantageous and will rise in frequency in the population until fixation, as A_2 carriers are better adapted and favoured by the positive selection (8). If $s < 0$, then the newly occurred mutation is deleterious and will be purged from the population by purifying (or negative) selection as A_2 carriers are less fit (8). Random new mutations are more likely to be deleterious than beneficial and are constantly selected against and removed from the gene pool before achieving appreciable frequencies (a phenomenon called background selection) resulting in conserved genomic regions with little or no variation (10-12). Finally, in case of over-dominant selection acting on an advantageous allele ($s > 0$ and $h > 1$) the heterozygote has the highest relative fitness (so-called heterozygote advantage) (8). Selection of this kind is called balancing selection and multiple alleles are maintained within the gene pool (9, 10, 13). There are other types of selection defined by the phenotypic outcome rather than the underlying pattern of variability that are commonly used in the population-genetic literature e.g. diversifying and stabilizing selection. Diversifying (or disruptive) selection is described as a trend where extreme phenotypes are favoured over intermediate phenotypes; while stabilizing selection favours intermediate phenotypic values (10). Furthermore, another type of selection proposed by Darwin is referred to as sexual selection, driven by competition for mates, which explains sexually dimorphic features or increased prevalence of sexually attractive traits (1). It is important to realise that allele frequencies in a population (especially those of selectively neutral alleles that do not affect the organism’s fitness) are also subjected to random fluctuation known as genetic drift (9, 14, 15). New mutations that arise in the population may increase in prevalence due to genetic drift, even though they do not confer any selective advantage (9). Finally, selection efficiency depends critically on the effective population size, with small populations being more prone to genetic drift and thus experiencing less efficient selection (9).

1.1.2. Modes of positive selection

Here we define positive selection as any type of selection where new mutations or existing variants are advantageous (with positive selection coefficients) and there is no heterozygote advantage. There has been great interest in positive selection as it is the primary mechanism of adaptation and the evolution of novelty (9, 10). Selective episodes leave their signatures in the human genome and thus can be recognised from the pattern of nucleotide polymorphisms in a population sample due to genetic hitchhiking (16-18). The classical hard sweep model assumes that a new advantageous mutation rapidly spreading to fixation or high prevalence affects the pattern of linked variation (16, 18, 19). Its genetic characteristics include high-frequency long-range haplotypes with a concomitant reduced level of genetic variation, large allele frequency differences between populations, and changes to the allele frequency spectrum (e.g. increased fraction of derived common and rare alleles, depletion of intermediate-frequency variation) (Figure 1), although these features can also arise by genetic drift or purifying selection and are confounded by population demography (9, 16, 18-23). The size of the genomic region that is subjected to hitchhiking depends mainly on the local recombination rate and the selection coefficient, and its signature decreases with increasing distance from the selected allele (24). Ongoing, or incomplete sweep refers to any stage of selective sweep before reaching fixation, while fixed sweep is said to be complete (19).

However, it has been argued that hard sweeps were rather rare in recent human evolution and it is unusual for a new mutation to be rapidly driven to fixation (18, 20, 21). Just the opposite, it seems that most of the variants increase in frequency rather slowly and steadily, without reaching fixation and creating extensive LD patterns, because of limited dispersion over large geographic areas and low selection coefficients (18, 20, 21). Furthermore, the waiting time for new mutations can be extremely long and hard sweeps may be an inefficient response to a rapidly changing environment (18). Therefore, selection may more often operate on pre-existing variation that has evolved neutrally in the population until it becomes advantageous under certain conditions (‘selection on standing

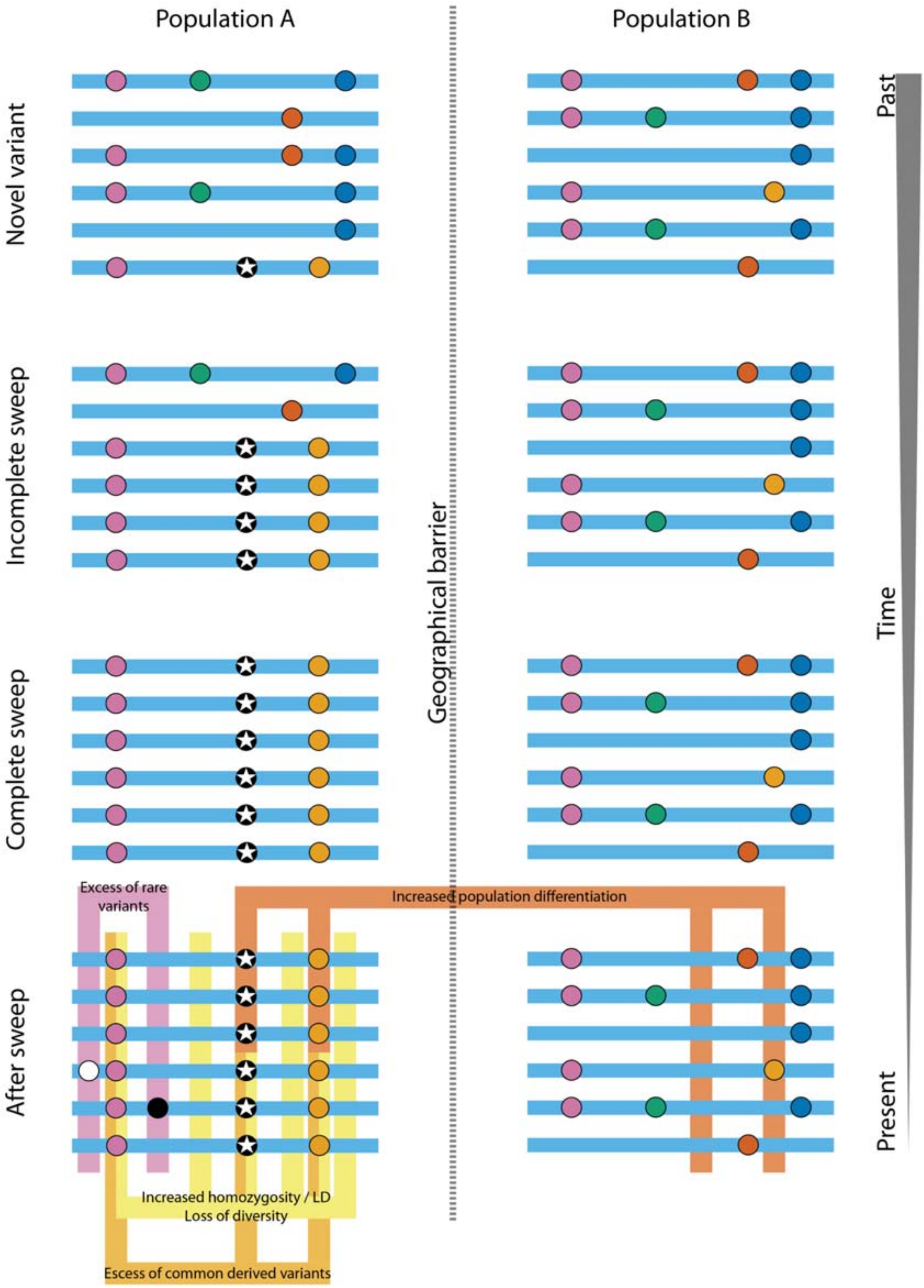

Figure 1. Hard sweep model. The blue lines indicate individual haplotypes, and derived SNP alleles are represented as circles. Population A and B with identical haplotype structure are geographically isolated with no gene flow between them. A new advantageous mutation (indicated by a star) appears *de novo* on one haplotype in population A and rapidly spreads to fixation bringing nearby linked derived alleles to high frequency. This creates a region of extended homozygosity (high LD) as there were not enough time for recombination to break it down. It also causes a population-wide reduction in genetic diversity around selected mutation as SNP-alleles that do not occur on the selected haplotype will be lost. After the sweep is complete, new mutations appear against a homogenous background creating an excess of rare alleles. Finally, differences in allele frequencies between population A and B reflects the population-specific adaptation.

variation') (18, 20, 21). Selection from standing variation is difficult to detect using most standard approaches, because the selected variant often exists on multiple haplotype backgrounds (so called 'soft sweep') and has weaker effects on closely-linked sites, so does not produce the classical selective sweep signatures of strong linkage disequilibrium (LD) and site frequency spectrum (SFS) changes, although it might exhibit an increased proportion of alleles at intermediate frequency (10, 17, 18, 20, 21, 25, 26). Similarly, the decrease in diversity around the standing variant is subtle (10, 25). Another alternative model of positive selection on standing variation is polygenic adaptation, defined as selection at many loci simultaneously affecting quantitative traits composed of hundreds of alleles of small individual effect sizes (17, 18, 21) e.g. selection on standing variants associated with greater height in Northern Europe (27-29). Polygenic selection could allow rapid adaptation (18). Signatures of selection on a complex trait are even more problematic to detect as they are composed of subtle shifts in allele frequencies at multiple loci while not producing classical sweep signatures (17, 18, 21). Finally, adaptive variation might have been acquired from archaic hominins in the process of admixture (so-called adaptive introgression), as modern humans were shown to have had limited interbreeding with archaic Eurasian hominins after out of Africa migration (30). The signature of adaptive introgression is the presence of a high frequency haplotype characterised by strong LD in a particular population that is also found in the archaic source population but is absent from populations depleted of archaic admixture (30).

It is difficult to estimate the proportions of hard sweeps, soft sweeps and polygenic adaptation, as well as adaptive introgression, in human evolution, but it seems that much of human adaptation may not have produced classical signatures of selective sweeps (18, 20).

1.1.3. Human population diversification

The out-of-Africa expansion ~60,000 years ago exposed humans to a diverse range of new environments and selective pressures including new pathogens, climatic conditions and diets (18, 23, 31). Genetic drift and local adaptations in spatially distant populations consequently led to geographically-structured genetic and phenotypic diversification, illustrated by the inter-population variation observed for numerous morphological and physiological traits, such as skin pigmentation (18, 21, 23). As gene flow between groups decreases with increasing distance, members of the same local group are usually more closely related to each other than to members of groups living in distant geographical areas (32). Traits that show extreme differentiation between populations are thus candidates for local adaptations (10). Pigmentation is not the only trait whose phenotypic values strongly associate with geography. Similar trends were proposed for hair shape (Figure 2) and body shape (e.g. larger, stockier body shape in cold climates due to thermal efficiency or the 'pygmy' phenotype in tropical rainforests) (18, 33, 34). Apart from the latitude pattern, altitude-associated adaptation has also been reported, i.e. physiological adaptations to low oxygen at high-altitudes (35, 36). We do not know, however, to what extent these phenotypic differences between populations are driven by selection. It is important to realise that genetic and morphological variation is often gradual, and phenotypic boundaries are not discrete but often show continuous clines correlated with geography (32). In addition, populations with similar physical characteristics can be genetically very different, partially due to convergent evolution (32). Finally, the genetic diversity of humans is relatively low compared with many other species (37-39) and the relationships between ethnicity, patterns of human genetic variation, and ancestry, are complicated (32, 40).

Not only are the genetic variants underlying differences between populations crucial for understanding recent human evolution and present-day human diversity, but they may also be clinically relevant, as the prevalence of some common diseases and disease susceptibilities and drug responses varies across regions e.g. the higher odds of developing hypertension in African Americans

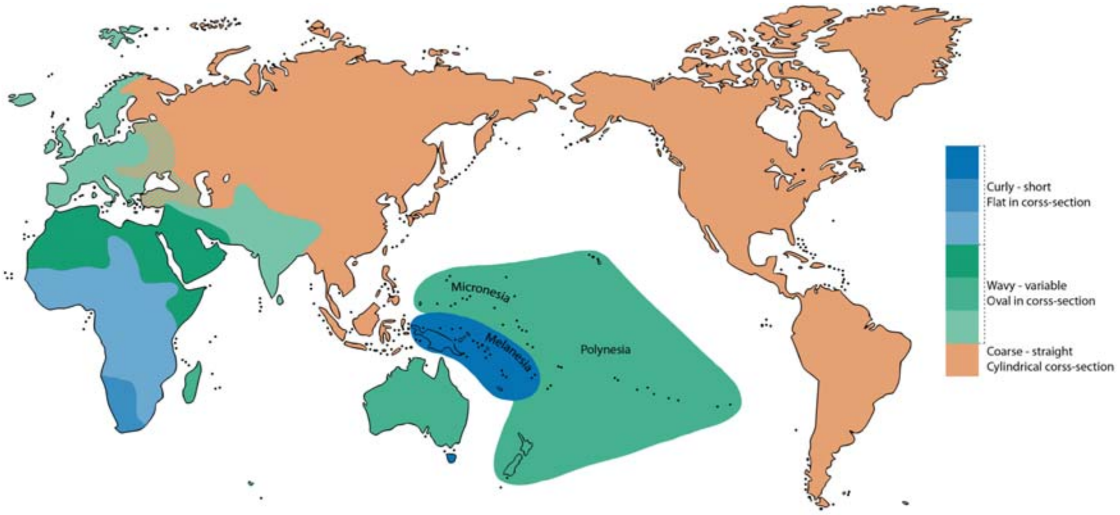

Figure 2. Global human hair texture distribution. Reproduced after (41).

compared with European Americans as a result of past positive selection favoring salt and liquid retention in hot African climate (21, 42, 43). Similarly, according to the thrifty genotype hypothesis, the high prevalence of obesity and type 2 diabetes might be a result of an adaptation to a harsh past environment and food shortages: an attractive hypothesis that nevertheless has some limited genetic support (42, 44-47). The field of evolutionary medicine (also called Darwinian medicine) argues that evolutionary biology and studies of natural selection could improve our understanding of the origins and causes of complex diseases (42, 48, 49). Medical implications of adaptive variation arise because natural selection can only act in a direct way on functionally-important variants driving phenotypic variation (14, 22); selected alleles usually confer protective effects (e.g. pathogen resistance *CASP12* (50), *CCR5* (51), *FUT2* (52) deficiency alleles), but paradoxically, may turn harmful in non-traditional environments (42, 44, 45, 53, 54) e.g. *CPT1A* (55, 56) and *APOL1* (57, 58), or in a homozygous state, e.g. sickle-cell alleles, Tay-Sachs disease, cystic fibrosis and Phenylketonuria (59-62). Regions targeted by positive selection might be disease-causing not only due to alleles that lost their advantage or balancing selection, but also through the effects of genetic hitch-hiking of moderately deleterious variants (19, 22, 42, 63). Some have argued that hitchhiking, rather than genetic drift, might be the primary force shaping the pattern of neutral variation (so-called genetic draft) (64, 65). All of the above might contribute to disease-causing mutations that segregate at relatively high frequencies (22, 66). Identification of the genetic variants that underlie regional adaptations and proving their functionality might thus sometimes facilitate disease-related research and shed more light on the diversification of modern humans and refine the human genotype-phenotype map.

1.2. Methods to detect positive selection using genomic data

Advances in genotyping and sequencing technologies laid the foundation for population genomics and enabled moving from hypothesis-driven candidate gene studies toward hypothesis-generating genome-wide screens for selection (8). Whole genome analyses provide a less biased way of searching for selection signatures, free from a priori assumptions regarding putatively selected loci (8). The common approaches applied in population genomics to distinguish neutral variation affected by genetic drift from variation subjected to selection involve:

1. i) Calculation of the summary statistic informative about selection in empirical data, ii) comparison of the results against data generated by model-based simulations of genetic drift, iii) rejection or acceptance of the null hypothesis of neutrality (10). A common problem with such an approach is that population-genetic models often make unrealistic assumptions about the demography of the populations, such as a constant population size and no population structure, and sometimes uniform distribution of recombination and mutation rates across the genome (9). Many neutrality tests have been shown to be highly sensitive to such unrealistic demographic assumptions (9, 67).
2. i) Calculation of the summary statistic informative about selection across the whole genome, ii) construction of the empirical distribution of this genome-wide statistic, iii) investigation of the top outliers in the extreme tail of the empirical distribution as selection candidates (8). Such a nonparametric outlier approach is based on the assumption that demographic history and stochastic processes affect the whole genome equally, while selection acts in a locus-specific manner placing selected targets in the extreme tails of the genome-wide distribution (8, 68). However, presence in the extreme tail of empirical distribution alone does not prove that a candidate was indeed targeted by selection, but rather that it shows unusual characteristics relative to the rest of the genome consistent with the hypothesis of selection

(8). In particular, we do not know how prevalent selection has been and what percentage of the genome should be considered an extreme outlier.

We will present the rationale behind the most commonly used approaches to detect positive selection in the following sections. An overview of the main classes of methods recovering selection events of different time-scales and modes is given in Table 1.

Table 1. Comparison of time-scale and modes of positive selection detected by different methods. ‘Relative-rate’ refers to comparative methods based on inter-species comparisons explained in the next section. *Diff* – population differentiation based methods; *SFS* – site frequency spectrum based methods; *LD* – linkage disequilibrium based methods; *Comp* – composite methods. + indicates that a method is sensitive to given type of selection; - indicates lack of power; (+) indicates low power. Old selection stands for species-wide adaptation that occurred during the divergence of species. Recent selection stands for recent or ongoing selection after out-of-Africa population split.

Selection time/mode		Relative rate	*Diff*	*SFS*	*LD*	*Comp*
Old selection (~200 kya – 6 Mya)		+	-	-	-	-
Recent selection (~5 – 100 kya)	Hard sweep	-	+	+	+	+
	Soft sweep	-	+	-	(+)	(+)
	Polygenic selection	-	-	-	-	-
	Adaptive introgression	-	+	(+)	(+)	(+)

1.2.1. Macroevolutionary relative-rate approaches

Relative-rate comparative methods are based on comparisons between different species and their relative rates of genetic substitution (10). Although they can in theory be applied to within-species comparisons, they are usually used to identify selective events that happened in the deep past at the macroevolutionary level, as their effect is stronger in divergence data than in polymorphism data and they are not suitable for recent selection (9, 10). Detecting positive selection using such methods involves comparison of homologous sequences across related taxa and searching for acceleration in the rate of evolution indicated by an excess of substitutions relative to the baseline mutation rate (10). Relative-rate methods have been widely used to identify genomic regions showing a significantly accelerated rate of substitution in the human lineage (69-72).

Probably the most common relative-rate method is the d_N/d_S ration, sometimes referred to as ω or K_a/K_s. This method compares the ratio of nonsynonymous mutations per nonsynonymous site to the number of synonymous mutations per nonsynonymous site in a multiple species alignment, and can be applied either to a region of interest or a single codon (9). Assuming that selection acts on nonsynonymous mutations, negative selection will reduce the number of nonsynonymous mutations, while continued positive selection will increase the number of nonsynonymous mutations, relative to the number of functionally neutral synonymous mutations that serve as a baseline substitution rate (9). In case of neutrality, synonymous and nonsynonymous substitutions should occur at the same relative rate, and therefore $d_N/d_S = 1$. If negative selection operates, $d_N/d_S < 1$, indicating a relative depletion of nonsynonymous substitutions. If region is under positive selection, then $d_N/d_S > 1$, indicating a relative excess of nonsynonymous substitutions. This method detects repeated selective fixations that occurred in the same gene or at the same site across taxa over long evolutionary time periods (9). One of its strengths is that it indicates directionality of selection (positive vs negative), but it is restricted to coding regions and nonsynonymous sites (9). Further improvements to this method were also proposed (73-77) and a similar comparative method was designed for non-coding regions (78). The rate of substitution in noncoding regions relative to the rate of synonymous substitution in coding regions is estimated by a parameter ζ. When a site in a noncoding region is evolving neutrally $\zeta = 1$, whereas $\zeta > 1$ indicates positive selection, and $\zeta < 1$ suggests negative selection (78). However, in practice this method only picks up highly variable or highly conserved noncoding sites, that may or may not be targets of selection (78).

Other relative-rate methods are based on the principle that selection modifies the levels of variability within and between-species (9). The MacDonald-Kreitman Test (*MKT*) employs between species variation ('divergence') and within-species variability ('diversity') (79). It calculates and compares two d_N/d_S values, one between species and one within species, which should be equal under neutrality (assuming constant mutation and substitution rates) (10). If one exceeds the other, then the null hypothesis can be rejected (79). Larger between-species values suggests positive selection between species, while a greater within-species ratio indicates balancing or weak negative selection within the species (10).

Similarly, the Hudson-Kreitman-Aguade (*HKA*) test compares the rate of divergence to polymorphisms for multiple genes (80). *HKA* calculates the ratios of fixed interspecific differences (D) to within-species polymorphisms (P) (10). The test assumes that under neutrality D and P should be proportional (given a constant mutation rate) and the deviation from the neutral D/P value allows rejection of the null hypothesis (80). The expected neutral D/P ratio for a lineage can be estimated by examining multiple sites (10). Relatively large D/P values indicate either directional selection between (accelerated speciation) or within (reduced diversity within the species) species (10). Relatively small values suggest balancing selection between species (10). In contrast to d_N/d_S based methods, it can be applied to both coding and non-coding regions (10).

1.2.2. Microevolutionary population genomic approaches

Population genomic approaches aim to detect microevolutionary selective events using within-species polymorphisms (9). They have been widely applied to uncover recent and ongoing selection underlying local adaptations in humans following the out-of-Africa migration (10). Most methods of this class rely upon the classical hard sweep model assumptions, and its effects on patterns of linked neutral variation (8, 18-20). Here, we present the basic strategies and some of their derivatives commonly used in the field of human population genomics that can be classified based on the type of selective signature they detect, as proposed by Vitti *et al*. (10).

1.2.2.1. Population differentiation-based methods

Local adaptation is manifested by a geographic gradient in the frequency of the selected allele within a geographical region (21). Any selection event, regardless of its mode, will eventually produce an excess of allele frequency differentiation between populations as long as (i) it has taken place in one population but not in another (and the allele was at low frequency when first favored) or/and (ii) there is variation in selection coefficient over space, (iii) migration and gene flow between the populations have been restricted, (iv) and there has been enough time for selection to act (20, 21). Even if an allele is equally advantageous in all environments, but its selection happened in a regionally-restricted manner, the selected variant will be concentrated around its geographic origin due to limited dispersal (21, 22). Therefore, larger than average allele frequency differences between populations may indicate local adaptation. This measure of selection is sensitive to many types of selection including classic sweeps, selection from standing variation and negative selection (17, 20, 21, 25, 26).

Using population differentiation as an indicator of geographically restricted positive selection was originally proposed by Cavalli-Sforza in 1966 (68). The first attempt to implement a population differentiation-based statistical test for positive selection was made by Lewontin and Krakauer, who used the variance of the F_{ST} parameter (Wright's fixation index) and proposed rejection of the neutral model for loci with F_{ST} values larger than expected by chance (81). F_{ST} is a common measure of population differentiation, defined as variance in the allele frequency between different subpopulations (weighted by the sizes of the subpopulations) divided by the variance of allele frequency in the total population (subpopulations combined) (82). Its values range from 0 to 1 (82). A zero value implies no population structure (a panmictic population), while a value of one implies no gene-flow between two populations (a fixed difference) (82). In practice, various F_{ST} estimators have been proposed, e.g. calculated using nucleotide diversity (π) or heterozygosity (H) (Equation 1).

Many F_{ST} derivatives have also been proposed (83-90), including the locus-specific branch length metric (*LSBL*) and the population branch statistic (*PBS*) which use pair-wise calculations of F_{ST} from three or more populations to isolate population-specific changes in allele frequency relative to a broader genetic context (91, 92), and the cross-population composite likelihood ration (*XP-CLR*) of allele frequency differentiation that extends F_{ST} to many loci (93). Some statistics explore differentiation of haplotypes instead of individual alleles (94). Another common summary of population differentiation is difference in derived allele frequency between populations *(ΔDAF*) (95). Finally, one can directly compare allele frequencies in ancient human genomes with modern samples, although the availability of well-preserved ancient DNA is limited (96, 97).

Equation 1. Wright's fixation index (F_{ST}). π_T - average number of pairwise differences between two individuals sampled from different sub-populations (nucleotide diversity within the total population). π_S - average number of pairwise differences between two individuals sampled from the same sub-population (nucleotide diversity within subpopulations). H_S - mean expected heterozygosity within subpopulations. H_T - expected heterozygosity in the entire population (*2pq*).

$$F_{ST} = \frac{\pi_T - \pi_S}{\pi_T} \qquad or \qquad F_{ST} = \frac{H_T - H_S}{H_T}$$

1.2.2.2. Site frequency spectrum-based methods

Selection is known to distort the site frequency spectrum (SFS) within a population (9). The SFS distribution of alleles in the population sample can be defined as a count of the number of mutations of a given frequency class, for each class (9). Negative selection increases the proportion of variants segregating at low frequencies in the sample (9, 10). Positive directional selection tends to increase the proportion of high frequency variants (selected alleles and linked neutral sites), but also the proportion of low frequency variants in the case of a hard selective sweep causing a population-wide reduction in the genetic diversity around the selected allele (Figure 3) (9, 10). Balancing selection increases the proportion of intermediate frequency variants (9). Such distortions can persist for thousands of generations (10).

The most common neutrality test summarizing the SFS is Tajima's *D*, comparing the average number of nucleotide differences between pairs of sequences with the total number of segregating sites in a population sample (98). If the difference between these two measures of variability is larger that expected under neutrality, then the null hypothesis is rejected. The rationale behind this method is that the low-frequency alleles contribute less to the average number of pairwise nucleotide differences (as most haplotypes in the selected region are the same or very similar, therefore there are few differences between them on average), but they do contribute to the total number of segregating sites. As a result, the excess of rare alleles drives smaller/more negative values of *D*, which might be indicative of selection (both positive or negative) or population expansion (98). Variations on this theme have been proposed with further extensions (99-101).

Another commonly used test is Fay & Wu's *H*, which compares the number of pair-wise differences between individuals to the number of individuals homozygous for the derived allele (102). As selective sweeps increase the frequency of derived alleles near the causal allele hitchhiking to high frequencies, small values of *H* indicate an excess of high-frequency derived alleles and possibly positive selection (10). Similarly, Kim and Stephan's composite likelihood ratio (CLR) test detects an excess of derived alleles across multiple sites (103).

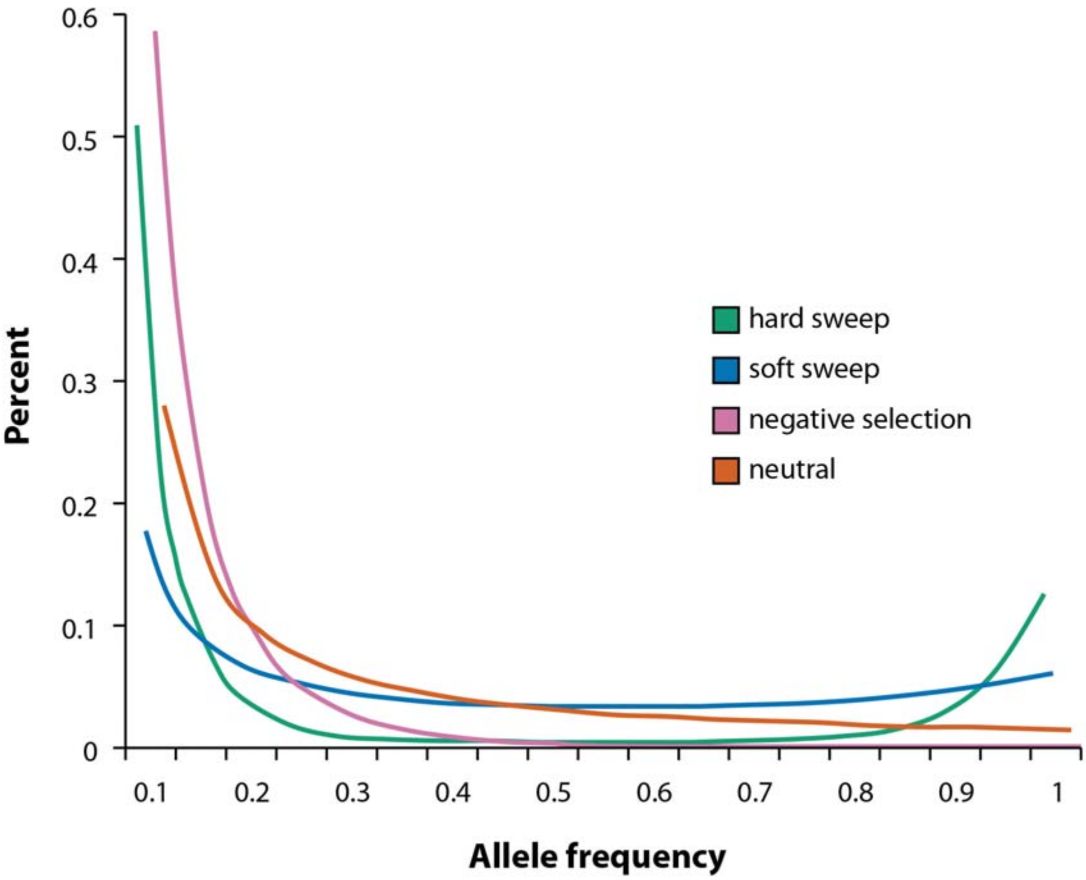

Figure 3. Site frequency spectrum under selection and neutrality.

1.2.2.3. Linkage disequilibrium-based methods

The level of linkage disequilibrium (LD) is defined as the correlation among alleles from different loci (9). A beneficial mutation spreading rapidly to a high frequency in the population brings nearby linked hitchhiker variants along, creating a region of high LD (or, equivalently, long haplotype) quickly enough that recombination has not had time to break it down (10, 16, 18, 19). Selection occurring less than 400 generations ago should leave a clear LD pattern (104). Many statistical methods for detecting regions of strong LD relative to their prevalence within a population, or a consequent reduction in haplotype diversity, have been proposed (104-107). This approach is commonly used to detect recent incomplete sweeps, but also has the potential to detect soft sweeps at lower power (10).

One class of such methods is based on the extended haplotype homozygosity (*EHH*) statistic that measures LD at a distance *x* from the core haplotype (a given haplotype at a locus of interest) (104). *EHH* is defined as the probability that two randomly chosen chromosomes carrying the core haplotype are identical by descent (homozygous at all SNPs) for the entire genomic interval from the core region to the point *x* (104). In fact, *EHH* detects the transmission of an extended haplotype without recombination (104). *EHH* ranges from 0 (no homozygosity, all extended haplotypes are different) to 1 (complete homozygosity, all extended haplotypes are the same) and decreases with increasing distance from the core region (104). Relative *EHH* is the ratio of the *EHH* on the tested core haplotype compared with the *EHH* on all other core haplotypes at the region, and ranges from 0 to infinity (104). The long-range haplotype (*LRH*) test compares a haplotype's frequency to its relative *EHH* at various distances, looking for core haplotypes that are extended as well as common, compared with other core haplotypes at the locus (104). Modified versions of this test have been proposed (108, 109), including integrated haplotype score (*iHS*) capturing extreme *EHH* for short distances and moderate *EHH* for longer distances, increasing power to detect incomplete sweeps (110), and cross-population extended haplotype homozygosity (*XP-EHH*) as well as *EHHS* comparing haplotype lengths between populations (111, 112).

LD decay (*LDD*) test is an alternative LD-based method that does not need phased data, as it operates on homozygous sites and looks for large differences in LD around the ancestral and derived alleles of a given SNP (assuming the derived allele arose on a single haplotype) (8). The fraction of inferred recombinant chromosomes (*FRC*) at polymorphisms surrounding the homozygous site (*S*) is then computed within a certain physical distance (8). Neighbouring sites are binned according to the separation distance from the site *S*. The calculated *FRC* associated with the distance from the *S* is informative about the LD decay at various distances and is compared to the genome average (8). Strong local LD around the new high-frequency allele in comparison with the alternative allele indicates a selective sweep (8).

Alternative methods detecting long identical-by-descent DNA stretches and reduced haplotype diversity or reduce heterozygosity with increased proximity to a selected mutation in a population have also been proposed (113-116).

1.2.2.4. Composite methods

Methods combining multiple complementary metrics based on distinctive signatures into one composite test might provide greater power and/or resolution in pinpointing drivers of selection (10). For instance, methods combining information from different SFS tests assessing the distribution of different frequency classes of variation (117, 118) or methods merging SFS inferences with the Ewens-Watterson homozygosity test of neutrality (*DH* test) were proposed (119-121). Nielsen *et al*. combined population differentiation-based signatures with site frequency distortion measurements (excesses of high-frequency derived alleles and low-frequency alleles) (122). The composite of multiple signals (*CMS*) combines multiple signatures of selective sweeps taking into account three features: (i) haplotype length (measured by *iHS*, *XP-EHH* and *ΔiHH*), (ii) population differentiation (F_{ST}) and (iii) differences in derived allele frequencies between populations (*ΔDAF*) (123). Finally, Pybus *et al*. applied a machine-learning hierarchical classification framework (*boosting* algorithm) that exploits scores of

11 different selection tests to classify genomic regions into specific adaptive scenarios considering selective sweeps' completeness and time-frame (124).

1.2.2.5. Methods to detect adaptive introgression

Detecting adaptive introgression from archaic humans is a two-step process comprised of detecting a signature of selection and a signature of introgression, as currently there is no method detecting both signatures jointly (30). Any previously introduced population genomic approaches could be used to detect selection on introgressed DNA (30). However, as introgression alone (not necessarily adaptive) changes the pattern of LD and distribution of allele frequencies, methods relying on LD and the SFS, as well as composite methods, can lead to false inferences of selection (30). Therefore, it seems that population differentiation methods detecting the high frequency of archaic haplotype in a specific population relative to other populations are rather robust in this scenario (assuming a low level of introgression and low starting frequency of introgressed alleles) (30).

Detection of introgressed DNA tracts requires whole-genome sequences of modern and archaic humans and is usually based on the number of uniquely shared sites (sites containing high-frequency derived alleles in a particular population, which are also present in a distantly related population but absent in other more closely related populations) (30). Such methods are based on the assumption that the gene flow from archaic to modern humans happened after the out-of-Africa migration and that non-African haplotypes shared with archaic hominins but not with present day Africans might indicate introgression (125). The most commonly used method to identify overall introgression from genome-wide data is Patterson's *D* statistic (or the so-called *ABBA-BABA* statistic) based on differential sharing of derived alleles among different pairs of individuals or populations (125-127). *D* measures the excess of shared derived alleles between each of two populations in a pair (in-group populations) and an out-group population. In a scenario of a strict population phylogenetic tree with no admixture or migration, either of the in-group populations have had any gene flow from the out-group population and each of the two in-group populations should share approximately the same number of derived alleles with the out-group population (30). The significant deviations from the symmetrical pattern suggest introgression (30).

A challenge is to distinguish introgression from shared ancestral genetic variation (30). However, recent introgression should increase long-range LD, therefore introgressed tracts should be longer than shared ancestral variation or incomplete lineage sorting (30). The $S*$ statistic is a summary statistic based on patterns of LD and divergence which can be used locally to identify highly divergent haplotypes harbouring variants in strong LD shared with archaic hominins (128, 129). The expected length of the archaic tracts can be estimated using three parameters: the recombination rate, the number of migrants and the time since the admixture (130, 131).

Finally, an introgressed haplotype should have low sequence divergence from the putative archaic source population, but high sequence divergence from other present-day human individuals (30). This can be estimated by comparison of the time of the most recent common ancestor (TMRCA) of the test haplotype and the archaic haplotype with the TMRCA of the test haplotype and another modern human haplotype (132). A test human haplotype that has a recent TMRCA with an archaic population, but an ancient TMRCA with other human haplotypes is a candidate for introgression (30).

Probabilistic models based on the principles described above have also been employed to identify introgressed DNA fragments (130, 133, 134). Candidate adaptive introgressed segments are thus those that overlap between these two steps.

1.3. Thesis structure

This thesis presents results of a meta-analysis of previous selection scans, our computational approach to fine-map and prioritize candidate positively-selected variants, as well as results of functional follow-up of several candidates *in vitro* and *in vivo*. The general introduction to positive selection presented here is further extended in the following sections, each of which contains its own introduction and more specific relevant background information.