1 Introduction

1.1 The evolution and population history of modern humans

1.1.1 *Homo sapiens* and their close relatives

Homo sapiens, i.e. modern humans, is a unique species on the planet. We are the most populous and widespread, compared to other species with comparable body size, yet we have an exceptionally low genetic diversity among populations and are therefore a single species, while other comparable widespread species usually have sub-species in different geographical locations. An understanding of our evolutionary history can help us understand how this situation arose.

We have close relatives among living species that share a lot of common features, either morphologically or genetically. We are one member of the apes (Hominoidea) superfamily. Within this, there are two families: lesser apes, or Hylobatidae (gibbons), and great apes, or Hominidae, which are further divided into two subfamilies: Pongidae (orangutans), and Homininae (chimpanzees, bonobos, gorillas, and humans) (Figure 1.1). Apes share features such as higher level of dexterity of their upper limbs providing a wider range of movement, and no tail, compared to monkeys. Great apes are commonly believed to be the closest living relatives to humans, though which great ape is the closest to us was for a long time contentious. Morphological data were not enough to clearly establish the relationships between humans and other great apes, as we share some derived morphological features in an inconsistent way, from which the evolutionary relationship cannot be inferred. For example, modern humans have the thickest tooth enamel among great apes, and gorillas the thinnest, while the tooth enamel thickness of chimpanzees and orangutans lies in the middle[1]. The morphology of wrist and hand among great apes, however, are far more complex, which resulted in many years of debate on whether human bipedalism evolved from a knuckle-walking ancestor or from an arboreal ape ancestor[2,3].

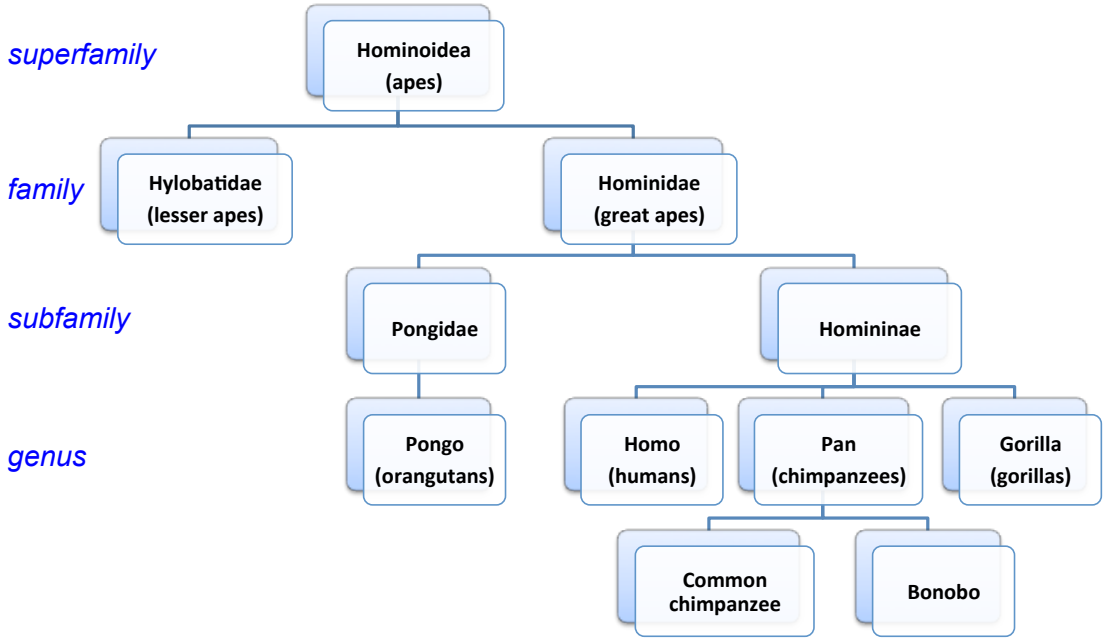

Figure 1.1 The species tree of apes. Please note that this tree is not a complete tree and some branches of the species tree of apes are not shown. Emphasis is on great apes and only extant genera are shown.

Genetic approaches allowed us to investigate evolutionary relationships between humans and great apes in much greater detail. Before being able to examine genetic materials at the molecular level, karyotypes, i.e. the structural characteristics of chromosomes revealed by staining and observation under the microscope, showed similarities as well as obvious differences between the chromosomes of humans and other great apes. Humans only have 46 chromosomes, while chimpanzees, gorillas and orangutans have 48. Despite the difference in number of chromosomes, the G-banding patterns are very similar among the four species[4]. The difference in chromosome number results from an end-to-end fusion of two small great ape chromosomes, which form the large metacentric chromosome 2 in humans. Alignments of G-banded chromosomes suggested the chimpanzee as the closest relative to humans, with chimpanzee and human being a sister-group to gorilla, and chimpanzee–human–gorilla a sister-group to the orangutan.

The investigation of genetic information at the molecular level has proven to be the most powerful tool to unveil the evolutionary relationships between the apes, as well as to estimate the time scale of their speciation. In 1967, Sarich and Wilson presented the first use of molecular methods to estimate a date for the

great ape-human split[5], where they measured the structural differences of serum albumins between old world monkeys, great apes and humans, using an immunological method called microcomplement fixation. Although this work estimated a date of great ape-human split as 5 million years ago (MYA), which contrasted with much older estimates from fossils, it was subsequently supported by similar results from other molecular methods. But perhaps due to the limitation of examining only a single locus, they were not able to resolve the gorilla-chimpanzee-human split. Another molecular approach used was DNA–DNA hybridization[6], which compares the entire single-copy components of two genomes, avoiding the biases of single-locus comparison. However, this method is only effective in comparing species that have diverged for more than 10 million years, so for closely related species, like gorillas, chimpanzees and humans, the small differences can be masked by random experimental errors and the conclusions were much debated.

DNA sequencing brought our understanding of the evolutionary relationships between humans and other great apes to a new era. By comparing the sequences of the same locus from two or more species, gene trees can be constructed, which should accurately show the evolutionary relationships among species for that particular locus. However, gene trees do not necessarily have the same topology as the species tree. There are different factors that contribute to the shapes of gene trees. For example, coding regions in the genome usually have more selective constraints; for instance, positive selection drives the frequency of advantageous haplotypes up rapidly in a particular population or an entire species, which may affect the shape of the gene tree on this locus. So, in the presence of differing selective pressures, the topology of the gene tree may not reflect the relationships between the species. Some other loci in the genome, for example within Human Leukocyte Antigen (HLA), have undergone balancing selection, with the result that a certain proportion of very ancient alleles is maintained in the genome. This results in the HLA loci in some humans being more related to chimpanzees than to other humans, or more closely related to gorillas than to chimpanzees, which again does not reflect the species phylogeny. In addition, incomplete lineage sorting in the ancestral species leads to random

differences in topology. As the founding populations of the species were only subsets of the ancestral population, and thus might not have all its genetic diversity, some alleles might not be transmitted to the next species. This would result in the topology of the phylogenetic trees of some loci differing from the species phylogeny. Therefore, in order to construct a species tree based on genome sequences, multiple neutral, single-copy loci across the genome need to be examined, and a predominant topology identified, which will most likely be the same as the species phylogeny[7]. Gene trees from haploid mitochondrial and Y-chromosomal sequences generally better reflect the species phylogenies, due to their single sex inheritance and the lack of recombination, which result in a smaller effective population size (N_e) and shorter coalescence times.

The draft reference sequences of chimpanzee[8] and gorilla[9] provided great insights into the evolution of these two closest relatives to humans. 70% of the loci showed human-chimpanzee as a clade, while the other 30% showed that gorilla is closest to either the human or chimpanzee genome[9]. These studies also concluded that, making reasonable assumptions about the mutation rate, chimpanzees, as the closest living relative to modern humans, split from the common ancestor of the two species about 6-7 MYA, while the human–chimpanzee–gorilla speciation happened about 10 MYA. However, these genome sequences also revealed the complexities of the genetic similarities and differences among these species, demonstrated by various chromosomal rearrangements, deletions and insertions, gene losses and gains, and so on. Apart from the whole genome sequences, several research groups have also analyzed particular genetic loci in multiple great apes, aiming to understand the divergence and diversity of these species at a deeper level, including a better understanding of the subspecies within the great apes. One example of these studies is the genomic sequence analysis on multiple loci from 20 bonobos and 58 chimpanzees[10], which revealed the close evolutionary relationship between bonobos and chimpanzees, with bonobos lying within chimpanzee variation.

Although we are the only extant *Homo* species on the planet, there were other archaic hominin groups existing until tens of thousands of years ago, which are believed to be sister groups of modern humans. Evidence of these archaic

hominin groups was first provided by fossil records. Neandertals, the fossils of which have been discovered in Europe and western Asia, lived in those areas from at least 230 thousand years ago (KYA), before *Homo sapiens* arrived in Europe and Asia from Africa, and disappeared about 30 KYA[11]. In southern Siberia, a distal manual phalanx of a juvenile hominin was found in 2008 at the Denisova Cave[12], and later DNA analysis suggested that this hominin must be a distinct species from Neandertals or humans. The mitochondrial DNA (mtDNA) of Neandertals was the first DNA to be extracted from the fossils and sequenced[13-15]. These studies showed that the mtDNA of Neandertals share a common ancestor with the mtDNA of present-day humans about 500 KYA[15]. Then the mtDNA of the Denisova phalanx was sequenced[16], showing that this Denisovan mtDNA diverged about 1 MYA from the common lineage of modern human and Neanderthal mtDNAs. However, due to the small effective population size of the haploid, maternally inherited mtDNA, events like genetic drift or selection would affect the time to the most recent common ancestor (TMRCA) of mtDNAs dramatically, so this tree would not necessarily represent the species tree. The draft genome sequences of Neandertal and Denisova were recently published by the same group[12,17], providing more robust estimations of the evolutionary time scale. The study of the Neanderthal genome sequence estimated the split time of modern humans and Neanderthal populations as about 270-440 KYA, and also claimed evidence of gene flow from Neandertals to early modern humans in Eurasia ~50 KYA, before the split of the European and Asian human populations, which may have resulted in 1–4% of the genomes of people outside Africa being derived from Neandertals[17]. The analysis on the Denisovan genome sequence suggested that the ancestor of Denisovans and Neandertals diverged from the ancestor of present Africans about 804 KYA, and Denisovans diverged from Neandertals around 640 KYA[12]. Although the Denisova hominin did not make genetic contributions to the Eurasian human group as broadly as Neandertals, there was evidence that they may still have contributed 4-6% to Melanesian genomes, as well as to the ancestors of New Guineans and Bougainville Islanders[12,18]. However, a recent study suggested that using geographic patterns of shared polymorphism is not an effective way to infer archaic admixture; population structure should be taken into account, as it

can generate similar genetic patterns as those caused by interbreeding[19]. Therefore, whether or not ancient modern humans had interbred with Neanderthals and Denisovans is still debated.

1.1.2 Modern human origins and demographic history

As mentioned, the human lineage diverged from the chimpanzee lineages about 6-7 MYA. During the long period of time until anatomically modern human emerged about 200 KYA, there were many ancient hominin groups, some of which are ancestors of modern humans. However, the classification of these fossils and their relationships with *Homo sapiens* are much debated. The boundaries of modern humans and other hominin species are also not clear, based on the fossil records and very limited ancient DNA analyses. The earliest hominin fossils, dating back to as early as 6.8-7.2 MYA, till about 4.2 MYA, are *Sahelanthropus tchadensis, Orrorin* and *Ardipithecus*. There is uncertainty about whether these species should be classified within the human lineage and the relationships between them, as they all have considerable morphological similarities with chimpanzees, e.g. body size, while they also showed signs of hominin characteristics[20], e.g. up-right walking. Most fossils dated after about 4.2 MYA and before the appearance of the *Homo* genus belong to the genus *Australopithecus*. Fossils of various *Australopithecus* species were found in multiple sites in east and southern Africa, dating from around 4 MYA to 1.8 MYA. The most well-known fossil of *Australopithecus* is the partial skeleton "Lucy", dated to 3.2 MYA, as well as the Laetoli footprints[21], dated to 3.5 MYA. These belong to the species *Australopithecus afarensis*. The significance of these findings is the unequivocal illustration of bipedal locomotion, which is an important characteristic of modern humans. Due to the small body sizes, they are called gracile (lightly built) Australopithecines. Robust (heavy built) hominins, notable for their small brains and large jaws and chewing teeth, belong to the genus *Paranthropus.* A few fossils, including the rather complete "Black Skull" from Lake Turkana, were found in several sites in South Africa, dating to around 1-2 MYA. It is still under debate about which species or fossils of *Australopithecus* represent the ancestor of our own *Homo* genus, but *afarensis* and *africanus* are candidates.

Homo erectus is sometimes considered to be the first *Homo* species (although others consider the earlier species *habilis* to belong to this genus). The earliest *erectus* fossils, dated to around 1.8-1.9 MYA, were found in Africa, which indicates the origin of our genus in Africa. The most complete *erectus* fossil that has been found is the Nariokotome Boy[22], dated to about 1.6 MYA. His body size and shape was very similar to modern humans, though his brain size was much smaller. *H. erectus* is also the earliest hominin found outside of Africa. Fossils have been found in Indonesia ("Java man"), China ("Peking man"), and Georgia (Dmanisi), dated back to as early as 1.6-1.8 MYA. Another *Homo* species, *H. floresiensis*, found in Indonesia, was much smaller (about 1 meter tall). It was believed that they were descendants of *H. erectus* living in areas with poorer resources, and thus selected for dwarfism. A later *Homo* species, *H. heidelbergensis*, found in Africa and Europe, have larger brains (~1,200 cc) than *H. erectus* (~900 cc). Fossils of this species were dated to as widely as around 200-800 KYA. Thus it is considered to be a widespread and variable species that emerged after *H. erectus* and gave rise to more recent *Homo* species, including Neandertals and modern humans.

Anatomically modern humans are believed to emerge around 200 KYA in Africa, though it is difficult to define modern human morphology unambiguously, so as to distinguish them from the archaic hominins discussed earlier. The widely accepted criteria for modern human morphological features are focused on the extent of the globular shape of the skull and the degree of retraction of the face. The earliest known modern human fossil is a skull found in Omo-Kibish, Ethiopia, dated to about 195 KYA. Later crania fossils, dated to 154–160 KYA, showed many modern human morphological features, such as large brain size and globular braincase, but retained some archaic features, such as protruding brows. The earliest modern human fossils found outside Africa in Europe, East Asia and Australia are all dated later than 45 KYA, suggesting the much later appearance of *Homo sapiens* in areas outside Africa.

Archaeological evidence, much more common than the fossil remains, provides insights into hominins and modern human behavior. Hominins from as early as 2.5 MYA started to construct and use artifactual stone tools, in contrast to

natural tools, which were also used by apes and earlier hominins. Stone tools, such as symmetrical teardrop-shaped bifaces, flake tools and choppers, dated as early as about 1.76 MYA onwards, are widely found throughout Africa, in Europe, and in parts of Asia except eastern Asia. More sophisticated tools, such as flakes described as side-scrapers and points, appear in the record around 300 KYA. In the Later Stone Age/Upper Paleolithic, blades instead of flakes, as well as tools from other materials such as wood and bone, became more common. Although these tools are often associated with modern humans, there is often no clear correspondence between tool type and species.

Although fossil records and archaeological evidence both suggest the first appearance of modern humans in Africa, the relationship between modern humans and those who expanded out of Africa earlier has been much debated. There were two basic simple models: (1) the multiregional model, which proposes that modern human ancestors lived in multiple regions in the Old World, and the human characteristics arose in parallel or at different times in different parts of the world; and (2) the out-of-Africa model, which proposes that all modern humans are descended from the ones who emerged in Africa and gradually expanded to other parts of the world, while their contemporaries from other continents did not contribute to our ancestry. Of course there are also possibilities of intermediate models, i.e. gene flow between archaic humans in other continents and our ancestors from Africa, and this debate, according to some interpretations, may have partially been resolved by the sequences of the Neandertal and Denisova genomes mentioned earlier, providing quantitative measures of the amount of gene flow from earlier species and confirming a minor contribution.

Fossil records and archaeological evidence of modern humans were sought to provide direct insights into the dating of the appearance of modern humans in different parts of the world and their origins. Modern human fossils are rare and can be difficult to date. However, all fossils found outside Africa are now dated to around or after 40-45 KYA, indicating that modern humans moved to Eurasia by this time, though this conclusion is subject to revision by future discoveries due to the incompleteness of the fossil records obtained so far. In addition, it is still

unclear what routes the out-of-Africa migrations followed. Archaeological evidence is of limited usefulness because, as mentioned before, it can be difficult to distinguish the archaeological remains left by modern humans and archaic hominins, or sometimes even natural objects. Stone tools, bone tools and artificial ornaments that are considered as "art", which is associated with modern human behavior, are identified as representing different cultures in different geographical regions. In Africa, the Middle Stone Age (MSA) refers to archaeological remains dated from about 250 KYA to 40-80 KYA, while the Later Stone Age (LSA) describes subsequent remains until the emergence of agriculture. Outside Africa, the equivalents are termed the Middle Paleolithic (MP) and Upper Paleolithic (UP), respectively. Although the dating of the archaeological deposits is often disputed, various evidence supports the conclusion that the transition from MSA to LSA humans may have begun in southern Africa as early as ~80 KYA, and in east Africa around 50 KYA. Outside Africa, the transition from MP to UP appears to have happened first in West Asia in around 47 KYA, and a few thousand years later in Europe, and subsequently in Siberia. The migration of people to the Americas from Siberia, and to the Pacific islands from the nearby landmasses, were more recent, occurring ~15-20 and ~5 KYA, respectively.

Around 10 KYA, the emergence of agriculture independently in several regions of the world allowed dramatic expansions of human populations, as well as cultural and social revolutions. Unsurprisingly, extensive changes to tool usage occurred along with the agricultural revolution. This period is designated the Neolithic (New Stone Age). Archaeological evidence suggested that farming practices originated independently in multiple regions in the world, and then these practices spread to surrounding areas. Some of the earliest evidence of agriculture was found in the Near East, dating to about 10 KYA, the earliest Neolithic archaeological sites became younger towards the northwest of Europe. The earliest appearance of agriculture in northern and southern China is also dated to around 10 KYA, and is believed to have an independent origin. In Africa, it is widely believed that agriculture spread from the Near East into Egypt between 9.5 and 7 KYA. In Sahara, evidence of cattle herding is dated back to

around 8 KYA, and cereal agriculture was widespread throughout the belt of savanna south of the Sahara by 3.5 KYA. In sub-Saharan Africa, there was a series of population movements from around 3 KYA, known as the Bantu expansion, linked to the spread of Bantu languages from West Africa into much of east, central and southern Africa. Archaeological, linguistic and genetic evidence has been largely consistent in support of it; however, the details of this complex expansion are far from clear (Figure 1.2).

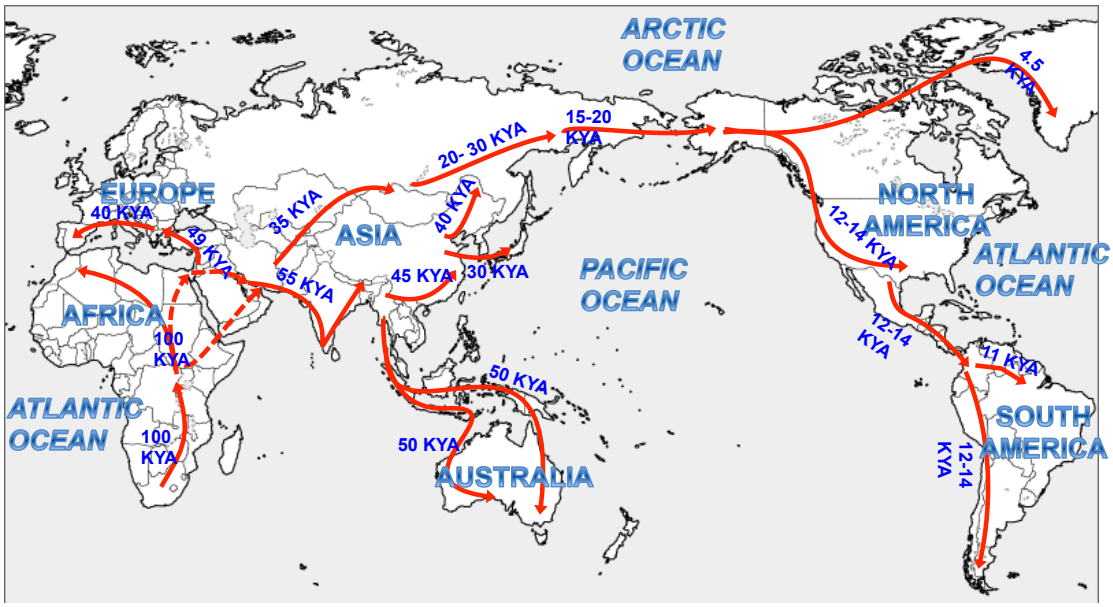

Figure 1.2 Map of human expansions. This map shows the putative migration routes and dates of early modern human migrations from Africa to other parts of the world. Red arrows indicate the possible routes, and estimated dates of migration are shown in blue text (KYA: thousand years ago). Note that the migration routes and dates are still under debate and further investigation, so are subject to updating by new findings.

There are two basic demographic models to explain the expansion of agriculture. One is called acculturation (or cultural diffusion), which proposes a movement of farming technology and ideas, without the migration of early farmers. In contrast, the second model, demic diffusion (or wave of advance), proposes that the farmers moved due to the growth of the population and local migrations. In this model, two scenarios could have occurred: (1) gene flow between the farmers and hunter-gatherers when the former moved to the pre-existing hunter-gatherer populations; or (2) the migrating farmers replaced the gene pool of the indigenous Europeans without interbreeding. While the demic diffusion model

described by Ammerman and Cavalli-Sforza[23] has provided that basis for many subsequent genetic studies, the expansion may be better described by a more complex model.

Genetic approaches have made it possible to test models of human expansions over many timescales. By looking at patterns of genetic diversities and building genetic phylogenies, we can trace back the root of our lineages in different parts of the world. mtDNA and the Y-chromosome were the first to be used to build human phylogenies, because of their simple single-sex inheritance and haploid nature. These studies generally supported the out-of-Africa model, with evidence showing near-complete separation of African and non-African lineages, deepest braches in African, and a star-like structure in out-of-Africa lineages[24,25]. Phylogenetic studies of autosomal loci also largely supported the out-of-Africa model, but due to the complication of recombination in diploid regions, phylogenies of specific loci can be more difficult to reconstruct. Having said that, genome-wide studies of genetic diversity and variation patterns do provide insights into the evolutionary relationships between modern human populations that cannot be obtained from other evidence. If the out-of-Africa theory of human origin is correct, we should expect the highest human genetic diversity in Africa, with populations in other areas containing a subset of African variation, together with their unique variants gained after moving out of Africa. Analyses of the genetic variation of multiple human populations have confirmed that this is largely the case in real genetic data. Furthermore, the advancement of computational modeling approaches plus the availability of large-scale genetic diversity data, yield dramatic increase in power for revealing human population histories.

It is worth noting that human populations have never been completely isolated. Admixture, i.e. the formation of hybrid populations whose genetic pool was derived from two or more ancestral populations, happened at different levels during different stages throughout modern human history, perhaps including with Neandertals and Denisovans as noted earlier. Various historical, linguistic and archaeological records as well as genetic studies have helped understand past admixture events and the degrees of admixture. However, we should note a

number of complexities regarding human admixture. For example, under many admixture scenarios, the contributions of males and females in the ancestral populations may be very different. Therefore, the estimation of the degrees of admixtures from autosomes, X chromosome, Y chromosome or mtDNA can vary. Also, human population admixture, especially those events that happened during the last few thousand years, was greatly affected by different social practices, for example, endogamy. Therefore, studies of recent human demographic events should be considered in the context of societal and economic conditions.

Simplified demographic models have been developed based on population genetic theories and empirical genetic data to mimic modern human population structures and their changes over time. These models seek to best explain the genetic diversity and variation patterns observed in current human populations, and largely support the out-of-Africa model. Two types of demographic models are widely used. One consists of "best-fit" models, which propose a single exit from Africa to Europe and East Asia, followed by subsequent bottlenecks and expansions. These models only include three main continental populations, i.e. African, European and Asian, which are greatly simplified but sufficient for many purposes in global genetic studies. They include parameters such as effective population sizes at different times, migration rates, expansions and bottlenecks. One of the most widely used best-fit models was developed by Schaffner et al.[26], which could generate simulated data that closely resembles empirical genetic data in many characteristics (Figure 1.3). The other type of demographic model consists of "serial founder" models, which propose a subset of an initial population as the founder of a subsequent population, and after expansion, a subset of this second population founds another population[27-30]. This type of models can accommodate more populations than the "best-fit" models, but with fewer parameters being considered. Details of some population models and their use will be considered further below.

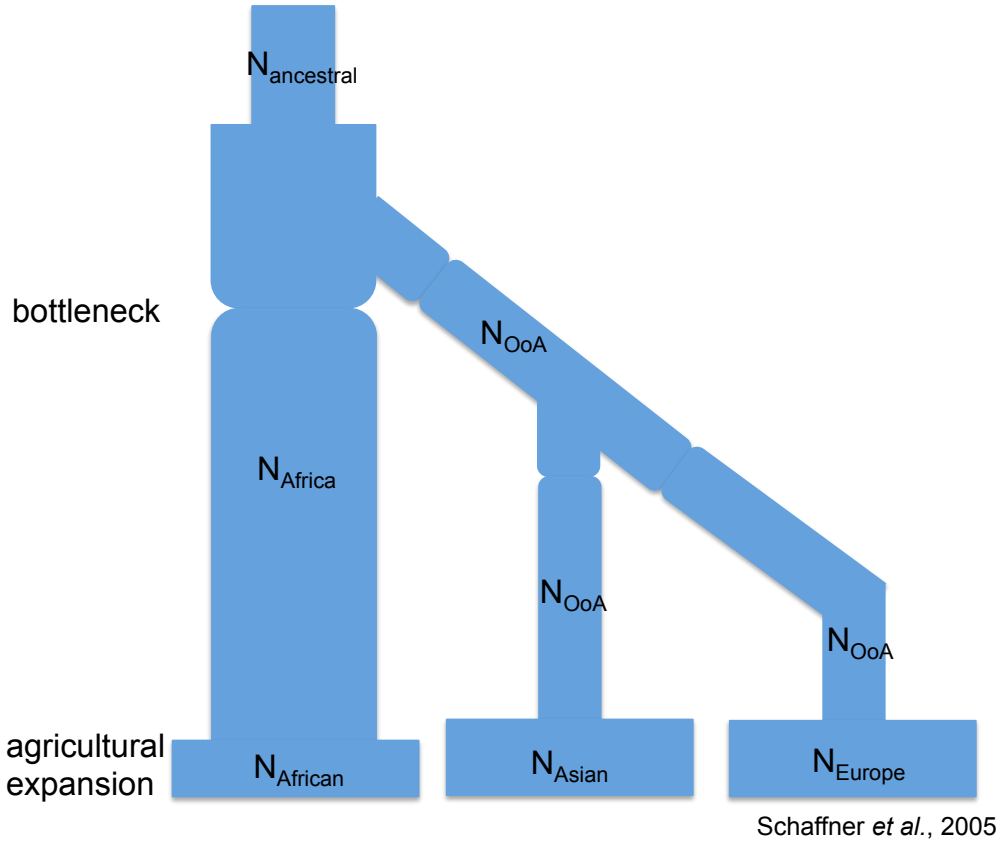

Figure 1.3 A best-fit demographic model. The widths of the bars represent relative population sizes (noted as N in the figure). Bottlenecks are represented by dents in the bars. This figure was adapted from Schaffner et al. 2005.

1.2 Human genome variation

1.2.1 Types of genomic variation

Any two randomly chosen people in the world share about 99.9% of their alignable DNA sequences, which means that there is on average 0.1% sequence difference between two human genomes. These genomic differences make major contributions to the phenotypic variability among people, the genetic basis of which we have not yet fully understood. The sequencing of our DNAs has helped us to understand, at least at the genotype level, how people differ. There are many types of genomic variation in healthy individuals, ranging from single base pair substitutions to rearrangements of tens of megabases. Here we categorize the genomic variation by size into three main types: (1) single base pair substitutions, known as SNPs (single nucleotide polymorphisms) or SNVs (single nucleotide variants); (2) one to hundreds of base pair structural variants (SVs), including small to medium sized insertions and deletions, variation in the

number of microsatellite units (repeats of 2-6 base pairs of DNA) and minisatellites (repeats of 10-100 base pairs of DNA); (3) a few kilobase to a few megabase structural variants, including large insertions and deletions, macrosatellites, inversions, and copy number variants (CNVs). Please note that there is no clear boundary between the last two types of variation; this categorization is only for the purpose of helping the description and understanding of our genomic variation (Figure 1.4).

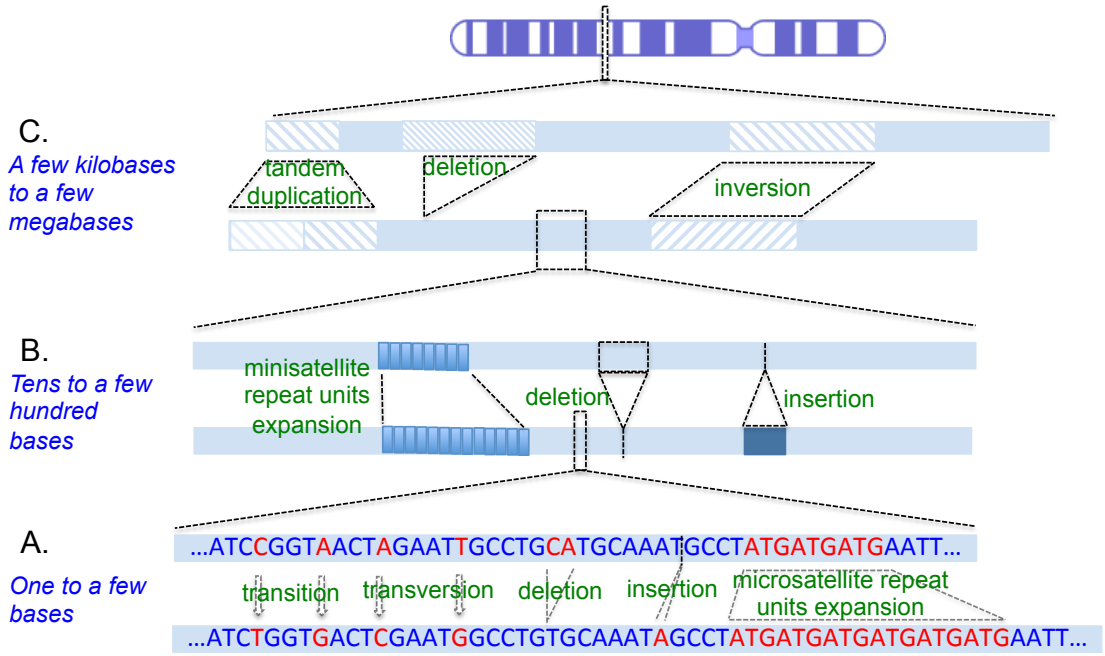

Figure 1.4 Types of genomic variation. A. Examples of transitions, transversions, a single base insertion, two-base deletion, and mutation of repeat unit number of a three-base microsatellite. B. Examples of minisatellite repeat unit number mutation, deletion and insertion of segments of DNA. C. Examples of tandem duplication, large region deletion, and inversion.

Base substitutions, here referred to as SNPs, are the most common and well-studied type of variation in the human genome. There are two types of base substitution: transitions, which are the substitution of a pyrimidine base for another pyrimidine (i.e. C to T or T to C), or a purine for another purine (i.e. A to G or G to A); and transversions, which, in contrast, are when a purine is exchanged for a pyrimidine, or vice versa (e.g. A to T). Transitions are more than twice as frequent as transversions, perhaps because chemically a purine (or a pyrimidine) can be altered to the other purine (or pyrimidine), while it is impossible to alter a purine to resemble a pyrimidine, and vice versa, or the replication and correction enzymes find them more difficult to correct. Base

substitution mutations are caused mainly by two basic processes: (1) the misincorporation of nucleotides during DNA replication, and (2) mutagenesis caused by chemical modifications of bases, or physical damage induced by ultraviolet, ionizing radiation or other harmful physical or chemical exposure. The mutation rate of single nucleotide substitutions has been estimated from several studies. Although the estimates vary when different data or methodologies are used, it is widely accepted that the neutral genome-wide average base substitution rate is in the order of 10^{-8} per base per generation[31-33]. However, it is worth noting that local mutation rates can vary up to an order of magnitude. For example, the CpG dinucleotide is a mutation hotspot, with a mutation rate about ten-fold higher than other sites, and a strong tendency of mutating to TpG or CpA.

Small insertions and deletions (often called "indels") are another common type of variant, though the number per genome is about 10 times less than SNPs. Deletion or insertion of one base pair was sometimes considered as a SNP, but because the mechanisms and frequencies of the single nucleotide indels are more similar to multi-base indels than to single base substitutions, here we categorize them as indels rather than SNPs. Indels often occur in repetitive sequences, the typical forms of which are microsatellites and minisatellites (Figure 1.4). Numbers of copies of micro- or minisatellite repeat units are very variable and have high mutation rates. Such loci are sometimes called variable number tandem repeat loci, or VNTRs. Microsatellite unit numbers can range from a few to tens, and typical mutation rates can be around 10^{-3} to 10^{-4} per locus per generation. Interestingly, although overall mutation rate increases as array length increases, with a small bias towards increases, this is counteracted by the contraction rate becoming higher when the number of repeats is large, which results in very large microsatellites (>50 repeats) being very rare. Minisatellites not only have larger sizes, but also have a larger range of repeat unit copy numbers (from as few as 5 to as many as 1000). They also show a higher level of diversity, so it is rare to find two alleles the same in the population. VNTR mutations are mainly caused by three mechanisms. (1) Replication slippage: this happens when one or more units in the template

strand of the DNA misalign during replication, resulting in the loss of the longer strand (deletion) or the shorter strand (insertion). This is because repetitive sequences can easily mispair during DNA replication. (2) Unequal crossing over events: this also often happens in repetitive sequences, as recombination happens unequally between the two homologous loci, causing deletions or duplications. (3) Gene conversion: this is the nonreciprocal transfer of genetic information, where one allele does not change, whereas the other allele converts to the state of the unchanged allele. It is a result of homologous recombination via the four-stranded intermediate, known as the "Holliday junction". Gene conversion is one of the major mechanisms of mutations in minisatellites.

Larger structural variation in the human genome has been extensively studied recently[34-36]. These studies revealed a remarkable abundance of structural variation. Many of the large structural variants are caused by non-allelic homologous recombination (NAHR); non-homologous end joining (NHEJ) and more complex replication-associated mechanisms such as FoSTeS (fork stalling and template switching) are other major mechanisms. Some inter-chromosomal segmental duplications are caused by retro-transposition[36].

Due to the diploidy of autosomes (and the X chromosome in females), for every heterozygous variant, there is a question of which allele lies on which of the two copies of the chromosome in one individual. A haplotype is the combination of polymorphic alleles that locate on the same DNA molecule, i.e. on the same chromosome. Knowing the haplotypes is often very important in evolutionary studies, as it provides valuable information about ancestry and inheritance. Determining haplotypes experimentally can be very difficult, time-consuming and expensive. Therefore, haplotypes of large genomic data sets are often inferred by computational algorithms, and the widely used ones are based on the Bayesian approach incorporating Markov chain Monte Carlo methods[37]. Apart from mutations, recombination is the main cause of haplotype diversity. Like mutation rates, recombination rates are very variable at different genomic locations. There are recombination hotspots and coldspots along the genome, where recombination rates can be several magnitudes higher or lower than the average, respectively. This creates blocks of genomic sequences where a certain

set of alleles is often linked on the same chromosome, known as linkage or haplotype blocks. Gene conversion also contributes to haplotype diversity by converting part of one haplotype at a locus into the state of the other.

1.2.2 Identification of genomic variation

As the most common and simplest type of variation in the genome, SNPs are the most well-typed and widely used genomic variants in many genetic studies. There have been quite a few widely used methods to discover or type SNPs in genomes, which can be broadly described in three categories: (1) enzyme based methods; (2) hybridization based methods; and (3) sequencing. An early method to detect SNPs was an enzyme-based approach called Restriction Fragment Length Polymorphism (RFLP) analysis. RFLP study uses restriction endonucleases that cut specific restriction sites with high fidelity. By using endonucleases that cut sites containing a SNP of interest to digest the DNA samples amplified by the polymerase chain reaction (PCR) technique and then running a gel electrophoresis assay to determine the lengths of DNA fragments after digestion, samples that were or were not cut at certain sites will be detected, indicating the presence of alternative alleles. Although this method is simple and straightforward, it has great limitations, for example, it requires specific endonucleases, and the specific base of the alternative allele may not be determined from the experiment, and it is very expensive and time-consuming to run multiple electrophoresis assays. Some other enzyme-based methods apply the PCR technique in other ways, some of which are used in several commercialized arrays that can detect multiple SNPs in one assay[38]. Other enzyme-based methods use 5'-nuclease, Flap endonuclease or DNA ligases in the process of SNP detection.

Hybridization-based methods detect SNPs by hybridizing complementary DNA probes to the SNP locus. This type of method is used in the currently most widely used genotyping technology - high-density SNP microarrays, where hundreds of thousands of probes are arrayed on a small chip, enabling large-scale detection of SNPs. Many commercial microarrays designed to detect different sets of SNPs are available in the market and are widely used in various large-scale genetic

studies. The International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) genotyped more than three million SNPs in more than 200 individuals from four populations, using SNP microarrays and related techniques[39], which significantly enriched the database of human SNP variation. One genotyping technology used in this project was the GeneChip® Mapping 500K Array set from Affymetrix Inc. This array set contains about 500,000 human SNPs and can genotype 100 samples per week per instrument. Another company, Illumina Inc., developed a series of SNP arrays that are able to genotype up to 5 million SNPs per sample, with a high level of customizability. These arrays are based on Illumina's BeadArray technology, where SNP-specific oligonucleotides are generated by PCR amplification using fluorescently labeled universal primers, with a particular address sequence complementary to sequences attached to beads just downstream of the SNP, which can be translated to a specific locus. These fluorescent products are subsequently hybridized to beads either on a solid matrix or in solution, depending on the specific platform, and the fluorescence on each bead is then quantified, resulting in a signal of the SNP genotype associated with the particular address sequence.

Most of the methods above can only detect known SNPs. The emergence of DNA sequencing technologies, especially the Next Generation Sequencing (NGS) technologies, brought the discovery of all SNPs in a target region, both known and new, as well as other types of variation to a new era. The sharp drop of the costs and increase of speed in whole genome sequencing have made it possible to sequence whole genomes of multiple individuals. The 1000 Genomes Project (http://www.1000genomes.org/) is aiming to provide a deep catalog of human genomic variation by sequencing whole genomes of 2,500 individuals in 27 populations around the globe. The pilot project, published in 2010, discovered about 15 million SNPs by the whole genome sequencing of 179 individuals from four populations, and limited targeted exon sequencing of 697 individuals from seven populations[40]. It is expected that the main project, consisting of three phases, will reveal far more variants. Phase 1 of the main project, sequencing just over 1,000 individuals and completed in the summer of 2012, has discovered ~40 million variants.

Before sequencing technologies were widely used, detection of tandem repeats (micro- and minisatellites) was mostly done by PCR-based assays. These assays use primers closely flanking the repeat locus, so that one or more differences of the number of repeats could be detected by the variation in length of the PCR products. This has a few limitations. Firstly, some tandem repeat variants have sequence variation within the repeats, which cannot be identified by PCR. Secondly, the resolution of PCR methods is relatively low, so some variants that consist of a large number of small repeats may not be well distinguished. Thirdly, PCR has limitations on the length and base composition of the sequence to be amplified. So some large minisatellites may not be detectable. Some arrays were also developed to detect marker microsatellites that are common and typical, in a relatively large scale.

Structural variation, especially copy number variants (CNVs) were under-investigated until recently, due to the complexity and lack of large-scale assays. Array-Comparative Genomic Hybridization (known as aCGH) allowed large-scale and moderate-resolution detection of CNVs in the genome. In this assay, DNA fragments from samples and a reference genome are labeled by different fluorophores, and then these fragments are mixed and hybridized to thousands of probes on the array chip. After washing off un-hybridized fragments, the intensity of fluorophores from the sample and the reference is measured, and then the ratio of the intensity is calculated to detect the copy number differences between the sample and the reference on the particular locus. Current aCGH assays can achieve a resolution of less than 100 base pairs at breakpoints. A good example of large-scale studies of CNVs using aCGH is the study in 2009 by Conrad et al.[36], providing a comprehensive map of CNVs in the human genome. Various algorithms, for example, CNV-seq[41] and BIC-seq[42], have been developed to detect CNVs from NGS data, aiming to achieve a higher resolution than aCGH. The 1000 Genomes Pilot Project comprehensively mapped CNVs based on 185 whole-genome sequences[43].

Compared to all these assays targeting the detection of different types of genomic variation, genome sequencing has the obvious advantage of detecting all sorts of variation in one go, as well as being able to discover novel variants.

NGS technologies have undoubtedly introduced a new sequencing era, with possibilities of sequencing targeted regions or whole genomes in tens or hundreds of samples rapidly and relatively cheaply. There are several widely used NGS platforms in the marketplace, including Illumina/Solexa, Roche/454, Life Technology's SOLiD, Complete Genomics platforms, and others. The dominant platform during my PhD project was the Illumina/Solexa Genome Analyzer IIx, with the capacity of sequencing up to 95 Gb per run (http://www.illumina.com/systems/sequencing.ilmn). The company introduced the HiSeq system in 2011, which can sequence up to 600 Gb per run. The Illumina/Solexa sequencing systems are all based on the sequencing by synthesis (SBS) technology. The sequencing process includes three steps: (1) template preparation, (2) sequencing and imaging, and (3) genome alignment or assembly. During template preparation, genomic DNA is firstly broken into smaller sizes from which either fragment templates or, more generally, mate-pair templates are created by ligating appropriate primers to their ends, and then randomly distributed, clonally amplified clusters are produced on a glass slide, which acts as a solid support to immobilize millions of spatially separated template sites, allowing sequencing reactions on all these templates to be performed simultaneously. The Illumina slide is partitioned into eight lanes, allowing independent samples to be run simultaneously. During sequencing, the cyclic reversible termination (CRT) process takes place, which uses reversible terminators in a three-step cyclic process, including nucleotide incorporation, fluorescence imaging and cleavage of the terminating group and the fluorescent dye. In SBS technology, four nucleotides are labeled with four different dyes and are present during the sequencing cycles at the same time. During each cycle, the colours are detected by total internal reflection fluorescence (TIRF) imaging using two lasers. Errors and biases may be introduced during the template preparation and sequencing processes. For example, studies showed that Illumina sequencing data have an underrepresentation of AT-rich[44] and GC-rich regions[45]. A common feature of NGS technologies is that the reads generated are very short, usually ranging from tens to hundreds of base pairs, as they only sequence a fraction of the DNA molecule at either one end or two ends, which produces two types of reads: single-end reads and paired-end reads. Paired-end

reads help dramatically in the alignment and the detection of SVs, as the approximate sequence length between two ends is often known.

The last step, which is probably the most challenging one, is the alignment and/or assembly of the genome sequences, and subsequent variant calling. Here we only consider alignment without assembly, as when sequencing multiple human genomes, we only need to align the reads to the human reference sequence, so that genomic variants can be called. The accuracy and reliability of variation detection by sequencing is highly dependent on the sequencing and mapping quality. Random sequencing errors can be largely solved by simply increasing the read depth, i.e. sequencing the same DNA region multiple times, so that one or two substitution errors can be ignored at one locus, although this may introduce higher costs and longer sequencing time. However, due to the error-prone nature of NGS, for a single-base variant, sometime it's still ambiguous whether a particular locus is homozygous or heterozygous. For example, if there are 20 reads at a locus, 5 of them read A and 15 of them read C, it would be difficult to tell whether the genotype is AC or CC, as the possibility of 5 A's being misread as C's may be similar to 5 C's misread as A's. There are several ways to resolve this issue. One is to ignore or assign lower weight on reads with low quality, such as those reads where the SNP in question lies at either end of the read. If there are multiple samples being sequenced, one can also calculate the likelihood of the genotype of the individual in question by looking at the genotypes of other individuals at the same locus. If haplotype information is known or can be inferred, it will be very helpful in inferring the correct genotype at ambiguous sites. While single-locus substitution errors are relatively easy to resolve, due to the short lengths of reads produced by NGS technologies, correct alignment is a challenge, especially in regions with indels, repetitive regions or copy number variable loci. For example, if a locus has a 2-base deletion, reads that contain this locus towards the two ends may be aligned without a gap and the two mismatches may be called as SNPs instead of deletion. In repetitive regions, reads may be able to align to multiple loci with similar numbers of mismatches. Apart from increasing the read depth, we may choose to ignore reads that map to multiple loci or reads that have mismatches at the two

ends, in order to avoid possible false calls (Figure 1.5). Various bioinformatics tools have been developed to align NGS reads to the reference sequence and call variants, such as MAQ[46], ELAND[47] and SSAHA2[48], aiming to achieve a minimum level of misalignment and high accuracy in variant calling. Target assembly tools, for example TASR[49], were also developed to help alignments and variant calling at loci with indels. While none of them is perfect, each algorithm demonstrates certain strengths in different conditions[50]. Therefore, choosing the appropriate alignment algorithm is critical in getting the best quality in aligning the sequencing data and calling variants.

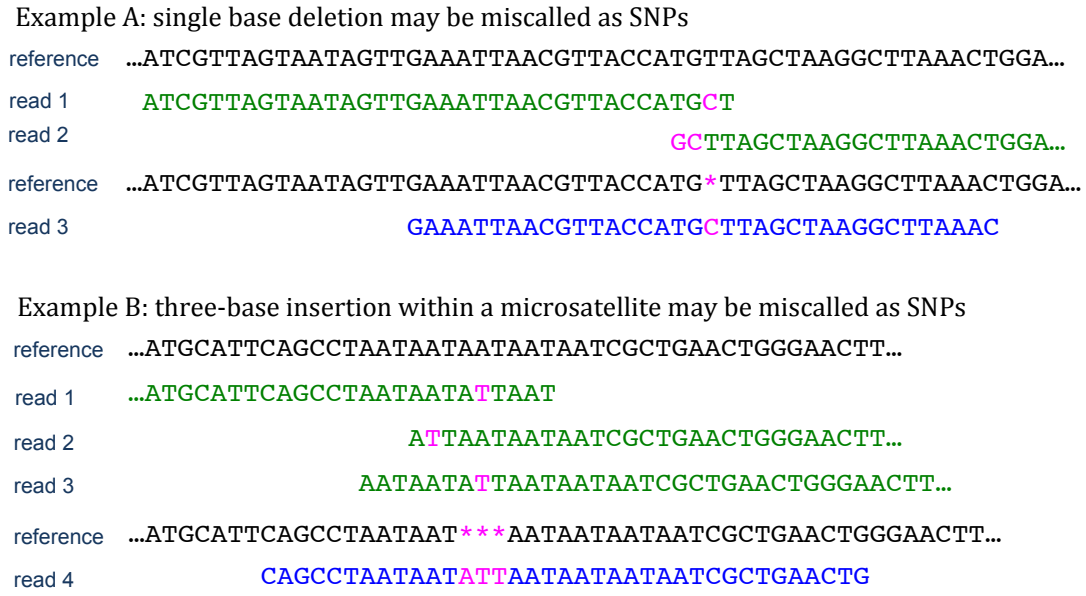

Figure 1.5 Examples of misalignment and miscall. In both examples, black letters are reference sequences, green letters are the reads where miscalls occur, and blue letters are the reads where variants are called correctly. Magenta letters are the variants called. If there is insertion, stars are used to fill the bases in reference sequences. In example A, a single-base insertion 'C' is called as single-base substitutions in read 1 and 2, because the base is near the end of the reads. The insertion is correctly called in read 3, because the base is in the middle of the read, there is more context for alignment. In example B, a three-base insertion is called as SNPs in reads 1, 2 and 3, because the insertion has only one base difference from the microsatellite unit, and the reads do not extend beyond both sides of the microsatellite. Read 4 is correctly aligned and the insertion is called, because it extends to non-repetitive sequences on both sides of the microsatellite.

1.2.3 Functional impact of genomic variation

One of the most important yet challenging questions for geneticists is: which pieces of the human genome are functional? In the early stages of genetic research decades ago, researchers focused mainly on protein-coding genes, which have obvious functional products – proteins. As these genes only make up

~1.5% of the genome, it was believed that 98.5% of our genome consisted mainly of "junk DNA". However, more and more studies have demonstrated functions of inter-genic or intronic sequences in the genome, and there are also a large number of transcribed non-coding RNAs, more and more of which have shown evidence of functionality. In order to understand how genomic variation contributes to the phenotypic differences of modern humans, we will look at the potential impact of different types of genomic variants in four types of genomic regions: (1) exons, i.e. sequences that determine the amino acids of proteins; (2) non-coding transcribed regions, i.e. sequences with RNA products that are not translated into proteins; (3) intronic regions, i.e. sequences between exons; and (4) inter-genic regions, i.e. sequences that do not contain any gene.

DNA sequences in exons code for proteins. Three consecutive nucleotides specify one of the 20 kinds of amino acids, or a stop codon, which is a signal of the end of the protein or polypeptide. Because there are four types of nucleotide, 64 types of codons can be formed by three nucleotides. Therefore, the genetic code is redundant, which means that multiple codons can represent the same amino acid. SNPs in protein coding sequences, therefore, can have two different consequences: one is to change the amino acid encoded by the codon containing the SNP, which we describe as non-synonymous; and the other is not to change the amino acid, i.e. the codon is still encoding the same amino acid, so we describe this SNP as synonymous. It seems obvious that non-synonymous SNPs should have a functional impact on the protein, while synonymous SNPs should not. Although in most cases this is true, one should note at least two exceptions: on one hand, change in amino acid does not always change the structure or function of the protein. It is possible that the changed amino acid has very similar physical and chemical features to the original amino acid, thus would not affect the function, or that parts of the protein are tolerant of variation. On the other hand, although synonymous SNPs do not change the amino acid, they may affect the structure of the DNA or RNA, or the binding of enzymes during the transcription or translation process, or create a new splice site, and thus may still have functional impacts. However, as this kind of situation is not common, in evolutionary studies, we normally consider non-synonymous SNPs as functional,

while synonymous ones as not. Small indels in coding sequences can sometimes have bigger functional impact than SNPs. Insertion or deletion of one or two nucleotides (or any number that cannot be divided by three) in an exon causes reading frame shift, which results in a complete change of amino acids of the protein from the variable site onwards, and will also be likely to change the position of the stop codon. Therefore, in most cases, the protein product of such a mutation will not be functional. As exons are usually short and separated by longer introns, larger SVs or gene conversions in exons may result in the removal or addition of several exons or even the entire gene, or imbalanced dosage of a gene.

Although we have not yet known how many RNA genes are there in our genome, tens of thousands of them have been discovered by either experimental or computational approaches, yet the majority of them have poorly understood functions. Functions of non-coding RNAs (ncRNAs) seem to be very diverse and are involved in multiple molecular processes, many of which are still poorly understood. There are many types of ncRNAs based on their functional roles. Here I list the relatively well-understood ones. (1) Transfer RNA (tRNA): tRNA is involved in translation, and plays a role of transferring the right amino acid to the growing polypeptide chain during protein synthesis. (2) Ribosomal RNA (rRNA): rRNA is part of the RNA-protein complex called ribosome, which is the protein-producing organelle in the cytoplasm. rRNA is the most abundant RNA in a cell, and its genes are highly repetitive, because a large number of ribosomes are needed for protein synthesis. (3) Small nuclear RNA (snRNA): snRNA is present in the nucleus of eukaryotic cells. It is involved in a few different regulatory processes, including RNA splicing, chemical modifications, e.g. methylation or pseudouridylation of rRNAs, tRNAs and snRNAs, RNA biosynthesis and regulation of transcription factors. (4) microRNA (miRNA): miRNA is the reverse complement of part of another gene's mRNA, and it changes the expression levels of one or several genes by RNA interference. miRNAs are single-stranded and generally 21-23 bases long when they are in their mature form. (5) Small Interfering RNA (siRNA): siRNA plays a similar role to miRNA, but is double-stranded and derived from long double-stranded RNAs

or small hairpin RNAs. (6) Piwi-interacting RNA (piRNA): this forms a RNA-protein complex with piwi proteins, and the complex functions in transcriptional gene silencing in germ line cells. piRNAs are found in mammalian testes and somatic cells, and are 29-30 bases long. Apart from these ncRNAs, there are also bifunctional RNAs that have two different functions, for example, some mRNAs also act as ncRNAs, and some ncRNAs play roles in two different categories above. Variants within the unprocessed or immature ncRNAs can still have functional impacts, for example, altering the splicing sites, altering which strand is functional in miRNAs, or changing the binding target of the ncRNAs. It is worth noting that the functional impact of variants in ncRNAs is often not obvious and difficult to identify, due to the complexity of the functional mechanisms of ncRNAs.

Intronic regions in the human genome are those sequences between two exons are usually removed from the transcribed RNA before translation, to generate the mature RNA. Although the majority of introns seem to have no function, more and more studies have revealed various functions for some introns. For example, some sequences of introns adjacent to exons can determine the splicing sites, which in turn affect the protein products. Some introns themselves can be further processed to generate non-coding RNA molecules, and some even encode proteins. Some introns are transposons, which copy themselves and insert the copies into other locations in the genome. Some intronic sequences may regulate nucleosome or transcriptional factor binding, which will affect the expression level of the gene. Therefore, variation in some intronic sequences may have functional impacts, and the most obvious one is to generate alternative splicing sites, which is a common mechanism of generating multiple protein products from one gene. Some intronic variants may also have an impact on the regulation of gene expression.

Intergenic regions are sequences located between genes, and were sometimes considered as non-functional. However, many studies have shown evidence of regulatory functions of intergenic regions. Although it is often difficult to distinguish regulatory regions from non-functional regions in intergenic areas, conserved non-coding sequences (CNS) are believed to be likely to contain

regulatory regions[51], so most studies of regulatory elements in the genome are centered on CNS, along with other sequence features such as known regulatory motifs and transcription factor binding sequences[52-55]. These studies have discovered several types of regulatory regions, including promoters, transcription factor binding sites, enhancers, insulators, and so on. Variants within these regulatory regions may have functional impact on the expression level of certain genes. The positioning and structural changes of nucleosomes also regulate gene expression levels. Although the variation of this type of regulation is mostly by the modification of histones, variants of the DNA sequences within or nearby a nucleosome may also alter the positioning of nucleosomes, which may have regulatory impacts. Strikingly, many Genome Wide Association Studies (GWAS) have identified a large proportion of hits associated with certain diseases or traits that are in intergenic regions, which implies unknown functionality of these intergenic sequences. However, for most of these variants, it is difficult to study their functions experimentally, and we are yet to understand their real impacts on human traits or diseases. The ENCODE (Encyclopedia Of DNA Elements) project, launched in 2003 by the National Human Genome Research Institute (NHGRI), is aiming to identify all functional elements in the human genome[56]. The project develops technologies to enable large-scale and systematic identification and characterization of functional elements, and has yielded fruitful results in its pilot project[57].

1.3 Footprints of natural selection on genomic variation

1.3.1 The theory of genetic drift

Most genomic variants are believed to be neutral, i.e. they have no biological effect on the fitness of the carrier. In this case, genetic drift plays a major role in determining the fate of a particular allele of a variant in the genome. The concept of genetic drift was first introduced by Sewall Green Wright, one of the founders of population genetics. It refers to the changes in frequency of an allele in a population due to random sampling, where only chance determines which allele is inherited by the offspring[58]. Genetic drift eventually causes one allele to either disappear or being fixed in the population, and thus reduces the level of genetic

diversity (Figure 1.6). The effect of genetic drift is closely related to the effective population size (N_e). This concept was also first introduced by Wright, and was defined as the minimum size of a Wright-Fisher population that shows the same level of genetic variation as the population in question. N_e is usually much smaller than the actual population size, and can be determined either from the variance of allele frequencies from one generation to the next, or the probability of two alleles within an individual being descended from a common ancestor. The smaller the effective population size, the shorter time it takes for genetic drift to either eliminate or fix the allele in the population, and vice versa (Figure 1.7). Although the effective population size is related to the actual size of the population (N), there are many factors that influence the relationship between N_e and N. For example, most populations experience fluctuations in the actual population size over time, which has a great impact on the effective population size. Other factors, such as the variation of number of offspring among individuals, and the level of randomness in mating, all affect the effective population size.

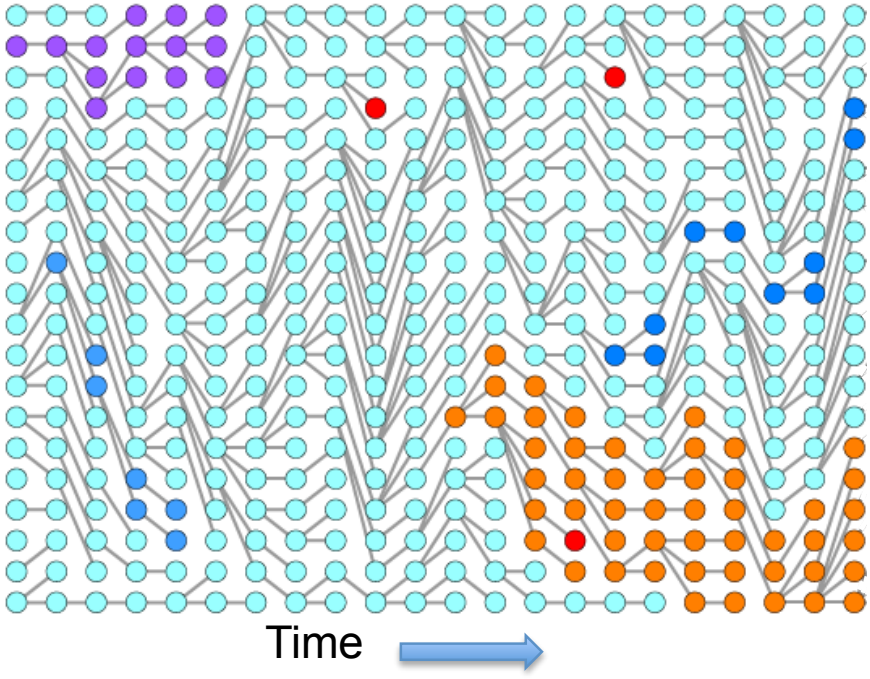

Figure 1.6 Genetic drift in a population. Different colored circles represent different variants in the population. In a Wright-Fisher population, genetic drift drives frequencies of variants up and down by chance, and a variant will eventually disappear or get fixed in the population.

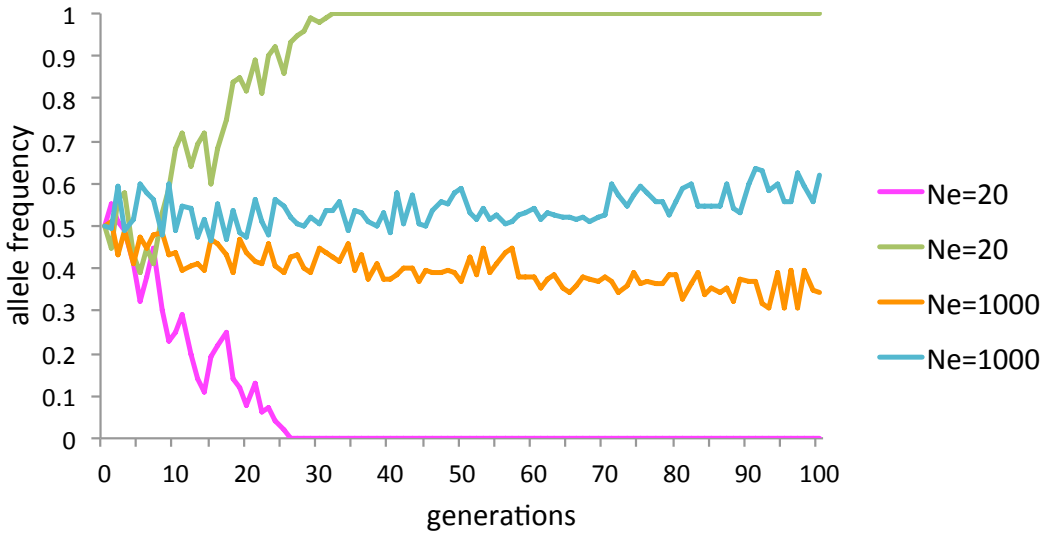

Figure 1.7 Genetic drift in populations with different effective population sizes. This figure shows the change of the frequency of one allele with an initial frequency of 0.5, in populations with effective population sizes of 20 or 1000. In a population with a smaller effective population size, it takes less time for the variant to disappear or reach fixation, and the frequencies of alleles tend to change more dramatically from one generation to the next.

One of the fundamental models of genetic drift is the Wright-Fisher model, developed by Wright and Sir Ronald Aylmer Fisher. This model describes the effect of genetic drift on allele frequencies. It assumes that the generations do not overlap, the population size is constant, and the population is randomly mating. If the frequency of one allele of the variant is q, and that of the other is p, then the probability of obtaining k copies of the allele that had frequency p in the last generation is:

$$\frac{(2N)!}{k!(2N-k)!}p^k q^{2N-k}$$

Although this model is widely used in population genetics, its assumptions are not at all realistic for human populations. However, for most populations, this model is a good approximation to start with.

1.3.2 Positive (Darwinian) selection

Although genetic drift plays an important, and often dominant, role in evolution, it is not the only force that drives the changes in allele frequencies in a

population. Since Darwin set out his theory of natural selection as a means of speciation and adaptation in 1859 in his book *On the Origin of Species*[59], Darwinian, or positive, selection has been considered as one of the most important driving forces of evolution. On the phenotypic level, Darwin's concept is very straightforward: if a new inheritable trait is useful, it will be preserved by nature. Here "useful" refers to advantages in either survival or reproduction. Individuals who have certain advantages, compared to other individuals with a different phenotype who are competing on the same resources, in surviving to the reproductive age, attracting mates, having better ability to fertilize, or producing more offspring for other reasons, will be more likely to preserve their traits in the population and have progeny that share the same traits. As time goes on, the advantageous phenotypic trait will become more common, and finally become a shared trait in the whole population. On the genetic level, frequencies of the alleles that determine the advantageous trait will go up rapidly in the population, and finally reach fixation (i.e. 100% frequency).

The effect of positive selection on the frequency of the advantageous allele in a population depends on two factors: the strength of the selection, i.e. the relative level of fitness of the advantageous genotype, and the number of generations since the selection started. We use the selection coefficient parameter (s) to measure the strength of a positive selection event. s is defined as the increased percentage of offspring that the individual carrying the advantageous genotype produces per generation, compared to individuals carrying the other genotypes. For example, if the genotype AA has a selection coefficient of 0.1 compared to genotype aa, and if the aa individual has 10 progeny, then the AA individual would have 11. The higher the selection coefficient, the shorter time it takes for the advantageous allele to reach fixation in the population. Also, the speed of allele frequency increase tends to become slower when the allele frequency gets higher. Therefore, the frequency of the advantageous allele is also dependent on the number of generations since the allele started to undergo a selective sweep, but in a non-linear fashion.

The most well-studied type of positive selection is known as a "hard" selective sweep, where a single new mutation occurs in one individual, and this new allele

results in some advantageous trait, so that positive selection favors the new allele immediately after it emerges, and it increases in frequency until reaches fixation. Another type of positive selection acts on standing variants, which means that the allele does not have an advantage at the beginning, so its frequency initially depends only on genetic drift. However, due to a change in the environment or other factors, the allele becomes advantageous at some stage, and then starts to be positively selected. This is called a "soft" selective sweep. In the case of a soft sweep, the frequency of the selected allele also depends on the starting frequency of the allele in the population before selection starts to act, in addition to the other two parameters mentioned earlier. A more complicated type of positive selection is that the advantage only happens if a combination of certain alleles is present together within the individual. Some of these alleles could be new mutations, while others could be standing variants. Among these three types of sweeps, hard sweeps are the easiest to detect, due to their simple process and clear pattern on the genetic variation. Soft sweeps are harder to detect, especially when the standing variant had reached a relatively high frequency before selection starts, as this will lead to the increase of frequencies of several haplotypes, which will make the genetic pattern difficult to recognize. The complex type of selection is the most difficult to detect, and we do not yet know whether, or to what extent, it has influenced the history of modern humans.

There has been debate about what proportion of our genome has been positively selected. Apart from some genome-wide analyses (discussed in section 1.4) that have yielded rather variable results, there are some positively selected genes in modern humans that have been widely studied and confirmed by functional evidence. One example is the Duffy blood group locus, which has three classical alleles: FY*A, FY*B and FY*O. FY*O has been found at high frequency in sub-Saharan African populations, but not elsewhere. People carrying the FY*O allele are highly resistant to *Plasmodium vivax*, a cause of malaria, which is a disease common in sub-Saharan Africa and responsible for many early deaths. The FY*O variant is a SNP in a transcription factor binding site that abolishes expression in red blood cells and thus blocks entry of the parasite[60]. Studies have shown some evidence of positive selection on FY*O allele in sub-Saharan African

populations[61], though the pattern is complex because the variant appears to have arisen independently more than once[62]. However, there are very few such compelling examples of positive selection in humans supported by functional evidence (Table 1.1).

Table 1.1 Examples of positively selected genes supported by functional evidence

Gene	Location	Selected function	Selected population(s)	Reference
FY	1q21–q22	malaria resistance	African	Hamblin & Di Rienzo (2000)
EDAR	2q13	hair/teeth/sweat gland development	Asian	Sabeti et al. (2007)
LCT	2q21	lactase persistence	European	Bersaglieri et al. (2004)
SLC45A2	5p13.3	skin pigmentation	European	Sabeti et al. (2007)
CYP3A5	7q21.1	salt sensitivity	European, Asian	Thompson et al. (2004, 2006)
FOXP2	7q31	language/speech	worldwide	Enard et al. (2002)
HBB	11p15.5	malaria resistance	African	Ayodo et al. (2007)
CASP12	11q22.3	sepsis resistance	worldwide	Xue et al. (2006)
SLC24A5	15q21.1	skin pigmentation	European	Lamason et al. (2005)
ABCC11	16q12.1	earwax secretion	Asian	Xue et al. (2009)
G6PD	Xq28	malaria resistance	African	Tishkoff et al. (2001)

1.3.3 Negative (purifying) selection

Mutations that reduce the fitness of the individual carrying them will be negatively selected, as contrasted with beneficial alleles being positively selected. This type of selection is also known as purifying selection, as the selection acts to eliminate harmful alleles, and thus "purifies" the genetic locus. Purifying selection is believed to be widespread in functionally important genes or regulatory elements, as mutations in these elements may often be deleterious.

Due to the linkage of nearby loci, purifying selection can result in a reduction of variation in regions surrounding the selected locus. Negative selection is responsible for the high level of conservation among species and low level of variants within species in exons of many functionally important protein-coding genes[63].

1.3.4 Balancing selection

Diploid individuals have two alleles at each locus, which together may contribute to the fitness of the individual. An individual heterozygous for the beneficial allele often has half of the advantage in fitness of an individual homozygous for the beneficial allele, but this is not always the case. Sometimes the heterozygous genotype has the highest level of fitness, in which case selection would act to maintain heterozygosity in the population. This, of course, will result in maintaining a moderate frequency of the allele in the population, instead of driving one of the alleles to fixation or elimination. This type of selection is referred to as a form of "balancing selection", where alleles are maintained at an intermediate frequency. Another type of balancing selection is not due to the higher fitness of heterozygous individuals, but to the low frequency allele having a higher level of fitness. Therefore, over time, an equilibrium with intermediate frequency will be maintained. An example of balancing selection in humans is the major histocompatibility (MHC) locus, a large and complex region that determines the histocompatibility of an individual and carries many genes involved in defense against pathogens. The cell-surface proteins that are known as the human leukocyte antigens (HLA) are encoded by genes in this locus. This locus has shown an exceptionally high level of diversity among humans, and some of the alleles are very ancient, even predating the chimpanzee-human split. It is believed that this high level of diversity is caused by balancing selection. However, it is not entirely clear whether the selection is to maintain a high level of heterozygosity in each individual, or to maintain low or intermediate frequencies of many alleles in the population. If the former is the case, it may be that a large number of heterozygous MHC loci provide the individual with a broader spectrum of antigen binding specificities, which results in a higher ability to resist infectious diseases. If the latter case is true, relatively low

frequencies of many alleles may prevent pathogens from evolving to evade immune detection of those antigens encoded from high frequency alleles. It is also possible that these two types of balancing selection both act on the HLA genes. Again, however, there are few other examples of balancing selection in humans supported by strong functional evidence.

1.4 Statistical approaches to detect signatures of positive selection in the human genome

1.4.1 Linkage disequilibrium-based neutrality tests

As mentioned earlier, due to the difference in recombination rates, there are blocks of certain variants in the genome that are often linked together on one haplotype, known as linkage or haplotype blocks. Linkage disequilibrium refers to the non-random associations of alleles at different loci. For two loci from different linkage blocks in a neutral situation, we are able to calculate the expected frequencies of any combination of alleles at these loci if we know the frequencies of the alleles. For example, if the frequencies of allele A_1 and allele B_1 at locus 1 are a_1 and b_1, and the frequencies of allele A_2 and allele B_2 at locus 2 are a_2 and b_2, then the expected probabilities of the four possible combinations of the two loci would be:

	A_1	B_1
A_2	a_1a_2	b_1a_2
B_2	a_1b_2	b_1b_2

If the actual frequencies of the four combinations are as expected, we say that these two loci are in linkage equilibrium. However, in many cases, the actual frequencies of the four combinations are less or more than the expected values. In this case, we say that the two loci are in Linkage Disequilibrium (LD).

There are many factors that can influence the level of LD at a locus in the genome. First of all, the variation of recombination rates causes some loci to be in higher

LD than others. For example, loci within a recombination cold region would be more likely to be linked than those within a recombination hot region, even if they have similar physical distances. As linkage information is critical for many genetic studies, genetic linkage maps, often known simply as genetic maps, have been generated to show the position of genomic variants relative to each other in terms of recombination frequency. The most widely used human genetic map was produced by the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/), and provides the genetic distances based on more than three million SNPs across the human genome[39]. LD can differ between populations, and population structure or non-random mating can also have impacts on the LD structure of the genome, but this effect is more likely to be genome-wide than locus-specific. Natural selection, especially positive selection, can have a high impact on the LD of the selected locus, and more specifically, will cause the locus to have unusually high LD compared with neutral loci of similar frequency.

As described earlier, if a new mutation turns out to be advantageous in fitness for the individual carrying the mutation, the frequency of that advantageous allele will go up rapidly in the population, and finally reach fixation or near-fixation. Due to the linkage of surrounding alleles with the selected allele, their frequencies will often go up along with the selected allele. As this process takes a much shorter time compared to random drift, it often does not allow sufficient time for recombination to break down the linkage. This will result in a long LD block at the locus, centered on the selected allele (Figure 1.8). Therefore, by measuring the level of LD of one particular locus in a population, a selective sweep can be detected if the level of LD at this locus is high compared with other frequency-matched haplotypes in the same or different populations.

As mentioned above, if genetic markers are in linkage equilibrium, their frequencies should match the expected frequencies calculated based on the allele frequencies. However, if the markers are in LD, their actual frequencies will be different from expectation. To measure the level of LD, we use *D* to represent the deviation of the observed frequency of one combination of the two loci in question from what is expected. Based on the example of locus 1 and locus 2

above, if the frequency of A_1A_2 is f_1, then $D = f_1 - a_1a_2$. Obviously, if the two loci are in linkage equilibrium, $D = 0$. The value of D is dependent on the frequencies of the alleles, so to measure the level of LD, we use a normalized D', which is (D/D_{max}), where D_{max} is the maximum theoretical value of D[64]. The most common measure of LD, however, is $r^2 = D^2/[a_1a_2\ b_1b_2]$, where r is called the correlation coefficient of two loci.

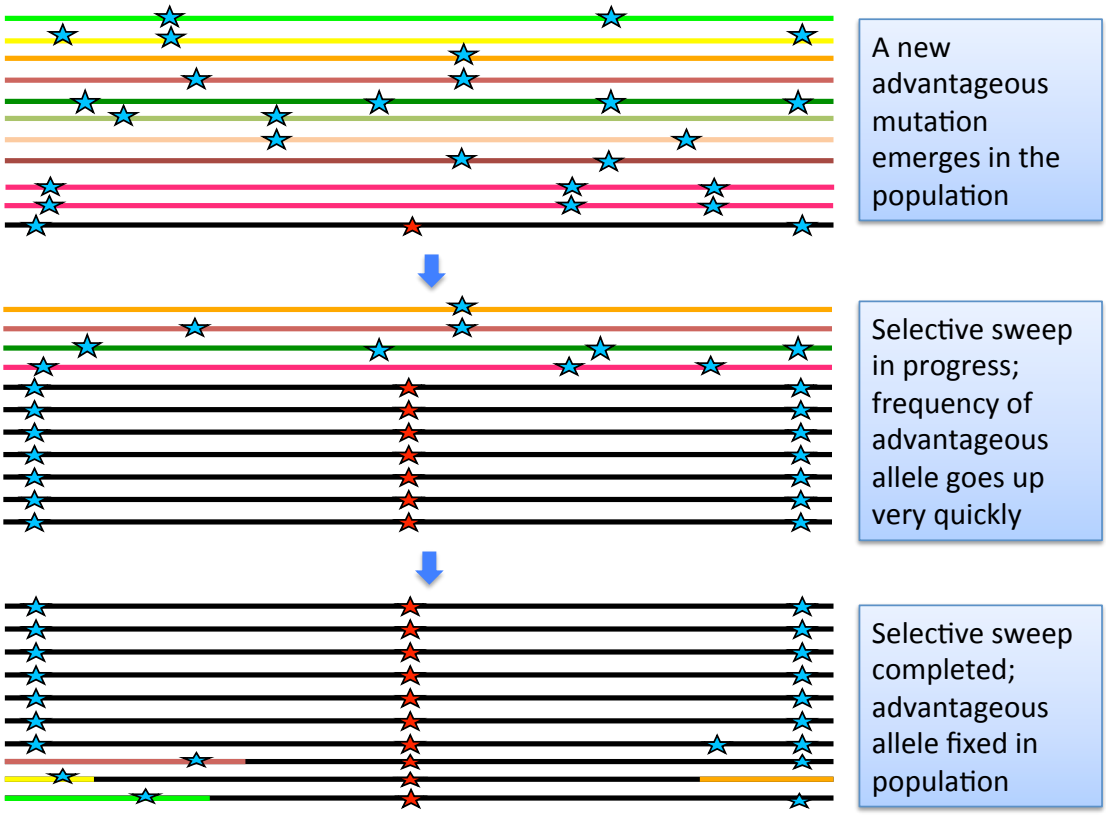

Figure 1.8 A selective sweep. Different colored lines represent different haplotypes in the population. Blue stars are neutral mutations, and the red star is the advantageous mutation under positive selection.

Simple measurements of LD at loci are not sufficient to detect signals of positive selection. Other factors that may influence the level of LD need to be considered and their effects need to be removed in order to isolate the long LD signal left by a selective sweep. Also, the pattern of LD scores along the region of interest needs to be considered, in order to identify the most likely selection target site. Based on these principles, several statistical tests have been developed to detect signals of positive selection by measuring the decay of LD scores over long genetic distances. One of the earliest such tests is the Extended Haplotype Homozogosity (EHH) test[65], which detects long-range haplotypes with a high

frequency in the population. Several other tests were then developed based on EHH, for example, the XP-EHH test calculates EHH scores in one population with another population as a reference, which provides power to detect population-specific positive selection[66]. Another test, iHS, calculates integrated EHH on haplotypes carrying the ancestral allele or derived allele, then generates a score based on the ratio of these two EHH scores[67]. This test seems to have a higher power for detecting selective sweeps that have not yet reached the near-fixation stage. Although these LD-based tests have a reasonable power for detecting signals of selective sweeps, due to the nature of LD-based tests, the regions they detect are often a few hundred kb to a few Mb in length, so they are generally not able to localize the selection signals into a small enough region in order to identify the causal variants. The later developed Composite of Multiple Signals (CMS) test, which combines multiple EHH-based tests and measures of derived allele frequency differentiation (XP-EHH, iHS, F_{ST}, ΔDAF and ΔiHH) to generate a composite score, is able to increase the resolution significantly in some cases[68].

Several research groups applied LD-based tests to genotype data like those from the HapMap project to perform genome-wide scans of positive selection. As mentioned earlier, Sabeti et al. identified ~300 candidate positively-selected regions from the HapMap2 data using the EHH test, including 22 strong candidate regions, from which they further identified putative selection targets[66]. Voight et al. identified ~250 strong signals of recent positive selection using data from the HapMap project, and generated a set of SNPs that tag these candidate regions[67]. Wang et al. developed the LD decay (LDD) test, which looked at the expected decay of adjacent SNP by sorting homozygosity of each high-frequency allele, avoiding the inference of haplotypes, and used this test on the 1.6 million SNP genotype data set from Perlegen Sciences[69]. They identified ~1800 genes with signals of positive selection[70].

1.4.2 Frequency-spectrum-based neutrality tests

One of the most important genetic effects of positive selection is that it drives the frequency of the beneficial allele to a high frequency or even fixation. Due to the linkage of surrounding alleles with the selected allele on the same haplotype, the

frequencies of those alleles will also go up. On the other hand, the corresponding alleles on the other non-selected haplotypes will go down rapidly or even disappear from the population. Therefore, alleles in the region surrounding the advantageous allele will differentiate into either very high or very low frequencies (Figure 1.8). In contrast, frequencies of neutral alleles are only driven by genetic drift, so they fluctuate randomly and are not likely to have the highly differentiated patterns. If we compare the allele frequency distributions of a region that has undergone a selective sweep with a neutral region, then three main differences may occur: (1) the selected region has a higher proportion of extremely low-frequency alleles than the neutral region; (2) the selected region has a higher proportion of extremely high-frequency alleles than the neutral region; and (3) the selected region has a lower proportion or even absence of intermediate-frequency alleles (Figure 1.9).

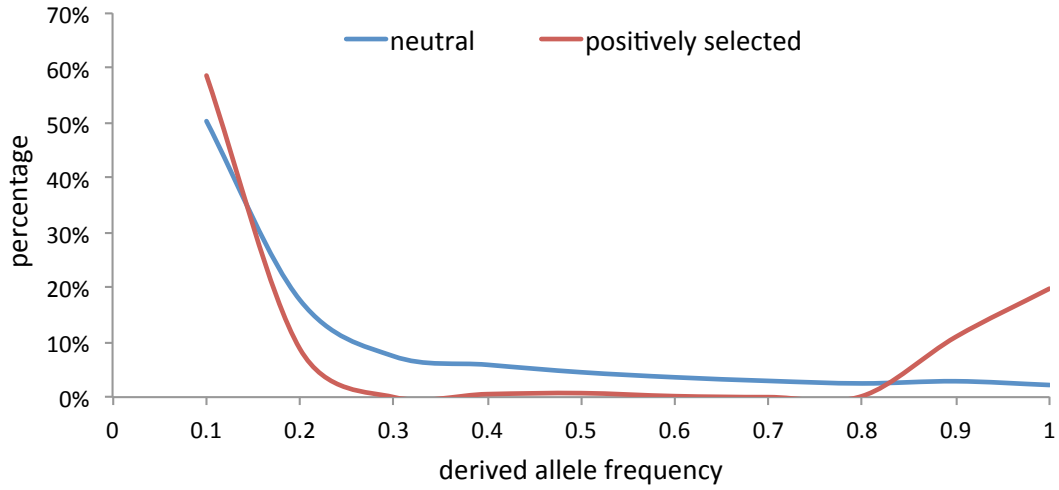

Figure 1.9 Derived allele frequency spectrum of a positively selected region versus a neutral region.

Several statistical tests have been developed to detect one or more of these three features, which, although strictly tests of neutrality, are often interpreted as evidence of selection. One of the earliest and still most widely used such tests is the Tajima's *D* statistic[71], which compares two estimates of $\theta = 4N\mu$, one of which uses the number of segregating sites (S), and the other the average pairwise differences (π), i.e. $d = \hat{\theta}_\pi - \hat{\theta}_S$. Then the *D* statistic is calculated by dividing *d* by its standard deviation. In theory, if the sequence fits the neutral

model and the alleles are in equilibrium, we expect $d = 0$. If the absolute value of *D* statistic is larger than expected by chance (i.e. the different is statistically significant), the neutral hypothesis is rejected. However, the rejection of neutral model by Tajima's *D* can be caused by several factors, including positive selection, negative selection, balancing selection, population expansion or bottleneck, non-random mating, and so on. A positive Tajima's *D* value suggests a low level of both low and high frequency alleles in the region, indicating either balancing selection or a decrease in population size, or both. In contrast, a negative Tajima's *D* suggests an excess of low and high frequency alleles in the region, indicating positive selection, or population expansion. In order to use Tajima's *D* to detect a selective sweep, we need to (1) measure the significance of the negative *D* value, and (2) eliminate the possibility of demographic factors (e.g. population expansion after a bottleneck). There are two commonly used ways to gauge the level of significance. One is to simulate a large set of regions that mimic the real genetic data in a neutral scenario, and then calculate the *D* statistic on the simulated regions. A p value can be obtained from the distribution of the *D* statistic in the simulated neutral regions. The other way is to obtain an empirical p value, in which case data on a large number of comparable regions in the genome need to be obtained, and by ranking the *D* statistic of the empirical data, outliers with significant empirical p values will be identified. There are pros and cons of both approaches. The first method has the advantage of independency, so is free from potential bias in the empirical data themselves. However, it cannot rule out the possibility of being influenced by demographic effects, as the simulated data may not take into account population structure and changes. The second approach can effectively eliminate the demographic factors, as usually population expansions or bottlenecks would affect the whole genome or at least a large fraction of it, so is not likely to affect the empirical rankings. However, the second approach cannot be strictly treated as a measure of statistical significance, since it is unknown what fraction of the empirical data should be the target of selection, and in this method we assume that the empirical data set as a whole is neutral, which may not be true and therefore may introduce false positive or false negative results. In practice, both approaches may be used to

measure the significance, and the best way to measure the level of significance in a certain study should be judged based on the specific conditions of the study.

Another widely used statistic is Fay and Wu's *H*[72], which measures an excess of high frequency derived alleles. The *H* statistic is similar to Tajima's *D* in the sense that it also compares two estimates of θ, but differs by taking into consideration of whether a particular allele is derived or not when looking at pairwise differences. Therefore an outgroup species is needed in order to determine the derived alleles. Here $h = \hat{\theta}_{\pi} - \hat{\theta}_{H}$, where θ_H is the estimate of θ weighted by the homozygosity of derived variants. Another difference between the *H* and *D* statistics is that Fay and Wu's *H* measures departures from neutrality by mainly looking at the difference between high frequency and intermediate frequency alleles, whereas Tajima's *D* mainly looks at the difference between low-frequency and intermediate frequency alleles. This makes Fay and Wu's *H* less sensitive to population expansion than Tajima's *D*; therefore, by comparing the two statistics on the same region, we may be able to distinguish the effects of population expansion from selection.

More recently developed frequency-spectrum based tests use more sophisticated algorithms to increase the robustness to demographic factors. These methods aim to capture the comprehensive spatial patterns of allele frequencies in the region, instead of focusing on just one aspect[73-76]. Although some of these methods are relatively computationally expensive, they to some extent have higher power and sensitivity in detecting selective sweeps. One example of this new generation of tests is the Composite Likelihood Ratio (CLR) test developed by Nielsen et al.[76]. The CLR test calculates a composite likelihood ratio by dividing the maximum composite likelihood under a neutral model by that under a model with a selective sweep. Instead of using a pre-set neutral model with certain demographic parameters, the null model in the CLR test is derived from the background frequency spectrum pattern of the data set in question. This approach has two advantages: (1) it avoids biases introduced by simplified or unrealistic demographic models, so minimizes the effects of demographic factors of the population in question; and (2) it eliminates the ascertainment biases of the variant discovery process, as this kind of bias would

occur across the whole data set and thus have been taken into account in the neutral model. This algorithm is also faster than previous likelihood ratio-based tests, which made it feasible to apply the test to whole-genome data sets with large sample sizes.

Although frequency-spectrum-based tests are best used on sequencing data, they can also be applied to genotype data in a genome-wide scale. Kelley et al. used Tajima's *D* statistic to look for outliers using the Perlegen Sciences SNP genotype data, and found 385 genes with signals of positive selection[77]. Williamson et al. applied a composite likelihood ratio (CLR) approach based on site frequency spectrum to the same set of data, and identified 101 regions with evidence of positive selection[78].

1.4.3 Population differentiation based tests

When a population moves to a new environment, adaptation may take place, and positive selection may act on mutations that help the individual better adapt to their new environment. Human populations moving to different parts of the world have experienced distinct climates and natural resources. Therefore, some genetic changes may be favored in one particular population but not the others. If one or more alleles at a particular genomic locus have highly differentiated frequencies in different populations, or are even population-specific, positive selection may have acted on the particular locus in one or more of the populations. The fixation index, F_{ST}, first introduced by Wright, is often used to measure such population differentiation[79]. F_{ST} is often defined as the relative difference of the average number of pairwise difference between and within two populations at one locus:

$$F_{ST} = \frac{\pi_{between} - \pi_{within}}{\pi_{between}}$$

The value of F_{ST} ranges from 0 to 1, with a value of 0 implying complete panmixis (i.e. no differentiation), compared with a value of 1 indicating a complete separation between the two populations.

F_{ST} is often used in the detection of population-specific selective sweeps, with higher values indicating a higher probability of selection. However, this method is often criticized, as the value of F_{ST} is highly influenced by population structure and demographic history, as well as the ascertainment biases of variant discovery in different population samples. Therefore, F_{ST} values are often evaluated by comparing to the genome-wide or multi-locus distribution, as demographic factors or data biases will most likely affect the whole data set equally. Akey et al. estimated locus-specific F_{ST} compared with genome-wide distribution, and identified over a hundred loci showing "signatures of positive selection" with high levels of differentiation among populations[80]. However, by examining the Perlegen (~1 million SNPs) and HapMap phase I (~0.6 million SNPs) data sets, Weir et al. showed that locus-specific estimates of F_{ST} are too variable to be used in detecting selection[81]. Nevertheless, when multiple independent background loci along with appropriate criteria are used to detect outliers, F_{ST} can be a good indicator of population specific selection[82].

Population differentiation was often used along with LD-based tests or other approaches to identify positive selection in one population versus another. For example, the HapMap project used LD-based tests in combination with F_{ST} to identify regions that have undergone population-specific positive selection[83]. Oleksyk et al. used a set of 183,997 SNPs in European and African American population samples to look at population differentiation, and identified 180 regions with evidence of positive selection in either population, validated by LD, population divergence and other methodologies[84].

1.4.4 Functional-annotation based neutrality tests

A certain allele at a genomic locus can be positively selected only if it has functional consequences that are beneficial for the carrier. Therefore, non-functional variants should be neutral and their frequencies should only be affected by genetic drift or demographic factors. By comparing patterns of functional variants versus non-functional variants in a gene or functional element, one could potentially identify signatures of selection at this locus. The K_a/K_s ratio (also known as ω, or dN/dS), for example, is often used for this

purpose. It is the ratio of the number of non-synonymous substitutions per non-synonymous site (K_a) to the number of synonymous substitutions per synonymous site (K_s) in a protein-coding gene. In the simplest analysis, a K_a/K_s ratio greater than 1 indicates a sign of positive selection, since a K_a/K_s ratio of 1 is expected for a neutral gene. However, more sophisticated statistical analysis needs to be performed to determine the significance of the K_a/K_s ratio as an indicator of positive selection, especially when the number of substitutions is low. Simulations or maximum likelihood analysis may be applied to distinguish between a neutral model and a significant K_a/K_s ratio.

The K_a/K_s ratio is a simple yet powerful tool to identify signatures of positive selection in protein-coding genes, as it uses few assumptions and has a strong functional foundation. However, it has complications and limitations. First of all, mutation rates of different base substitutions are variable, and the codon usage is often biased, which may result in a higher probability of certain non-synonymous or synonymous changes. Secondly, certain synonymous changes may have functional impact on the gene, and certain non-synonymous changes may result in similar amino acids and thus have no functional impact on the protein. Thirdly, the K_a/K_s ratio can only be applied, of course, to protein-coding genes, so functional non-coding genes or regulatory elements, which constitute a probably larger proportion of functional loci in the genome, are out of its radar. Lastly, this method requires a rather strong signal of selection leading to multiple amino acid changes in the same protein, and the two lineages being compared need to be distant enough to allow for this accumulation of non-synonymous substitutions.

A good example of using functional annotation to identify positively selected genes is the study by Bustamante et al., in which the authors examined the patterns of synonymous and non-synonymous variants in over 11,000 human genes using sequencing data of these genes in 39 humans, as well as the divergence from the chimpanzee genome. They identified 304 genes with evidence of rapid amino acid evolution[63].

1.4.5 Time to coalescence

Most of the statistical tests discussed above are aimed at detecting recent selective sweeps, i.e. those that nearly reached or just reached fixation. These selective sweeps are likely to have started after ~50 KYA, when human populations from Africa had already started migrating to other parts of the world. As mentioned earlier, anatomically modern humans first appear in the fossil record around 200 KYA. Therefore, in order to understand which, if any, genes or loci were selected during the earliest stages of modern human evolution (~50-400 KYA), thus contributing to the features that make humans unique as a species, we need to identify positive selection events happening around that time period. These events apparently cannot be detected by the above statistical tests, as they are by definition complete in modern humans, so new mutations and recombination events will have erased most of the footprints on allele frequency spectra and LD patterns left by any early selective sweeps.

By estimating coalescence times, i.e. the time to the most recent common ancestor (TMRCA), of genomic loci among all humans and picking out genomic regions that coalesce less than 400 KYA, we will identify loci in the human genome that have spread through all human populations as modern humans emerged, which would indicate that these loci might have undergone positive selection in our lineage. The estimation of coalescence times is based on coalescent theory, developed in early 1980s by John Kingman[85]. It is a retrospective model using mathematics to describe the characteristics of the joining of lineages back in time to the most recent common ancestor (MRCA), which is referred to as coalescence (Figure 1.10). This theory provides the foundation of many neutral genetic models, as well as the estimation of many population genetic parameters, including the relationship between coalescence and effective population size, and TMRCA. Designating the effective population size of a certain population as N_e, the probability of two gene copies coming from the same parent in the preceding generation is $1/2N_e$, so the coalescence time of the sampled lineages through previous generations follows a geometric distribution with E = $2N_e$. Likewise, for k copies of the gene, the probability of k copies reducing to (k -1) copies in the preceding generation is k (k -1)/$4N_e$, and

the expectation for the time interval is E = $4N_e/k(k$ -1). According to these equations, four conclusions can be drawn about the coalescence: (1) the larger the sample size (k), the greater the rate of coalescence ($k(k$ -1)$/4N_e$); (2) the larger the effective population size (N_e), the slower the rate of coalescence; (3) the time to coalescence gets longer as the process moves toward the most recent common ancestor, as when k gets smaller, $4N_e/k(k$ -1) gets bigger; and (4) even small samples sizes have a high probability of including the MRCA of the population, as the probability of the MRCA of the samples being the same as that of the population is $(k - 1)/(k + 1)$.

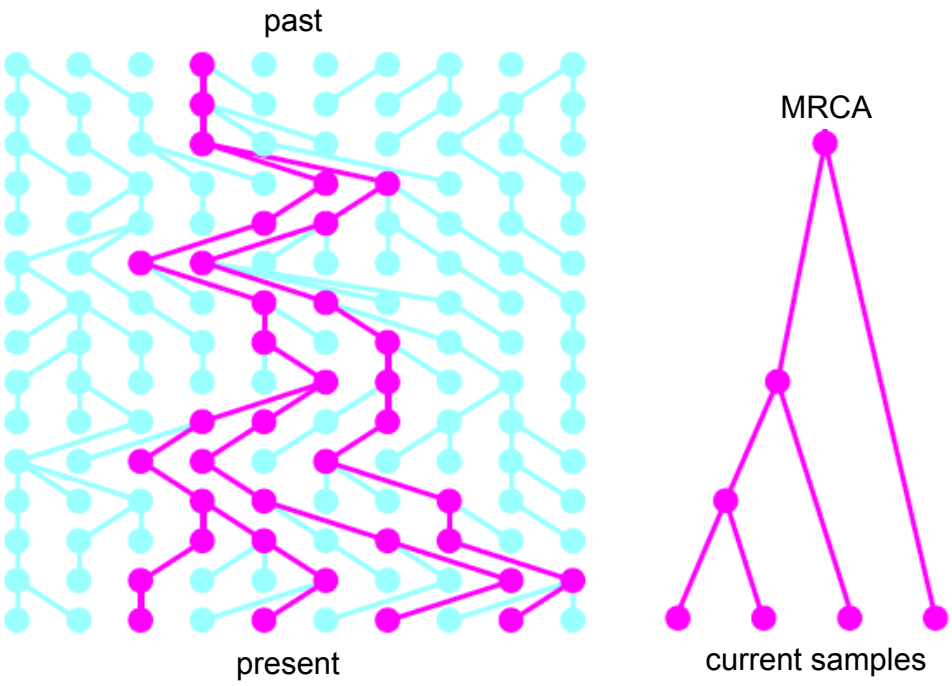

Figure 1.10 The coalescent. Purple circles in each generation are those being traced backwards in time until reaching the common ancestor.

The GENETREE algorithm, developed by Griffiths and Tavaré, uses coalescent theory and Monte Carlo Markov Chain simulation to estimate likelihoods of genetic data under the infinitely-many-sites model. The population mutation parameter $\theta = 4N_e\mu$ and the TMRCA of the locus and given samples can be estimated[86]. It is worth noting that GENETREE assumes no selection and recombination, so it can only be applied to relatively short genetic regions. Previous evolutionary studies have applied this method, yielding fruitful results[87].

1.5 Validation and evaluation of candidate positively selected regions

1.5.1 Simulation as a means of assessing and validating genome-wide scans

As discussed earlier, statistical approaches applied to large genetic data sets are powerful tools to investigate different types of selection and demographic events that occurred in the modern human evolutionary history. However, statistical analyses based on the empirical data alone, in most cases, are not sufficient to lead to scientific conclusions. Values of the statistics are often "relative" rather than "absolute", and various uncertainties, biases and data-specific factors may skew the statistics. For example, we could use Tajima' *D* statistic to perform a genome-wide scan on 20 human genome sequences aiming to identify regions under positive or balancing selection. After we have got the *D* values across the genome, two questions will arise: (1) what significance threshold should we use to choose the interesting low and high *D* values? (2) Does a significant *D* value reflect a real signal of selection? One way to answer the first question is to rank all the *D* values and pick 0.5% or 2.5% (or other percentages) at each end of the ranking as "significant" values. The main drawback of this approach is the pre-set assumption about the proportion of outliers. If we pick 1% as significant, we are assuming that 1% of the genomic regions under investigation are under selection. This is rather arbitrary and will most likely introduce false positive or false negative results, and will not answer the scientific question of what proportion of the genome or regions under investigation are under selection, which is often an important question for researchers in genome-wide studies. To answer the second question, we need to eliminate all other factors that may contribute to the statistical results. One way to attempt this is to use various independent data sets from different sources, which ideally may not have been influenced by the same factors that could result in a significant p value, to see whether the results are replicable. This would require more time and resources, and is subject to availability of data.

Since the development of coalescent theory and the advancement of the computational capacity of computers, simulations have become a powerful tool

in population and other genetic studies. By simulating genetic data that mimic the real evolutionary process and population demographics, one can generate large sets of independent data with all features accurately known, which can then be used to assess the statistical results from empirical data. Simulation approaches can potentially answer the above two questions convincingly without any more empirical data or experimental studies being required. For example, to figure out the best significance threshold for the statistical results on a particular empirical data set, we may simulate corresponding sets of genetic data under a neutral model and appropriate demographic parameters to see what the data would look like without selection, and then a significance threshold can be set based on the distribution of the simulated neutral data. In this case, any biases of the empirical data are eliminated. If we want to figure out whether the significant statistics are real indicators of selection, we may simulate data under selection along with the neutral scenario, and compare the statistics from the two conditions to assess the power and reliability of the statistics.

Coalescent simulation was the first widely adopted approach to simulate genetic data at the sequence level. As the name suggests, this approach is based on coalescent theory, and it traces only the observed samples from the present backwards in time, ignoring the rest of the population. This provides the biggest advantage of coalescent simulation – computational efficiency. Several coalescent simulation programmes have been developed, and examples include *ms*[88], *SelSim*[89], *cosi*[26], *CoaSim*[90], and *FastCoal*[91]. Most of these programmes can simulate genetic variant data covering a few megabases or longer regions in tens or hundreds of samples, usually within a few seconds and with a reasonable amount of computational resource. Therefore, thousands or even millions of simulated data sets can be generated in a speedy manner, which is very important when p values need to be generated from the distribution of the statistics in simulated data.

However, there are some limitations of coalescent simulations. One is that the number of recombination and gene conversion events as well as the level of complexity of recombination patterns that can be incorporated into the

simulation is currently very limited. Therefore, although large genomic regions can be simulated by assuming over-simplified recombination pattern and very few recombination events, if a realistic recombination map is to be used, only a few megabases can be simulated, with a much lower speed. Another limitation is the ability to model selection events. Some of the coalescent simulation programmes cannot incorporate selection scenarios, and those that can, for example, *SelSim*, are only able to simulate the event with a single locus under selection, and this programme is restricted to conditions like a relatively short genomic region and small sample size, a constant population size, and a uniform recombination rate.

These limitations can be resolved by a forward simulation approach, which simulates genomic data forward in time from an ancestral status. Tracking the evolutionary process forward in time allows a high level of flexibility; therefore, complex recombination patterns and demographic parameters can be incorporated. This approach obviously requires the simulation of the whole population, so is computationally very expensive. Even with large computer clusters, the speed and computational resource requirement of forward simulations have prevented this approach from being used in generating large data sets. However, its high flexibility is still appealing for certain studies. A few pieces of forward simulation software have been developed. One example is *simuPOP*[92], which was designed as an interactive programme, allowing users to manipulate the models and parameters during the evolutionary process and enabling highly flexible simulations. Later-developed forward simulation tools incorporated rescaling techniques to enhance the computational efficiency. Basically, these algorithms allow the user to divide population sizes and numbers of generations by a small factor x (usually 5-10), and increase the mutation and recombination rates by that same factor. By doing this, the parameters at the population level (e.g. $\theta = 4 N_e \mu$) remain unchanged, while the speed of the simulation can increase up to x^2 fold. The simulation programmes *FREGENE*[93] and *mpop*[94] are examples of this type. The increased computational efficiency of these programmes allows large-scale forward simulations with selection scenarios and complex recombination patterns and demographic

models.

1.5.2 Validation by independent data sets and/or approaches

Although simulation is a powerful tool in assessing the overall effectiveness of statistical approaches in large data sets, after candidate regions or genes are shortlisted, more validations are needed to verify the signals of selection. One intuitive way is to use alternative data sets or approaches to investigate the same question, and if the results are replicated independently, they are more likely to be reliable. Three approaches can be taken in this type of validation: (1) using different statistical methods on the same data; (2) using the same statistical methods on different data; and (3) using different statistical methods on different data. The decision of which approach to use is of course restricted by the availability of alternative data or methods, and also depends on the purpose of the study as well as the reliability of the data and methods that have been used. The first approach is best suited when a new, comprehensive and high-quality data set becomes available, which can be used in different ways, or when there are multiple methods that capture different aspects of the features under study. For example, the HapMap project provided a highly reliable and comprehensive data set of human SNPs and haplotypes, which enabled genome-wide studies of natural selection in the human genome. Voight et al. first developed a new LD-based statistical method to detect positive selection, and applied it to the HapMap data[67]. This study generated a genome-wide map of recent positive selection, though most of the regions were not validated by other approaches. Sabeti et al. then applied three LD-based statistical tests to the ~3 million SNPs from HapMap2 data[66], yielding fruitful results with a high-confidence list of positively selected regions showing strong signals in multiple tests. The second approach is suitable if the methods used are potentially powerful but new and/or untested, and if there are multiple sets of data available to test the robustness of the methods from different angles. For example, Nielsen et al. applied their newly-developed CLR methods on both Seattle SNPs data and the HapMap data, which are two independent data sets, to test their methods[76]. The third approach is most desirable if a scientific conclusion is to be drawn from the study, yet all evidence is based on limited statistical investigations on limited

data, thus more evidence is needed. This approach can be the most powerful among the three, since if a candidate gene shows signals multiple times in completely independent investigations of different data sets using different methods, it will be most convincing and less likely to be a false positive. A good example of such a candidate is the Duffy blood group locus mentioned earlier. Multiple independent studies revealed signals of positive and possibly other types of selection acted on this locus[61,65,78], making it a good example of recent positive selection on disease resistance in a human population, and also attracted interest from clinical researchers. However, caution needs to be taken in choosing the data and methods when applying this approach, so that the results are comparable and free from biases that may jeopardize the validity of the comparison and validation.

1.5.3 Validation by functional studies

One of the main purposes for all the efforts made in the identification of positively selected regions in the human genome is to aid a better understanding of human genomic functions, as well as provide insights into studies in human diseases and healthcare. Therefore, the real functional targets of positive selection must be sought after candidates are identified by statistical approaches. If a plausible functional target is identified within the candidate region, and the function is likely to affect the carrier's fitness, it is more plausible that positive selection may have acted on this candidate than if no function is related to the candidate. Therefore, looking for functional targets of positive selection within or near the candidate regions is the ultimate way to validate statistical results. For example, a few pigmentation-related genes showed strong signals of positive selection in non-African populations in several studies[66,95,96]. This can be explained by the climate differences between areas in the world. In areas with higher temperature and more exposure to sunshine, darker skin is selected to prevent sunburn, while in colder and less sunny areas, the skin can become lighter in colour, perhaps to allow production of vitamin D or because of sexual selection[97,98]. A functional study on the SLC24A5 gene revealed its critical role in human pigmentation, and a functional coding polymorphism with highly

differentiated frequencies between African and other populations[99] was identified, which provided strong functional evidence for selection in this gene.

If a candidate region contains one or more protein-coding genes, intuitively one of the genes would be thought as the most likely selection target. However, a large proportion of the candidate regions from genome-wide scans of positive selection are either too large so that functional targets cannot be pinpointed, or lie in intronic or intergenic regions in the genome where there is no obvious functional element. This can be seen as both a challenge and an opportunity. The challenge is, on the one hand, the difficulty of identifying putative selection targets in the "non-functional" region, and on the other hand, the lack of validation of whether the statistically-significant candidates are true or false. However, "no known function" is not equal to "no function". The signals of positive selection in "non-functional" regions may be seen as a sign of unknown functional importance of the genomic regions, and thus worth pursuing further by functional investigations. Statistical analyses can serve as a means of identifying candidates for experimental biologists to study potential functions, which will lead to a better understanding of functional elements in our genome. One should also note that experimental studies often take years and require huge amounts of resources; therefore, a high-quality list of candidates will be tremendously helpful for enhancing the efficiency of such research.

1.6 Aim of this thesis

The main goal of this dissertation is to detect regions in the human genome that have been positively selected during the course of modern human evolution, taking advantage of the abundance of genome sequencing data, and to localize the selective target to a small genomic region, so that putative functional variants under selection can be identified. Within this general goal, this thesis is aiming to answer three fundamental questions: (1) can sequencing data help better detect positively-selected regions and localize selection targets when frequency-spectrum based statistical tests are applied? (2) If the answer to the first question is yes, can novel positively selected regions be identified and selection targets be localized if such an approach is applied on whole-genome sequencing

data from worldwide populations? (3) By calculating time to the most recent common ancestor (TMRCA) from sequencing data, can we identify regions that were selected during the early stage of modern human evolution, which are not detectable by available statistical neutrality tests?

Three studies will be presented in this dissertation to answer these questions.

(1) Exploration of signals of positive selection derived from genotype-based human genome scans using re-sequencing data. The aim of this project was to localize selection targets in candidate regions identified by LD-based tests on genotype data, by applying frequency-spectrum based tests (Tajima's *D*, Fay and Wu's *H*, and a Composite Likelihood Ratio test) to targeted resequencing data. Two candidate regions from the HapMap2 scan for positive selection[66] were resequenced, and likely selection targets in both regions were narrowed down from ~300 kb to ~30 kb. Plausible biological targets of selection could be proposed for both regions.

(2) A genome-wide scan of selective sweeps using frequency-spectrum based tests on 1000 Genomes Project low-coverage Pilot whole-genome sequencing data. The aim of this project was to provide a map of positively-selected regions in the human genome, with a higher power of detection and better resolution. Comprehensive simulations were performed to understand the power of our combined score of frequency-spectrum tests for detecting and localizing selection targets. A high-confidence list of positively selected genes was produced in each of the three populations (African, European and Asian), with highlights of some strong candidates with clear functional implications. Bioinformatic functional analyses were performed to reveal the general features of selected genes, as well as detailed understanding of the likely selection targets in the strongest candidates.

(3) A genome-wide scan for regions with recent common ancestry among all humans. This project aimed to identify regions in the human genome that have been positively selected during early modern human evolutionary history, as regions with shared recent coalescent times indicate positive selection affecting all modern humans, which has an older age than the recent positive selection

identified by neutrality tests. Coalescence times were calculated using the GENETREE package[86] in 5kb windows across the genome from high-coverage whole-genome sequencing data of 54 unrelated samples from 11 populations around the world, produced by Complete Genomics Inc.. Simulations showed that there might not be an excess of recently-coalesced regions in all humans, although there are some regions with recent TMRCAs. Regions with a TMRCA of less than 400,000 years were identified, and variants within those regions were compared with the sequence of the Denisovan genome. Phylogenetic network analyses were performed on some of the regions with recent TMRCAs.

These three studies together build up a basic yet comprehensive investigation of positive selection in the human genome using sequencing data, and provide an understanding of how the availability of multi-population, large-scale sequencing data will propel and enable insightful human evolutionary studies that could not be done before.