Chapter 2
Positive Selection in Human Populations: Practical Aspects and Current Knowledge

Pierre Luisi, Marc Pybus, Hafid Laayouni, and Jaume Bertranpetit

Abstract Natural selection targets a heritable trait that provides greater or lower chances for an organism to reproduce, and/or to survive, in a given environment. This evolutionary process is therefore directional: while an advantageous trait will be selected for and, thus, increase in frequency in the population, a prejudicial phenotype will be selected against and purged from the population. This concept, introduced in 1858 simultaneously by Charles R. Darwin and Alfred Wallace ((Darwin and Wallace J Proc Linnean Soc London 3:46–50, 1858); (Darwin On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London, 1859)), has been at the core of the study of

P. Luisi
Institute of Evolutionary Biology (UPF-CSIC), CEXS-Universitat Pompeu Fabra, Barcelona, Spain

Microbial Paleogenomics Unit, Institut Pasteur, Paris, France

Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba, Córdoba, Argentina

M. Pybus
Institute of Evolutionary Biology (UPF-CSIC), CEXS-Universitat Pompeu Fabra, Barcelona, Spain

Fundació Puigvert, Barcelona, Spain

H. Laayouni (✉)
Institute of Evolutionary Biology (UPF-CSIC), CEXS-Universitat Pompeu Fabra, Barcelona, Spain

Bioinformatics Studies, ESCI-UPF, Barcelona, Spain
e-mail: hafid.laayouni@upf.edu

J. Bertranpetit (✉)
Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
e-mail: jaume.bertranpetit@upf.edu

N. Saitou (ed.), *Evolution of the Human Genome II*, Evolutionary Studies,
https://doi.org/10.1007/978-4-431-56904-6_2

evolution and biological research. However, since then there has been passionate debate concerning its relative importance among other evolutionary processes, the prevalence of adaptive traits, and how they are originated in natural populations.

Since the recent wealth in genomics data, population and evolutionary geneticists have been able to interrogate the genome to understand the molecular basis of natural selection. In this chapter, we will focus on a particular mode of natural selection: positive selection also referred as adaptive selection or Darwinian selection. We describe statistical approaches to identify signals of positive selection and their practical challenges using genomics data. Then, we give a review on the current knowledge on positive selection in the human genome.

Keywords Positive selection · Adaptive selection · Hard sweep · Haplotype · Polygenic adaptation · Genome-wide selection scans · Complex adaptive traits · Selection on regulatory elements

2.1 Statistical Approaches to Identify Signals of Positive Selection

Charles R. Darwin and Alfred Wallace introduced at the mid-nineteenth century the concept of natural selection, focusing on phenotypic variation (Darwin and Wallace 1858; Darwin 1859). Since then, natural selection has been also widely studied at the genomic level, with a particular interest for positive selection. Positive selection refers to the process through which an allele that determines an advantageous trait will increase rapidly in frequency in the population, potentially until it reaches fixation. The allele frequency trajectory in the population through the action of positive selection depends on two main factors: the strength of the selective pressure and the number of generations since it started. The strength of positive selection is measured by the selection coefficient defined as the increased percentage of offspring of the individuals carrying the advantageous genotype in each generation as compared to individuals with alternative genotypes. A higher selection coefficient allows the advantageous allele to increase quicker in frequency, and thereby, to reach fixation in a shorter time. The speed of the increase tends to decline with the rise of frequency of the advantageous allele in the population since the relative advantage of individuals carrying the advantageous genotype declines with the frequency of their competitors. As a consequence, the allele frequency trajectory is non-linear and depends on the number of generations since the allele began to increase in frequency through the action of positive selection.

The shift in allele frequency comes with some typical molecular footprints used to detect selective events in the genome. Usually, we distinguish between two method families according to the kind of data analyzed:

- **Using divergence data**, i.e. sequences from different species, one can identify substitutions in the genome that are different across the species due to past selective events that contributed to the species divergence.
- **Using polymorphism data**, i.e. sequence or genotype data from different populations within the same species, to explore the nucleotide and haplotype diversity within and among populations.

The different molecular patterns left by a selective event are not maintained forever in the genome, and those footprints allow inferring how many generations have passed since the selective events occurred. In this chapter we will focus only on methods of detecting positive selection using polymorphism and the recent advances in methods developed to detect selection using both genotyping and sequencing data.

2.1.1 Using Polymorphism Data

In 1974, Maynard-Smith and Haigh (Maynard-Smith and Haigh 1974) proposed a model to explain the molecular mechanisms at play when positive selection acts on a variant. In this model, now referred to as the hard sweep model, they described the phenomena of genetic hitchhiking which results from positive selection driving a quick increase in frequency of an initially rare and beneficial allele toward fixation. This selective sweep occurs so quickly that recombination is not efficient to cut the haplotype where the selected variant arose, and thus, most of the variants carried by this haplotype also increase in frequency (Fig. 2.1). Therefore, under the hard sweep model, one expects a decrease in genetic diversity in the surrounding genomic region. The size of the region affected by such a sweep is proportional to the ratio

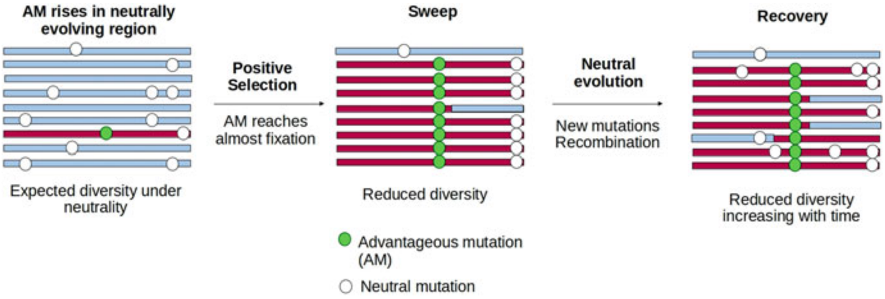
AM rises in neutrally evolving region
Expected diversity under neutrality
Positive Selection
AM reaches almost fixation
Sweep
Reduced diversity
Neutral evolution
New mutations Recombination
Recovery
Reduced diversity increasing with time
Advantageous mutation (AM)
Neutral mutation

Fig. 2.1 Molecular patterns in a genomic region suffering from a selective sweep. In a neutrally evolving region (before the selective sweep), an adaptive mutation (green circle) arises on one chromosome. During the selective sweep the frequency of the adaptive allele and its linked variants rapidly increase in frequency. After the sweep, the adaptive and linked alleles are fixed, and variability in the region is lost. During the recovery phase, new mutations begin to appear in different chromosome backgrounds by recombination and mutation restoring the diversity patterns

of the strength of selection and the rate of recombination (Barton 1998; Kaplan et al. 1989; Maynard-Smith and Haigh 1974). Thus, the reduction in levels of diversity within the genome is determined by the distribution of selection coefficients and the number of selective events in unlinked genomic regions. A selective sweep drives a quicker shift in allele frequency than what is expected under genetic drift. However, if recombination occurs, neutral alleles far from the selected site may not be driven to fixation, resulting in a temporary excess of high-frequency derived alleles at intermediate distance from the selected site (Fay and Wu 2000; Kim 2006; Przeworski 2002). Once the sweep is over, the genomic region enters a recovery phase during which it returns to neutral diversity levels through new mutations leaving a strong skew towards low frequency alleles persisting for many generations (Braverman et al. 1995; Kim 2006; Przeworski 2002). The strength and occurrence of sweeps can allow hitchhiking to dominate genetic drift, especially in large populations, and become the source of stochasticity for neutral alleles (Gillespie 2000; Kaplan et al. 1989; Maynard-Smith and Haigh 1974); this concept is known as genetic draft (Gillespie 2000). Maynard-Smith and Haigh formulated the theoretical background for most of the tests implemented thus far to detect signatures of selection at a molecular level using polymorphism data. A recently implemented database (Pybus et al. 2014) reports genome-wide scores for most of those tests ran on 1000 Genomes data in worldwide populations (The 1000 Genomes Project Consortium 2012), the latest publicly available polymorphism data. Those tests rely on three main features expected to be present in a genomic region surrounding a selected allele: long linkage disequilibrium (LD) haplotypes, a skewed Site Frequency Spectrum (SFS), and an excess of genetic differentiation among populations. The list of tests reported in this database (Pybus et al. 2014) is given in Table 2.1.

2.1.1.1 Tests Based on Long Haplotypes

Positive selection creates high levels of LD in the region surrounding the selected variant due to a quick shift in allele frequencies. For a given shift in allele frequency, less recombination events take place when there is a selective sweep than under genetic drift since the shift in allele frequency is much quicker in the former case. The Long Range Haplotype (LRH) test is commonly used to detect this signal (Sabeti et al. 2002a). However, this test does not take into account the recombination rate heterogeneity across the genome. To overcome this limitation, other tests have been implemented and are based on the Extended Haplotype Homozygosity decay (EHH, Sabeti et al. 2002a), which measures the decay of the haplotype homozygosity observed when moving away from the selected variant; this is caused by hitchhiking of a neutral allele (see Fig. 2.2 for a schematic representation of EHH decay calculation). The Cross-Population Extended Haplotype Homozygosity (XPEHH) compares the EHH decay observed in a population of interest to a reference (Sabeti et al. 2007). The integrated Haplotype Score (iHS; Voight et al. 2006) compares within the same population the EHH decay for the derived and ancestral alleles. Those two comparisons correct for recombination rate

Table 2.1 Statistics implemented by (Pybus et al. 2014) and are available in as UCSC tracks in the 1000 Genomes Selection Browser 1.0 at http://hsb.upf.edu/

Method family	Method	Reference
Site frequency Spectrum	Tajima's *D*	Tajima (1989)
	CLR	Nielsen et al. (2005)
	Fay and Wu's *H*	Fay and Wu (2000)
	Fu and Li's *D*	Fu and Li (1993)
	Fu and Li's *H*	Fu and Li (1993)
	R2	Ramos-Onsins and Rozas (2002)
Long haplotypes	XPEHH	Modified from Sabeti et al. (2007)
	ΔiHH	Modified from Voight et al. (2006)
	iHS	Modified from Voight et al. (2006)
	$\text{EHH}_{\text{average}}$	Modified from Sabeti et al. (2002a)
	EHH_{max}	Modified from Sabeti et al. (2002a)
	Wall's *B*	Wall (1999)
	Wall's *Q*	Wall (2000)
	Fu's *F*	Fu (1997)
	DH	Nei (1987)
	Za	Rozas et al. (2001)
	ZnS	Kelly (1997)
	ZZ	Rozas et al. 2001)
Population differentiation	F_{ST}	Weir and Cockerham (1984)
	XPCLR	Chen et al. (2010)
	ΔDAF	Hofer et al. (2009)
Descriptive statistics	Segregating sites	
	Singletons	
	π (nucleotide diversity)	Nei and Li (1979)
	DAF (Derived allele frequency)	
	MAF (Minor allele frequency)	

heterogeneity across the genome. Only recent selective sweeps (<30,000 years ago) can be characterized by the presence of long haplotype blocks because older sweeps have had time to shuffle the haplotype blocks, and are therefore not identifiable through this method.

2.1.1.2 Tests Based on Site Frequency Spectrum

The SFS is the representation of the number of alleles observed in a sample belonging to different frequency classes for a given set of polymorphic sites. Genetic hitchhiking around a selected allele will drive neutral alleles located nearby to high frequency leading to a reduced diversity, an excess of rare and derived alleles, and a

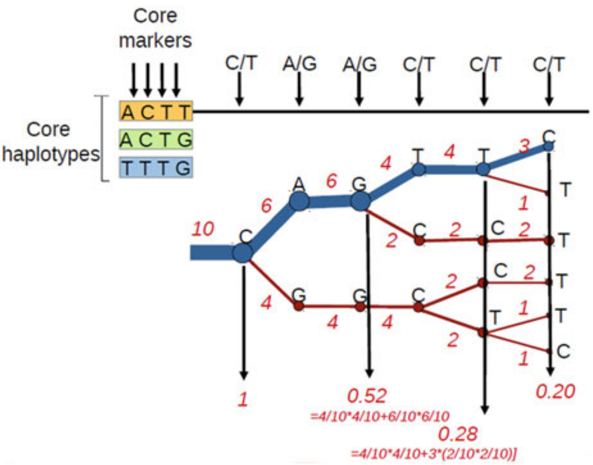

Fig. 2.2 Extended Haplotype Homozygosity decay. Moving away from the variant of interest, the haplotypes bifurcate and the haplotype carrying the core markers are less and less frequent. Thicknesses of the lines represent the frequency of the haplotype (haplotype counts in red). The haplotype homozygosity is given at the bottom

scarcity of alleles at intermediate frequency as compared to what is expected under neutrality (Fig. 2.3). The excess of rare alleles which persists for a long time during the recovery phase (up to ~250,000 years) can be formally tested by the famous statistic Tajima's *D* (Tajima 1989). Moreover, if the ancestral state of the variants is available, one can also test for the expected excess of high-frequency derived alleles (Fig. 2.3), with the Fay and Wu's *H* test (Fay and Wu 2000). This excess of rare alleles vanishes more rapidly as recombination allows neutral variants to evolve under genetic drift. This pattern can be detected for up to ~80,000 years after the sweep has occurred.

2.1.1.3 Tests Based on Genetic Differentiation

When a population faces a change in environment, positive selection may act on mutations that help the individual adapt better to this new environment. To detect the alleles responsible for local adaptation, one approach is to study genetic differentiation among populations. Traditionally the most used statistic is the fixation index, F_{ST}, first introduced by Sewall Wright which has been reformulated by multiple researchers. Using Cockerham and Weir's formula (Weir and Cockerham 1984), F_{ST} can be viewed as the proportion of genetic diversity due to allele frequency differences among populations:

$$F_{ST} = \frac{\sigma_{a^2}}{\sigma_{w^2} + \sigma_{b^2} + \sigma_{a^2}}$$

${\sigma_w}^2$, ${\sigma_a}^2$, and ${\sigma_b}^2$ are the intra-individual, inter-population, and within population inter-individual variances, respectively.

F_{ST} ranges from 0 to 1, with 0 signifying no differentiation (complete panmixia) and 1 indicating complete differentiation of the populations. Although high F_{ST} can

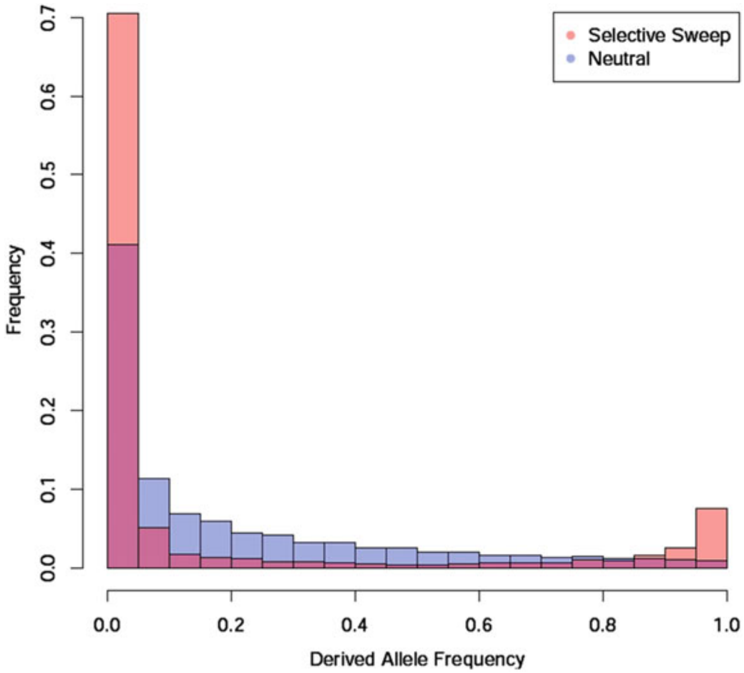

Fig. 2.3 Site Frequency Spectrum under different evolutionary models. The Unfolded SFS represents the number of derived alleles observed within different frequency classes. A region that has evolved under positive selection presents an excess of rare variants and of derived alleles at high frequency (red). During the recovery phase, the former pattern will remain due to new mutations arising in the region while the latter is lost more rapidly. Based on coalescent simulations of 100Kb regions evolving under neutrality (3000 neutral replicates in blue) and with a recent selective sweep driving an advantageous mutation up to fixation (300 selective sweep replicates in red), in an European-ancestry demographic model using COSI software (Schaffner et al. 2005)

putatively be attributed to the action of positive selection in one population, this approach is often criticized because of its sensitivity to population structure, demographic history, ascertainment bias, sample size, and minor allele frequency (for a review, see Holsinger and Weir 2009). The ΔDAF score (the differences of derived allele frequency between one population and a reference; Hofer et al. 2009) is another genetic differentiation index which suffers the same limitations. However, the use of the derived allele state allows identification of the population where positive selection has occurred. Further methods using genetic differentiation pattern have been developed. For example, the Cross-Population Composite Likelihood Ratio test (XPCLR) developed by Chen et al. (2010) relies on the comparison of a null model of genetic drift to one with a selective sweep by taking advantage of the genomic context around the selected allele in order to detect genomic regions with SFS differentiation among populations due to hitchhiking. This makes XPCLR more robust to demography and ascertainment bias than individual single nucleotide polymorphism (SNP) based methods such as F_{ST} and ΔDAF.

2.2 Practical Challenges in Detecting Positive Selection Using Polymorphism Data

Until recently, positive selection studies have been limited to sequence data from a restricted number of genes covering only a few thousands nucleotides. Now that detailed genetic maps are available in many human populations, it is possible to measure the signature of positive selection on a genomic scale using polymorphism data. Therefore, it is interesting to describe some potential challenges in detecting positive selection using polymorphism data and approaches to overcome them. First, detecting the different genomic footprints left by positive selection may be difficult in chip-array data. Second, those footprints may result from other mechanisms.

2.2.1 Distortions Due to Ascertainment Bias

Most genotype data used to study population diversity contain relatively important ascertainment bias. Ascertainment bias is the systemic distortion of the allele frequency spectrum due to a priori discovery of the polymorphisms segregating in a reduced sample. Thus, when genotyping individuals from other populations, especially those distant from the reference sample (the one where the initial genetic variants are described), it is not possible to catch all the genetic variation present in these populations.

Ascertainment bias is an intrinsic feature of genotyping technologies which are extensively used because they are simpler, cheaper, and much faster than sequencing approaches. The resulting genotype information for the population of interest will not be accurately produced for all the segregating sites but only for those present in the discovery sample. The probability of identifying a SNP is a function of its frequency, and as a consequence common SNPs are easier to detect. For example, many arrays use SNPs discovered in European samples, and, when used worldwide, the positions on the array are not polymorphic for the population of interest. Populations do not share all variation and some SNPs are private to particular populations (Casals and Bertranpetit 2012). The SNPs for newly designed arrays have been selected from public databases such as HapMap (www.hapmap.org) which in turn present an ascertainment bias of their own.

Usually, SNPs are selected to be genotyped in a population of interest with some of the following criteria: (1) having a Minor Allele Frequency (MAF) above a given threshold, usually relatively high in discovery samples representing either one or several populations of interest; (2) SNPs that are distant from one another by a given number of base pairs; (3) SNPs within targeted regions of interest; (4) SNPs maximizing the tagging of additional common SNPs that are in LD with them.

The criteria used affect the ascertainment bias, and it is difficult to assess a posteriori its extent when using genotyping arrays designed by others. Arrays with reduced ascertainment bias have been developed, for example, the Omni Family of Microarrays from Illumina which includes up to five million markers per sample and extensive coverage of new variants identified by the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012), i.e. SNPs discovered through Next Generation Sequencing (NGS) in samples from worldwide populations; and Patterson and collaborators designed the Affymetrix Human Origins array with clearly documented ascertainment specifically for the study of population genetics (Patterson et al. 2012).

Ascertainment bias has a direct effect on many statistics that detect positive selection using polymorphism data (Thornton and Jensen 2007). First, and the most straightforward, SFS-based statistics are distorted by the artifact of the excess of common variants in genotyping arrays. Second, the tests based on genetic differentiation, such as the F_{ST} index, rely on a measure of genetic variance within and among the populations. Hence, if the SNPs genotyped within different populations present different ascertainment bias, the distribution of the index of genetic differentiation will be distorted. Haplotype-based statistics were developed in the first decade of this century with the goal to implement other methods less sensitive to ascertainment bias (Sabeti et al. 2007, Voight et al. 2006). These methods rely on an accurate estimate of LD patterns within a genomic region in order to infer whether there is a pattern of EHH (Granka et al. 2012). If the genotyping array only contains common variants and particularly chosen to *tag* the variability from another population (González-Neira et al. 2006), the observed LD patterns in the studied population are unlikely to be real. For example, in the data from the Human Genetic Diversity Panel (HGDP; Cann et al. 2002), for African populations the genotyping tag only 67% of SNPs with MAF above 5% and the power to detect positive selection is lower than for European samples, where 90% of such SNPs are tagged. It has been proved that haplotype diversity is more representative than individual SNP heterozygosity in the HGDP data (Conrad et al. 2006), suggesting that the ascertainment schemes affect more individual variants than haplotypes.

Nowadays, more studies obtain genotype information through NGS which does not suffer any ascertainment bias. However, the SFS is highly dependent on the coverage (read depth) used for sequencing. The power to detect rare variants increases with coverage (The 1000 Genomes Project Consortium 2012). Moreover, the genotype information may also depend on the sequencing center, its technology, and the SNP calling algorithm used. Therefore, for population genetics studies, one should be cautious when merging data from different datasets and control for the coverage across the genome.

2.2.2 The Confounding Factor of Background Selection

Background selection (BGS) is a process by which neutral variation is removed from the population when linked to deleterious variants (Charlesworth et al. 1993). Therefore, BGS reduces levels of polymorphisms in regions with many functional elements and low recombination. The lower level of polymorphisms in an extended region is often attributed as a result of positive selection because it is a molecular pattern expected under the hard sweep model. It is consequently important to correct for BGS. One straightforward approach when analyzing protein-coding regions is to look for lower levels of neutral variation near functional substitutions, i.e. at functional sites where a mutation has been fixed in a set of species, which is evidence for positive selection while not being expected under BGS. However, this approach is biased towards protein-coding regions, and would just detect events of positive selection acting on mutations with a priori known function. An alternative to this approach would be to correct for several genomic variables that correlate with BGS, such as levels of recombination rate and functional constraint. Measuring functional constraint is not straightforward but one can use the density of coding sequences (CDS), conserved coding sequences (CCDS), conserved non-coding sequences, and untranslated regions (UTRs). Moreover, Enard et al. recently found that GC content presents a strong correlation with levels of neutral diversity (Enard et al. 2014). Although BGS has been seen as mimicking positive selection at a molecular level, after the article by Charlesworth, Morgan, and Charlesworth (1993), tests based on LD—namely XPEHH and iHS—show insensitivity to BGS (Enard et al. 2014; Fagny et al. 2014), and therefore, their extreme deviations may directly be attributed to recent hard sweeps.

2.2.3 Demography Can Mimic Positive Selection

Many neutral mechanisms can affect the genetic diversity in populations or species, among which several demographic processes can lead to molecular patterns expected under a positive selection scenario (Table 2.2).

2.2.3.1 Migration and Structure

The neutral model assumes that any cross-gender individual pair has the same probability to reproduce in the population. However, there may be population subdivision due to geographic distance, social, linguistic, or economical barriers (e.g., in India with the caste system). Barriers to random mating are likely not to be absolute, and a number of migrants can move between subpopulations each generation. When hidden population subdivision is occurring and panmixia is improperly assumed, the genetic variability is higher than expected with an excess of variants at

Table 2.2 Some demographic processes can leave molecular patterns expected under positive selection

Process	Description	Molecular pattern
Migration	Individuals move from one population to another	Increased genetic variability within each population and lower genetic differentiation among populations
Isolation	One population is isolated from the others and drifts on its own	Increased genetic differentiation among populations
Population structure	The studied population is actually structured into several subpopulations	Higher variability than expected
Population expansion	The population increases rapidly in size	Increased number of rare variants and decreased variability
Population bottleneck	The population decreases rapidly in size and rebounds to its original size after several generations	Increased number of rare variants and derived alleles at high frequency with decreased variability
Founder effect	A new population is founded by a small number of individuals from a larger population and the new population then increases in size	Gene surfing: Mutations that occur on the frontier of a growing population are more likely to expand and get fixed since only a few individuals are founding the population

intermediate frequency. Migration from an external population causes a higher variability with an excess of rare variants is expected.

2.2.3.2 Population Expansion

During population expansion, a new generation has a greater number of individuals than the previous one. A well-described human population expansion event occurred after the Neolithic transition. One possible cause is that the agriculturist way of life may have provided a more reliable mode of sustenance and allowed settlements to increase in size. A population with expansion will show an excess of singletons at low frequency as compared to a population with constant size due to recent mutations which have not increased in frequency through genetic drift, and remain almost individual specific (Keinan and Clark 2012). This also implies a lower genetic variability than expected for the population size.

2.2.3.3 Population Bottleneck

A bottleneck is the phenomena through which population size decreases suddenly, followed by a recovery, or increase, of the original population size in a few generations. One striking example is the Black Death plague faced by Asian and European populations in the fourteenth century. Plague is thought to be responsible for several large epidemics with death rates of up to 30–50% of the European population and lingering thereafter in Europe for several centuries (McEvedy 1988). Many alleles from the original population, mostly at low frequency, will either disappear or become very frequent during the decreasing size phase, thus reducing the genetic variability. During the recovery phase, as in population expansion, an excess of rare variants will arise.

2.2.3.4 Founder Effect

A founder effect occurs when a small subpopulation leaves its former habitat to establish a new one. This can be seen as a particular case of a bottleneck. Modern humans likely colonized geographic areas out-of-Africa through several founder effects (Reich et al. 2001). One more recent example would be the colonization of Quebec, Canada ~400 years ago by ~8500 French settlers. Such event allows variants to rapidly reach fixation through genetic drift, a phenomena called gene surfing (Hallatschek and Nelson 2008), which mimics genetic hitchhiking.

2.2.4 Has a Region of Interest Evolved Under Positive Selection?

One major challenge in assessing whether a region of interest has evolved under the action of positive selection is to circumvent the confounding factors of past demographic processes as well as data ascertainment bias. For that purpose, one can compute the statistic designed to detect footprints of positive selection and estimate its significance by comparison to a reference distribution. This reference distribution must reflect the expected score under selectively neutral evolution with the data used. Indeed, values of statistics are not absolute but are relative to the studied population and to the kind of data analyzed. There are two main approaches to defining reference distributions: simulations and the outlier approach.

2.2.4.1 Using Simulations Accounting for Demography

Since the development of coalescence theory (Hudson 1991; Kingman 1982; Wakeley 2008) and the recent wealth in computational capacity, simulations have become a powerful approach in population genetics. It is now possible to generate large independent data sets through simulations of genetic data that mimic population demographics. Those data sets are, in turn, used to assess the statistical significance of empirical data accurately. Particularly, one can simulate sets of genetic data under a neutral model with appropriate demographic parameters to infer what the empirical data would look like without the action of positive selection, and then, a significance threshold at a given false positive rate (FPR) can be estimated. In this case, any putative biases from empirical data are eliminated. Furthermore, in order to evaluate the reliability of the estimated threshold, simulations can incorporate selective events to the neutral model to infer the power of the approach.

The simulation software that has been implemented so far can be divided into those based on coalescent theory and on forward simulation. Coalescent simulation is the first approach widely used to simulate genetic data at the sequence level and, as the name suggests, is based on coalescent theory. First introduced by John Kingman in 1982 (Kingman 1982), it relies on a backward model describing the characteristics of joining lineages back in time to the most recent common ancestor (MRCA). It represents the theoretical background for most neutral genetic models, as well as the estimation of many population genetic parameters. The coalescence theory provides computational efficiency with several coalescence simulation software available, such as *FastCoal* (Marjoram and Wall 2006), *CoaSim* (Mailund et al. 2005), *SelSim* (Spencer and Coop 2004), *cosi* (Schaffner et al. 2005), *ms* (Hudson 2002), and *msms* (Ewing and Hermisson 2010).

For many of the underlying coalescent models, parameters have been calibrated to fit empirical data in order to retrieve the past demographic history of human populations. For example, Schaffner et al. used HapMapIII data to infer the demographic history of three populations through the calibration of their model by making the simulated data match empirical data for pairwise F_{ST} values, LD decay (how LD for pairwise SNPs decreases with physical distance in the genome), and SFS (Schaffner et al. 2005). Further implementations used more complex empirical data features, such as the joint SFS across populations (Gravel et al. 2011). Those programs simulate genomic regions spanning a few megabases in hundreds of samples without large computational costs in time or resources. This is particularly useful when computing large simulated distributions of the statistics to estimate the statistical significance for a genomic region. However, coalescent simulations present several limitations. Most importantly they have limited accuracy in simulating the number of recombination and gene conversion events, and the ability to implement possible recombination patterns. As a consequence, a realistic recombination map incorporated into the model increases the computational cost and therefore reduces the size of the simulated region. With a simplistic recombination map, the simulated genomic regions can be longer but the model is unlikely to be accurate. Another traditional issue with coalescent simulations is the incorporation of selective events. Attempts to improve coalescent simulations (Ewing and Hermisson 2010; Grossman et al. 2010; Spencer and Coop 2004) have usually come at the cost of over-simplifying other aspects of the model such as recombination map, population changes, sample size, and length of the simulated genomic regions.

To circumvent the limitations, the forward simulation approach has been proposed as an alternate. Genomic data is simulated forward in time from an ancestral status, allowing more flexibility to the model including complex recombination patterns and other genomic features (gene content, background selection; for an example, see *SFS_CODE,* Hernandez 2008). The demographic processes included in the model can also present a much higher layer of complexity (e.g., see *dadi,* Gutenkunst et al. 2009; Uricchio and Hernandez 2014). However, these approaches require the simulation of whole populations and, therefore, are very computationally expensive, preventing the generation of large data sets. For a neutral model of human demography, Excoffier and colleagues implemented a coalescent model, *fastsimcoal2*, which allows for a high level of demographic complexity, with serial founder effects, range expansions, and admixture among populations (Excoffier et al. 2013). This model overpasses forward simulation models such as *dadi* *(*Gutenkunst et al. 2009) which is arguably the reference in the field. The models are calibrated to make the simulated data fit the empirical data. Therefore, when the empirical data contains ascertainment bias, it is important to either correct for it (Nielsen et al. 2004) or to take it into account in the estimation procedure (Pickrell et al. 2012; Wollstein et al. 2010). Although it is not an easy task the calibrated model can inaccurately reflect past demography if ignored (Excoffier et al. 2013). In addition, most models rely on a priori assumptions on demographic events and therefore accurate models are available for a reduced number of well-studied populations.

2.2.4.2 Outlier Approach

As mentioned before, constructing a neutral model using simulations is computational expensive and the model is not likely to incorporate all the layers of demographic and genomic complexity. One may prefer to use the outlier approach: an empirical distribution of statistics to detect positive selection built from a large number of loci across the genome. The loci located in the extreme tail(s) of the distribution, i.e. outliers, are considered as possible targets of positive selection. The assumption behind this framework is that demography stochastically affects the whole genome evenly while positive selection, a deterministic process, affects only a few loci and does not distort the distribution. This approach also allows correction for ascertainment bias and the confounding effect of background selection, as long as the reference loci are accurately sampled. It is important to note that the genome can be seen as a mosaic of several chunks, each with its own history, and although the population definition is accurate, the chunk demographics may be very different with some specific genomic regions exhibiting extreme molecular patterns that mimic positive selection. This may be inaccurately identified as under positive selection resulting in false positives (Kelley et al. 2006). Inaccuracies occur particularly in the case of positive selection targeting recessive alleles, standing variation, and population bottlenecks (Teshima et al. 2006). Another difficulty of the outlier approach is the arbitrary threshold used to consider a score's significance. Setting thresholds require a priori definition of the proportion of the genome expected to be under positive selection. For example, if the 5% most extreme scores are considered to be under putative positive selection, the underlying assumption is that 5% of the genome is expected to be under selection. However, no accurate estimate is available for many organisms and it remains one of the main questions in studying positive selection. Finally, the outlier approach only identifies the most extreme case of positive selection, many of the selected alleles, especially those with a relatively low selection coefficient, are likely to be false negatives.

2.2.4.3 Combination of Different Tests

Assessing statistical significance for a given score through either simulations or the outlier approach is necessary to determine whether a genomic region has been evolving under positive selection. However, it is delicate to make sure that a significant score is not actually a false positive and especially difficult when drawing conclusions from only a single method that a locus has been targeted by positive selection. To reduce the risk of false positives, it is wise to use different methods developed to detect the impact of positive selection at a molecular level. Particularly, one may use methods based on different kinds of molecular footprints left by a selective sweep (SFS, LD, and genetic differentiation). This way, the false discovery rate is likely to be reduced: the false positives from individual methods are unlikely to overlap, since each method is sensitive to different demographic processes. Zeng et al. implemented two compound tests, DH (Zeng et al. 2006) and DHEW (Zeng et al. 2007), which combine the SFS-based methods Fay and Wu's H and Tajima's D specifically the Ewens–Watterson test (Watterson 1978) for DHEW. The underlying idea of DH is that Fay and Wu's H and Tajima's D are sensitive to population bottlenecks and expansions, respectively (Zeng et al. 2006), while insensitive to the other demographic process. Thus, combining the two tests is robust to both demographic processes. The idea is very simple; using neutral simulations, a significance threshold for both tests is set for a given FPR. Afterwards, if a region of interest is significant for both tests, it is identified as a target of positive selection. The original method relies on neutral simulations with rather simplistic demography using *ms*, but the framework suggested by Zeng et al. can extend it to an outlier approach as in Luisi et al. (2015) where Fay and Wu's H and Tajima's D are computed in a large number of genomic regions to make the reference distribution and estimate of the join threshold significant.

A more simplistic method is to use any combination test, i.e. a test that combines K individual test's P-values, such as the Fisher combination's test:

$$\mathrm{Z_F} = -2\sum_{\mathrm{i=1}}^{\mathrm{K}} \log \mathrm{P_i}$$

where P_i is the P-value associated to the score of the i^{th} test.

Following this idea, Grossman et al. implemented a Composite Multiple Score (CMS; Grossman et al. 2010) which multiplies P-values of five individual tests based on long haplotypes—XPEHH, ΔiHH, and iHS—and genetic differentiation—F_{ST} and ΔDAF. The main improvement from a rather simplistic combination score is that they computed P-values from simulations using the demographic model calibrated by Schaffner et al. (2005) under a neutral scenario and with a selective event. Then, the CMS is obtained as the following:

$$\mathrm{CMS} = \frac{\prod_{i=1}^{5} P(s_i \mid \text{selected}) \times \pi}{P(s_i \mid \text{selected}) \times \pi + P(s_i \mid \text{unselected}) \times (1 - \pi)}$$

where s_i is the score of the i^{th} method, the P-values are obtained from reference distributions from simulations under either neutral (*unselected*) or selective scenarios and π is the uniform prior probability of selection.

CMS and other combination tests (e.g., Fisher's combination) cannot use any kind of individual tests since they rely on the assumption of the independence among tests. Moreover, they attribute equally to the combined score. In Pybus et al. (submitted), an alternative framework, *Boosting*, incorporates the information from different methods. Based on *Boosting* functions (Lin et al. 2011), this framework allows detection and classification of selective events. *Boosting* is a Support Vector Machine (SVM; Schapire 1990) which is trained on simulated data to estimate the best regression function of scores from different individual methods to distinguish between two scenarios. The algorithm begins with a neutral demographic model (Schaffner et al. 2005) to which a selective sweep scenario can be incorporated (Grossman et al. 2010); thousands of genomic regions have been simulated under a selectively neutral scenario and 45 selective ones. Then, two *Boosting* functions have been trained to distinguish among the scenarios, (1) evolution under either pure genetic drift or with a partial selective sweep (where the selected mutation reaches a final allele frequency (FAF) of 0.2 or 0.4); (2) evolution with an incomplete selective sweep (FAF = 0.6 or 0.8); and (3) evolution with a complete sweep (FAF = 1). Two further boosting functions have been built to classify regions evolving under a complete or incomplete sweep into recent or ancient sweep categories (Fig. 2.4). Those functions are included in a classification tree as shown in Fig. 2.5. This framework uses the combination of different, although relatively correlated, tests to classify the mode of positive selection for the detected selective events. As seen in Fig. 2.6, the standardized coefficients for each test give valuable insight into the methods that contribute the most when distinguishing between two given scenarios, and thus, on their ability to detect a given selective event. Moreover, the boosting coefficients are quite similar for the three populations analyzed (African (AFR), European (EUR), and Asian (ASN)), and thus seems quite robust to demography.

2.2.5 *Selection Not Only by Hard Sweep*

On the one hand, clear evidence of morphological and physiological adaptations in modern human populations exists, such as pigmentation for solar radiation, body size for thermal condition, and blood flow and oxygen delivery for high altitude. On

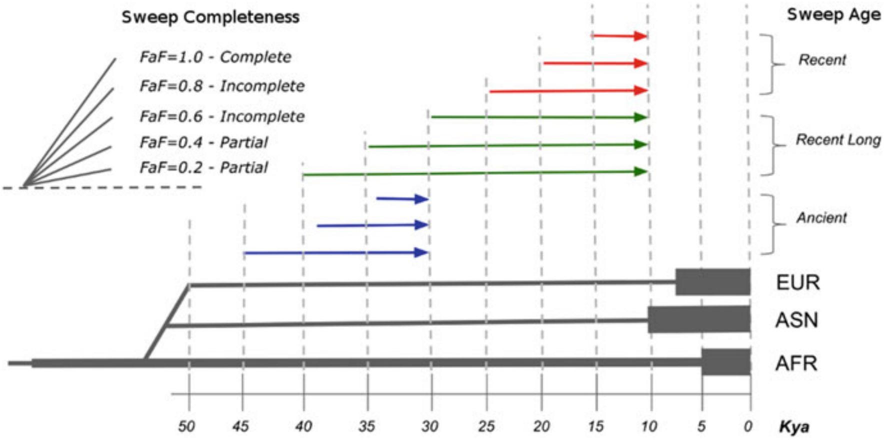

Fig. 2.4 Simulation scenarios. Simulations were run following a calibrated human demography that resembles population genetic data from three reference continental populations (European (EUR), Asian (ASN), and African (AFR), from left to right) [45]. Nine different time-spanning selective sweeps were simulated (grouped as Neutral, Recent, Recent Long, and Ancient) allowing for five different final allele frequencies (FaF = 0.2, 0.4, 0.6, 0.8, and 1.0)

the other hand, there are few examples of fixed, or almost fixed, genetic differences among populations and/or validated cases of adaptive mutations (see Sect. 3 for an overview). Moreover, Hernandez et al. (2011) showed that hard sweeps may have been rare during human evolution (but see Sect. 3.3.3). This striking inconsistency between the number of known phenotypic and genotypic adaptive examples may be explained by the simplistic way that positive selection has been researched. Indeed, until now, most studies of natural selection relied on the hard sweep model making use of methods designed to detect molecular patterns expected to remain in the genome under this model. In order to have a complete picture of adaptation and its genomic processes, it is important to consider other modes of positive selection. The other types of positive selection do not leave the same molecular footprints as a hard sweep. These alternate modes of positive selection require theoretical development but are beginning to be studied after being overlooked for many decades (Pritchard et al. 2010).

2.2.5.1 Soft Sweep

Recently, empirical (Colosimo et al. 2005; Hamblin and Di Rienzo 2000; Jeong et al. 2008; Scheinfeldt et al. 2009; Tishkoff et al. 2007) and theoretical (Hermisson and Pennings 2005; Innan and Kim 2004; Orr and Betancourt 2001; Pennings and Hermisson 2006; Przeworski et al. 2005) studies indicate the importance of soft sweeps which can occur through two different modes of adaptation:

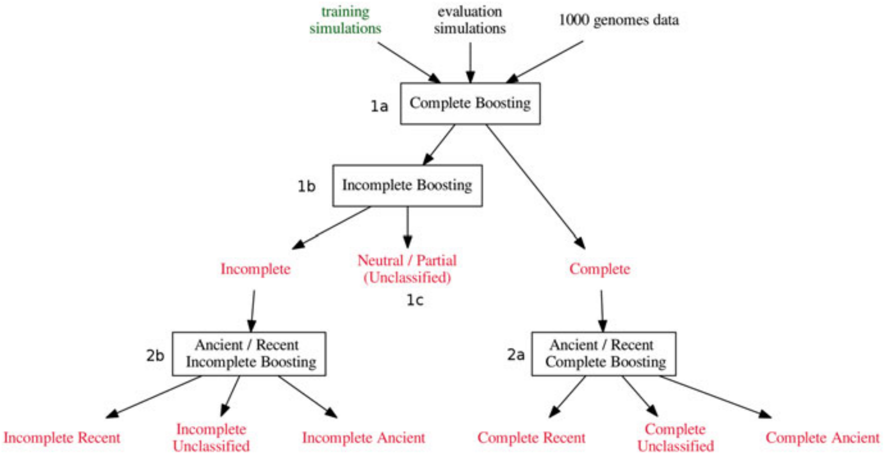

Fig. 2.5 Implemented classification tree. The implemented classification tree was organized in two levels: an unknown genomic region is firstly classified according to the completeness of the sweep, as being Complete, Incomplete, or Unclassified. In the second step, it is then classified according to the age of the sweep, being Ancient, Recent, or Unclassified. The algorithm can be described as following: (1a) If the Complete Boosting score is above the 99th percentile of the distribution of the Complete Boosting scores for the training simulations under the Neutral, Partial, and Incomplete scenarios, the region is classified as Complete and go to step 2a, otherwise go to step 1b. (1b) If the Incomplete Boosting score is above the 99th percentile of the distribution of the Incomplete Boosting scores for the training simulations under the Neutral and Partial scenarios, the region is classified as Incomplete and go to step 2b, otherwise go to step 1c. (1c) If not classified at iteration 1a or 1b, the genomic region is left unclassified and the algorithm stops. (2a) If the Ancient/Recent Complete Boosting score is above the 99th percentile of the distribution of the Ancient/Recent Complete Boosting scores for the training simulations under the Complete Recent scenario the region is classified as Complete Ancient, while if it is below the 1st percentile of the distribution of the Ancient/Recent Complete Boosting scores for the training simulations under the Complete Ancient scenario the region is classified as Complete Recent, otherwise the region remains only classified as Complete. (2b) If the Ancient/Recent Incomplete Boosting score is above the 99th percentile of the distribution of the Ancient/Recent Incomplete Boosting scores for the training simulations under the Incomplete Recent scenario the region is classified as Incomplete Ancient, while if it is below the 1st percentile of the distribution of the Ancient/Recent Incomplete Boosting scores for the training simulations under the Incomplete Ancient scenario the region is classified as Incomplete Recent, otherwise the region remains only classified as Incomplete

- *Selection on a standing variant*. In opposition to a hard sweep, selection on a standing variant does not rely on the appearance of an advantageous mutation to arise in the population, but rather targets a variant already segregating at a relatively important frequency when a change of environment occurs.
- *Selection on recurrent mutation*. For selection on recurrent mutations to occur, the derived and advantageous allele arises in the population several times independently, as a result of recurrent mutations or gene flow from another population. All copies of the derived allele increase in frequency until the allele reaches fixation. However, if all copies of the derived allele have similar selective

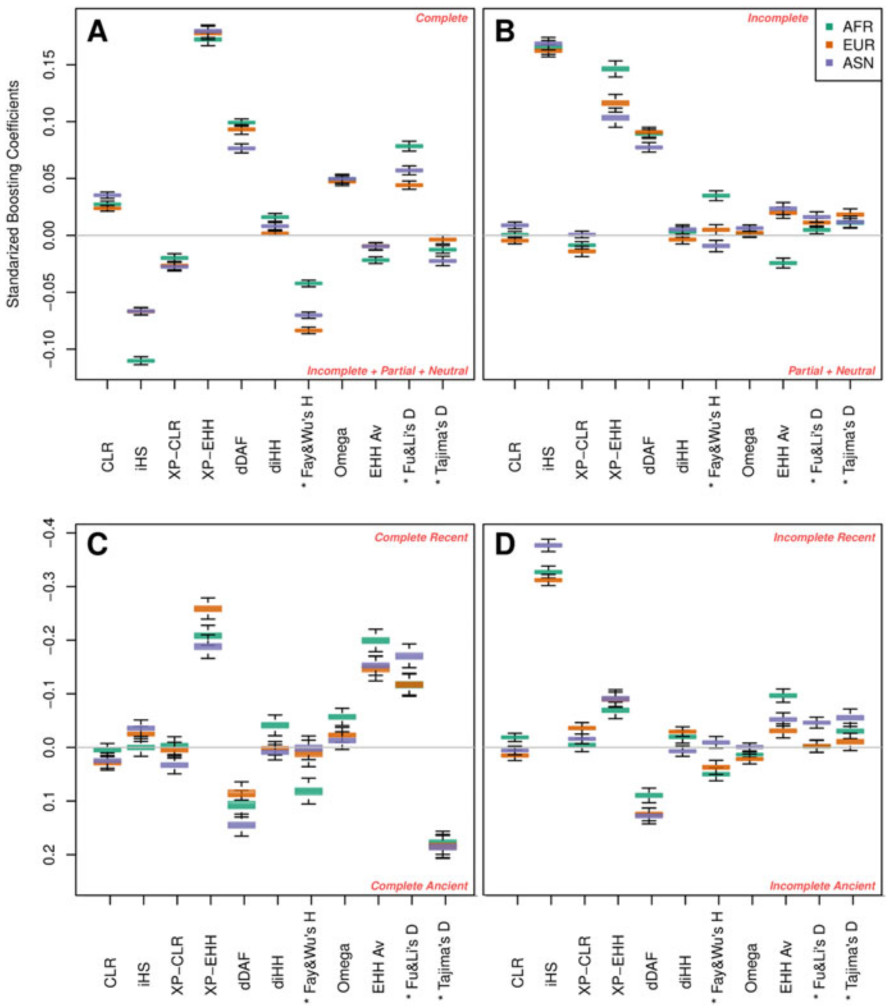

Fig. 2.6 Standardized coefficients for the European (EUR), Asian (ASN), and African (AFR) populations and for each implemented boosting function. Estimated coefficients for each population in the four boosting functions used in the classification tree: Complete (**a**), Incomplete (**b**), Complete Recent/Ancient (**c**), and Incomplete Recent/Ancient (**d**). The relevance of the positive selection tests to classify the different scenarios is given by the strength of its standardized coefficient

coefficients (because the genetic background has no affect through, for example, intragenic epistasis), none of the haplotype carrying one of these copies will fix during the selective event (Hermisson and Pennings 2005; Pennings and Hermisson 2006). Actually, diferent haplotypes, each carrying one copie of the advatageous allele, will increase in frequency until the allele has fixed in the population.

In both cases different copies of the selected allele may belong to different haplotypes: in the case of standing variation it was already segregating on different haplotypes before the selective event, while in the recurrent mutation case, it arose on different haplotypes. In both cases tests based on long haplotypes are not suited to detect this mode of adaptation. However, if the selective pressure is population specific, methods based on genetic differentiation may be able to detect it. In addition, other haplotype patterns, beyond the EHH, can be informative (see below).

2.2.5.2 Polygenic Adaptation

Recent genome-wide association studies (GWASs) confirm the view of classic quantitative genetics that many phenotypes are encoded by several dozens, hundreds, or even thousands of genes, rather than a unique one (Fu et al. 2013). This drastically contrasts with the idea that positive selection acts on a single advantageous mutation to drive phenotypic adaptation. Therefore, more focus on polygenic adaptation is required. Such a mode of adaptation would simultaneously cause a limited shift in allele frequencies at several variants located in different genomic regions and have small effects on fitness. This pattern is extremely difficult to distinguish from pure genetic drift.

2.2.5.3 Recent Methodological Advances in Detecting Alternative Sweep Scenarios

The molecular patterns expected to be left by soft sweeps and polygenic adaptations are not as evident as those left by hard sweeps (Fig. 2.1). Therefore, a lack of methods designed to detect such selective events at the genetic level exists. However, ongoing methodological development is in progress. Some already existing methods can be used to detect soft sweeps. Indeed, as mentioned above, if the selective pressure is population specific, a locus-based statistic of genetic differentiation (e.g., F_{ST}) may be powerful provided the variant is segregating at low frequency in the reference populations. iHS shows sensitivity when positive selection acts on a standing variant that was segregating at low frequency before the selective event (Ferrer-Admetlla et al. 2014). Two other methods relying on specific haplotype patterns have been recently developed (Ferrer-Admetlla et al. 2014; Garud et al. 2014). First, nS_L (Ferrer-Admetlla et al. 2014) is based on the comparison between EHH for derived and ancestral alleles, as in iHS, but also takes into account the length of the segment of haplotype homozygosity between a pair of haplotypes. Besides showing greater power than iHS for scenarios where the advantageous allele was already present in the population at frequency $> 3\%$, it does not need any genetic map and is robust to recombination rate and mutation rate. Second, the H12 and H2/H1 statistics (Garud et al. 2014) also rely on homozygosity of multiple haplotypes. H12 use the combined frequency of the first and second most frequent haplotypes observed in a genomic region as the following:

$$\mathrm{H12} = (p_1 + p_2)^2 + \sum_{i>2} p_{i^2}$$

where p_i is the frequency of the i^{th} most common haplotype in the sample.

The H12 statistic has power to detect hard sweeps and—not so—soft sweeps, i.e. when the starting frequency is below 0.1%. In order to distinguish between those two scenarios Garud et al. further developed the H2/H1 statistics (Garud et al. 2014):

$$\mathrm{H2/H1} = \frac{\sum\limits_{\mathrm{i} \geqslant 2} p_{1^2}}{\sum\limits_{\mathrm{i} \geqslant 1} p_{1^2}}$$

where p_i is the frequency of the i^{th} most common haplotype in the sample.

While H1 is expected to be higher under the hard sweep model, H2 is expected to be higher under the soft sweep scenario. Therefore H2/H1 increases with the softness of the sweep, i.e. the number of haplotypes on which the advantageous mutation is segregating prior to the selective event.

Those two recent methods demonstrate that accurate theoretical implementation allows detection of soft sweeps despite difficulty in recognizing the molecular patterns. Further theoretical work is required to increase the power to detect even softer sweeps. Despite the fact that the reduced shift in allele frequency expected under polygenic adaptation leaves very weak footprints in the genome, it could be argued that increasing the sample size would increase the power, and implementing methods using only genetic information seems a losing battle. For this reason, the few methods that have been proposed include other kinds of information. First, the BayENV (Coop et al. 2010; Günther and Coop 2013) method uses environmental variables. It is based on the correlation between allele frequency and an environmental variable observed in many populations. For each locus, it provides a Bayes Factor which is the ratio between two Bayesian posterior probabilities:

- Under the null (neutral) model, the correlation we observe in allele frequencies between different populations is just explained by demographic factors (genetic drift, migration, and population size changes).
- Under the model where a specific environmental variable has caused a selective pressure in (a) population(s) it may have caused an imbalance in the allele frequency spectrum across the populations.

Therefore, this method detects variants that shifted similarly in allele frequency in populations facing the same environmental pressures compared to their neighboring populations. Parallel selection, recently theoretically analyzed by Ralph and Coop (2010), is more likely to occur on ancient variants that are shared among worldwide populations. Note that the signal of selection is driven by the shift in allele frequency across populations rather than by its amplitude. This method corrects for population structure and therefore is less sensitive to demography than a simple correlation analysis. Indeed, the genetic differentiation among populations is directly related to their geographic distance (Gutenkunst et al. 2009) due to the isolation by distance phenomena. However, retrieving environmental variables from many populations may be challenging, especially because it relies on representative geo-localization.

An approach suggested by Mendizabal et al. (2012) includes phenotypic information rather than selective pressure. More precisely, the authors have analyzed the covariance between allele frequencies and height measurements to detect genetic variants allowing Pygmy adaptation to the rainforest climate by better thermo regulating with size reduction which is known as Bergmann's rule. This approach would require extensive phenotypic measurements but the authors implemented a permutation procedure that only requires the average and variance of the phenotype found in literature. This method can detect advantageous variants only if the phenotype is hypothesized to be the result of an adaptive process.

H. Allen Orr suggested a sign test (Orr 1998), to determine whether the observed number of plus (or minus) alleles at Quantitative Trait Loci (QTLs) is different in two groups of individuals with different phenotypes, instead of being similar as expected under genetic drift. Orr's sign test has recently been used for expression QTLs (eQTLs) where polygenic adaptation can be indicated by the accumulation of many eQTL even if each eQTL has low effect on the phenotype (Fraser et al. 2011). Similarly, an alternative is to use a set of SNPs associated with a given phenotype, e.g. height in European populations (Turchin et al. 2012), and show systematic allele frequency differences between populations with different phenotypic values that better fit a model of adaptive evolution than genetic drift. Finally, Berg and Coop (2014) have implemented a test using the mean additive genetic value, Q_X, estimated from the additive effect size of loci associated with a given phenotype (GWAS loci). The test is an extension of the BayENV method and determines whether the genetic value (instead of the allele frequency) covaries with a given environmental variable. They further developed a generalization of the Q_{ST}/F_{ST} comparison (Leinonen et al. 2013). The Q_{ST}/F_{ST} test of neutrality contrasts whether there is an excess of quantitative trait differentiation (as measured by the Q_{ST} index) to the genetic differentiation among populations (as measured in a large set of loci by the F_{ST} index), to identity traits that have evolved adaptively. In their implementation Berg and Coop (2014) use the estimated Q_X instead of Q_{ST}.

The theoretical development to identify variants with a small effect on fitness but the basis of phenotypic adaptation through polygenic adaptation is progressing. However, most of the methods rely on GWAS loci, and as a consequence, are still limited. First they assume that the associated loci act in a strictly additive manner, ignoring the putative dominance or epistasis among them. Second, GWAS loci are unlikely to be the causal ones, but rather tag the true positives; since the LD patterns are variable among populations, the GWAS loci may not be a good proxy of the causal variant in all the studied populations. Third, the genetic values are relatively accurate when calculated in a population where the association studies were performed, but the GWAS loci may not be portable to all genetic backgrounds.

2.2.6 From Putative Advantageous Mutation to Increased Fitness

Most studies attempt to identify advantageous mutations. This goal may be reached if, at least, the four following steps are completed.

1. *Identify candidate adaptive loci*. The main issue is to disentangle whether a strong statistical signal for detecting positive selection is truly due to positive selection or alternative processes aforementioned.
2. *Identify the underlying functional variant*. Strong LD within a genomic region with hitchhiking must be removed in order to pinpoint the variant targeted by positive selection.
3. *Quantify the phenotypic consequences of the candidate adaptive allele* by performing experiments in vivo with model organism (mouse, zebrafish, etc...), in vitro using cell cultures, or genotype–phenotype association studies. An alternative is to use the wealth of functional public databases to retrieve information from the literature.
4. *Clarify the relationship between phenotype and reproductive fitness* in the population and environment where the allele has increased in frequency. This is a complicated task because one must infer the relevant environment which selected the variant in the ancestors of the studied population, and whether the phenotypic change encoded by the functional variant is fitter than the ancestral one.

Few studies present conclusive results from the four steps together. Particularly, the fourth step may result in story-telling and it is impossible to formally test such relationship in humans. Therefore, it is important not to dismiss the possibility that a locus is adaptive despite the inability to determine the past selective pressures and to demonstrate that the phenotypic change resulted from an increase in fitness in past populations.

In the future, the recent wealth in *omics* data will most probably allow partially to bridge the gap between genotype and phenotype when studying adaptive evolution. Indeed, thanks to NGS data, functional data has been produced in the past few years in epigenomics, metabolomics, transcriptomics, and interactomics, among others. For example, the Encyclopedia of DNA Elements (ENCODE) project (Dunham et al. 2012) has identified functional elements across the genome, in coding and non-coding regions. In order to identify the underlying functional variant, one may use this emerging functional data, for example, through an integrative genomics approach, along with results from population genetics of positive selection (Barrett and Hoekstra 2011; Scheinfeldt and Tishkoff 2013).

2.3 Current Knowledge on Positive Selection in the Human Genome

The previous sections emphasize the practical challenges in (1) detecting positive selection in the genome, (2) confirming the adaptive loci, and (3) linking the genotype to the phenotype. Although research into human adaptation has many challenges there have been several striking success stories since the beginning of the genomic era one decade ago (Table 2.3). Studies of the impact of positive selection can be divided between candidate gene studies and genome-wide scans.

Table 2.3 Examples of positively selected genes supported by functional evidence. Caution: a unique article is cited while for many genes, several studies were required to conclude about the impact of positive selection and on the function of the putative selective allele

Gene	Selected function(s)	Adapted population	Approach	Reference
ABCC11	Ear wax secretion	Asian	Genome-wide scan	Xue et al. (2009)
CASP12	Sepsis resistance	Worldwide	Candidate gene	Xue et al. (2006)
CCR5	Bubonic plague or smallpox resistance	European	Candidate gene	Sabeti et al. (2005)
CD5	Pathogen recognition	East Asian	Candidate gene	Carnero-Montoro et al. (2012)
DARC	Malaria resistance	African	Candidate gene	Hamblin and Di Rienzo (2000)
EDAR	Hair/teeth/sweat gland development	Asian	Genome-wide scan	Sabeti et al. (2007)
EGLN1	Response to hypoxia	Tibetan and Sherpa	Genome-wide scan	Jeong et al. (2014) and Simonson et al. (2010)
EPAS1	Response to hypoxia	Tibetan and Sherpa	Genome-wide scan	Beall et al. (2010) and Jeong et al. (2014)
G6PD	Malaria resistance	African	Candidate gene	Tishkoff et al. (2001)
HBB	Malaria resistance	African	Candidate gene	Ayodo et al. (2007)
HERC2	Eye pigmentation	European	Candidate gene	Wilde et al. (2014)
LCT	Lactase persistence	European and African	Candidate gene	Bersaglieri et al. (2004) and Tishkoff et al. (2007)
SLC24A5	Skin pigmentation	European	Candidate gene	Lamason et al. (2005)
SLC45A2	Skin pigmentation	European	Genome-wide scan	Sabeti et al. (2007)
TLR5	Bacterial flagellin	African	Genome-wide scan	Grossman et al. (2013)
TNFSF5	Malaria resistance	African	Candidate gene	Sabeti et al. (2002a, b)
ZIP4	Zinc uptake	West Africa	Candidate gene	Engelken et al. (2014)

2.3.1 Candidate Gene Studies of Positive Selection

Candidate gene studies are driven by an a priori hypothesis about the implication of a gene in a putatively adaptive phenotype. Before the recent wealth of genomic data, this approach was most commonly used to detect positive selection. These studies show the impact of positive selection on specific genomic regions, identify candidate adaptive loci, and provide informative insights into the molecular basis of phenotypic adaptation across human populations. For example, several genes have been identified as targets of positive selection with supporting functional evidence for and a link to a phenotypic change conferring a fitness increase (Table 2.3): *G6PD*, *DARC*, *TNFS5*, and *HBB* which provide malaria resistance in Africa (Ayodo et al. 2007; Hamblin and Di Rienzo 2000; Sabeti et al. 2002b; Tishkoff et al. 2007); *LCT* which proffers lactose metabolism in populations with herder ancestors in Europe (Bersaglieri et al. 2004) and Africa (Tishkoff et al. 2007); *CASP12* which increases resistance to sepsis (Xue et al. 2006); and *CD5* which allows better pathogen recognition (Carnero-Montoro et al. 2012).

Although the aforementioned successes in detecting variants that have been selected, the candidate gene approach suffers from the three following main drawbacks:

1. An a priori hypothesis is required about which genes have been under positive selection, as well as knowledge of the relationship between genotype and phenotype. A candidate gene approach aims to pinpoint the functional variant, but the goal is rarely reached. Furthermore, when the function of the adaptive allele is established, it is difficult to determine how it confers a selective advantage to its carriers.
2. The adaptive variant can be located far from the region spanning the gene either within the coding or flanking region. In that case, if no previous knowledge on the gene regulatory regions exist, it would be impossible to detect the adaptive locus within a candidate gene framework.
3. In general, no sufficient biological knowledge on the molecular basis of adaptive phenotypes (or even diseases) across most of the genome can make a good a priori hypothesis of the underlying molecular bases of traits. Thus, a candidate gene approach is reduced to the study of annotated genes encoding relatively simple phenotypes.

For those reasons, and with the recent wealth of polymorphism data, an alternative approach has been developed: the genome-wide scan approach.

2.3.2 Genome-Wide Scans for Positive Selection

During the last decade, impressive technological progress in genotyping has been made, from high-throughput genotyping arrays to NGS, resulting in the bulk of genotype data needed to perform population genetics analyses. Now, large catalogs of genetic variability in worldwide human populations are publicly available allowing the study of the impact of natural selection on our genome. A large number of genome-wide scans of positive selection in different populations have been published recently (reviewed in Akey 2009; Fu and Akey 2013; Scheinfeldt and Tishkoff 2013). A top-down approach, with no a priori hypothesis on the adaptive phenotype, avoids the limitations of candidate-gene studies. The first genome-wide scan for positive selection in human populations was performed by Akey et al. (2002) and was rapidly followed by more than 20 others (Akey 2009). Since 2002, the number of individuals and markers available increases constantly and there has been theoretical development and implementation of several new methods for hard sweeps and alternative modes of positive selection. The boom of data and statistical methods to detect positive selection has revealed many more genomic regions that have putatively evolved in at least one population. In 2009, more than 5000 regions in the genome spanning a total of 400 Mb and encompassing more than 4000 protein-coding genes were reported in a review of 21 genome-wide scans published at that time (Akey 2009). Those 21 scans used methods designed to detect the molecular patterns left by a hard sweep. They also relied on the outlier approach and, therefore, established an a priori proportion of the genome under positive selection in the studied populations, likely leading to a high FPR. In his review (Akey 2009), Joshua Akey looked at the overlap of the genomic regions reported by 10 studies using the same data, but different statistics. Strikingly, only 14.1%, 5.3%, and 2.5% of the overall regions were reported in two, three, or four studies, respectively. Besides the FPR issue, it is clear that those genome-wide scans can also miss real events of selection as suggested by the fact that neither *G6PD* nor *DARC* has been reported by such studies.

Although the overlap among individual scans is low, more than 700 regions have been identified encompassing previous candidate adaptive loci and new well-supported ones (Table 2.3). Moreover, it appears that most signals of putative positive selection are not shared among populations from different geographic regions (for example, see Pickrell et al. 2009; Voight et al. 2006). This is expected when considering that the scans mostly relied on the hard sweep model, and therefore detected advantageous mutations that appear in the population just before being selected for. Indeed, geographically distant populations present different genetic backgrounds and have to adapt to very heterogeneous environmental conditions.

Genome-wide scans can map the signals of putative positive selection and will give great insights into how natural selection has shaped the human genome. They will also continue to aid in the discovery of functional elements. However, it remains challenging to extract the relevant information in the bulk of signals of positive selection from genome-wide scans in order to understand how the human population really evolved and what is at the molecular basis of phenotypic adaptation. Indeed, although the genome-wide approach circumvents some limitations of the candidate gene approach, it presents its own ones.

1. Large scale studies do not allow the extensive control for many layers of complexity. Indeed, in opposition to candidate gene approach, in a genome-wide scan it is extremely difficult to build an accurate model including both demographic and genomic processes that describe the evolution of a specific genomic region or to investigate in depth the molecular mechanisms affecting the genetic variability. Therefore, most scans rely on the outlier approach, and as already mentioned, only detect the most extreme cases of positive selection as well as suffering a likely high FPR (Teshima et al. 2006). As described before, one solution to reduce the FPR is to cross the results from different scans performed with different methods and/or on different populations.
2. Regions reported by genome-wide scans are usually large, spanning hundreds of kilobases and containing several contiguous genes and regulatory regions. Sometimes signals can be located in intergenic regions where no function has been reported yet. Therefore, it is often difficult to follow-up on the signals to identify whether the selected variant and the phenotype putatively increase the fitness.
3. For most genes, a certain amount of speculative discussion (story-telling) is necessary to determine which could be the adaptive phenotype.

For those reasons, most genome-wide scans focus on a reduced set of signals of putative selection based on biological information for a follow-up analysis. This practice is often referred as cherry picking. Hence, most of the signals already reported remain unexplained.

The recent scan performed by Grossman et al. (2013) developed new standards to overcome the aforementioned limitations and represents an important step toward the identification of putative adaptive variants as well as the underlying phenotypes increasing the fitness. This study made progress in several areas: (1) they used CMS which pinpoints more accurately the selected variant (Grossman et al. 2010); (2) they performed their analysis on the 1000 Genomes Project Pilot 1 re-sequencing data (The 1000 Genomes Project Consortium 2010); and (3) they analyzed the putative phenotypic implications of the selective variants by interrogating the ENCODE database and the GWAS catalog (Hindorff et al. 2009).

2.3.3 Insights from Published Studies of Positive Selection in Humans

All the studies aforementioned allowed the identification of putative adaptive loci, but also provide interesting insights in the nature of the genomic regions that have been preferentially targeted by positive selection in human populations. Allowing exploration of the phenotypic differences among populations and species that are induced from adaptation to new environments and which were the underlying biological functions at play.

2.3.3.1 Functional Categories for the Selected Protein-Coding Genes

A functional enrichment analysis is almost always performed after a genome-wide scan for positive selection. Such analysis tests whether the set of variants located within the regions of a signal for positive selection enrichment is in a biological process or functional pathway by contrasting whether more variants belong to a given functional class or pathway than expected by chance. To perform a functional enrichment analysis, these following databases are available:

1. *Gene Ontology* (GO; The Gene Ontology Consortium 2000) groups genes according to the features of the gene product. There are three main domains: (1) cellular component, i.e. the parts of the cell or its extracellular environment where the gene product is active; (2) molecular function, i.e. the elemental activities of the gene product at the molecular level (e.g. binding, catalysis, etc...); and (3) biological process, i.e. operations and sets of molecular events with a defined beginning and end that is pertinent to the functioning of integrated living units.
2. *PANTHER* (Protein Analysis Through Evolutionary Relationships; Mi et al. 2013) relies on annotation from GO among others and classifies proteins (and the encoding genes) as one of the following: (1) family, i.e. groups of evolutionarily related proteins and subfamilies (related proteins that also have the same function); (2) molecular function of the protein by itself or with directly interacting proteins at a biochemical level; (3) biological process, i.e. the function of the protein in the context of a larger network of proteins that interact to accomplish a process at the level of the cell or organism (e.g., mitosis); or (4) pathway which explicitly specifies the relationships between the interacting molecules.
3. *KEGG* (Kyoto Encyclopedia of Genes and Genomes; Kanehisa and Goto 2000) is a collection of manually curated databases integrating genomes, biological pathways, diseases, drugs, and chemical substances.
4. *Reactome Pathway Database* (Croft et al. 2011) contains curated functional pathway annotations that cover a diverse set of topics in molecular and cellular biology.

Genome-wide scans of positive selection using polymorphism data in human populations pointed to different categories enriched for genes that have evolved under a selective scenario: skin pigmentation, immunity, hair density, sweat gland, etc. (Kelley et al. 2006). Scans based on comparative genomics have revealed categories such as immunity and pathogen defense or sensory perception (Kosiol et al. 2008; Marques-Bonet et al. 2009).

However, functional enrichment analyses using such databases are biased toward protein-coding genes. In addition, an assumption of these databases is that all genes are independent and that all genes have the same level of importance within a pathway or a functional category. Although functional enrichment analysis has shed light on important functions and pathways that are preferentially targeted by positive selection, it does not provide a formal test for selection on a function. The current approach commonly used for large genome-wide analysis of positive selection is to detect signals at individual genes or regions. However, selected loci are just at the molecular basis of positive selection acting on the phenotypic level. Thus, single mutations rarely act in isolation to improve a function or to contribute to the acquisition of new ones. To overcome those limitations, Serra et al. created a new method called the Gene Set Selection Analysis (GSSA) to detect significant differences in scores of natural selection over functionally related genes (Serra et al. 2011). The method was applied genome-wide to coding regions of five mammals. But it still has never been used to interrogate non-coding elements or for polymorphism data.

2.3.3.2 Complex Adaptive Traits

The studies listed above describe the first attempts to move from individual genes to the biological modules they belong to. These studies start from individual genes or loci to then integrate the information on functional systems. The idea is that many loci will be involved in phenotypic adaptation, excluding Mendelian traits. This implies that polygenic adaptation is likely to be the main adaptive force acting on the human genome. First, Daub et al. used a gene-set enrichment test based on the F_{ST} statistic (*SUMSTAT*) to test for functional pathways or gene sets enriched in differentiated loci among populations (Daub et al. 2013). They found pathway enrichment in immune response confirming the general idea that response to pathogens has been a major selective force for human populations (for reviews, see Barreiro and Quintana-Murci 2010; Quintana-murci and Clark 2013). They also observed evidence of epistatic interactions between members of the same pathway. Specifically, a genome-wide scan detected several signals of selection for genes involved in the hypoxia-inducible factor 1 (HIF1) pathway which is involved in physiological response to hypoxic conditions (Simonson et al. 2010).

In order to examine polygenic adaptation and soft sweeps, several studies used methods better suited to study small shifts in allele frequency (Fumagalli et al. 2011; Hancock et al. 2010, 2011). When looking at covariation of diet, subsistence, or ecoregion, Hancock et al. found that pathways involved in starch and sucrose metabolism are enriched with signals of polygenic adaptation to a diet rich in roots and tubers, as well as an over-representation of signals associated to polar climate in genes involved in energy metabolism pathways (Hancock et al. 2010). Applying the same method with other environmental variables, they also described an enrichment of signals in gene sets related to UV radiation, infection, and immunity (Hancock et al. 2011). Conversely, Fumagalli et al. (2011), using a similar method, showed that local adaptation has been driven by the diversity of the local pathogenic environment while climate played a relatively minor role.

Berg and Coop, using the mean additive genetic value Q_X described several complex traits likely to have evolved through the action of polygenic adaptation (Berg and Coop 2014): height, pigmentation, and body mass index.

2.3.3.3 The Importance of Regulatory Elements

Although the method proposed by Berg and Coop (2014) is limited by relying on GWAS loci and the problem of portability among populations, it represents a major shift in the field. It is becoming clear that focusing only on protein-coding elements is not enough to understand adaptive evolution in humans. Although protein-coding sequences are very well annotated, they only represent around 1.2% of the human genome. Furthermore, the similarity between humans and chimpanzees in their protein-coding gene sequences cannot explain the observed phenotypic differences. In 1975, King and Wilson (King and Wilson 1975) suggested that differences in gene regulation may largely account for those phenotypic differences among species and populations. Since 1975, the relative contribution of variants located within protein-coding genes and regulatory regions has been debated. Evidence of the functionality of non protein-coding regions is the amount of conservation among species across the genome. For instance, 5% of the genome has been estimated to be largely conserved since the MRCA of mouse and human through the action of purifying selection. Hence, the conserved proportion of the genome is likely to be functional (Siepel et al. 2005). Since the proportion of conservation is higher than the proportion of protein-coding sequences in the genome, a large fraction of the elements with relevant biological function is non-coding.

Until recently, technical limitations have barred the exploration of the adaptive role of non-coding elements. Annotation outside gene regions has been lacking making it difficult to distinguish functional evidence of putative adaptation. This makes comparative genomic studies difficult as they rely on the comparison of the rate of substitution on functional versus non-functional elements and struggle to find an equivalent to the non-synonymous and synonymous changes in these badly annotated regions. In recent years, evidence indicates the role of regulatory elements in adaptive evolution. Using putatively neutral elements as a reference, Haygood et al. found that variants located in promoter regions had signatures of positive selection in the human and chimpanzee lineages (Haygood et al. 2007). Strikingly, they found an enrichment of signals of selection in nervous-system functions. Recent population genetics studies also indicate similar findings. First, Kadaravalli et al. using a genome-wide set of eQTLs and the statistic iHS they found that SNPs showing signals of selection are more likely than random to be associated with gene expression levels in *cis* (Kudaravalli et al. 2009). Second, with a similar study design but taking advantage of the recent wealth in eQTL databases and the ENCODE project, Hunter B. Fraser uses BayENV scores for polygenic adaptation to perform the first genome-scale study on the hypothesis that changes in gene expression have driven human adaptation (Fraser 2013). Third, Enard et al. observed a greater correlation in the observed signatures of positive selection (as inferred by iHS, XPEHH, and CLR) with the presence of regulatory sequences from ENCODE than with amino acid substitutions (Enard et al. 2014). Fourth, Arbiza et al. (2014) found a substantial amount of adaptive changes during human evolution affecting transcription binding sites.

All those studies suggest the functional importance of regulatory regions, their implication in adaptive evolution, and thus a substantial proportion of adaptive changes responsible for biological diversity are both inter and intra specific in regulatory regions. The aforementioned observation that hard sweeps were rare during human evolution (Hernandez et al. 2011) was based on a study design focusing on protein-coding regions. Thus, although the relative scarcity in hard sweeps pointed in this study is usually mentioned as a genomic trend and used against the hard sweep model, we think that generalizing those results to the whole genome is groundless.

2.4 Concluding Remarks

Identifying the molecular basis of phenotypic adaptation is a major challenge in evolutionary biology. The insights from population genetics are paramount to understanding human evolution through adaptive changes. However, most remain to be discovered. An exhaustive detection of selected variants will only be possible with tests for positive selection and in particular beyond the hard sweep model. We have discussed in this chapter that other scenarios further than positive selection must be considered. This is particularly true for the genes from the immune system which demonstrates that balancing selection has been impacting genome variability. Moreover, the examples of recently discovered regulatory adaptations and their importance in human adaptive evolution strongly suggest that only considering variants located within protein-coding regions is outdated. As genome annotation is getting more precise every day we are able to discover more targets of natural selection in non-coding regions. Moreover, while in this chapter we mostly focused on studies of point mutation (i.e., SNPs), other kinds of mutations segregate in the genome in large parts (i.e., structural variants) and have been overlooked by population geneticists.

The identification of variants encoding phenotypic selective changes relies on the downstream implementation of accurate models of neutral evolution that account for the complex human demography which have affected the genetic variability within and between populations. Those models must also integrate genomic mechanisms influencing the molecular patterns across the genome (e.g., mutation, recombination, and gene conversion).

Furthermore, the recent advances in many biological areas with the advent of the *omics* (e.g., transcriptomics, metabolomics, epigenomics, proteomics, and genomics) promise future groundbreaking discoveries. Although the rate at which data is currently generated may seem overwhelming, it allows many layers of complexity to come together which reduces the gap between the genotype and the phenotype. Therefore, population genetics must now work within a multidisciplinary framework in order to achieve its final goal of understanding the fitness consequences of selective variants.

Acknowledgements This work is a part of a research program developed in Jaume Bertranpetit's group with funding by grant PID2019-110933GB-I00/AEI/10.13039/501100011033 awarded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades (MCIU, Spain) and with the support of Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 702). Part of the "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M). We are very grateful to Jessica Nye for her comprehensive proofreading of the manuscript. We want also to thank Elena Carnero-Montero, Brandon Invergo, and Ludovica Montanucci for sharing valuable material.

References

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19(5):711–722

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814

Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, Siepel A (2014) Genome-wide inference of natural selection on human transcription factor binding sites. Nat Genet 45 (7):723–729

Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, Orago ASS, Reich D (2007) Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am J Hum Genet 81(2):234–242

Barreiro LB, Quintana-Murci L (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11(1):17–30

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 12(11):767–780

Barton N (1998) The geometry of adaptation. Nature 395(6704):751–752

Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, Knight J, Li C (2010) Natural selection on EPAS1 (HIF2 α) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci U S A 107(25):11459–11464

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. PLoS Genet 10(8): e1004412. https://doi.org/10.1371/journal.pgen.1004412

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111–1120

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140(2):783–796

Cann H, de Toma C, Cazes L, Legrand M, Morel V, Piouffre L, Cavalli-Sforza L (2002) A human genome diversity cell line panel. Science 12(296(5566)):261–262

Carnero-Montoro E, Bonet L, Engelken J, Bielig T, Martínez-Florensa M, Lozano F, Bosch E (2012) Evolutionary and functional evidence for positive selection at the human CD5 immune receptor gene. Mol Biol Evol 29(2):811–823

Casals F, Bertranpetit J (2012) Human genetic variation, shared and private. Science (New York, NY) 337(6090):39–40

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134(4):1289–1303

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20(3):393–402

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, Kingsley DM (2005) Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science (New York, NY) 307(5717):1928–1933. https://doi.org/10.1126/science.1107239

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg N a, Pritchard JK (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38(11):1251–1260

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. Genetics 1423(August):1411–1423

Croft D, Kelly GO, Wu G, Haw R, Gillespie M, Matthews L, Stein L (2011) Reactome : a database of reactions, pathways and biological processes. Nucleic Acids Res 39(Database issue):691–697

Darwin CR (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London

Darwin CR, Wallace AR (1858) On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. J Proc Linnean Soc London 3(9):46–50

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-murci L, Robinson-rechavi M, Excoffier L (2013) Evidence for polygenic adaptation to pathogens in the human genome article fast track. Mol Biol Evol 30(7):1544–1558. https://doi.org/10.1093/molbev/mst080

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Lochovsky L (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74

Enard D, Messer PW, Petrov D a (2014) Genome-wide signals of positive selection in human evolution. Genome Res 24(6):885–895

Engelken J, Carnero-Montoro E, Pybus M, Andrews GK, Lalueza-Fox C, Comas D, Bosch E (2014) Extreme population differences in the human zinc transporter ZIP4 (SLC39A4) are explained by positive selection in sub-Saharan Africa. PLoS Genet 10(2):e1004128

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26(16):2064–2065

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. PLoS Genet 9(10):e1003905

Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G (2014) Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol Biol Evol 31(7):1850–1868

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155(3):1405–1413

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol 31(5):1275–1291

Fraser HB (2013) Gene expression drives local adaptation in humans. Genome Res 23 (7):1089–1096

Fraser HB, Babak T, Tsang J, Zhou Y, Zhang B, Mehrabian M, Schadt EE (2011) Systematic detection of polygenic cis-regulatory evolution. PLoS Genet 7(3)

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147(2):915–925

Fu W, Akey JM (2013) Selection and adaptation in the human genome. Annu Rev Genomics Hum Genet 14:467–489

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133(3):693–709

Fu W, O'Connor TD, Akey JM (2013) Genetic architecture of quantitative traits and complex diseases. Curr Opin Genet Dev 23(6):678–683

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Ferrer-Admettla A, Pattini L, Nielsen R (2011) Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet 7(11):e1002355

Garud NR, Messer PW, Buzbas EO, Petrov DA (2014) Recent selective sweeps in Drosophila were abundant and primarily soft. *arXiv*

Gillespie JH (2000) Genetic drift in an infinite population : the Pseudohitchhiking model. Genetics 155:909–919

González-Neira A, Ke X, Lao O, Calafell F, Navarro A, Comas D, Cann H, Bumpstead S, Ghori J, Hunt S, Deloukas P, Dunham I, Cardon LR, Bertranpetit J (2006) The portability of tagSNPs across populations: a worldwide survey. Genome Res 16(3):323–330

Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW (2012) Limited evidence for classic selective sweeps in African populations. Genetics 192(3):1049–1064

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Bustamante CD (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108(29):11983–11988

Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Sabeti PC (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327(5967):883–886

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Sabeti PC (2013) Identifying recent adaptations in large-scale genomic data. Cell 152(4):703–713

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. Genetics 195(1):205–220

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5(10):e1000695

Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. Theor Popul Biol 73 (1):158–170

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66(5):1669–1679

Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A (2010) Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philos Trans R Soc Lond Ser B Biol Sci 365(1552):2459–2468

Hancock AM, Witonsky DB, Alkorta-aranburu G, Beall CM, Sukernik R, Utermann G, Di Rienzo A (2011) Adaptations to climate-mediated selective pressures in humans. PLoS Genet 7(4): e1001375

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet 39(9):1140–1144

Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169(4):2335–2352

Hernandez RD (2008) A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24(23):2786–2787

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Przeworski M (2011) Classic selective sweeps were rare in recent human evolution. Science 331:920–924

Hindorff LA, Sethupathy P, Junkins H a, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362–9367

Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet 73(1):95–108

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. Nat Rev Genet 10(9):639–650

Hudson RR (1991) Oxford surveys in evolutionary biology. Oxford University Press, Oxford, pp 1–44

Hudson RR (2002) Bioinformatics applications note. Bioinformatics 18(2):337–338

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci U S A 101(29):10667–10672

Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB (2008) The evolution of gene regulation underlies a morphological difference between two drosophila sister species. Cell 132 (5):783–793

Jeong C, Alkorta-Aranburu G, Basnyat B, Neupane M, Witonsky DB, Pritchard JK, Di Rienzo A (2014) Admixture facilitates genetic adaptations to high altitude in Tibet. Nat Commun 5:3281

Kanehisa M, Goto S (2000) KEGG : Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

Kaplan NL, Hudsont RR, Langle CH (1989) The “hitchhiking effect” revisited. Genetics 899:887–899

Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science (New York, NY) 336(6082):740–743

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16(8):980–989

Kelly JK (1997) A test of neutrality based on Interlocu associations. Genetics 1206:1197–1206

Kim Y (2006) Allele frequency distribution under recurrent selective sweeps. Genetics 1978:1967–1978

King M, Wilson AC (1975) Humans and Chimpanze es. Science 188(4184):107–116

Kingman JFC (1982) The coalescent. Stoch Process Appl 13:235–248

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genet 4(8):e1000144

Kudaravalli S, Veyrieras J-B, Stranger BE, Dermitzakis ET, Pritchard JK (2009) Gene expression levels are a target of recent natural selection in the human genome. Mol Biol Evol 26 (3):649–658

Lamason RL, Mohideen M-APK, Mest JR, Wong AC, Norton HL, Aros MC, Cheng KC (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science (New York, NY) 310(5755):1782–1786

Leinonen T, McCairns RJS, O’Hara RB, Merilä J (2013) Q(ST)-F(ST) comparisons: evolutionary and ecological insights from genomic heterogeneity. Nat Rev Genet 14(3):179–190

Lin K, Li H, Schlötterer C, Futschik A (2011) Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics 187(1):229–244

Luisi P, Alvarez-Ponce D, Pybus M, Fares MA, Bertranpetit J, Laayouni H (2015) Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. Genome Biol Evol 7(4):1141–1154. https://doi.org/10.1093/gbe/evv055

Mailund T, Schierup MH, Pedersen CNS, Mechlenborg PJM, Madsen JN, Schauser L (2005) CoaSim: a flexible environment for simulating genetic data under coalescent models. BMC Bioinformatics 6:252

Marjoram P, Wall JD (2006) Fast “coalescent” simulation. BMC Genet 7:16

Marques-Bonet T, Ryder O a, Eichler EE (2009) Sequencing primate genomes: what have we learned? Annu Rev Genomics Hum Genet 10:355–386

Maynard-Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23 (1):23–35

McEvedy C (1988) The bubonic plague. Sci Am 258(2):117–123

Mendizabal I, Marigorta UM, Lao O, Comas D (2012) Adaptive evolution of loci covarying with the human African pygmy phenotype. Hum Genet 131(8):1305–1317

Mi H, Muruganujan A, Thomas PD (2013) PANTHER in 2013 : modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41 (Database issue):377–386

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York, NY

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A 76(10):5269–5273

Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168:2373–2382

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15:1566–1575

Orr HA (1998) Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. Genetics 149:2099–2104

Orr HA, Betancourt AJ (2001) Haldane ' s sieve and adaptation from the standing genetic variation. Genetics 157:875–884

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Reich D (2012) Ancient admixture in human history. Genetics 192(3):1065–1093

Pennings PS, Hermisson J (2006) Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol 23(5):1076–1084

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Pritchard JK (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19 (5):826–837

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, Pakendorf B (2012) The genetic prehistory of southern Africa. Nat Commun 3:1143

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20(4):R208–R215

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 1189:1179–1189

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. Evol Int J Organic Evol 59(11):2312–2323

Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, Engelken J (2014) 1000 genomes selection browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res 42(Database issue):D903–D909

Quintana-murci L, Clark AG (2013) Population genetic tools for dissecting innate immunity in humans. Nat Rev Immunol 13(4):280–293

Ralph PL, Coop G (2010) Parallel adaptation: one or many waves of advance of an advantageous allele? Genetics 668(October):647–668

Ramos-Onsins SE, Rozas J (2002) Statistical properties of new neutrality tests against population growth. Mol Biol Evol 19(12):2092–2100

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411(6834):199–204

Rozas J, Gullaud M, Blandin G, Aguadé M (2001) DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. Genetics 158(3):1147–1155

Sabeti P, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Lander ES (2002a) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Sabeti P, Usen S, Farhadian S, Jallow M, Doherty T, Newport M, Pinder M, Ward R, Kwiatkowski D (2002b) CD40L association with protection from severe malaria. Genes & Immun 3(5):286–291

Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Lander ES (2005) The case for selection at CCR5-Delta32. PLoS Biol 3(11):e378

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Stewart J (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913–918

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11):1576–1183

Schapire RE (1990) The strength of weak learnability. Mach Learn 5:197–227

Scheinfeldt LB, Tishkoff SA (2013) Recent human adaptation: genomic approaches, interpretation and insights. Nat Rev Genetics 14(10):692–702

Scheinfeldt LB, Biswas S, Madeoy J, Connelly CF, Schadt EE, Akey JM (2009) Population genomic analysis of ALMS1 in humans reveals a surprisingly complex evolutionary history. Mol Biol Evol 26(6):1357–1367

Serra F, Arbiza L, Dopazo J, Dopazo H (2011) Natural selection on functional modules, a genome-wide analysis. PLoS Comput Biol 7(3):e10001093

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15(8):1034–1050

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Ge R (2010) Genetic evidence for high-altitude adaptation in Tibet. Science 329(5987):72–75

Spencer CCA, Coop G (2004) SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics 20(18):3673–3675

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16(6):702–712

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65

The Gene Ontology Consortium (2000) Gene ontology : tool for the. Nat Genet 25:25–29

Thornton KR, Jensen JD (2007) Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175(2):737–750

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293(5529):455–462

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Deloukas P (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39 (1):31–40

Turchin M, Chiang CWK, Palmer CD, Sankararaman SRD, Hirschhorn JN, GIANT consortium (2012) Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet 44(9):1015–1019

Uricchio LH, Hernandez RD (2014) Robust forward simulations of recurrent hitchhiking. Genetics 197:221–236, 1–33

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4(3):e72

Wakeley J (2008) Coalescent theory: an introduction. Roberts & Company Publishers, Greenwood Village

Wall JD (1999) Recombination and the power of statistical tests of neutrality. Genet Res 74:65–79

Wall JD (2000) A comparison of estimators of the population recombination rate. Mol Biol Evol 17 (1):156–163

Watterson GA (1978) The homozygosity test of neutrality. Genetics 88(2):405–417

Weir BS, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Burger J (2014) Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A 111(13):4832–4837

Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, Kayser M (2010) Demographic history of Oceania inferred from genome-wide data. Curr Biol 20(22):1983–1992

Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, Tyler-smith C (2006) Spread of an inactive form of Caspase-12 in humans is due to recent positive selection. Am J Hum Genet 78:659–670

Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, Macarthur DG, Tyler-Smith C (2009) Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics 183(3):1065–1077

Zeng K, Fu Y, Shi S, Wu C (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174:1431–1439

Zeng K, Shi S, Wu C-I (2007) Compound tests for the detection of hitchhiking under positive selection. Mol Biol Evol 24(8):1898–1908