# Positive Natural Selection in the Human Genome

Min Hu

Darwin College

University of Cambridge

August 2012

This dissertation is submitted for the degree of Doctor of Philosophy

UNIVERSITY OF CAMBRIDGE

# Declaration

This thesis describes my work undertaken at The Wellcome Trust Sanger Institute, in fulfillment of the requirements for the degree of Doctor of Philosophy, at Darwin College, University of Cambridge. This dissertation is the result of my own work and contains nothing that is the outcome of work done in collaboration, except where specifically indicated in the text. The work described here has not been submitted for a degree, diploma, or any other qualification at any other university or institution. I confirm that this thesis does not exceed the word limit set by the Biology Degree Committee.

Min Hu

Cambridge, August 2012

# Acknowledgements

I can hardly express in words how thankful I am to people who made my past four years full of growth, joy and cheer, and who made this thesis possible. In 2008, excited but a little unsure, I entered the world of cutting-edge science in genomics and genetics, and for the first time, I started my life in a country with a different language, unfamiliar culture and perhaps many more opportunities. Big thanks to Matt Hurles and Alex Bateman, who kindly provided me with opportunities to rotate in their labs when I knew very little about the subjects, and taught me the basics of being a good scientist. Great thanks to Don Conrad and Ni Huang in Matt's group, who patiently taught me all the important and trivial things I needed to know to start coding and using the powerful Sanger computing farm.

The most grateful thanks go to my supervisor, Chris Tyler-Smith. He has always been extremely generous in devoting his time to teach and guide me in my research and scientific writing. There had been countless idea exchanges in our weekly meetings during the last three years, and lots of encouragements when I was facing challenges in my projects. Big thanks to Yali Xue, who held my hands throughout my very first project in the group and took care of me in every aspect; Qasim Ayub, who conducted excellent experiments for me in the lab; and Yuan Chen, who helped me a lot in retrieving data from the databases. Great thanks to previous colleagues Daniel MacArthur and Bryndis Yngvadottir, who shared with me lessons they learned from their PhD studies, and provided me with lots of support. Also thanks go to other members in team 19, who had made my life joyful with banters, parties and BBQs, and who had always been caring and helpful in difficult times.

Great thanks to Toomas Kivisild, my external supervisor, and to Jeff Barrett and Alex Bateman in my thesis committee, who provided me with valuable suggestions and ideas, as well as kind support and praise. Also thanks to Annabel Smith and Christina Hedberg-Delouka, who provided support on each critical

step in my PhD. Great thanks to Wellcome Trust for the excellent PhD programme and generous scholarship.

I was lucky enough to get involved in the 1000 Genomes Project, collaborating with excellent scientists from allover the world. I enjoyed all the meetings, conferences and phone calls, and learnt a lot from this perhaps one of the world's biggest research projects in life sciences. Big thanks to all the participating scientists in the 1000 Genomes Project, who had been extremely helpful in providing data and exchanging ideas.

Finally, I would like to thank all my friends and my family, without whom I would not have been where I am today. Thanks to my fellow PhD students at Sanger, who formed a fun and supportive community around me. Thanks to all the friends I made during my time in Cambridge, who had made my life abroad more enjoyable than I could have ever hoped for. Special thanks to my parents, who had always been supportive to every decision I have made in my life.

# Publications

Publications arising during the course of the work described in this thesis by the time of submission:

**Hu M**, Ayub Q, Guerra-Assunção JA, Long Q, Ning Z, Huang N, Romero IG, Mamanova L, Akan P, Liu X, Coffey AJ, Turner DJ, Swerdlow H, Burton J, Quail MA, Conrad DF, Enright AJ, Tyler-Smith C and Xue Y (2012). Exploration of signals of positive selection derived from genotype-based human genome scans using re-sequencing data. *Human genetics*, 131;5;665-74.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, **Hu M**, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, 1000 Genomes Project Consortium, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB and Tyler-Smith C (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* (New York, N.Y.), 335;6070;823-8.

1000 Genomes Project Consortium (including **Hu M** and Tyler-Smith C) (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467;7319;1061-73.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, **Hu M**, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW and Hurles ME (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464;7289;704-12.

# Abstract

The detection of positive natural selection in the human lineage is of great interest for the understanding of modern human phenotypes and adaptations to different environmental conditions. Although extensive genome-wide scans for signatures of positive selection have been performed using genotype data, these have significant limitations, illustrated by the low overlap among different studies. Thanks to the Next-Generation Sequencing technology, near-complete sequence data for both the whole genome and targeted regions are now available, allowing a nearly unbiased genome-wide scan for positive selection as well as the possibility of localizing the specific variants selected.

The theme of this PhD thesis is to detect and localize positive selection targets in the human genome using sequencing data. This includes three projects:

(1) Localizing 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 regions were resequenced at high coverage and putative selection targets were identified.

(2) A genome-wide scan of selective sweeps using frequency-spectrum based tests on 1000 Genomes Project low coverage Pilot data. Candidate positively selected regions and genes were identified and some interesting examples and their plausible selected functions are discussed.

(3) A genome-wide search for regions with very recent ancestry among all humans. Regions with shared recent coalescence times indicate positive selection affecting all modern humans, which has an older age than the recent positive selection identified by neutrality tests. We calculated the Time to the Most Recent Common Ancestor (TMRCA) of low diversity/divergence regions in the human genome, with the aim of identifying regions with very recent common ancestor, which may have been positively selected during early modern human evolution.

These three projects altogether demonstrated the value and impact of low-coverage or high-coverage, targeted or whole-genome sequencing data on providing new insights into positive natural selection in the modern human history, and built up the first steps of the exciting new sequencing era for the exploration of human evolution.

# Abbreviations

| | |
|---|---|
| aCGH | array-comparative genomic hybridization |
| ASW | African ancestry in Southwest USA |
| CEU | Utah residents with European ancestry |
| CGI | Complete Genomics Inc. |
| CHB | Chinese Han in Beijing |
| CLR | composite likelihood ratio |
| cM | centimorgan |
| CMS | composite of multiple signals |
| CNV | copy number variant |
| CRT | cyclic reversible termination |
| DAF | derived allele frequency |
| DBP | diastolic blood pressure |
| DNA | deoxyribonucleic acid |
| EHH | extended haplotype homozygosity |
| ENCODE | encyclopedia of DNA elements |
| eQTL | expression quantitative trait loci |
| FDR | false discovery rate |
| FoSTeS | fork stalling and template switching |
| Gb | gigabases |

| | |
|---|---|
| GIH | Gujarati Indian in Houston |
| GWAS | genome wide association studies |
| HLA | human leukocyte antigen |
| IQR | interquartile range |
| JPT | Japanese in Tokyo |
| kb | kilobases |
| KYA | thousand years ago |
| LD | linkage disequilibrium |
| LSA | later stone age |
| LWK | Luhya in Webuye, Kenya |
| MAF | minor allele frequency |
| Mb | megabases |
| MHC | major histocompatibility complex |
| miRNA | micro RNA |
| MKK | Maasai in Kinyawa, Kenya |
| MP | middle Paleolithic |
| MRCA | most recent common ancestor |
| mtDNA | mitochondrial DNA |
| MXL | Mexican ancestry in Los Angeles |
| MYA | million years ago |
| NAHR | non-allelic homologous recombination |

| | |
|---|---|
| ncRNA | non-coding RNA |
| NCS | non-coding sequences |
| NGS | next generation sequencing |
| NHEJ | non-homologous end joining |
| NHGRI | National Human Genome Research Institute |
| OoA | out of Africa |
| PCR | polymerase chain reaction |
| piRNA | piwi-interacting RNA |
| PUR | Puerto Rican in Puerto Rico |
| PWM | position weight matrix |
| RFLP | restriction fragment length polymorphism |
| RNA | ribonucleic acid |
| rRNA | ribosomal RNA |
| SBS | sequencing by synthesis |
| siRNA | small Interfering RNA |
| SNP | single nucleotide polymorphism |
| snRNA | small nuclear RNA |
| SNV | single nucleotide variant |
| SV | structural variant |
| TF | transcription factor |
| TIRF | total internal reflection fluorescence |

| | |
|---|---|
| TMRCA | time to the most recent common ancestor |
| tRNA | transfer RNA |
| TSI | Toscans in Italy |
| UP | upper Paleolithic |
| VNTR | variable number tandem repeat |
| XP-EHH | cross-population extended haplotype homozygosity |
| YRI | Yoruba in Ibadan |

# Table of contents

# 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].

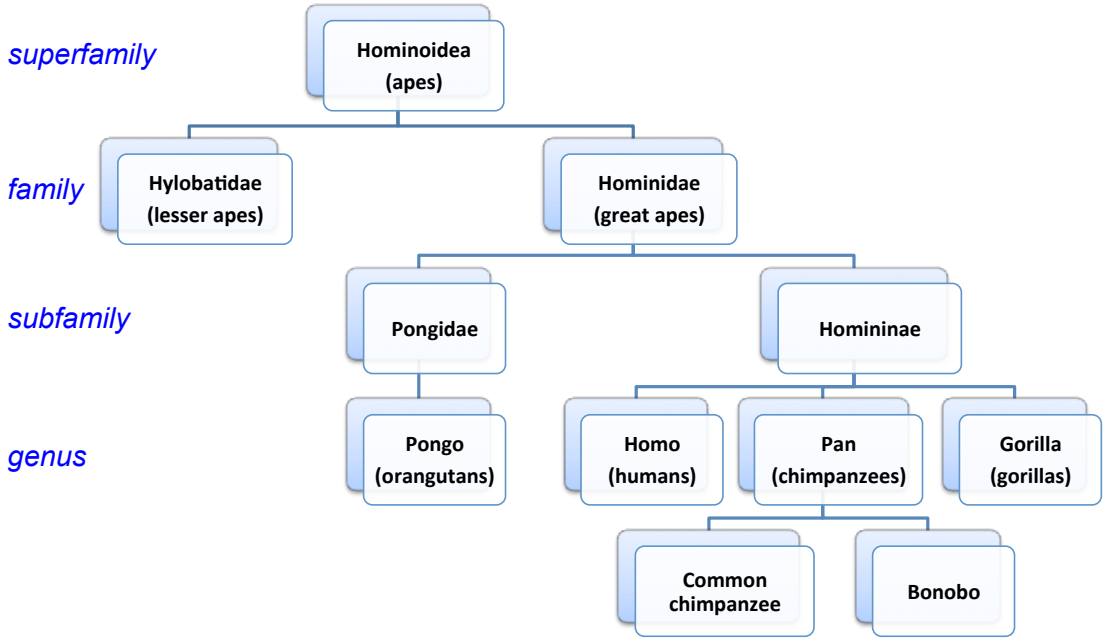


**Figure 1.1 The species tree of apes.** Please note that this tree is not a complete tree and some branches of the species tree of apes are not shown. Emphasis is on great apes and only extant genera are shown.

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).

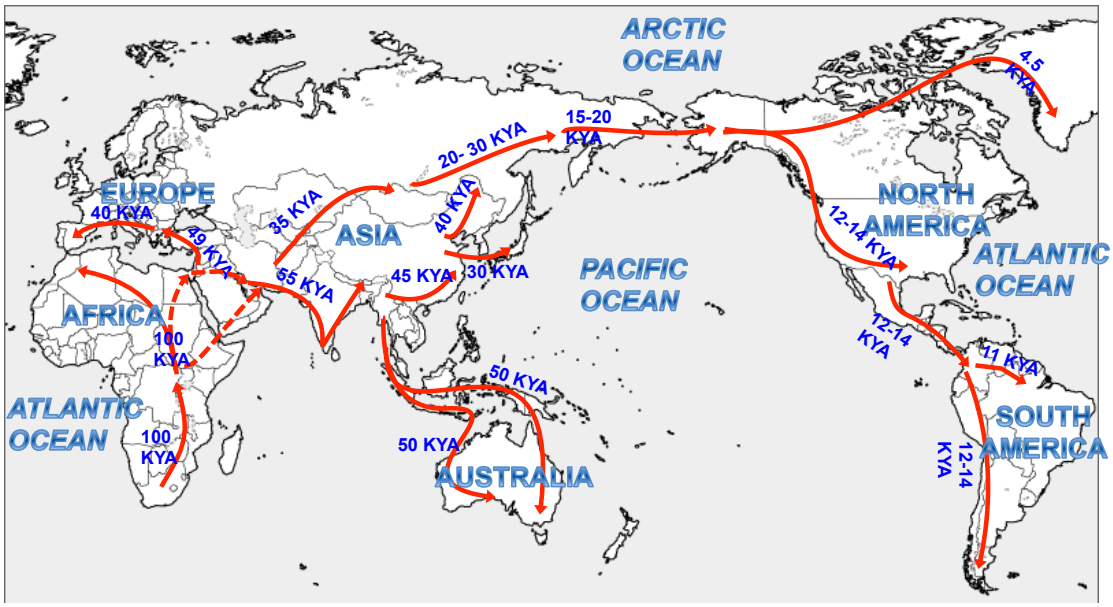


**Figure 1.2 Map of human expansions.** This map shows the putative migration routes and dates of early modern human migrations from Africa to other parts of the world. Red arrows indicate the possible routes, and estimated dates of migration are shown in blue text (KYA: thousand years ago). Note that 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.

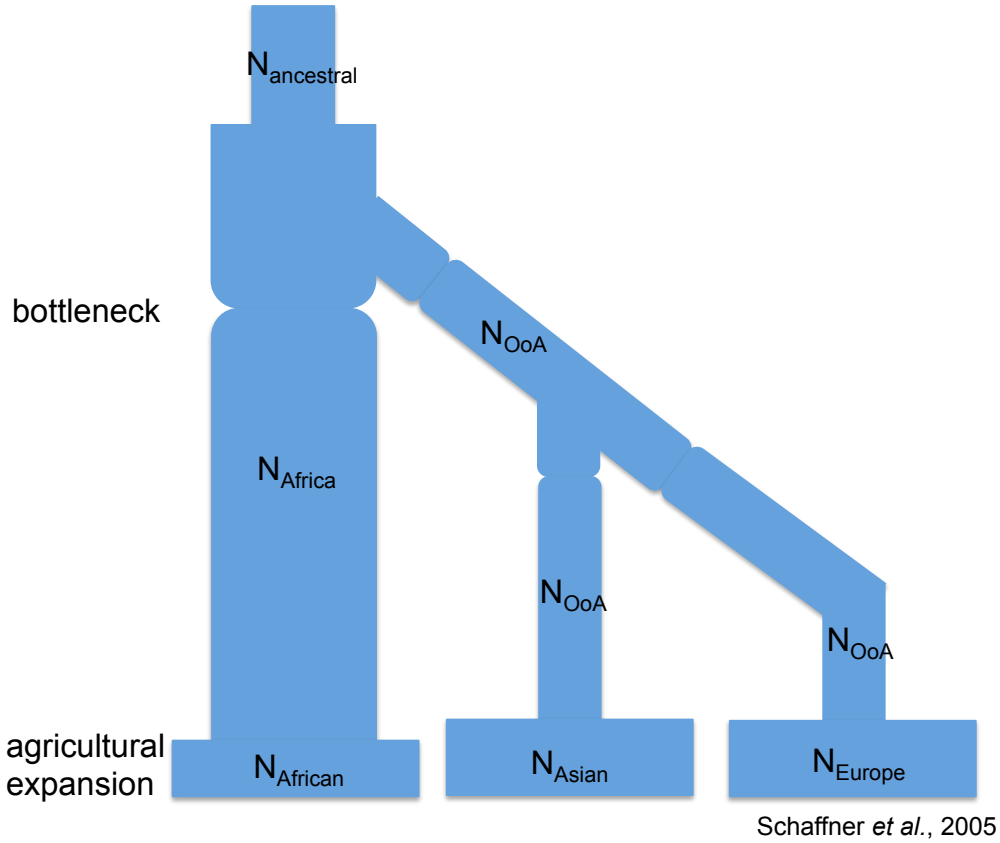


**Figure 1.3 A best-fit demographic model.** The widths of the bars represent relative population sizes (noted as N in the figure). Bottlenecks are represented by dents in the bars. This figure was adapted from Schaffner et al. 2005.

## 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).

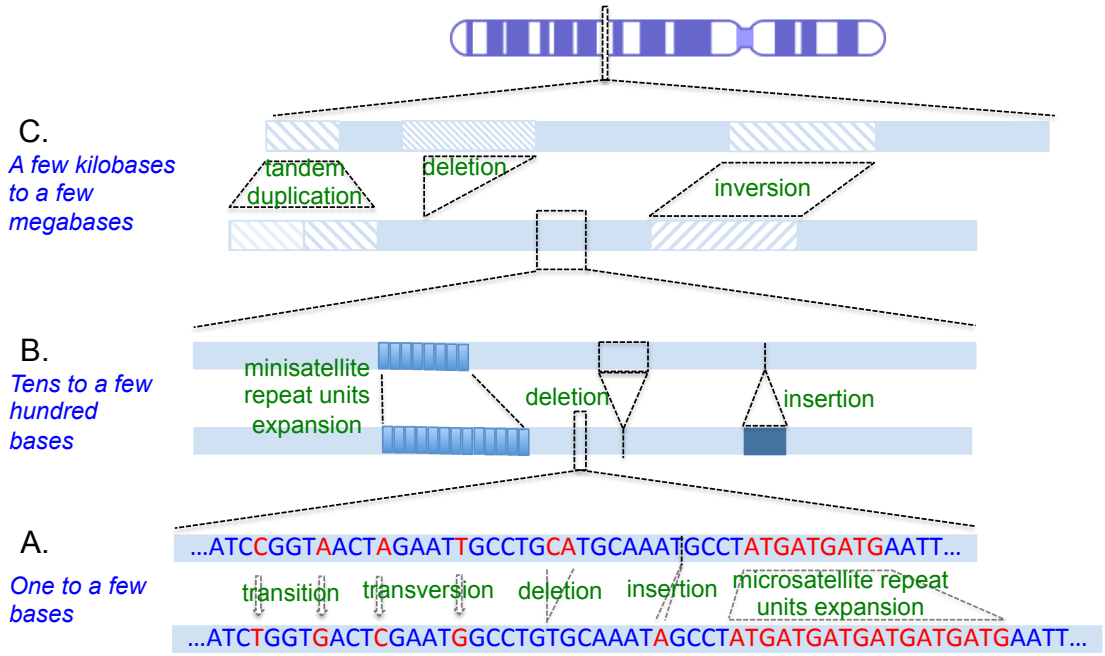


**Figure 1.4 Types of genomic variation.** A. Examples of transitions, transversions, a single base insertion, two-base deletion, and mutation of repeat unit number of a three-base microsatellite. B. Examples of minisatellite repeat unit number mutation, deletion and insertion of segments of DNA. C.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

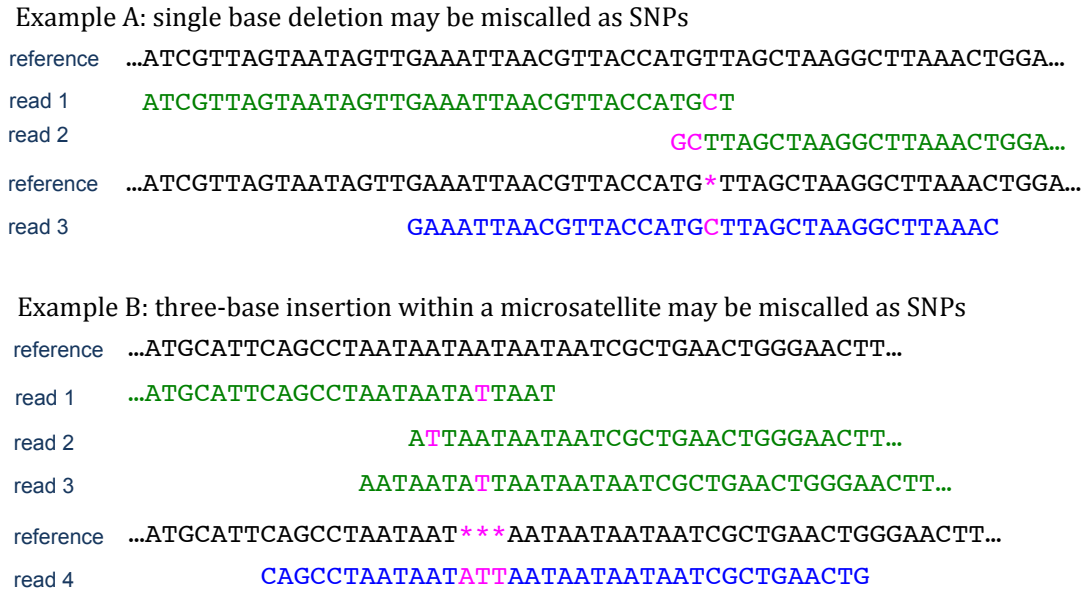


**Figure 1.5 Examples of misalignment and miscall.** In both examples, black letters are reference sequences, green letters are the reads where miscalls occur, and blue letters are the reads where variants are called correctly. Magenta letters are the variants called. If there is insertion, stars are 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.

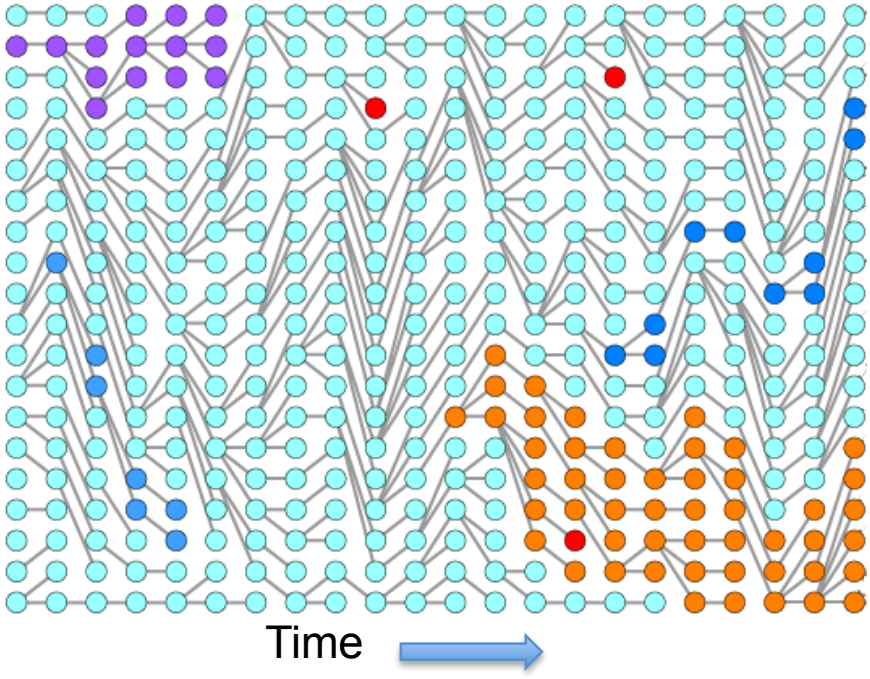


**Figure 1.6 Genetic drift in a population.** Different colored circles represent different variants in the population. In a Wright-Fisher population, genetic drift drives frequencies of variants up and down by chance, and a variant will eventually disappear or get fixed in the population.

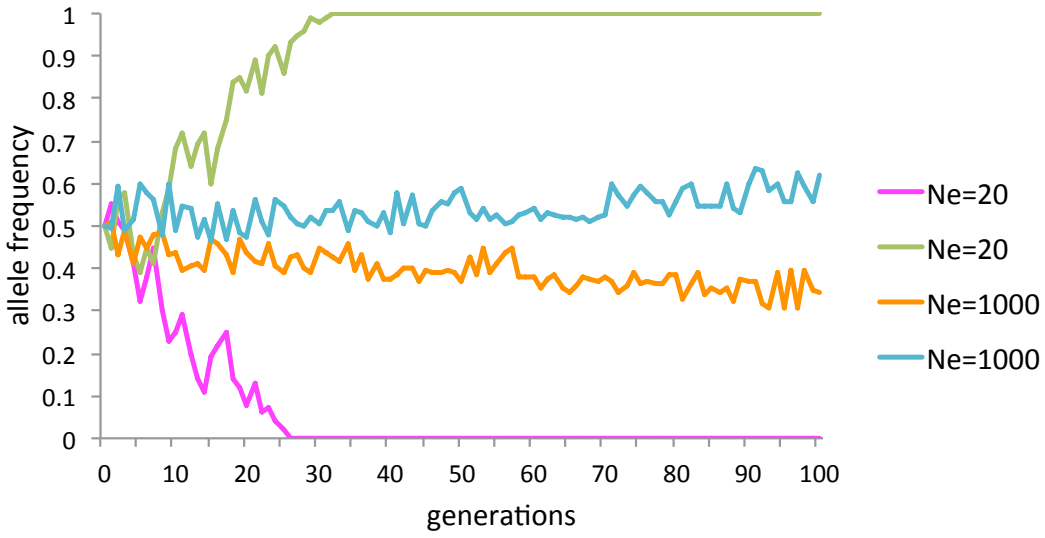


**Figure 1.7 Genetic drift in populations with different effective population sizes.** This figure shows the change of the frequency of one allele with an initial frequency of 0.5, in populations with effective population sizes of 20 or 1000. In a population with a smaller effective population size, 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.

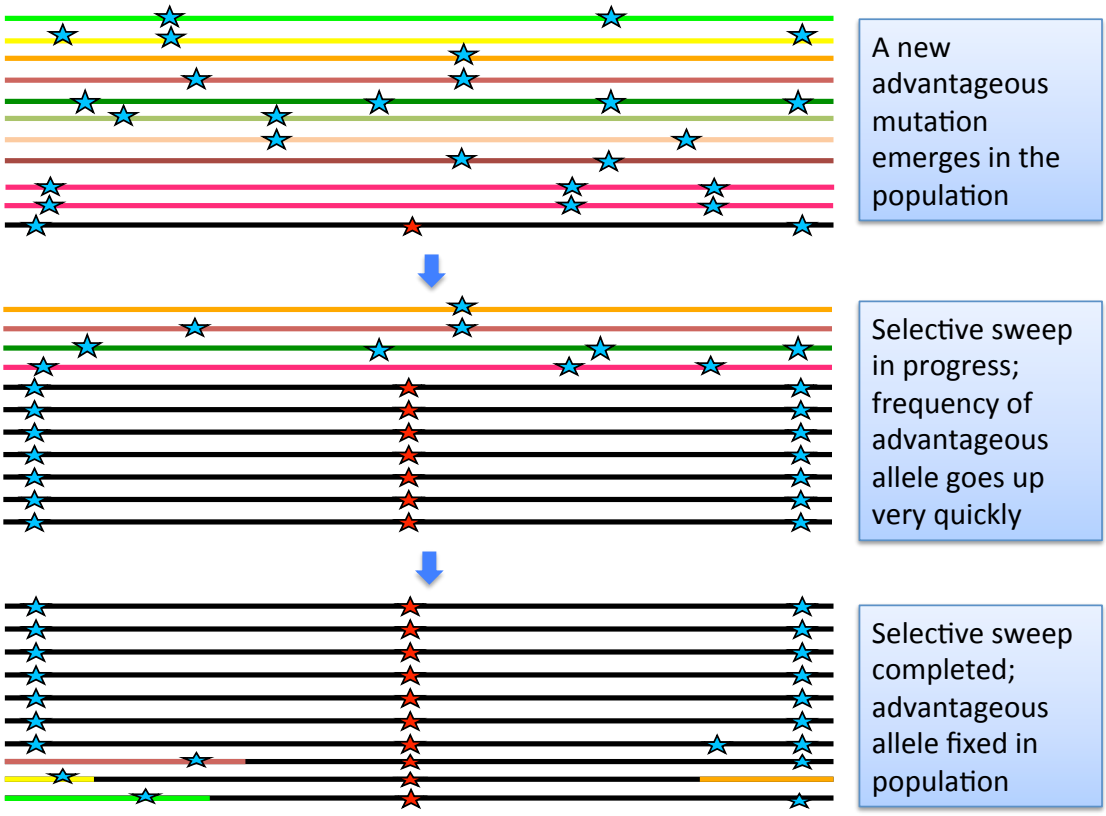


**Figure 1.8 A selective sweep.** Different colored lines represent different haplotypes in the population. Blue stars are neutral mutations, and the red star is the advantageous mutation under positive selection.

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).

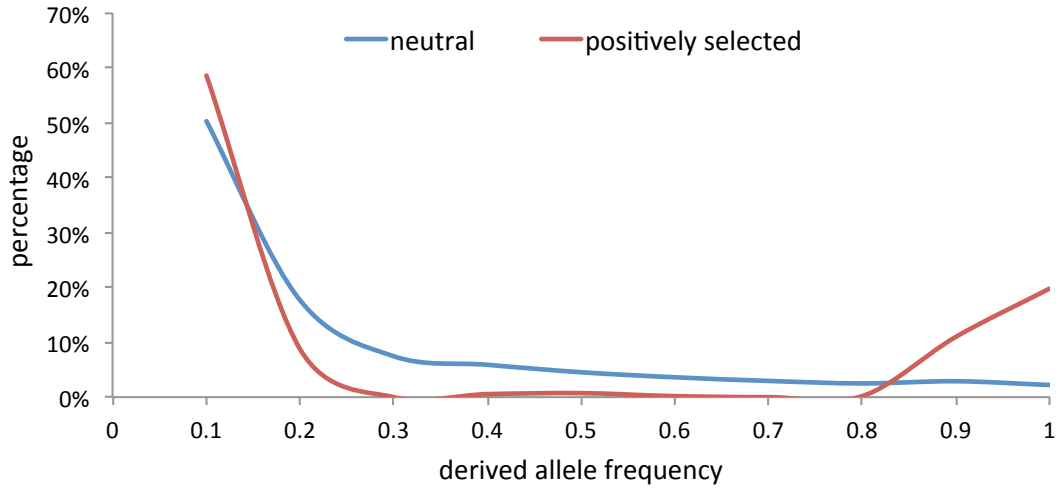


**Figure 1.9 Derived allele frequency spectrum of a positively selected region versus a neutral region.**

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 ($\pi$), 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 $\theta$, 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 $\theta_H$ is the estimate of $\theta$ 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 $(k - 1)/(k + 1)$.

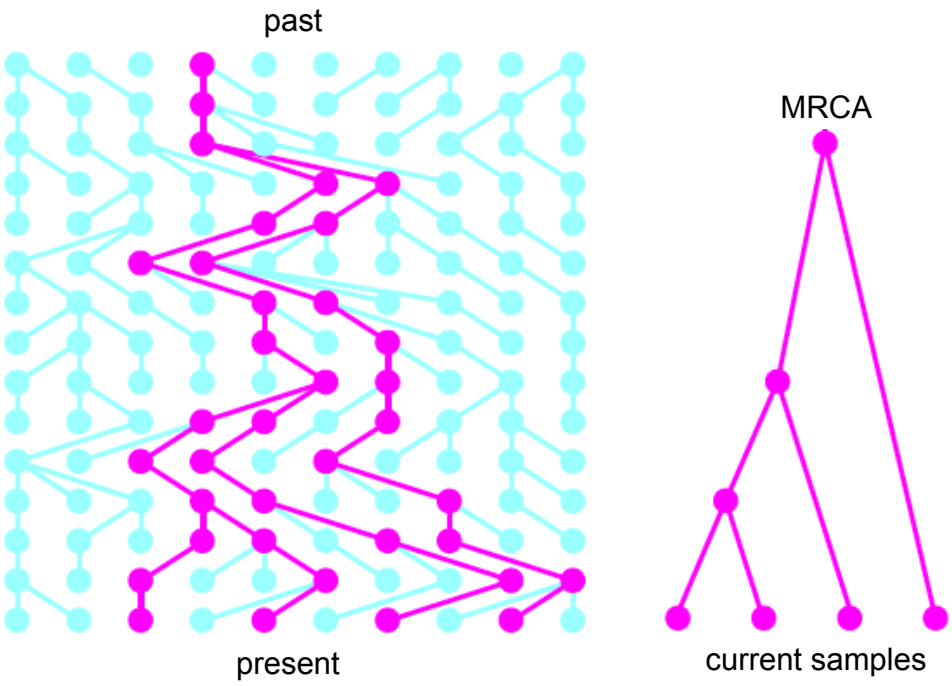


**Figure 1.10 The coalescent.** Purple circles in each generation are those being traced backwards in time until reaching the common ancestor.

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.

# 2 Exploration of signals of positive selection derived from genotype-based human genome scans using re-sequencing data

## 2.1 Introduction

A genome-wide scan of positive selection, in which the entire genome is examined, has been used in several studies. In some scans, such as when non-synonymous amino-acid substitutions showing high levels of population differentiation were chosen[83], there has been a limited prior hypothesis about the target of selection. But genome scans can also be carried out in the absence of any such hypothesis. Such unbiased scans have the attractive feature that they can potentially lead to entirely unsuspected insights into the evolutionary history, but in order to derive full benefit from them, the target of selection must be identified. In practice, most genome scans have been based on SNP genotyping, and methods for detecting potential selection have been primarily based on searching for unusual LD or population differentiation patterns. Such scans have, in some senses, been highly successful. A review summarizing the combined results of nine such genome scans found that 5,110 distinct regions covering 14% of the genome and 4,243 (23%) RefSeq genes showed apparent evidence of positive selection[100]. However, although these findings are impressive for their yield of putatively selected regions, it was notable that there was limited overlap between the individual surveys and only 129 of the regions (2.5%) were identified in four or more studies. This poor concordance was described as "sobering"[101] and pointed to the need for a better understanding of the false positive and false negative rates in such scans. Indeed, other analyses have suggested that the classic selective sweeps detected by these approaches are unlikely to have been frequent enough to dominate overall patterns of human genome diversity[102]. A second feature of some of these scans, particularly those based on LD, is that the candidate regions identified can be very large. For example, the HapMap2 project listed 22 strong candidate regions with a

combined length of ~16.7 Mb and mean size of ~760 kb[66], making it difficult to identify the selected target and further investigate the biological implications of the selection.

We have set out to address three questions raised by genome scans that identify large candidate regions. First, do such candidates show evidence for selection if alternative criteria are used? Second, to what extent can the targets of selection be localized more precisely? And third, if more precise localization is possible, does this lead to increased insights into the possible biological basis of the selection? To achieve these aims, we reasoned that full re-sequence data would provide the most information. Indeed, only technical and cost limitations have previously hindered its use: re-sequencing complete genomes or even hundreds of kilobases (kb) to high accuracy in population samples has not been practical until recently[40]. We have thus explored experimentally the potential for enrichment of such regions followed by next-generation sequencing to generate suitable datasets. We chose for these trials two regions from the HapMap2 survey, which were of intermediate size (~300 kb each) and where there was no obvious target for selection[66]. We show using simulations that alternative tests for selection applied to sequence data from regions identified in such a way should readily distinguish between neutrality and likely selection, and will usually produce a more precise localization of the selected variant. We also show experimentally that suitable high-quality sequence data can be generated using next-gen technology, and finally that plausible biological candidates can then be proposed for these selective events.

This study is published in *Human Genetics*[103]. This chapter is based on this publication, with some modification of the contents. All simulations, statistical and bioinformatic analyses were performed by the author of this thesis, except the CMS calculation, which was done by Irene Gallego Romero. The PCR experiments were done by Qasim Ayub, and the sequencing work was done by the Sanger Institute Sequencing Team.

## 2.2 Materials and Methods

### 2.2.1 Simulations

Two-step simulations were performed to model both neutral and positively selected scenarios, and are summarized in Figure 2.1. In the first step, we carried out coalescent simulations using the *cosi* package to generate 1 Mb long haplotypes in a pair of ancestral populations 2,000 generations ago based on the best-fit demographic models for African and Asian populations[26]. These haplotypes were then used as input for the second step - forward simulations, using *mpop*[94]. In some of these forward simulations in the Asian population, one allele with an initial frequency of 0.0006 (default initial frequency for new variants in the package), which would be under selection, was added in the middle of the simulated Asian haplotypes. Four different selection scenarios with selection coefficients ($s$) of 0.001, 0.004, 0.007 and 0.01 were simulated, and the selection start time was set at 2,000 generations ago. In total, 1,000 independent simulations were performed for each set of conditions. These used the genome-average recombination rate of 1cM/Mb from the HapMap2, a mutation rate of $1.8\times10^{-8}$ per nucleotide per generation calculated from a comparison of human and chimpanzee sequences for the whole of chromosome 4, and a current effective population size of 100,000. The rest of the demographic parameters were as in Schaffner et al.'s best-fit demographic model[26] from the package *cosi*. For the purpose of computational efficiency, we re-scaled the parameters when performing the forward simulations: effective population sizes and times were reduced by a factor of 5, while mutation and recombination rates and selection coefficients were multiplied by 5 (see Appendix A for parameters and commands). Fifty chromosomes were sampled from each simulation. We call this set of data the "simulated re-sequencing data".

The SNPs in the "simulated re-sequencing data" were subsampled to mimic the frequency spectrum of HapMap2 genotype data by matching the proportion of the SNPs of HapMap2 data in each frequency bin (bin size 0.1). We call this set of subsampled simulation data the "simulated genotype data". XP-EHH scores[66] were calculated from the simulated genotype data and normalized using the

mean and variance of the XP-EHH scores from the simulated genotype data in the neutral simulation in the Asian population, using the African as the reference population. We only retained simulations with the XP-EHH score above the 95$^{th}$ neutral percentile continuously for at least 100kb surrounding the selected SNP, which mimics the experimentally-investigated candidate regions from the survey based on the HapMap2 data.

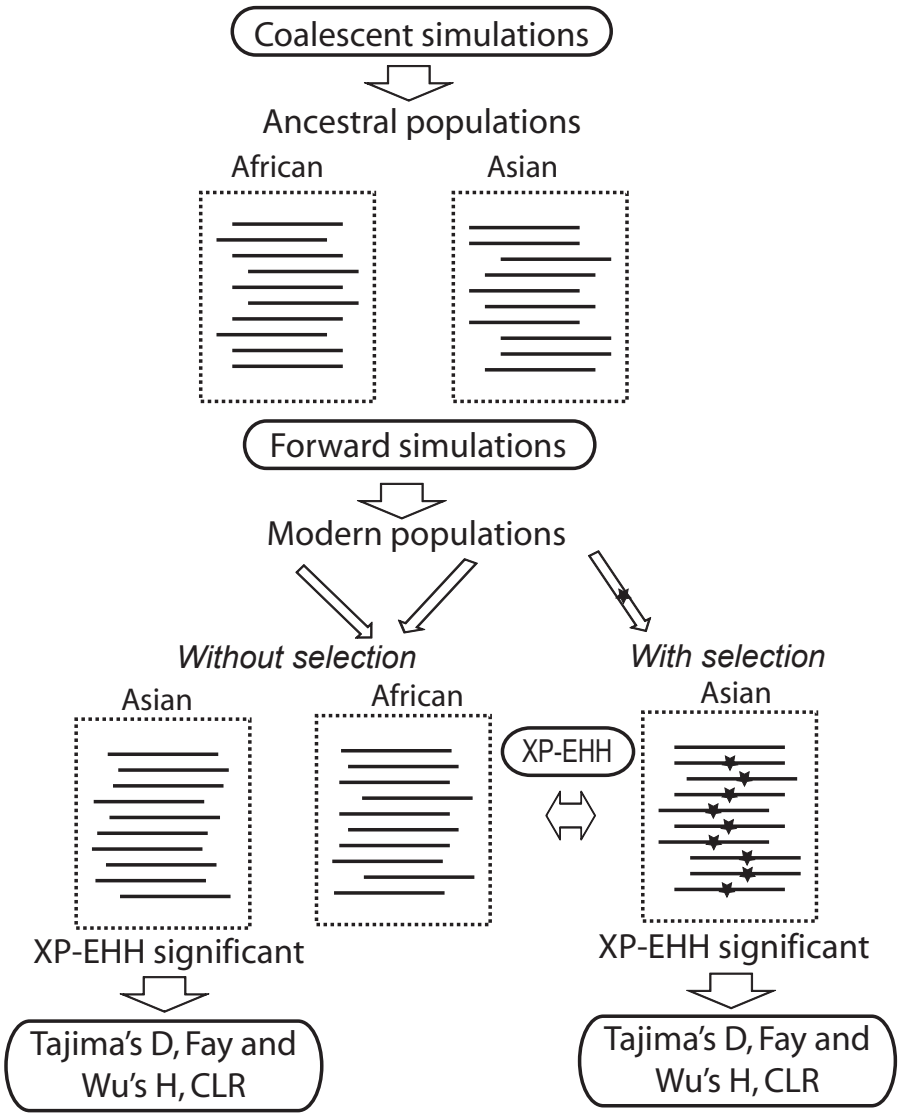


**Figure 2.1 Simulation design.** Dotted boxes represent simulated haplotype samples; the star indicates the presence of a positively selected SNP. Arrows show the performance of the analyses described in the oval boxes.

We then returned to the corresponding simulated re-sequencing data for the retained simulations and calculated Tajima's $D$[71], Fay and Wu's $H$[72] and Nielsen et al.'s CLR[76] statistics. These were calculated in 10 kb non-overlapping windows

across the whole 300 kb region centered on the selected SNP (or equivalent location in neutral simulations) in each individual set. The significance levels for each of the neutrality tests were estimated based on the percentile of the test values in the null distribution from 1,000 neutral simulations with the same demographic model. The background frequency spectrum required by the CLR analysis was calculated on the 1,000 independent neutral simulations with the same recombination and mutation rates. In order to combine signals from the three tests, we assessed the correlation coefficient between Tajima's *D* and Fay and Wu's *H* p values on the neutral simulated data, and found no correlation (r = 0.06); therefore, these two tests were treated as independent, and a combined p value from Tajima's *D* and Fay and Wu's *H* for each 10kb window was calculated using Fisher's method[104], and we use this combined p value to present the results below.

### 2.2.2 Target region resequencing

Two regions were picked from the HapMap2 list of 22 regions showing strong evidence of selection[66] using the following criteria: no obvious candidate for the selected SNP or gene; selection at least in the CHB+JPT population; moderate size (0.2-1 Mb). The coordinates of the chosen regions were (March 2006, NCBI 36 assembly; all genomic coordinates in Chapter 2 are based on this assembly) chromosome 4: 158,702,285-159,016,211 (314 kb, called chr4:158Mb) and chromosome 10: 22,587,453-22,850,110 (263 kb, called chr10:22Mb). We also included a set of control regions, including CASP12 (13 kb) for which we had the Sanger capillary sequencing data from a subset of the samples for the resequencing of this study[105] and 20 kb of unique sequence from the Y chromosome, where there should be no reads mapped in females and no heterozygote calls in males.

The target regions were then amplified from 28 CHB (Han Chinese in Beijing, China) and 2 YRI (Yoruba in Ibadan, Nigeria) samples from the HapMap collection in a series of long-range PCRs. In total, 49 pairs of PCR primers were designed for chr4:158Mb, 42 for chr10:22Mb and 4 pairs for the Y chromosome to amplify 5-11 kb PCR products with overlap of > 500 bp, using a Perl script

(http://droog.gs.washington.edu/PCR-Overlap.html). Two previous pairs for *CASP12*[105] were also used. The three base pairs at the 3' end of all primers were confirmed not to overlap with any SNP in dbSNP127 (http://www.ncbi.nlm.nih.gov/projects/SNP/). The primer sequences and PCR conditions are listed in Appendix B, PCR primers and protocols. 44 out of 49 fragments from chr4:158Mb, 37 out of 42 from chr10:22Mb and all from the Y chromosome and *CASP12* were successfully amplified in initial tests. These fragments were subsequently amplified in all samples. Three CHB provided poor quality data for chr4:158Mb, and four for chr10:22Mb, and were excluded from all subsequent analyses. Amplification was tested by agarose gel electrophoresis followed by ethidium bromide staining, and approximate quantification was performed from the band intensity. 39 out of 49 (~80%) long PCR primer pairs worked well for 22 or more samples for chr4:158Mb, and 32 out of 42 (~75%) for 20 or more samples for chr10:22Mb. The lab work of PCR enrichment was done by colleague Qasim Ayub. The PCR products from each individual sample were pooled, approximately equalizing the molar yield for the Illumina sequencing paired end library construction.

In order to avoid artifacts in tests results due to the missing data in PCR gaps, we used another set of data from a hybridization enrichment experiment based on a Nimblegen custom array or solution pulldown approach[106] on the same two regions in a subset of samples (19 CHB) to fill the missing data. For chr4:158Mb region, six gaps were filled:158,702,285-158,708,035, 158,770,931-158,783,816, 158,827,935-158,840,376, 158,880,521-158,900,211, 158,906,161-158,913,233 and 158,985,263-158,992,841. For chr10:22Mb region, six gaps were also filled: 22,624,537-22,630,034, 22,643,514-22,656,292, 22,662,169-22,675,644, 22,689,042-22,696,558, 22,761,012-22,769,435 and 22,801,106-22,813,376.

Illumina paired-end libraries of ~200 bp fragments were then constructed on the enriched regions, and 37 bp from each end sequenced on an Illumina GAII[107] platform, with one sample per lane. After filtering out duplicate reads, the amount of mapped data ranged from 322 Mb to 572 Mb, leading to a mean coverage per individual of ~500x to >1000x for the parts which PCR amplified and ~ 35x to ~ 250x for pulldown regions. The paired-end sequence reads were

mapped back to the target reference sequences or the whole genome by SSAHA2 and candidate SNPs were called by SSAHASNP[48] for the PCR amplified regions, while MAQ[46] and SAMtools[108] were used for the data from the pulldown-enriched regions. By comparing the SNP calls based on Illumina data from *CASP12* with the existing capillary sequence data and avoiding heterozygous Y chromosome SNP calls, we set filtering criteria to filter out unreliable calls. For the SSAHA2 candidate SNPs from PCR enrichment, we filtered out all SNPs which lay within the primers or SSAHASNP indel calls, had coverage less than 30, or showed a ratio of the second-highest:total read depth of < 0.30:1 for a heterozygous SNP call. We only consider SNPs since indel variants are not reliably identified by this approach. For the MAQ and SAMtools candidate SNPs from the pulldown enrichment, SNP calls were filtered individually based on coverage, SNP score and mapping quality using criteria set based on the *CASP12* and Y chromosome data. The quality of the filtered SNP data was assessed by comparing the overlapping calls from our data with the HapMap2 genotypes from the same individuals. There were 43 discrepancies out of 2,981 comparisons for the chr4:158Mb and 5 out of 857 for the chr10:22Mb region, which suggested a low error rate for both regions (98.6% and 99.4% concordance, respectively). To assess whether such error rates affect the quality of subsequent statistical analyses, random errors were introduced into the simulations described above, matching the error rates, and results were compared with simulations without errors. This analysis showed that such error rates would not affect the power of the sequence analyses (results shown in Section 2.3.1).

We inferred haplotypes and occasional missing data using PHASE 2.1[37]. Then the neutrality tests and Nielsen et al.'s CLR test were performed on non-overlapping 10-20kb regions containing two or three PCR fragments chosen based on the size of each PCR fragment and the SNP densities.

### 2.2.3 Bioinformatic analysis

All miRBase (Release 13) mature miRNA sequences were scanned against the selected regions of the human genome using the MapMi algorithm[109]. This

approach involves first scanning the regions for matches to mature miRNA sequences; regions with matches to known miRNAs (allowing one mismatch) were then excised and folded using RNAfold from the ViennaRNA package[110]. These candidate regions were scored and filtered according to how well they fitted the stem-loop precursor structure common to miRNAs. We ran the pipeline in stand-alone mode, using non-repeat masked genomic sequence for increased sensitivity. The chr10:22Mb region had no significant hits for any known miRNA; however, the chr4:158Mb region had two hits to the miR-548 family of miRNAs, discussed below. This analysis was done in collaboration with José Afonso Guerra-Assunção from the European Bioinformatics Institute.

## 2.3 Results

### 2.3.1 Simulation of the power to detect and localize positive selection using genotype-based and sequence-based tests

In order to understand whether sequencing data provide more power in detecting and localizing selection signals, we started by comparing the power of genotype-based and sequence-based analyses using simulations. We first modeled the genotype-based tests mimicking those in the HapMap2 study, and in particular, the selective events seen in the CHB+JPT by comparison with the YRI population. To do this, we performed forward simulations under neutrality using the YRI and CHB demographic models, and with selection coefficients of 0.001, 0.004, 0.007 and 0.01 using the CHB demographic model, as described in section 2.2.1. Of the 1,000 simulations in each neutral and selected CHB set, there were 16, 16, 233, 724 and 779, respectively, that met the XP-EHH filtering criteria. These were combined into 16 significant XP-EHH results under neutrality and 1,752 under a range of selective conditions that would reflect the data that might be obtained from a population experiencing a variety of selective pressures.

We next applied the sequence-based tests to the 16 neutral and 1,752 selected datasets. There were 2 simulations among the 16 retained neutral ones that showed at least one significant window for the combined p value (≤ 0.01), and 7

for Nielsen et al.'s CLR. These numbers represent the false positive rates for the two methods, and are significantly higher for the CLR (p = 0.048, Fisher exact test). In the retained selected simulations, 84% (1,469 out of 1,752) for combined p value and 85% (1,494 out of 1,752) for Nielsen et al.'s CLR showed at least one significant window. Thus there is good power to detect this form of selection using sequence-based tests.

To investigate the ability to localize the causal SNP using the sequence-based tests, we first examined the test statistics averaged over all retained simulations. The average values of both showed no pattern along the DNA in the neutral simulations, but a strong peak centered on the window containing the selected site in the selected set, with a gradual decrease on either side (Figure 2.2 A and B). This indicates that, on average, the frequency spectrum-based neutrality tests can correctly identify the location of the causal SNP, but that there is considerable variation between simulations.

We therefore investigated this variation further by counting the occurrence of the most significant signals in each window in different simulations. For the combined p value, the most significant window lay within the 40 kb region (i.e. ±20 kb) surrounding the selected allele in 46% of the simulations, compared with 68% for the CLR (Figure 2.2 C). These results show that Nielsen et al.'s CLR performs better for localizing the selection signal, as previously reported[76]. Although the combined p value of Tajima's *D* and Fay and Wu's *H* and Nielsen et al.'s CLR have similar power for detecting selection (84% and 85%), we saw a lower false positive rate on the combined p value but a better localization power in Nielsen et al.'s CLR. Therefore, we investigated the benefits of further combining these signals. We tried using the combined p value to detect selection and then the CLR to localize the signal. This approach did systematically increase the accuracy of localization, although only by a small amount (Figure 2.2 D). We also considered the subset of simulations where the combined p value and Nielsen et al.'s CLR signals lie within the same 10kb window. Although the proportion is low (11.3%, or 198 out of 1,752 simulations), these might represent a favorable situation with the best chance to localize the selection signal. Indeed, this subset of simulations has about 90% chance to localize the

selection to a 40kb region and 80% to 20kb. These results provide an overall view of the power for localizing the signals in different scenarios and can guide the search for the biological basis of the selection.

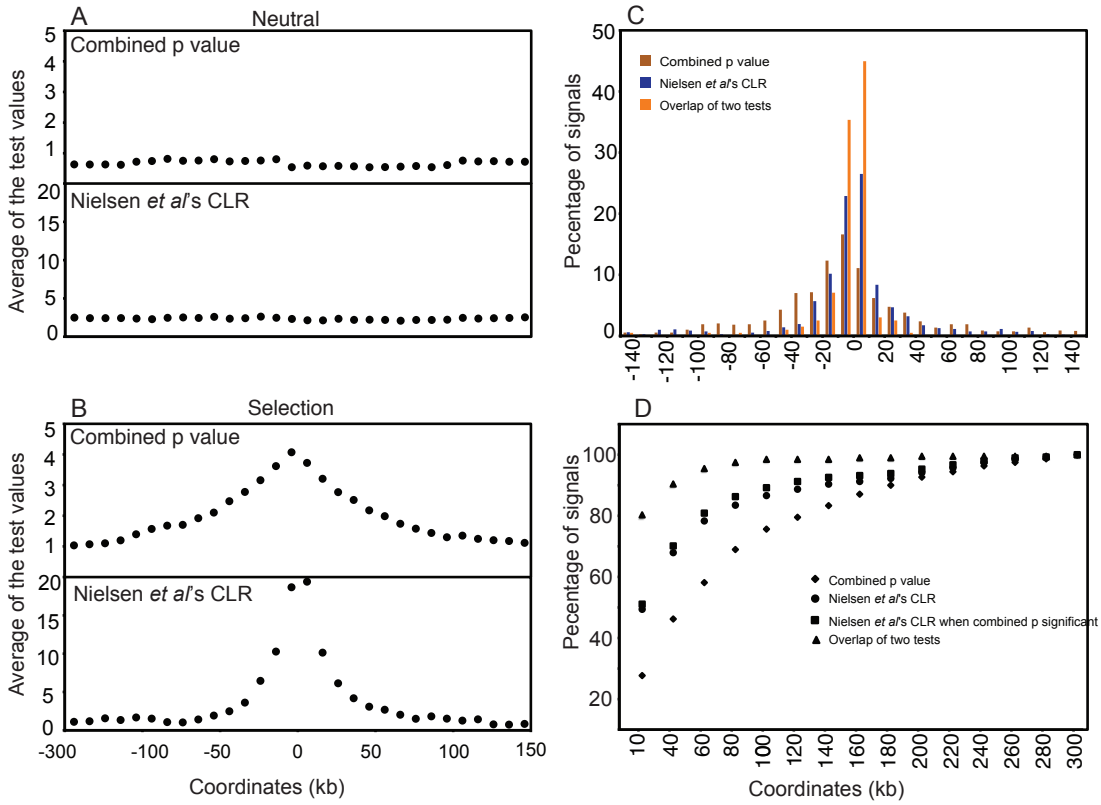


**Figure 2.2 Simulation results.** A. Simulations were carried out under neutrality, and tests for selection (-ln combined p values for Tajima's *D* and Fay and Wu's *H* (top) or Nielsen et al.'s CLR (bottom)) were calculated in non-overlapping 10 kb windows across 300 kb. Values of the test were averaged over 1,000 independent simulations. No departures from neutrality were seen. B. Simulations were carried out with selection (selection coefficient 0.007) and neutrality tests applied as in A. Departures from neutrality are seen most strongly in the window containing the selected SNP. C. The distribution of the top signal (lowest p value) in each simulation is shown across the 300 kb region. D. Probability that the known selected variant is found at each distance from the peak test value.

As mentioned above, there is ~1% error in the SNP calls from our sequencing data. In order to evaluate the effect of these errors on our analyses, we added 1% random base substitution errors to one of the datasets simulated with selection ($s$ = 0.007) and recalculated Tajima's *D* and Fay and Wu's *H* on the data with errors. Signals were overall slightly lower, but the pattern of signals was not affected (Figure 2.3). We therefore conclude that sequence errors at this level would not significantly influence our conclusions.

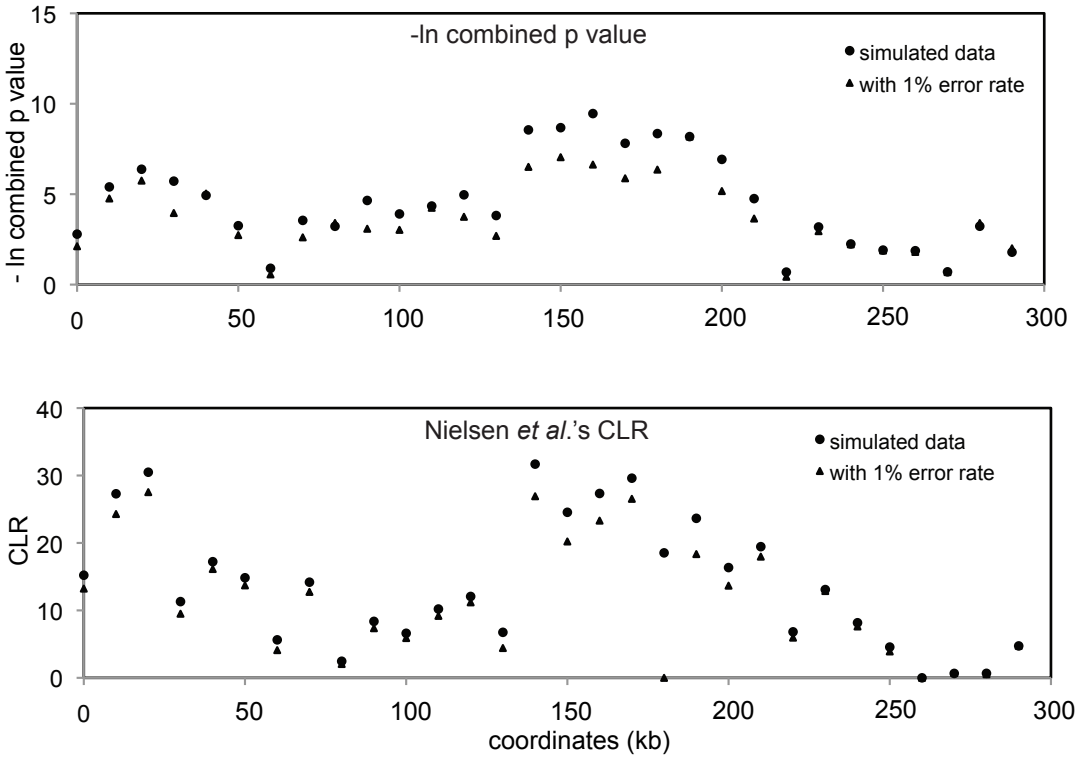


**Figure 2.3 Test results on simulated data with 1% sequencing error rate versus no error.** Dots represent results with no error in the simulated data, and triangles represent results with 1% random substitution errors introduced in simulated data.

### 2.3.2 Detection and localization of positive selection signals in experimental data

We re-sequenced two ~300 kb regions that had shown strong signals of positive selection in the HapMap2 study in 25 (chr4:158Mb) or 24 (chr10:22Mb) CHB individuals. The combined p value and Nielsen et al.'s CLR were calculated in chunks spanning either two or three PCR fragments, and are plotted in Figure 2.4 A and B. In both cases, a single window carries the most significant signal from each test: a combined p value of 0.00036 for chr4:158Mb, and 0.000015 for chr10:22Mb, and corresponding CLR values of 47 and 62. The two windows are located at 158,971,591-158,985,262 of chr4, and 22,755,918-22,776,116 of chr10, with sizes of ~13 kb and ~20 kb, respectively. Based on the simulations, this is a particularly favorable situation for localizing the selected variant, and we have 80% confidence that the target of selection should lie in a 20 kb region centered on these windows.

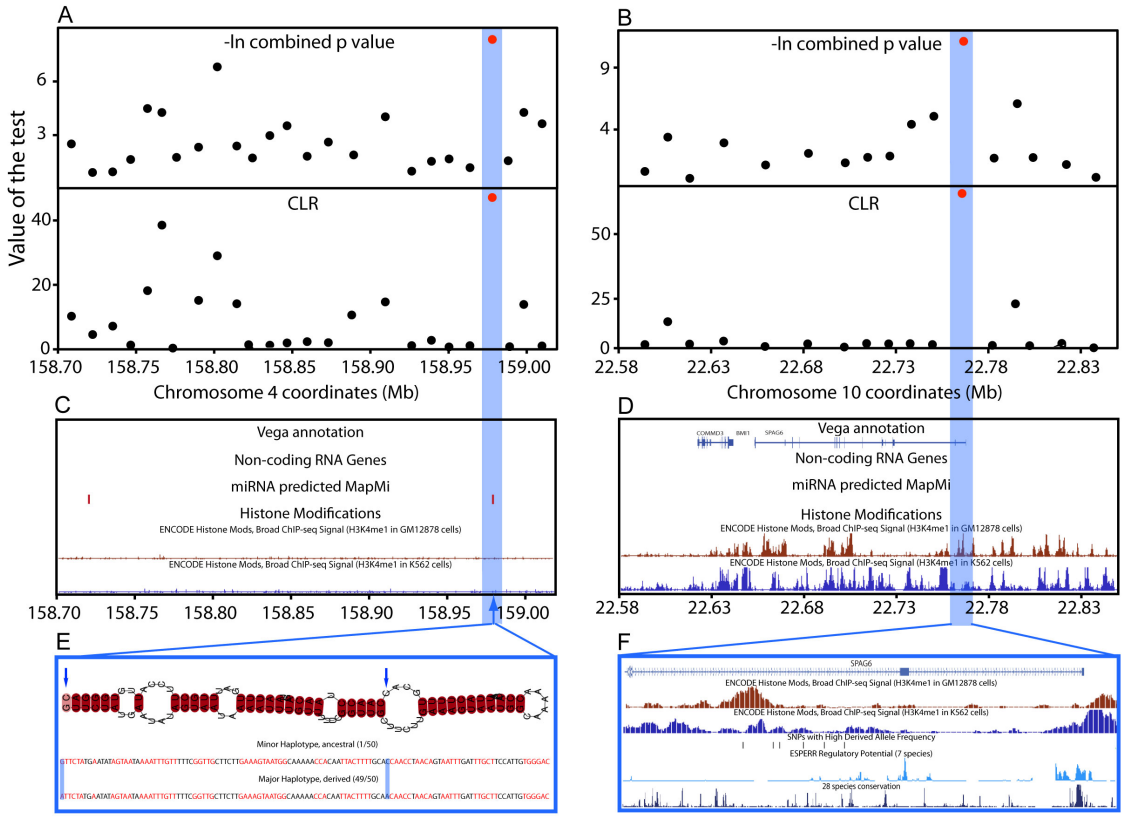


**Figure 2.4 Experimental results.** These figures show localization of likely selection targets in the chr4 and chr10 regions. A. -log e of combined p values from Tajima's *D* and Fay and Wu's *H* (top) and Nielsen et al.'s CLR (bottom) calculated from re-sequencing data in windows corresponding to two or three PCR fragments (10-20 kb). The most significant statistics are shown in red, and fall into the same window overlap at ~158.98 Mb (blue highlight). B. Corresponding analysis of the chr10:22Mb region, where the most significant signals again fall into the same window, this time at ~22.78 Mb. C, D. Protein-coding genes from the Vega annotation, non-coding RNA and miRNA genes, and relevant ENCODE chromatin modifications in the two regions. E. Predicted miRNA in the chr4:22Mb target region. Two SNPs are present, including a G>A at the end of the miRNA carried on the major haplotype (49/50 chromosomes, selected in CHB) that may influence the strand forming the mature miRNA. F. H3K4me1 chromatin modifications indicating enhancer regions in GM12878 (second) and K562 (third) cells, SNPs with high derived allele frequencies (fourth), predicted regulatory potential (fifth) and 28 species conservation (bottom). Three high-frequency derived SNPs lie within candidate enhancers in one or other of the cell lines, but high-frequency derived SNPs do not lie within regions with high predicted regulatory potential or conservation.

### 2.3.3 Biological targets of selection

The final stage of our analysis was to search for possible biological targets of selection. Such targets should most likely lie within the narrowed interval, and carry a biologically relevant difference between the selected and non-selected haplotypes. The 314 kb region on chromosome 4 consists entirely of intergenic sequence, and the nearest annotated protein-coding gene is located more than 50 kb outside this region. No histone modifications indicative of promoters,

insulators or enhancers were apparent in publically available data (Figure 2.4 C). However, using the MapMi approach[109], we found two predicted microRNAs (miRNAs) belonging to the mir-548 family (Figure 2.4 C). One of these lay far from the selection signal but the other, hsa-miR-548c, lay at 158.982 Mb, within the narrowed region (Figure 2.4 C). Strikingly, two SNPs are present within this predicted miRNA and both show high derived-allele frequencies in the CHB population. One of these SNPs lies within a loop in the predicted RNA and is not predicted to have functional consequences. However, the other is the first nucleotide of the miRNA precursor and could therefore determine which strand is processed to form the mature miRNA and consequently change the set of target genes (Figure 2.4 E).

The chromosome 10 region contains three annotated protein-coding genes, *COMMD3*, *BMI1* and *SPAG6*, and no miRNA genes (Figure 2.4 D). *SPAG6* transcripts (e.g. SPAG6-002, OTTHUMT00000047185: http://vega.sanger.ac.uk/Homo_sapiens/index.html) extend into the narrowed region (Figure 2.4 D). ChIP-seq experiments reveal extensive chromatin modification within the 263 kb region, including H3K4me1, H3K4me2, H3K4me3, H3K9ac, H327me3 and H3K27ac (http://genome.ucsc.edu/ ENCODE Histone Mods, Broad ChIP-seq; Figure 2.4 D), as would be expected for a region containing several protein-coding genes. The narrowed region contains two peaks of H3K4me1, which could indicate an enhancer[111]. Thus *SPAG6* provides a good candidate on the basis of its location relative to the signal of selection. Although *SPAG6* contains a relatively high-frequency derived non-synonymous SNP (rs7074847) in the YRI[66], there are no non-synonymous differences between the selected and non-selected CHB haplotypes, suggesting that selection is more likely to be acting on an aspect of transcription than on a change in the protein sequence.

In conclusion, based on our analyses, possible targets for selection can be identified in both regions and there is strong functional evidence for selection.

## 2.4 Discussion

### 2.4.1 Power of detection and localization

The first question we addressed was whether or not candidate regions identified in genome scans for positive selection using LD-based tests on genotype data, such as that performed by the HapMap2 project, would show supporting evidence for selection when frequency spectrum-based neutrality tests were applied to re-sequencing data. Such tests are sometimes considered most suitable for detecting complete sweeps, in contrast to the partial sweeps detected by LD-based methods, but are also highly effective in detecting partial sweeps[112]. The answer to this question, from both our simulations and the two experimental examples investigated, was a clear "yes". Significant departures from neutrality (combined p value from Tajima's *D* and Fay and Wu's *H*) were seen in 84% of the 1,752 simulations that passed the XP-EHH threshold, contrasted with just 2 out of the 16 neutral simulations that by chance passed (not significantly different from 0 out of 16, Fisher exact test). A similar result was seen with Nielsen et al.'s CLR, although the false positive rate was higher. This correspondence is unsurprising, given the similar underlying basis for the two tests, but there was value in combining the two (see Section 2.2). In the two regions investigated experimentally, significant values were seen in both with all the tests applied.

The second question was the extent to which targets of selection could be localized more precisely when using re-sequencing data. From the simulations, we found that re-sequencing data do provide valuable additional information about the localization of selection targets. Higher SNP density and the presence of more rare variants make a higher resolution of signals possible. One of the disadvantages of LD-based test is that they detect large LD blocks, which are often several hundred kb in length. Although some frequency spectrum-based tests can also be used on genotype data, for example Nielsen et al.'s CLR, the window size often has to be relatively large because information from many SNPs needs to be combined to get enough power. We applied Nielsen et al.'s CLR using HapMap2 genotype data, and for chr4:158Mb, it detected a signal of

selection localized to ~40 kb, while for the chr10:22Mb region, where the HapMap2 SNP density was high in the critical interval, the selected region was narrowed down to a similar length to the sequencing data (Figure 2.5). A method for combining multiple signals derived from genotype data has been described[113], which provides a median localization to a 55 kb interval. This method identified a chr4:158Mb interval spanning ~60kb (158,862,019-158,921,890, with top SNP at 158,904,521), but failed to find any significant signal at chr10:22Mb[68] (Figure 2.5). We repeated the CMS analysis using the HapMap2 genotype data and localized the chr4:158Mb signal to a similar ~58kb interval, although with a different peak SNP (158,862,019- 158,920,326, with top SNP at 158,920,326), and also found no signal in the chr10:22Mb interval (Figure 2.5 and Figure 2.6). In contrast, re-sequencing followed by the application of the tests used here provided localization to a < 20 kb interval in both cases.

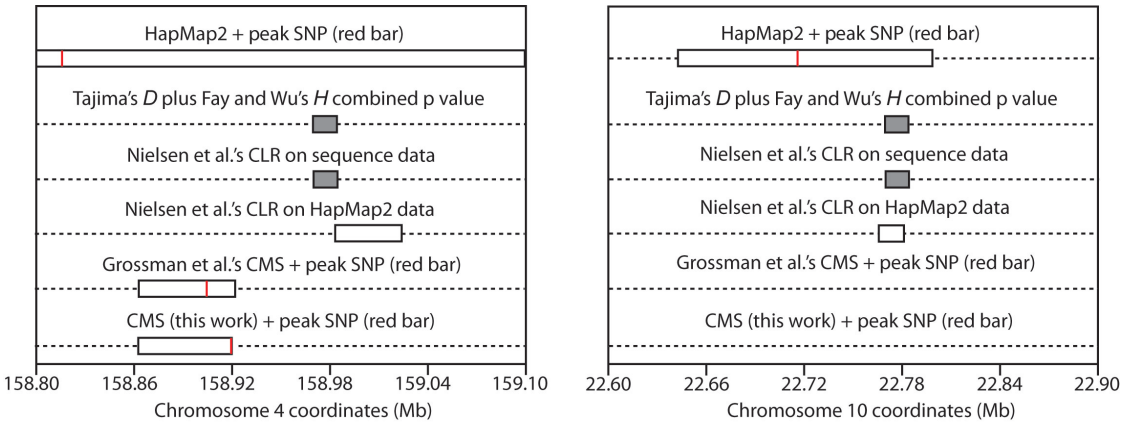


**Figure 2.5 Comparison of different approaches of signal localization.** These figures show localization of the signal of selection within the chr4 and chr10 regions using different approaches. The two starting regions are shown at the top (Sabeti et al. 2007), localizations using sequence data (grey bars) or HapMap2 genotype data (white bars) by this study in the middle, and the localization by the CMS statistic (Grossman et al. 2010 or this work) at the bottom.

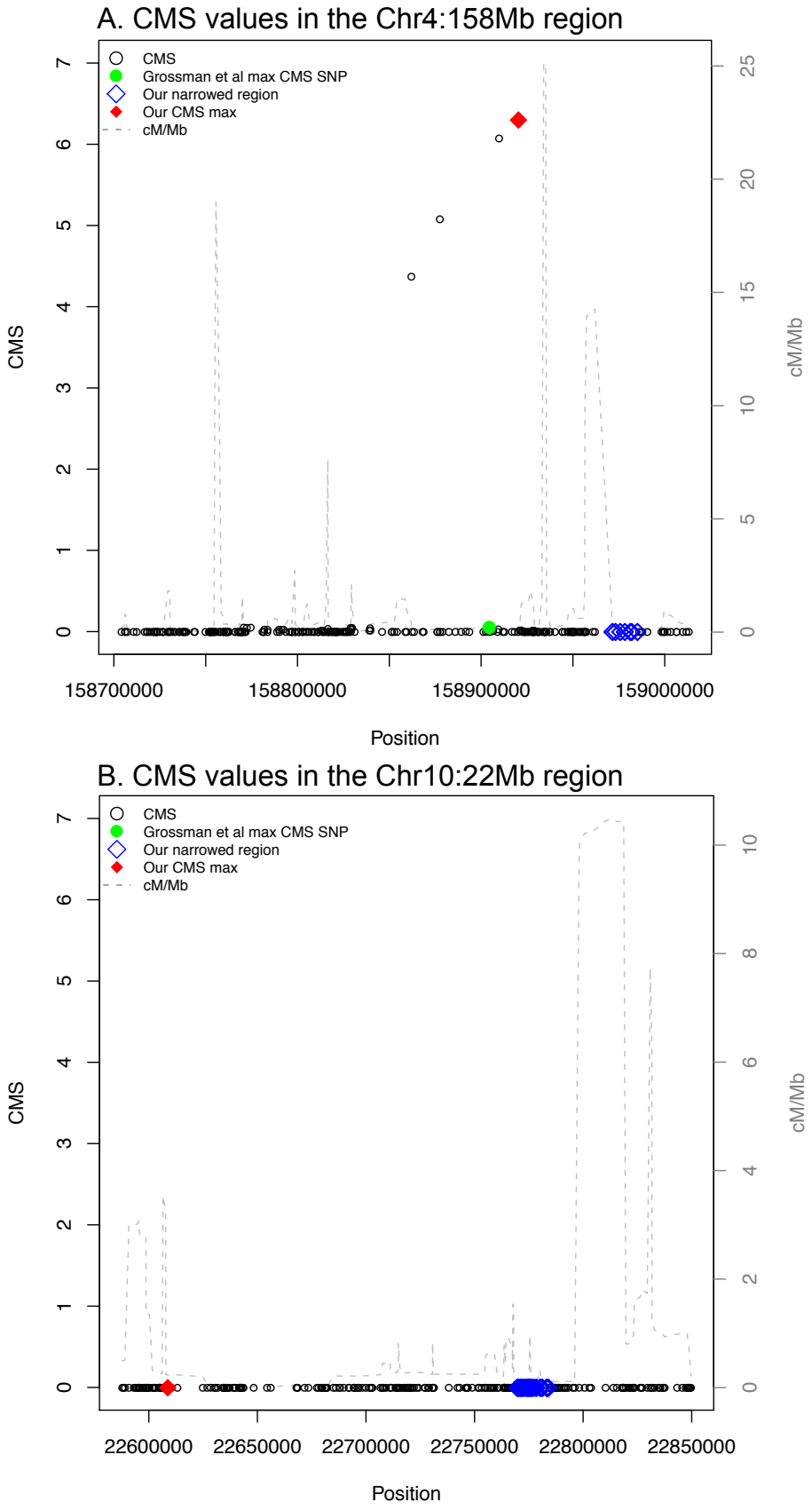


**Figure 2.6 CMS results on both regions.** Recombination intensities are shown as dashed lines.

### 2.4.2 Functional targets of selection

The final question we set out to address was whether increased insights into the possible biological basis for the selection could be obtained. Due to our inability to predict the phenotypic consequences of most DNA variants, particularly when these lie outside protein-coding regions, it is often still difficult to identify the causal variant. Nevertheless, the narrowed region provides the best starting point for further investigation. It is, in principle, possible that variants in a region could be acting on distant genes, but this in practice seems rare: a study of human eQTLs, for example, found that most lie either within or close to the genes they affect, with only 5% lying > 20 kb away[114]. On this basis, we therefore focus on targets close to the narrowed regions in the following discussion.

For chr4:158Mb, the above considerations and the lack of any annotated protein-coding genes in the vicinity make a direct effect on a protein-coding gene unlikely. Predicted miRNA hsa-miR-548c, however, provides an intriguing candidate. Members of the hsa-miR-548 family are derived from the transposable element Made1, present in multiple (~30) copies in the human genome[115]. Made1 elements are found only in primates, and hsa-miR-548 sequences have been documented only in the human, chimpanzee and macaque genomes, where they appear to be evolving rapidly. Since miRNAs function as post-transcriptional regulators by binding to partially complementary target sites in the 3' untranslated regions of mRNAs and inhibiting their expression, a change in the sequence of a mature miRNA could influence the expression of a large number of genes, and a change in the strand present in the miRNA could have even greater regulatory effects. More than 3,500 genes have been listed as predicted hsa-miR-548 targets, enriched in functions such as cell proliferation[115]. We can thus speculate that a variant hsa-miR-548c might have been selected because of altered target gene regulation, but the large number of hsa-miR-548 family members and potential targets makes it difficult to formulate or test more precise predictions. Nevertheless, a link to changes in gene regulation fits well with general thinking about the importance of regulatory mutations in human evolution[116] and the inference of recent positive selection acting on a miRNA-rich region on chromosome 14 devoid of annotated protein-coding genes[117] .

For chr10:22Mb, similar considerations lead to the suggestion that *SPAG6* is the most likely target of selection, and a change in the level, timing or location of its expression as the most likely mechanism. In support of this possibility, it was notable that a *SPAG6* transcription end site lay within the narrowed region, and Veyrieras et al.[114] had reported a strong enrichment of eQTLs in the 250 bp just upstream of the transcription end site. However, in the *SPAG6* data, the closest SNPs were 2,055 bp upstream and 843 bp downstream of the transcription end site. In contrast, two H3K4me1 signals indicative of enhancers are located within the narrowed region, and three high-frequency derived SNPs (rs16922285 at 22,773,002, rs11012996 at 22,773,902 and rs11012997 at 22,774,094) specific to the selected haplotype overlap with them (Figure 2.4 F). An altered enhancer activity thus provides the most plausible biological mechanism. SPAG6 is a component of sperm[118], and mouse knockout models have been investigated: 50% of *Spag6*$^{-/-}$ mice died within eight weeks due to hydrocephalus (fluid on the brain); males surviving to maturity showed abnormalities of sperm structure and mobility and were infertile[119]. Heterozygous *Spag6*$^{+/-}$ animals showed a much milder phenotype and were fertile, but their sperm swam more slowly, suggesting that a reduced level of SPAG6 protein can have a detectable effect on the sperm phenotype. The hydrocephalus phenotype, however, points towards a wider role of the protein in the function of cilia, and thus other potential modes of selection. Nevertheless, the best candidate remains an effect on reproduction, which would be consistent with both the inference of recent positive selection on another sperm protein gene, *SPAG4*, in the CHB among other populations[67], and the high frequency with which genes linked to reproduction are found more generally in surveys of positive selection[63,67].

There are two other protein-coding genes in the interval, both > 100 kb from the strongest selection signal. Little is known about *COMMD3* itself, but diverse functions have been ascribed to other COMMD family members, including copper metabolism and regulation of the activity of the transcription factor NF-κB and cell proliferation, perhaps through the ubiquitin pathway[120]. *BMI1*, in contrast, has been studied extensively. It is a polycomb protein, involved in DNA repair, chromatin remodeling and stem cell renewal, and its inappropriate over-

expression can lead to tumor formation[121-123]. Knockout mice are viable and homozygotes show hematopoietic, skeletal and neurological abnormalities, but phenotypic effects in the heterozygotes were not noted[124]. In humans, a cysteine to tyrosine substitution at position 18 leads to substantially lower levels of BMI1 protein, and is present in the general population, including in the YRI and CEU (but not CHB) HapMap samples[125]. Since increased expression of BMI1 leads to cancer, and a decreased expression phenotype is present in HapMap populations but has not been positively selected, both *COMMD3* and *BMI1* seem less strong candidates than *SPAG6* for the target of chr10:22Mb selection.

### 2.4.3 Conclusion

From these examples, we can conclude that the approach used here, of re-sequencing large target regions, refining the target location and making inferences about the biology of the selection events, is fruitful. However, it could be improved in several ways. Re-sequencing technology is still imperfect and data quality needs to be improved. This study required a combination of two enrichment strategies, PCR and pulldown, to generate adequate coverage, and such intensive effort is impractical for large-scale studies. Most urgently, however, better statistics for localizing the target of selection using re-sequencing are needed, and improved methods for interpreting the biological consequences of DNA variants discovered are especially needed. But even with the present tools, specific topics to follow up experimentally can be suggested, e.g. comparison of sperm mobility and other sperm characteristics between carriers of selected and non-selected haplotypes in the chr10:22Mb region. More generally, the availability of population-scale re-sequencing data from both the increasing number of personal genome projects[126] and projects such as the 1000 Genomes Project[40] will make the approach used here applicable across the genome.

# 3 A survey of positively selected regions using 1000 Genomes Project low-coverage Pilot data

## 3.1 Introduction

Whole genome sequencing of samples from multiple human populations provides powerful resources for studying evolution at the genomic level in an unbiased, holistic manner. Compared to genotyping, where only known variants, most of which have high or moderate frequencies in the population, are analyzed, sequencing reveals the whole set of variants in a particular genome without any ascertainment bias. This is beneficial in at least two aspects. One is the presence of rare variants in the data. In many neutrality tests, genetic diversity and allele frequency spectra are measured, which play important roles in the detection of selective sweeps. In genotype data, the majority of those rare variants (frequency less than 5%) are missing, which greatly reduces the power to detect selective sweeps that have nearly or already completed, where there may be an excess of rare alleles. The other aspect is the absence of bias in variant detection. Genotyping only detects a set of variants that are determined prior to the assay, regardless of what other variants may be present in the samples. This introduces bias, especially when the frequency spectrum needs to be measured in different populations. For example, if we use a certain SNP chip to measure the differentiation between populations, although we can measure the frequency differences of the SNPs included in this assay, we may miss a subset of population-specific SNPs or highly differentiated SNPs in certain population(s), depending on which population(s) the design of the SNP chip is based on. In this case, the measure of population differentiation may be highly biased. Sequencing data, however, can detect all these variants and thus provide the foundation of an unbiased measure of population differentiation.

The 1000 Genomes Project is an excellent example of such resources. The Pilot 1 (low-coverage) project sequenced 179 individuals from four populations: CEU (Utah residents with Northern and Western European ancestry from the CEPH

collection), CHB+JPT (Chinese Han in Beijing, China and Japanese in Tokyo, Japan) and YRI (Yoruba in Ibadan, Nigeria), with the average coverage of 2-4x[40]. 15 million SNPs were identified in the Pilot Project along with other types of genetic polymorphism, which greatly enriched the database of human genomic variation. As demonstrated in Chapter 2, a genome-wide survey of positive selection using frequency-spectrum based methods on such sequencing data would provide deeper insights into the extent to which positive selection has shaped modern human genomic variation, as well as the biological targets that may be selected during recent modern human evolutionary history.

In this chapter, neutral and positively selected simulations were performed to gauge the level of significance, as well as provide insights into the power of localizing selection targets, and how recombination affects the signals. A genome-wide scan of positive selection was then carried out on the 1000 Genomes low-coverage Pilot data, and bioinformatic analyses on both the general features of candidate genes/regions and the possible functional targets of selection in some strong candidates were performed. The data were generated in multiple centers as part of the 1000 Genomes Project. All the simulations, statistical calculations and data analyses in this chapter were done by the author of this thesis, with help from some participants in the 1000 Genomes Project. An early version of the results were published as part of the 1000 Genomes Project Pilot paper, and manuscript describing this work in more detail is under preparation.

## 3.2 Materials and Methods

### 3.2.1 Simulations

We first carried out coalescent simulations using the *msHOT* package[127] to generate 1Mb long neutral haplotypes in African, European and Asian ancestral populations 2,000 generations ago, based on the best-fit demographic models[26] for the three populations. Then these simulated haplotypes were used as seed haplotypes for the forward simulations using *mpop*[94], as described in section 2.2.1. In forward simulations, one neutral scenario (1,000 independent simulations) and sixteen selective sweep scenarios were simulated in each

population. Selection coefficients of 0.001, 0.004, 0.007 and 0.01, and the age of selective sweeps of 500 generations, 1,000 generations, 1,500 generations and 2,000 generations were used in the selective sweep scenarios, with 250 simulations for every combination of these two parameters. One allele with an initial frequency of 0.0006 under selection was added in the middle of the haplotypes at the starting time point of the selective sweep. The genome average mutation rate of $1.0 \times 10^{-8}$ per nucleotide per generation was used in the simulations. In addition, to mimic the real patterns of recombination in the genome, we used the HapMap recombination map[39] to generate a recombination hotspot map, and regions of 1 Mb were drawn randomly from the genome and the recombination hotspots they contained were assigned to the simulated regions. For the purpose of comparison and understanding of the effects of recombination hotspots on the signals of selection, we also did another set of simulations with all parameters being the same, except that a strong recombination hotspot (2,000-fold greater than the background recombination rate) with 0 kb, 10 kb, 20 kb, 30 kb or 40 kb distance from the selected allele was added into the simulated haplotypes. The rest of demographic parameters were as in Schaffner et al.'s best-fit demographic model for the European population[26]. For computational efficiency, we re-scaled the parameters by a factor of 5, as described in section 2.2.1. 120 chromosomes were sampled from each simulation, to match the sample sizes of 1000 Genomes Project low-coverage Pilot data (see Appendix C for parameters and command lines).

### 3.2.2 Neutrality tests on simulated data

In order to mimic the real situation of 1000 Genomes low-coverage Pilot data, where rare SNPs are still under-ascertained, we filtered the simulated data by matching the proportion of SNPs in each derived allele frequency bin (bin size 0.1) of the simulated data to the 1000 Genomes low-coverage Pilot data in each population (CEU, CHB+JPT and YRI). Then three frequency-spectrum based tests, Tajima's $D$[71], Fay and Wu's $H$[72] and Nielsen's CLR[76] were applied to the simulated data in 10 kb non-overlapping windows across the simulated regions. P values of each test were calculated based on the distribution of test values of 1000 neutral simulations in each population. In order to obtain a single score representing the

signals of all three tests, we calculated the correlations between the p values of every two tests in neutral simulations to see whether these tests are independent from each other. Results showed that the absolute value of the correlation of every pair of tests was less than 0.2. Therefore, we treated these tests as independent, and combined the p values of each test on the same window using Fisher's method[104].

### 3.2.3 Sensitivity and specificity analysis on simulated data

In order to understand the relationships between false positive rate, false negative rate and false discovery rate of our combined tests under different p value significance thresholds, we calculated the above rates under seven thresholds, with 10-fold decrease for each from $4\times10^{-3}$ to $4\times10^{-9}$. We obtained the false positive rate by calculating the percentage of neutral simulations that were detected as under positive selection. The false negative rates were obtained by calculating the percentage of 1,000 positive selection simulations with a selection coefficient of either 0.007 or 0.01, and the age of sweep of either 1,500 or 2,000 generations. We next counted the number of candidate regions from the 1000 Genomes low-coverage Pilot data across the genome under each significance threshold, and then calculated the false discovery rate based on the number of false positive regions, which was calculated by multiplying the false positive rate with the number of 300-kb regions in our empirical data, and divided by the total number of detected positively selected regions across the whole genome in each population.

### 3.2.4 Neutrality tests on 1000 Genomes low-coverage Pilot data

We segmented the whole-genome SNP data from CHB+JPT, CEU and YRI populations of 1000 Genomes low-coverage Pilot data into non-overlapping windows with a length of ~10 kb, where both the starting and ending point of each window were SNP positions. Windows that lay in regions with mapping gaps, low mapping quality or heavily filtered SNPs, were excluded (Table 3.1). The same neutrality tests were applied on these windows in each population as for simulations, and p values were obtained using the same approach as for the simulated data.

**Table 3.1 Total number of windows and total length scanned in each population.**

| Population | Total windows | Total length (bp) |
|---|---|---|
| CEU | 252,348 | 2,390,406,461 |
| CHB+JPT | 247,432 | 2,302,196,289 |
| YRI | 255,289 | 2,450,357,355 |

### 3.2.5 Identification of candidate regions and genes

After the genome-wide combined p values of our neutrality tests were obtained, we needed to decide which threshold of significance to use. As we aimed to get a confident list of candidate regions, we used the stringent Bonferroni correction[128]. We divided 0.01 by the total number of windows that we applied the tests to throughout the whole genome, which yielded a threshold of $\sim 4\times10^{-8}$ ($-\log_e$ value 17.0). We used this as a cutoff to identify significant windows in each population. Adjacent significant windows that are less than 150 kb apart were treated as likely to originate from the same selective sweep, and combined into a single candidate region.

As our simulations showed that there is ~75% chance that the selection target falls into the 100 kb region surrounding the peak signal, we identified candidate genes from the ~100 kb region around the most significant window in each candidate region. In regions where multiple genes were present, we treated the gene closest to the peak signal as the candidate gene. In a few cases where two genes either overlap with each other or have the same distance from the peak signal, we retained both of them as candidate genes for that region.

We also looked at positions of peak signals relative to the candidate protein-coding genes. We used three categories of positions: upstream of the gene, within the gene, and downstream of the gene. First of all, to determine which side of the gene is upstream or downstream, we obtained information about whether the gene is on the forward strand or reverse strand of the DNA sequence for each candidate protein-coding gene. Then we counted the number of peak windows falling into each category of position. For those peaks that cover more than one position, we used the proportion of the window in each

position as the count. For example, if 40% of the peak window is in the upstream sequence, and the other 60% is in the gene, we count 0.4 into "upstream" and 0.6 into "within gene" for that candidate.

### 3.2.6 Comparison with previous studies and bioinformatic analyses

We compared our lists of positively selected regions or genes with previous genome-wide scans of positive selection, as well as with functional annotations. We obtained annotations of synonymous and non-synonymous changes in the 1000 Genomes Pilot data. In order to see whether there was any enrichment or depletion of overlaps between our candidate regions/genes and those data sets being compared with, we randomly picked the same number of regions from the low-coverage Pilot data accessible genome matching the lengths of the candidate regions in each population, and counted how many of them overlap with regions from other studies. We did this 1000 times independently and obtained a distribution of number of overlaps in each comparison. Then we calculated p values of the enrichments of all the compared scenarios in our candidate positively selected region or gene lists, based on the percentile of the distribution of overlaps in random data sets that our candidate list falls into. In some of the comparisons and other analyses, we also looked at derived allele frequencies (DAF) of the variants. The ancestral alleles were identified by the 1000 Genomes Project from analysis on the sequences of human (NCBI36), chimpanzee (CHIMP2.1), orangutan (PPYG2) and rhesus macaque (MMUL_1) genomes[40] (The 1000 Genomes Project Consortium, *Nature* 2010, supplementary information 13.1).

In order to further understand the relationship between the functional consequences of non-synonymous changes and positive selection, we obtained the Condel scores[129] of high DAF (≥ 0.5) non-synonymous variants in the 1000 Genomes low-coverage Pilot data computed in Ensembl release 65 by combining the SIFT[130] and Polyphen2[131] scores. Non-synonymous variants with higher Condel scores are more likely to be deleterious. In order to investigate whether Condel scores of high DAF variants in positively selected genes tend to be higher than those in the random genes, we performed a Mann-Whitney test[132] on

Condel scores of high DAF non-synonymous variants in the candidate gene list versus those in the 1000 independent random sets of matched genes, using the built-in function in the R package. P values of each comparison between the candidate gene Condel scores and the random gene Condel scores were obtained from the test.

We also investigated non-coding functional variants within our candidate regions. We first obtained lists of variants with a high DAF (≥ 0.5) that are within one of four types of non-coding functional elements: UTR, non-coding RNA, enhancer, and transcription factor (TF) binding motif. The non-coding functional annotation was obtained from the 1000 Genomes Project Phase 1 and the ENCODE project[57]. For the TF binding motif variants, we further categorized them into two types: motif gain and motif loss. If the derived allele of a SNP has a higher frequency in the position weight matrix (PWM) of the bound motif than the ancestral allele, we call it motif gain. Likewise, if the derived allele of a SNP has a lower frequency in the PWM of the bound motif than the ancestral allele, we call it motif loss[133]. We then counted the number of high DAF variants within each of the five categories within our candidate regions, as well as within 1000 sets of random matched regions. We plotted the distribution of number of variants in each category in the random regions, in order to see if any of them was enriched by any of the functional elements.

We then used the online gene annotation clustering tool DAVID[134] to categorize our lists of candidate protein coding genes into functional clusters, and obtained Bonferroni-corrected p values of enrichments in each cluster from DAVID. We also identified genome-wide significant variants from Genome Wide Association Studies (GWAS) that fall into our candidate regions. The list of GWAS significant variants were obtained from the NHGRI "A Catalog of Published Genome-Wide Association Studies[135]" (http://www.genome.gov/gwastudies/).

## 3.3 Results from simulations

### 3.3.1 Sensitivity and specificity of selective sweep detection using low-coverage sequencing data

Balancing the false positive and false negative rates in the identification of statistical significance is a crucial step in a large-scale global survey of statistical tests. As mentioned above, we chose to use the most stringent p value cutoff (Bonferroni correction, $p = 4\times10^{-8}$) to identify significant windows. This, of course, sacrifices the sensitivity of detection. An alternative measure of the p value significance threshold is the false discovery rate (FDR). Since we are applying the statistical tests a large number of times, even a very small false positive rate can result in a large FDR. To measure this, we counted the number of candidate regions under different p value thresholds, and calculated FDRs accordingly. We found that even if the false positive rate is 0.6%, the FDR is still as high as 4%. In order to get a highly confident list of candidate regions, we would like the FDR to be less than 5%. A Bonferroni-corrected threshold of $4\times10^{-8}$ gives us 0% and 3% FDR in CEU and YRI, respectively (YRI, Table 3.2). Although in this case, we were only able to detect ~20% of the moderate-strength positive selection events, we are confident that the list of candidates we picked out is mostly real.

**Table 3.2 Sensitivity and specificity under different p value significance thresholds in the YRI population.**

| P value significance threshold | False positive rate | False negative rate | False discovery rate |
|---|---|---|---|
| 4E-03 | 30.0% | 27.3% | 49.2% |
| 4E-04 | 11.6% | 44.0% | 25.3% |
| 4E-05 | 2.5% | 56.7% | 9.4% |
| 4E-06 | 0.6% | 66.1% | 4.0% |
| 4E-07 | 0.3% | 73.9% | 3.8% |
| 4E-08 | 0.1% | 79.4% | 3.0% |
| 4E-09 | 0.0% | 85.0% | 0.0% |

### 3.3.2 Power of localizing positive selection targets

We found that in our simulations, although on average the most significant window was the one that contains the selected allele (Figure 3.1), in each individual simulation with positive selection, the peak signal can fall into any window across the 300 kb region with the selected allele in the middle. We found that in our selection simulations in YRI, 79% of the time the most significant signal is less than 50 kb away from the window with the selected allele, and this percentage in CEU is 72% (Figure 3.2). Based on this, in our candidate regions in the empirical data, we have more than 70% confidence that the selection target is within 50 kb distance from the peak signal.

### 3.3.3 Effects of recombination hotspots on localization of selection target

Recombination during the progress of a selective sweep can result in the breakdown of the selected haplotype, which thus disrupts the pattern of genomic variants in the selected region. In order to understand the effects of the position of recombination hotspots on the position of peak signals relative to the positively selected allele, we performed five sets of simulations with $s$ = 0.01, age of sweep = 1500 generations, and in each set, added an extremely strong recombination hotspot (2000-fold higher than background rate) with 0-5 kb, 10 kb, 20 kb, 30 kb and 40 kb distance from the selected allele, respectively. Our results showed that, in general, the closer the recombination hotspot to the selected allele, the more scattered the distribution of peak signals will be. When the recombination hotspot is 40 kb or more away from the selected allele, the effect on the localization power almost vanished. Not surprisingly, when there is a strong recombination hotspot at one side close to the selected allele, the peak signal tends to be on the other side of the selected allele (Figure 3.3). However, in most cases, the peak signal is still most likely to be within 50 kb distance from the selected allele. Moreover, in these simulations, we used an extremely strong recombination hotspot, in order to make sure that recombination happens in most of our simulated regions within the simulated period of time. In reality, most recombination hotspots are much more moderate, thus the effects may not be as dramatic. Therefore, when identifying selection target, choosing to use the region within 50 kb distance from the peak signal as the target is still reasonable

even if recombination hotspots are present. Having said that, it is still sensible to be more cautious about the location of the putative selection target when there is a recombination hotspot near the peak signal.

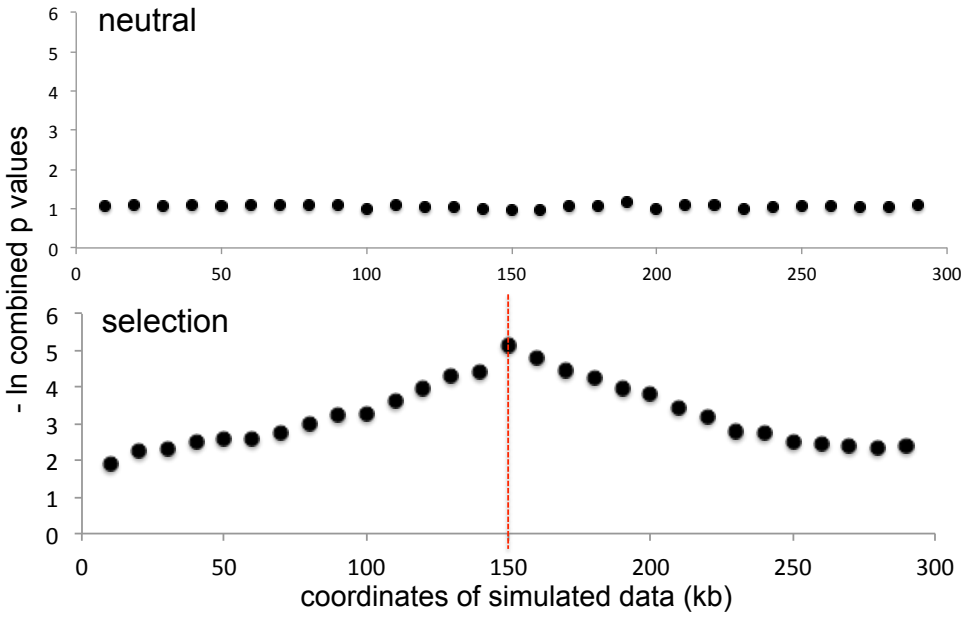


**Figure 3.1 Averaged scores in neutral and positively selected simulations.** The top plot shows average scores of each 10-kb window across the simulated neutral regions; the bottom plot shows the same but in simulated regions with selection. The red dashed line shows the position of the selected allele.

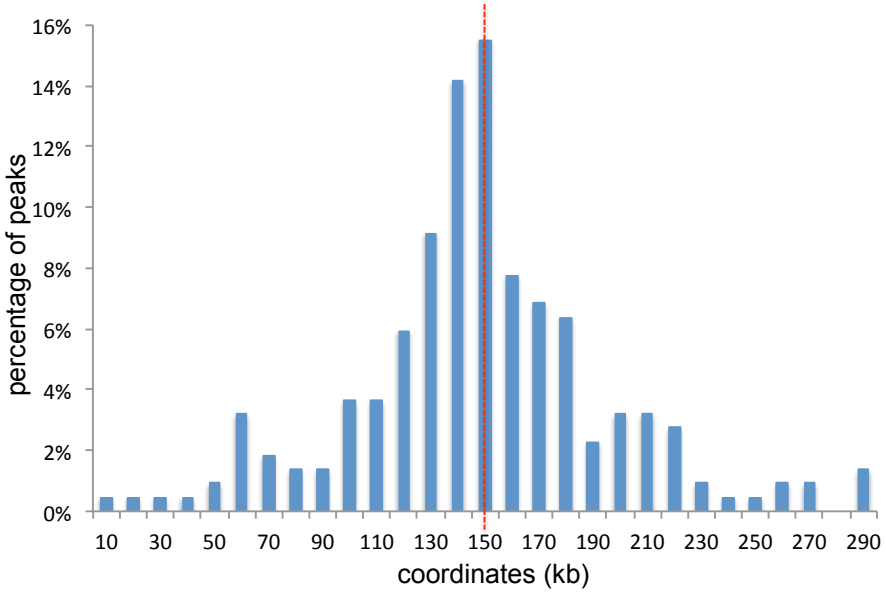


**Figure 3.2 Distribution of peak signals across the simulated regions with selection.** Each bar shows the percentage of peak signals falling in the particular window. The red dashed line shows the position of selected allele.

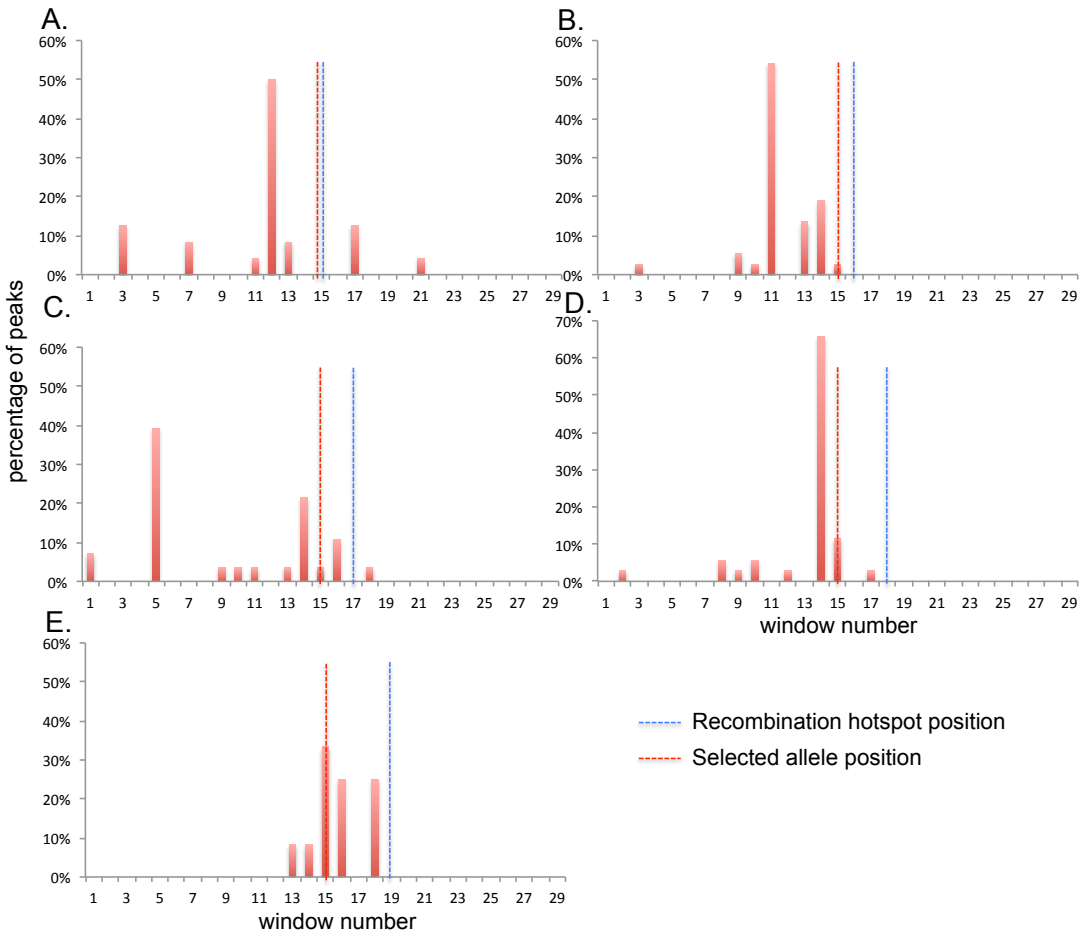


**Figure 3.3 Distribution of peak signals in simulations with single strong recombination hotspots.** Each plot shows the distribution of peak signals under the scenario with fixed distance between the selected allele and the recombination hotspot. The blue dashed line marks position of the recombination hotspot, and the red dashed line marks position of the selected allele. X-axis is the window number across the simulated region, and Y-axis is the percentage of peaks falling into each window.

## 3.4 Results from 1000 Genomes Project low-coverage Pilot data

### 3.4.1 Genome-wide scan on 1000 Genomes low coverage data

We applied the same tests and criteria to the 1000 Genomes Project low-coverage Pilot sequencing data in ~10-kb windows across the whole genome in CEU, CHB+JPT and YRI populations. We identified 477, 137 and 290 candidate regions in the three populations, respectively. In all populations, most regions only have one significant window, but CEU have more regions with larger numbers of significant windows than the other two populations (Figure 3.4). Among these candidate regions, 65%, 59% and 64% (308, 81 and 187 regions) in each of the three populations, respectively, overlap with genes (including

pseudogenes and non-coding RNAs) within the ~100 kb region around the peak signal, and among these, 258, 66 and 153 regions overlap with protein-coding genes in each population, respectively. The candidate regions are highly enriched with genes, when compared with that of randomly chosen regions across the genome ($p < 0.001$). They are also highly enriched in protein-coding genes compared to random regions ($p < 0.001$). Some candidate regions overlap with multiple genes, and as we believe that each candidate region should only have one selection target, we chose the gene(s) closest to the peak window as the candidate gene(s). We thus identified 275, 69 and 160 protein-coding genes that may have undergone positive selection in CEU, CHB+JPT and YRI populations, respectively (Table 3.3; Appendix D, candidate regions and protein-coding genes in each population). In a few cases, we identified two candidate genes in one region, either because these two genes have the same distance from the peak signal, or because these two genes overlap with each other. We then counted the number of peak signals at upstream to the candidate gene, within the candidate gene, or downstream of the candidate gene. We found that in all three populations, the biggest proportion of peaks is within the candidate genes, compared to upstream or downstream of the candidate genes (Figure 3.5).

**Table 3.3 Number of candidate regions and genes in each population.**

| | CEU | CHB+JPT | YRI |
|---|---|---|---|
| Candidate regions | 477 | 137 | 290 |
| Candidate coding genes | 275 | 69 | 160 |
| Candidate regions with non-coding genes | 120 | 35 | 89 |

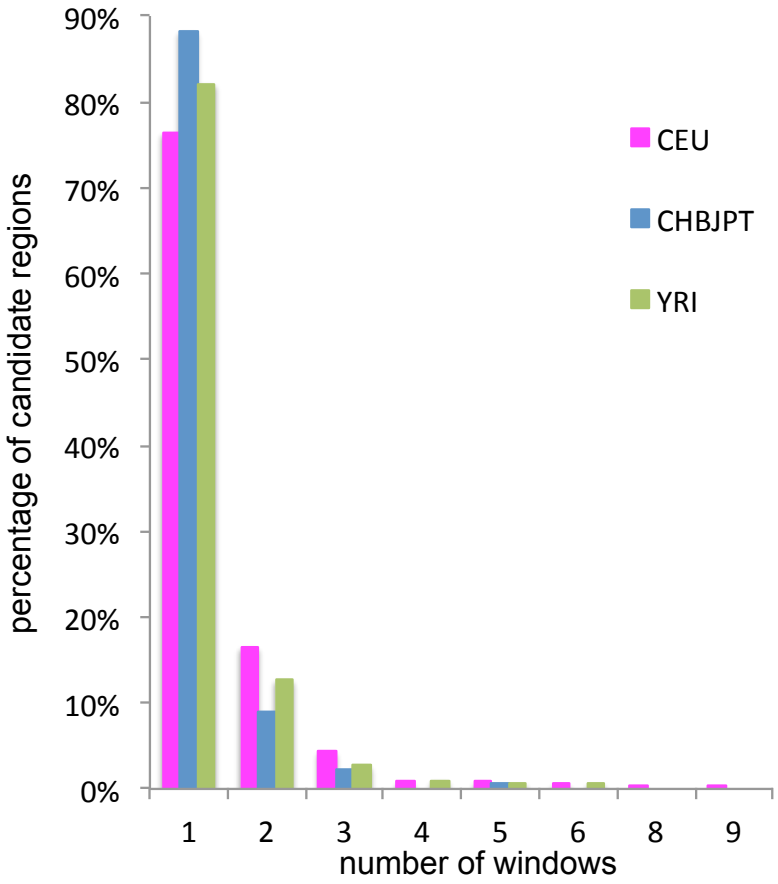


**Figure 3.4 Distribution of number of significant windows per candidate region in each population.**

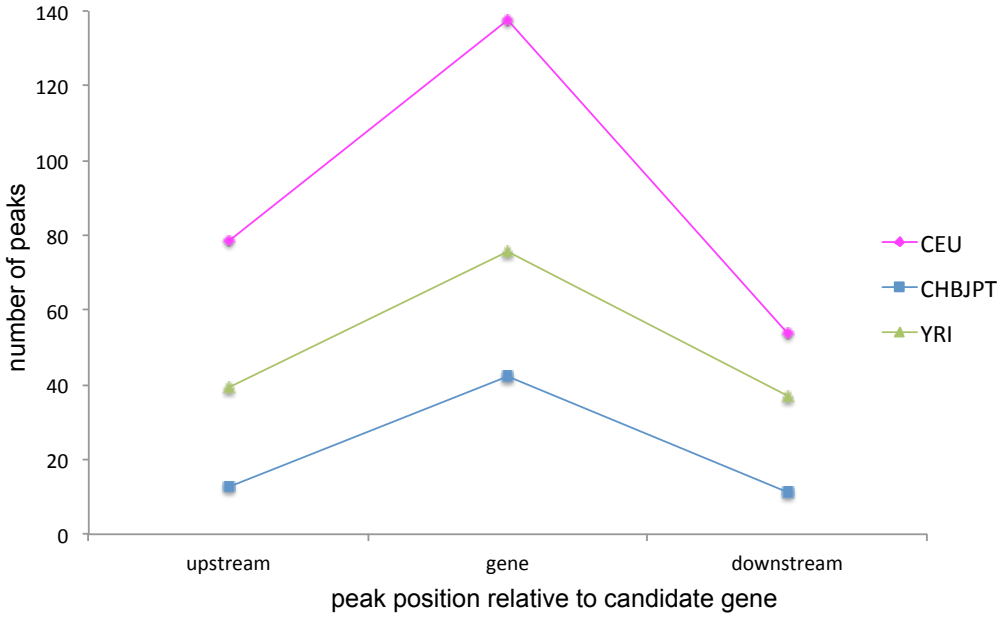


**Figure 3.5 Number of peak signals at each position relative to the candidate gene in each population.**

### 3.4.2 Comparison of candidate regions with previous studies

We compared our set of candidate regions with the list of 722 positively selected regions identified by at least two previous studies in Akey's review[100]. We found 100, 42 and 37 regions from those 722 regions that overlap with our list of candidate regions in CEU, CHB+JPT and YRI populations, respectively. Collectively there are 153 regions overlapping with our candidates (Appendix E). This is a high enrichment compared with randomly chosen regions from the genome ($p << 0.001$). Interestingly, we also found that within the candidate regions that overlap with Akey's list, a larger proportion was found to have evidence of positive selection in three or more previous studies (Figure 3.6). If we make a fair assumption that the more previous studies that have confirmed the candidate region, the more reliable the region is, then our list may represent a better set of candidate positively selected regions than the collection in Akey's review.

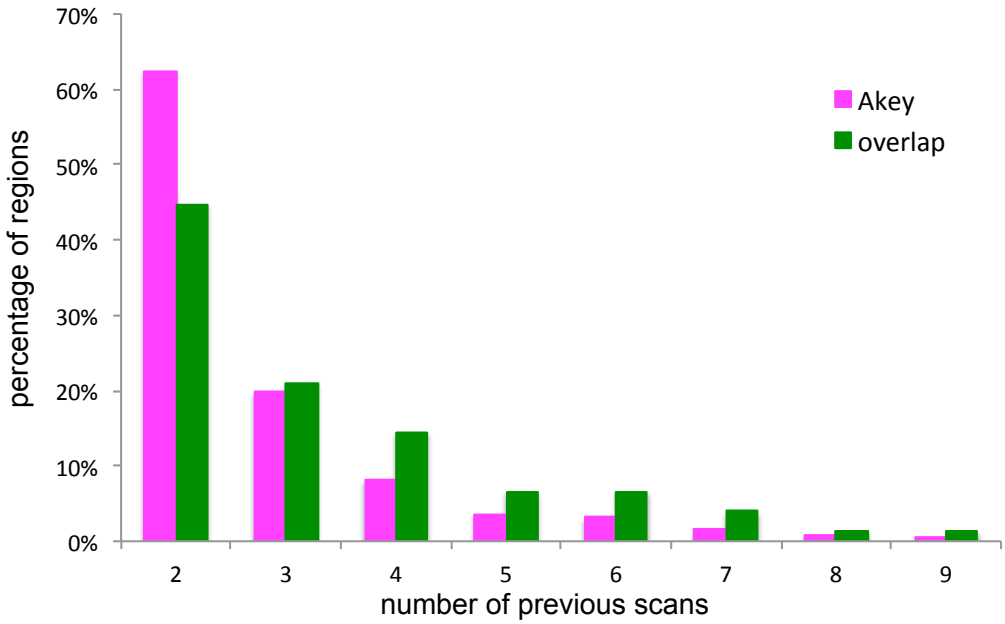


**Figure 3.6 Overlap of our candidate regions with Akey's review.** This plot shows the distribution of number of previous scans showing evidence of positive selection in all the candidate regions in Akey's review versus those overlap with our candidate regions.

### 3.4.3 Analysis of functional variants in candidate regions or genes

We then investigated whether or not our candidate genes were enriched with any particular type of functional variants. We looked at the overlap of our

candidate protein-coding genes with the synonymous and non-synonymous changes in 1000 Genomes Project low-coverage Pilot data[40]. We found that the percentage of non-synonymous changes with high derived-allele frequencies (DAF ≥ 0.5) overlapping with our candidate selected genes in CEU, CHB+JPT and YRI populations was 2.7%, 1.1% and 1.8%, respectively, while the percentage of synonymous changes with high DAF overlapping with our candidate genes is 3.0%, 0.8% and 1.4% in the three populations respectively. Interestingly, non-synonymous variants were enriched in all three populations ($p = 0.005$, 0.004, 0.001 in CEU, CHB+JPT and YRI, respectively), while in CEU and YRI populations, synonymous changes were also enriched ($p < 0.001$, $p = 0.005$, respectively) (Figure 3.7 A and B). In order to look further at the relationship between functional consequences of the non-synonymous changes and positive selection, we performed a Mann-Whitney test on Condel scores of high DAF (≥ 0.5) variants in our candidate genes versus the 1,000 random gene sets, and obtained 1,000 p values in each population. If the Condel scores in candidate genes are significantly higher, we should find a more-than-expected number of small p values in the distribution of the 1000 Mann-Whitney p values. However, our results showed that the distributions of p values are not skewed towards the lower end in all populations (Figure 3.8). This indicates that candidate genes may not be enriched in deleterious non-synonymous variants. It is worth noting that here "deleterious" does not necessarily mean "harmful" to the individual; it means that the variant can alter the structure and/or function of the protein that the gene encodes, and the impact on the individual can be either beneficial or harmful. Those deleterious variants with high frequencies in the populations, however, are highly likely to have some important functional impact and are thus worth further investigation.

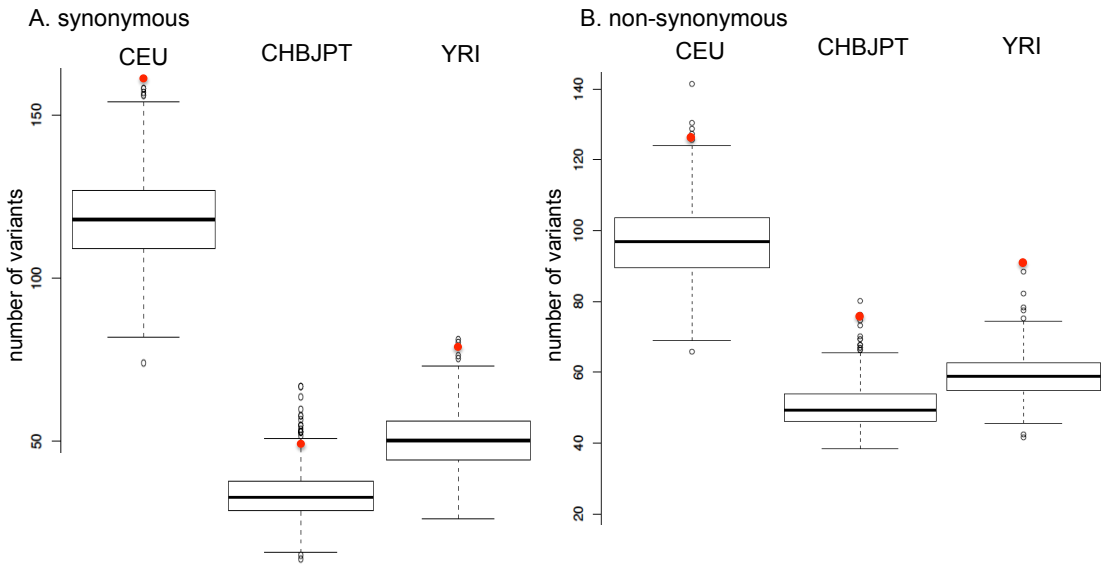


**Figure 3.7 Synonymous and non-synonymous variants in candidate regions.** These box plots show the distributions of the number of synonymous (A) or non-synonymous (B) changes in 1,000 sets of random genes that match the candidate genes. The upper and lower boundaries of the boxes show the 75$^{th}$ and 25$^{th}$ percentile, while the upper and lower lines show 1.5 times the IQR (interquartile range). The circles at each end of the box plots are data points that lie outside of 1.5IQR. The red dots are corresponding values of the candidate genes.

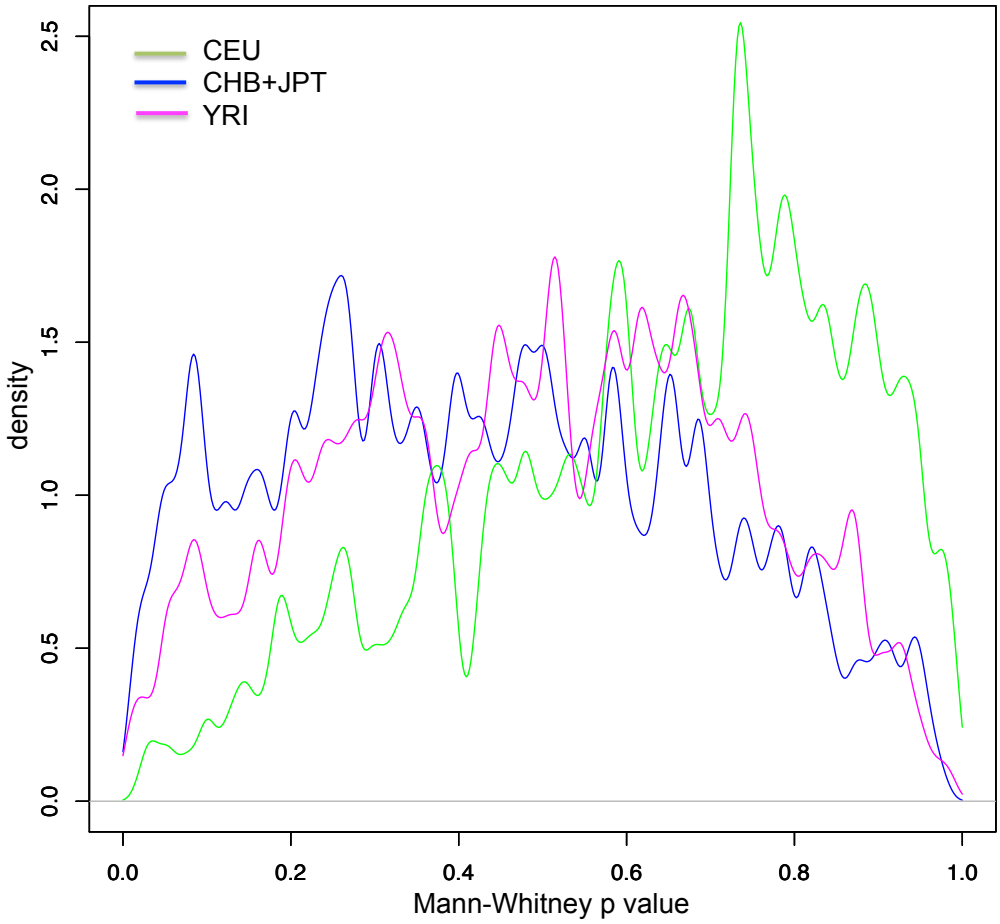


**Figure 3.8 Distribution of Mann-Whitney p values on Condel scores.**

We performed Gene Ontology clustering analysis on our candidate protein-coding genes in each population. Candidate positively selected protein-coding genes in the CEU population are highly enriched in proteins related to cell adhesion, signaling proteins and proteins with Ig-like C2-type 3 domain. Candidate protein-coding genes in YRI population are enriched in proteins with N-linked glycosylation sites, RhoGEF domains, and proteins involved in glutamate receptor activity (Table 3.4; see Appendix F for candidate genes within each enriched functional cluster). Perhaps due to the small number of candidate genes in the CHB+JPT population, there were no enriched functional clusters detected. Although functional clusters of candidate genes in each population are slightly different, they share some important similarities in terms of biological processes that they are involved in. All these enriched functional annotation clusters are involved in extracellular signal transduction and extracellular activities. More specifically, they are involved in the following three types of biological function: (1) Neurotransmission and synaptic plasticity, which are essential for learning and memory; (2) cell adhesion and migration, which plays important roles in the multicellular structure during early development, signal transduction and protein adsorption; and (3) immunological responses, which play an essential role in fighting with pathogens. These three areas are believed to play important roles in modern human evolution, thus it makes sense that they are highly enriched in genes that have undergone positive selection in the history of modern humans.

Apart from protein-coding genes, positive selection may also act on other functional elements in the genome. In order to investigate whether there is any enrichment of non-coding functional elements, we obtained annotation of variants within UTRs, non-coding RNAs, enhancers, and TF motif gains and losses. We calculated the distributions of number of such variants with higher than or equal to 50% DAF in the 1000 Genomes low-coverage Pilot data in each population in 1,000 sets of random regions matching our candidate regions, and looked at where the corresponding number in our candidate regions fall into these distributions. We found no significant enrichment of any of the five types of non-coding functional variants in our candidate regions (Figure 3.9). There

are three possible explanations for the lack of enrichment of non-coding functional variants. One is that selection on these regulatory elements might have been weaker and subtler in general, thus we were only able to identify a small proportion of them, which may not be representative of the whole set of positively selected non-coding functional elements. The second one is that our annotation of non-coding functional elements in the human genome has been very limited, in terms of both completeness and accuracy. The third one is that we did not categorize these functional elements based on their actual biological functions or processes. Positive selection may act on all types of non-coding functional elements, but favor certain types of biological function. However, due to our very limited understanding of the actual functions of those elements, we were unable to detect the enrichment.

**Table 3.4 Enrichments of functional clusters in the CEU and YRI populations.**

CEU

| Functional cluster | No. of genes | Bonferroni p-value |
|---|---|---|
| Cell adhesion | 27 | 0.001 |
| Signal | 74 | 0.002 |
| Ig-like C2-type 3 domain | 12 | 0.001 |

YRI

| Functional cluster | No. of genes | Bonferroni p-value |
|---|---|---|
| N-linked glycosylation site | 60 | 0.0007 |
| RhoGEF domain | 6 | 0.01 |
| glutamate receptor activity | 5 | 0.04 |

We then investigated published significant variants in Genome Wide Association Studies (GWAS) that fall into our candidate regions. We collected all the GWAS significant variants ($p \leq 5\times10^{-8}$) and identified those that are within our candidate regions in each population (Table 3.5). We found that a large number of HLA variants on chromosome 6 fell into our candidate regions in the YRI population, along with some other variants associated with infectious, autoimmune or inflammatory diseases. In the CEU population, skin/hair/eye

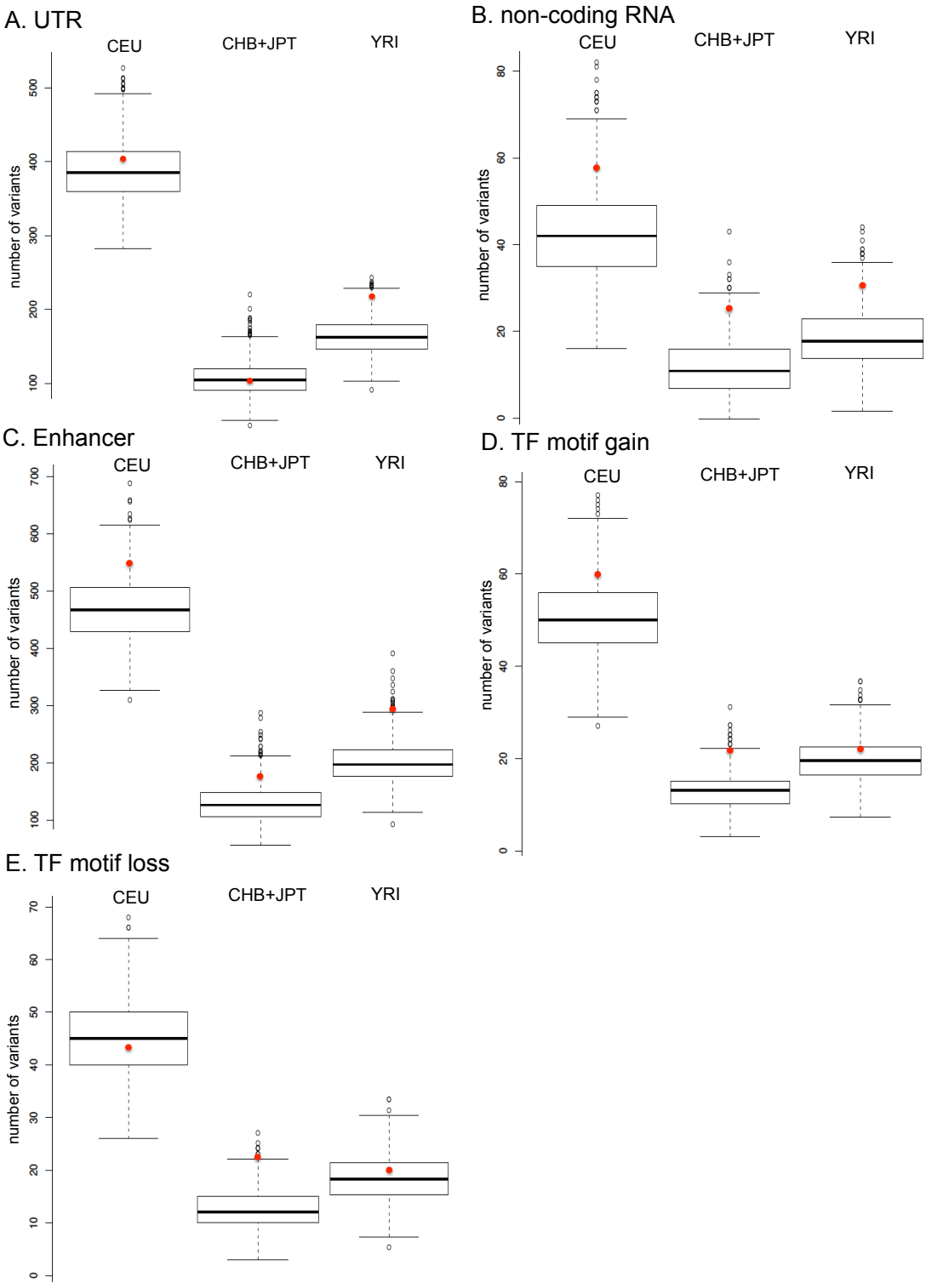


**Figure 3.9 Non-coding functional variants in candidate regions.** These box plots show the distributions of the number of UTR (A), non-coding RNA (B), enhancer (C), TF motif gain (D) and loss (E) variants in 1000 sets of random regions that match the candidate regions. The red dots are corresponding values of the candidate regions.

pigmentation variants overlap with our candidate regions. These reflect our general understanding of what types of traits are likely to be positively selected

in each continental population. However, we were not able to perform enrichment analysis on the GWAS significant variants in our candidate regions, for three reasons. First of all, the number of GWAS significant variants in each trait is small in most cases, and it varies substantially from one trait to another. So the power of detecting the enrichments in each trait is quite limited. Secondly, it is also not practical to categorize the traits that have been investigated by GWAS into a small number of meaningful types for enrichment analysis, as the traits are very diverse. Thirdly, the SNPs picked from previous GWAS studies might have some bias towards certain interesting traits, diseases or groups of genes, so they may not represent a whole-genome view of the functional variants. Having said that, the lists of GWAS significant variants overlapping with our positive selection candidates still provide valuable insights into what kinds of traits were under selection, and also give us some good candidate variants for further functional investigations.

**Table 3.5 GWAS significant variants in candidate regions in each population.**

A. CEU

| Chr | Position | rs ID | Gene(s) | Trait/disease | SNP risk allele | Frequency of risk allele | p value |
|---|---|---|---|---|---|---|---|
| 4 | 15,346,199 | rs11724635 | *BST1* | Parkinson's disease | rs11724635-A | 0.56 | 1E-16 |
| 4 | 15,347,035 | rs4538475 | *BST1* | Parkinson's disease | rs4538475-? | NR | 3E-09 |
| 6 | 30,026,078 | rs2517713 | *HLA-A* | Nasopharyngeal carcinoma | rs2517713-A | 0.62 | 4E-20 |
| 6 | 30,051,046 | rs6904029 | *HLA-A,HCG9* | Vitiligo | rs6904029-A | 0.29 | 1E-21 |
| 6 | 30,078,568 | rs7758512 | *ZNRD1, RNF39, HLA-A* | HIV-1 control | rs7758512-? | NR | 2E-08 |
| 8 | 19,863,608 | rs325 | *LPL* | HDL cholesterol | rs325-T | 0.89 | 8E-26 |
| 8 | 19,863,719 | rs326 | *LPL, C8orf35, SLC18A1* | Triglycerides | rs326-A | 0.78 | 5E-12 |
| 8 | 19,864,004 | rs328 | *LPL* | HDL cholesterol/Triglycerides | rs328-G | 0.09 | 2E-28 |
| 8 | 19,872,128 | rs10105606 | *LPL* | Triglycerides | rs10105606-C | 0.68 | 4E-26 |
| 8 | 19,875,201 | rs10096633 | *LPL* | Triglycerides | rs10096633-G | 0.88 | 2E-18 |
| 8 | 19,876,926 | rs17482753 | *LPL* | HDL cholesterol | rs17482753-T | 0.11 | 3E-11 |
| 8 | 58,468,572 | rs954295 | Intergenic | Longevity | rs954295-C | 0.39 | 4E-09 |
| 9 | 853,635 | rs755383 | *DMRT1* | Testicular germ cell cancer | rs755383-T | 0.62 | 1E-23 |
| 9 | 16,854,521 | rs2153271 | *BNC2* | Freckling | rs2153271-C | 0.41 | 4E-10 |
| 9 | 16,905,021 | rs3814113 | *BNC2, LOC648570, CNTLN* | Ovarian cancer | rs3814113-T | 0.68 | 5E-19 |
| 11 | 117,036,941 | rs10892151 | *APOA1, APOC3, APOA4, APOA5, DSCAML1* | Triglycerides | rs10892151-A | 0.028 | 3E-29 |
| 12 | 39,078,567 | rs11564258 | *MUC19, LRRK2* | Crohn's disease | rs11564258-A | 0.03 | 6E-21 |
| 15 | 26,039,213 | rs12913832 | *HERC2,OCA2* | Eye/hair color | rs12913832-A | 0.23 | 1E-300 |
| 15 | 46,179,457 | rs1834640 | *SLC24A5* | Skin pigmentation | rs1834640-G | 0.08 | 1E-50 |

B. CHB+JPT

| Chr | Position | rs ID | Gene(s) | Trait/disease | SNP risk allele | Frequency of risk allele | p value |
|---|---|---|---|---|---|---|---|
| 4 | 6,320,957 | rs4689388 | *WFS1, PPP2R2C* | Type 2 diabetes | rs4689388-T | 0.57 | 1E-08 |
| 4 | 6,353,923 | rs1801214 | *WFS1* | Type 2 diabetes | rs1801214-T | NR | 3E-08 |

C. YRI

| Chr | Position | rs ID | Gene(s) | Trait/disease | SNP risk allele | Frequency of risk allele | p value |
|---|---|---|---|---|---|---|---|
| 2 | 54,538,061 | rs11898505 | *SPTBN1* | Bone mineral density (spine) | rs11898505-A | 0.34 | 2E-08 |
| 4 | 1,068,187 | rs1670533 | *RNF212,SPON2* | Recombination rate (females) | rs1670533-C | 0.23 | 2E-12 |
| 4 | 1,085,281 | rs3796619 | *RNF212,SPON2* | Recombination rate (males) | rs3796619-T | 0.33 | 3E-24 |
| 4 | 88,994,267 | rs1471403 | *MEPE* | Bone mineral density (spine) | rs1471403-T | 0.34 | 2E-08 |
| 4 | 159,850,267 | rs8396 | *ETFDH* | Serum metabolites | rs8396-T | 0.3 | 4E-24 |
| 6 | 31,349,088 | rs13191343 | HLA | Psoriatic arthritis | rs13191343-T | 0.13 | 2E-72 |
| 6 | 31,360,375 | rs2524054 | *HLA-B* | CD4:CD8 lymphocyte ratio | rs2524054-A | 0.32 | 2E-28 |
| 6 | 31,360,904 | rs12191877 | *HLA-C* | Psoriasis | rs12191877-T | 0.15 | 1.-100 |
| 6 | 31,366,816 | rs9468925 | HLA | Vitiligo | rs9468925-? | 0.617 | 2E-33 |
| 6 | 31,371,730 | rs2894207 | *HLA-B,HLA-C* | Nasopharyngeal carcinoma | rs2894207-? | 0.82 | 3E-33 |
| 6 | 31,382,359 | rs9264942 | *HLA-C* | HIV-1 control | rs9264942-C | 0.34 | 3E-35 |
| 6 | 31,382,534 | rs10484554 | *HLA-C* | Psoriasis | rs10484554-T | 0.15 | 2E-39 |
| 6 | 31,420,305 | rs3134792 | *HLA-C* | Psoriasis | rs3134792-? | NR | 1E-09 |
| 6 | 31,430,538 | rs2523608 | *HLA-B* | HIV-1 control | rs2523608-G | 0.326 | 9E-20 |
| 6 | 31,435,043 | rs2523590 | *HLA-B* | HIV-1 control | rs2523590-C | 0.164 | 2E-13 |
| 6 | 31,444,079 | rs7743761 | MHC | Ankylosing spondylitis | rs7743761-? | NR | 5.-304 |
| 6 | 32,677,669 | rs477515 | *HLA-DQA1* | Inflammatory bowel disease | rs477515-? | 0.69 | 1E-08 |
| 6 | 32,681,607 | rs602875 | *HLA-DR-DQ* | Leprosy | rs602875-A | 0.68 | 5E-27 |
| 6 | 32,682,149 | rs615672 | *HLA-DRB1* | Rheumatoid arthritis | rs615672-? | NR | 8E-27 |
| 6 | 32,684,456 | rs9271100 | *HLA-DRB1* | Systemic lupus erythematosus | rs9271100-? | NR | 1E-12 |
| 6 | 32,685,358 | rs660895 | *HLA-DRB1* | Rheumatoid arthritis | rs660895-? | 0.21 | 1E-108 |
| 6 | 32,686,060 | rs674313 | *HLA-DRB5* | Chronic lymphocytic leukemia | rs674313-T | 0.26 | 7E-09 |
| 6 | 32,694,832 | rs9271366 | *HLA-DRB1* | Multiple sclerosis | rs9271366-G | 0.15 | 7E-184 |
| 6 | 32,700,715 | rs28421666 | *HLA-DQ,HLA-DR* | Nasopharyngeal carcinoma | rs28421666-? | 0.88 | 2E-18 |
| 6 | 32,710,985 | rs2040406 | *HLA-DRB,HLA-DQB1* | Multiple sclerosis | rs2040406-G | 0.26 | 1E-20 |
| 6 | 32,712,350 | rs9272346 | HLA | Type 1 diabetes | rs9272346-G | 0.61 | 5E-134 |
| 6 | 32,713,862 | rs2187668 | *HLA-DQA1, HLA-DQB1* | Celiac disease/Systemic lupus erythematosus | rs2187668-A | 0.26 | 1E-50/3E-21 |
| 6 | 32,733,847 | rs9273349 | *HLA-DQ* | Asthma | rs9273349-C | 0.58 | 7E-14 |
| 6 | 32,765,556 | rs7774434 | *HLA-DQB1* | Primary biliary cirrhosis | rs7774434-C | 0.371 | 3E-26 |
| 6 | 32,771,829 | rs6457617 | *HLA-DQA1, HLA-DQA2* | Rheumatoid arthritis/Systemic sclerosis | rs6457617-T | 0.49 | 5E-75/4E-17 |
| 6 | 32,771,977 | rs6457620 | *HLA-DRB1* | Rheumatoid arthritis | rs6457620-? | 0.5 | 4E-186 |
| 6 | 32,773,398 | rs10484561 | *HLA-DQB1* | Follicular lymphoma | rs10484561-G | 0.11 | 1E-29 |
| 6 | 32,775,888 | rs2647044 | *HLA-DRB1* | Type 1 diabetes | rs2647044-A | 0.13 | 1E-16 |
| 6 | 32,779,081 | rs13192471 | *HLA-DRB1* | Rheumatoid arthritis | rs13192471-G | 0.22 | 2E-58 |
| 6 | 32,786,977 | rs9275572 | *HLA-DQA2* | Alopecia areata | rs9275572-G | 0.59 | 1E-35 |
| 6 | 32,788,906 | rs7765379 | *HLA-DRB1* | Rheumatoid arthritis | rs7765379-? | NR | 5E-23 |
| 6 | 32,808,061 | rs2858884 | *HLA-DQA2* | Narcolepsy | rs2858884-A | 0.81 | 3E-08 |
| 6 | 122,187,733 | rs9398652 | *GJA1* | Resting heart rate | rs9398652-A | 0.1 | 4E-15 |
| 6 | 151,248,771 | rs11754661 | *MTHFD1L* | Alzheimer's disease (late onset) | rs11754661-A | 0.07 | 2E-10 |
| 6 | 160,601,383 | rs3127573 | *SLC22A2* | Serum creatinine | rs3127573-G | 0.13 | 7E-10 |
| 8 | 120,076,601 | rs2062377 | *TNFRSF11B* | Bone mineral density (spine) | rs2062377-T | 0.44 | 4E-16 |
| 8 | 120,081,881 | rs11995824 | *TNFRSF11B* | Bone mineral density (hip) | rs11995824-G | 0.55 | 7E-09 |
| 8 | 120,114,010 | rs6469804 | *OPG* | Bone mineral density (spine) | rs6469804-A | 0.51 | 7E-15 |
| 8 | 120,121,419 | rs6993813 | *OPG* | Bone mineral density (hip) | rs6993813-C | 0.5 | 3E-11 |
| 9 | 12,662,097 | rs1408799 | *TYRP1* | Blue vs. green eyes | rs1408799-C | 0.75 | 6E-17 |
| 9 | 138,251,691 | rs7849585 | *QSOX2* | Height | rs7849585-T | 0.33 | 5E-14 |
| 9 | 138,261,561 | rs12338076 | *LHX3, QSOX2* | Height | rs12338076-C | 0.34 | 2E-08 |
| 12 | 2,215,556 | rs1006737 | *CACNA1C* | Bipolar disorder and major depressive disorder (combined) | rs1006737-A | 0.36 | 3E-08 |
| 14 | 87,542,348 | rs8005161 | *GALC, GPR65* | Crohn's disease | rs8005161-T | 0.12 | 4E-18 |

## 3.5 Examples of strong candidate genes and their functions

In the final section of results in this chapter, we consider examples of individual selected genes of particular interest.

### 3.5.1 Examples of strong positively selected genes in a particular population

***CASP12***: previous studies have shown that a stop codon SNP, rs497116, which makes the protein non-functional, has been fixed or nearly fixed in European and Asian populations, but is less frequent in the African population. And this was believed to be due to positive selection acting on the inactive form of this gene[105,136]. If this stop codon allele is the selection target, it should have been selected in all three populations, as it has reached a very high frequency in all of them. In our genome wide scan, we found strong evidence of positive selection in the CEU population, as shown in Figure 3.10 A. In 1000 Genomes low-coverage Pilot data, the derived (stop codon) allele is fixed in both CEU and CHB+JPT populations, and has a frequency of 0.924 in the YRI population. However, we do not see strong signals in the other two populations. There are two possible explanations. One is data bias. As this selective sweep is likely to have already been completed in the CHB+JPT population and be nearly complete in the YRI population, the detection power largely relies on the presence of extremely low frequency alleles. As will be discussed later, due to the nature of low-coverage sequencing, the extent to which singletons were filtered out in each population was different. The variant data in the CEU population have a much higher percentage of singletons than the other two populations, so the detection power of this particular sweep may be higher in CEU. The other possible reason is that the selective sweep happened independently in these three populations, and thus the strengths and ages of the sweeps were different. This may have caused the sweeps in the other two populations to be undetectable by our tests. Nevertheless, it is encouraging that we have been able to obtain a very strong signal of positive selection in this known selected gene in exactly the same window as the selected allele, which was not detected by previous genome-wide scans using genotype data.

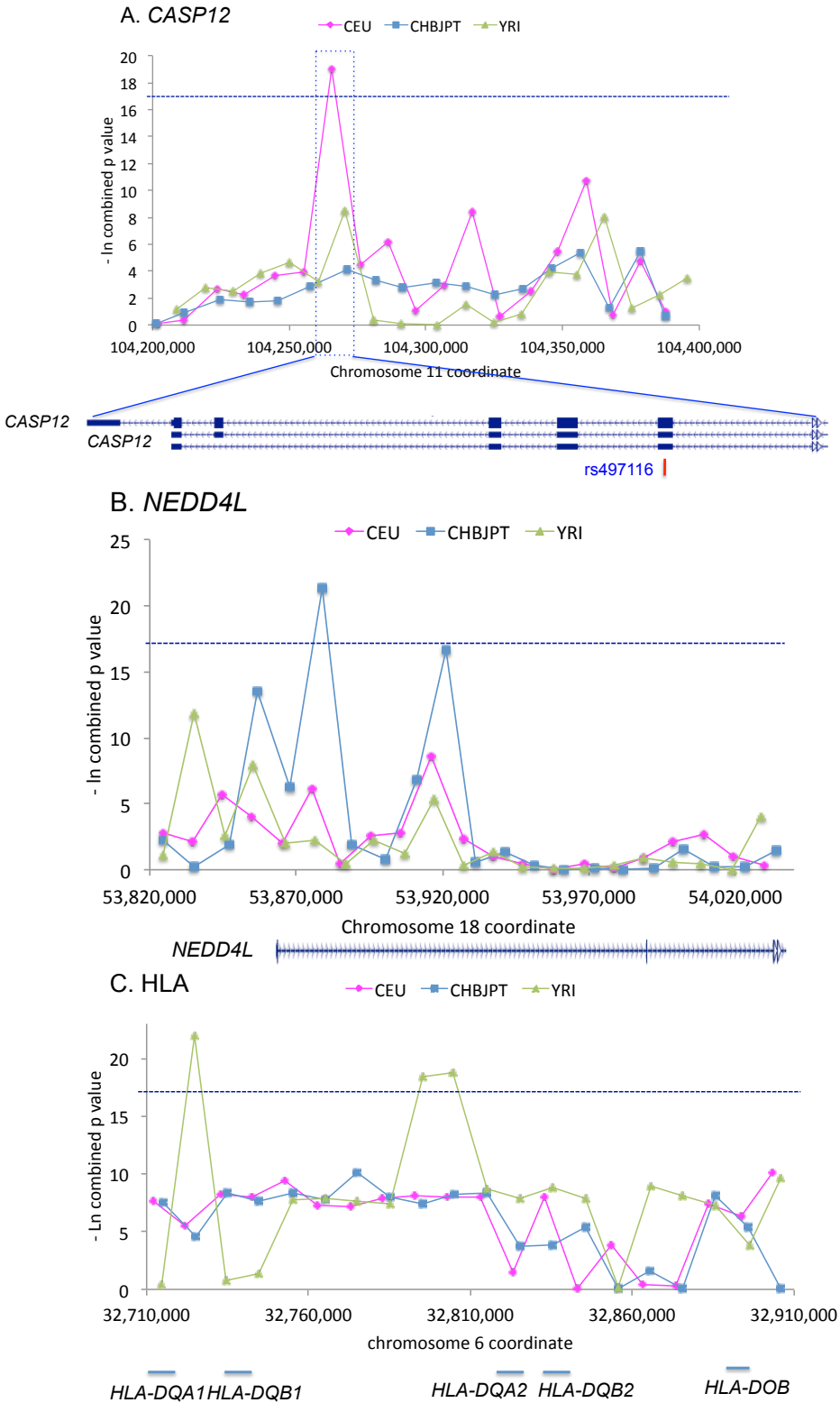


**Figure 3.10 Examples of positively selected genes with signals in only one population.** Blue dashed line marks the significance threshold. Candidate genes are shown and positions of putative selected SNPs are marked as red bars with the rs ID if applicable.

***NEDD4L***: This gene shows a very strong signal of positive selection in the CHB+JPT population, but not in the other two populations (Figure 3.10 B). The

gene encodes the enzyme E3 ubiquitin-protein ligase NEDD4-like, which is believed to regulate the expression and function of the epithelial sodium channel[137,138]. It plays a very important role in salt reabsorption. Studies have shown that this gene is associated with salt sensitivity[139], blood pressure[140], and essential hypertension[141]. Interestingly, it has been reported that African-Americans are more sensitive to salt than other groups in the US, and they develop hypertension at younger ages, with more severe consequences. So it appears that Africans are more sensitive to salt than other groups. Based on these facts, it is plausible that salt-insensitivity has been positively selected outside of Africa, due to the adaptation to the new environment. The climate was hot and dry in most human habitats in Africa, and salt was rare in ancient times, so retaining salt in the body was very important for the survival of humans. However, when our ancestors moved out of Africa, the climate was cooler, and salt was easier to access especially near the sea, so retaining salt in the body was no longer advantageous, and sometimes could be harmful, as it may cause high blood pressure. Therefore, there might have been a selective force favoring less efficient salt reabsorption in out-of-Africa populations. However, if this is the case, we should expect to see signals in both European and Asian populations. There are two possible reasons that we did not see signals in the European population. One is that the selective sweep might have happened earlier in Europe than in Asia, or the strength of selective force was much higher in Europe, so that the selective sweep had already been completed for a long time, therefore the footprint of positive selection had faded. The other explanation might be that the selection strength in Europe is very low, so the sweep has not reached to a detectable stage. All in all, the strong signal of positive selection plus the interesting functional implications of this gene makes it a very good candidate for further studies on its roles in salt sensitivity and blood pressure, and its association with hypertension. It may be worth doing functional analyses on highly differentiated alleles between African and other populations within this gene to find out which variant(s) is more likely to be the selection target.

**HLA gene cluster**: The HLA gene cluster on Chromosome 6 showed very strong signals of positive selection in the YRI population (Figure 3.10 C). The HLA

(human leukocyte antigen) system lies within the human major histocompatibility complex (MHC). This cluster contains a large number of genes related to the immune system of humans. There are different classes of HLA genes, and they play important roles in disease defense, may cause organ transplant rejections, and mediate autoimmune diseases. Many variants in this gene cluster are associated with various autoimmune or inflammatory diseases, including inflammatory bowel disease, HIV, Vitiligo, Ankylosing spondylitis, Rheumatoid arthritis and so on (Table 3.5). The positive selection signals in this locus may indicate the strong selective force of disease defense and immune functions in the African population.

### 3.5.2 Candidate genes selected in multiple populations and implications for the selected functions

***ITSN2***: This gene shows extremely strong signals in all three populations (ranked within the top 10 strongest signals in each population; Figure 3.11 A). Strikingly, the peak signals in all three populations fall into the same windows, which is the first exon and promoter region of this gene. There are two adjacent windows showing almost the same strength of signal. Within this ~20 kb region, we identified 49 variants with a DAF of more than 0.9 in all three populations, one of which is within the first non-coding exon of the gene, and others in either intron or 3' UTR regions (Table 3.6). This gene encodes Intersectin-2, which is involved in the regulation of the formation of clathrin-coated vesicles[142], and also plays a role in clathrin-mediated induction of T-cell antigen receptor (TCR) endocytosis[143], and may regulate T-cell mediated immune responses.

***NCAM2***: This gene, neural cell adhesion molecule 2, shows very strong signals in all three populations (Figure 3.11 B). The protein encoded by this gene belongs to the immunoglobulin superfamily. It is a type I membrane protein and may play important roles in selective fasciculation and zone-to-zone projection of the primary olfactory axons. It is primarily expressed in the brain, where it is believed to stimulate neurite outgrowth and to facilitate dendritic and axonal compartmentalization[144]. Interestingly, the peak signal of the CHB+JPT population is more than 400 kb away from the peak signals of the other two

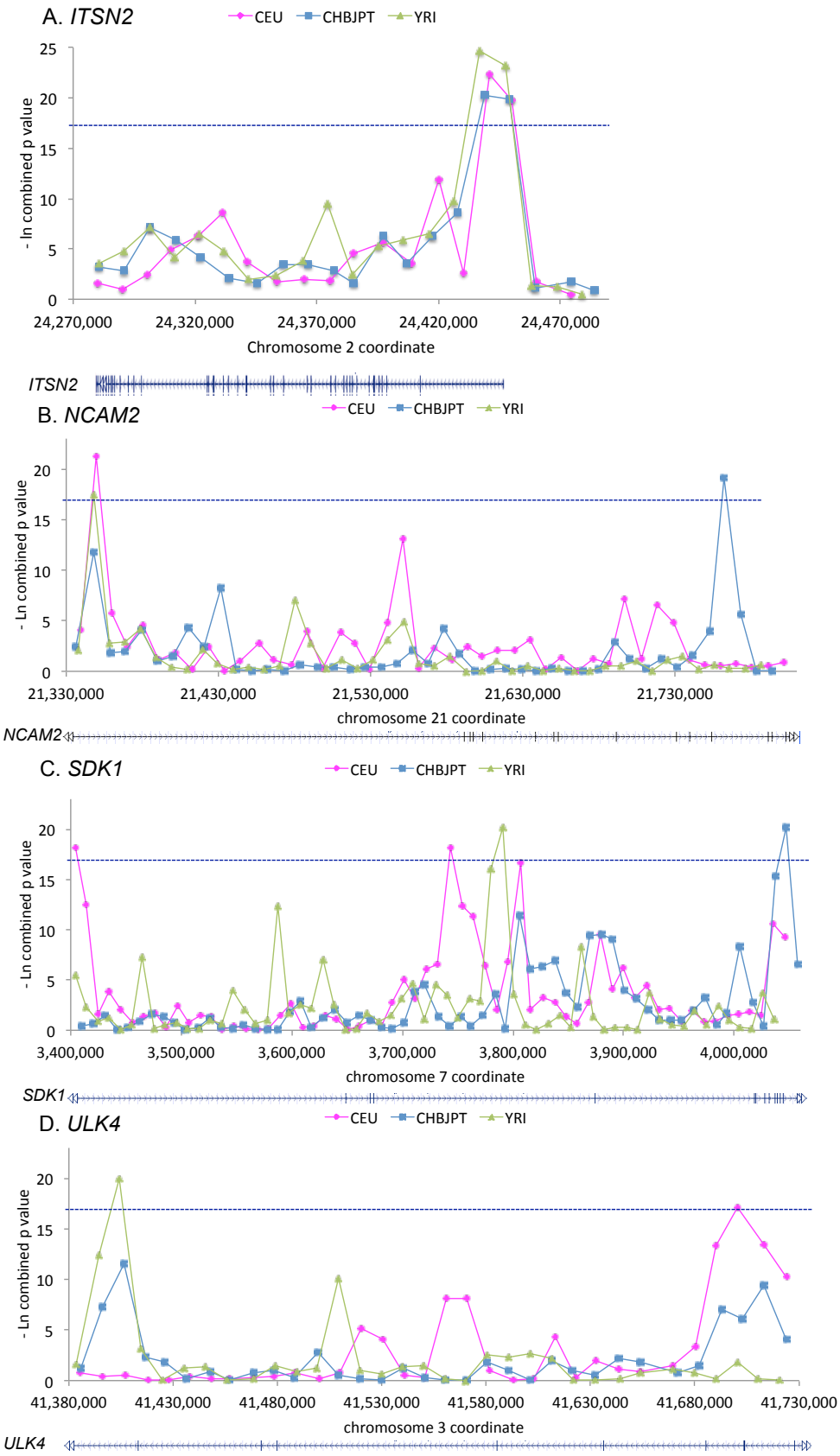


**Figure 3.11 Examples of positively selected genes with signals in multiple populations.**

**Table 3.6 High DAF variants in peak windows of ITSN2.** Chromosome coordinates are in March 2006, NCBI36.

| Chr | CEU position | ref allele | alt allele | ancestral allele | CEU DAF | CHBJPT position | CHBJPT DAF | YRI position | YRI DAF |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 24435849 | C | T | C | 0.983 | 24435849 | 0.967 | 24435849 | 0.966 |
| 2 | 24436130 | G | A | G | 0.975 | 24436130 | 0.967 | 24436130 | 0.975 |
| 2 | 24436273 | C | T | C | 0.983 | 24436273 | 0.967 | 24436273 | 0.966 |
| 2 | 24436426 | C | G | C | 0.983 | 24436426 | 0.975 | 24436426 | 0.966 |
| 2 | 24436979 | C | T | C | 0.975 | 24436979 | 0.967 | 24436979 | 0.966 |
| 2 | 24437367 | T | C | T | 0.983 | 24437367 | 0.967 | 24437367 | 0.966 |
| 2 | 24437522 | C | G | C | 0.983 | 24437522 | 0.967 | 24437522 | 0.966 |
| 2 | 24437726 | C | T | C | 0.908 | 24437726 | 0.975 | 24437726 | 0.966 |
| 2 | 24438162 | G | A | G | 0.967 | 24438162 | 0.95 | 24438162 | 0.966 |
| 2 | 24439534 | G | C | G | 0.983 | 24439534 | 0.975 | 24439534 | 0.915 |
| 2 | 24439653 | G | A | G | 0.983 | 24439653 | 0.967 | 24439653 | 0.966 |
| 2 | 24440355 | T | C | T | 0.983 | 24440355 | 0.967 | 24440355 | 0.966 |
| 2 | 24440851 | C | T | C | 0.992 | 24440851 | 0.967 | 24440851 | 0.966 |
| 2 | 24440929 | G | C | G | 0.975 | 24440929 | 0.967 | 24440929 | 0.966 |
| 2 | 24440930 | G | A | G | 0.975 | 24440930 | 0.967 | 24440930 | 0.966 |
| 2 | 24441809 | A | G | A | 0.983 | 24441809 | 0.975 | 24441809 | 0.966 |
| 2 | 24442311 | G | A | G | 0.975 | 24442311 | 0.967 | 24442311 | 0.966 |
| 2 | 24442435 | C | T | C | 0.983 | 24442435 | 0.967 | 24442435 | 0.992 |
| 2 | 24442604 | T | G | T | 0.992 | 24442604 | 0.967 | 24442604 | 0.966 |
| 2 | 24442639 | C | G | C | 0.983 | 24442639 | 0.967 | 24442639 | 0.966 |
| 2 | 24444362 | G | A | G | 0.983 | 24444362 | 0.992 | 24444362 | 0.966 |
| 2 | 24444623 | G | C | G | 0.983 | 24444623 | 0.967 | 24444623 | 0.975 |
| 2 | 24445579 | C | T | C | 0.992 | 24445579 | 0.967 | 24445579 | 0.975 |
| 2 | 24445841 | A | T | A | 0.983 | 24445841 | 0.967 | 24445841 | 0.949 |
| 2 | 24445880 | A | G | A | 0.983 | 24445880 | 0.967 | 24445880 | 0.966 |
| 2 | 24446357 | T | C | T | 0.983 | 24446357 | 0.967 | 24446357 | 0.966 |
| 2 | 24446367 | G | A | G | 0.983 | 24446367 | 0.967 | 24446367 | 0.966 |
| 2 | 24446904 | T | G | T | 0.983 | 24446904 | 0.967 | 24446904 | 0.975 |
| 2 | 24447399 | G | A | G | 1 | 24447399 | 0.967 | 24447399 | 0.966 |
| 2 | 24447452 | G | A | G | 0.992 | 24447452 | 0.967 | 24447452 | 0.966 |
| 2 | 24447481 | T | G | T | 1 | 24447481 | 0.967 | 24447481 | 0.966 |
| 2 | 24447753 | A | G | A | 0.975 | 24447753 | 0.967 | 24447753 | 0.966 |
| 2 | 24448832 | A | G | A | 0.983 | 24448832 | 0.967 | 24448832 | 0.966 |
| 2 | 24449141 | A | G | A | 0.958 | 24449141 | 0.967 | 24449141 | 0.992 |
| 2 | 24449259 | C | T | C | 0.983 | 24449259 | 0.967 | 24449259 | 0.966 |
| 2 | 24449274 | C | T | C | 0.983 | 24449274 | 0.967 | 24449274 | 0.975 |
| 2 | 24449318 | G | A | A | 0.942 | 24449742 | 0.933 | 24449318 | 0.949 |
| 2 | 24449992 | G | A | G | 0.992 | 24449992 | 0.967 | 24449992 | 0.966 |
| 2 | 24450279 | G | A | G | 0.992 | 24450279 | 0.975 | 24450279 | 0.966 |
| 2 | 24450287 | A | G | A | 0.983 | 24450287 | 0.975 | 24450287 | 0.966 |
| 2 | 24450338 | C | T | C | 0.983 | 24450338 | 0.975 | 24450338 | 0.966 |
| 2 | 24450541 | A | T | A | 0.983 | 24450541 | 0.967 | 24450541 | 0.966 |
| 2 | 24451714 | C | T | C | 0.983 | 24450866 | 0.033 | 24451714 | 0.966 |
| 2 | 24451783 | A | G | A | 0.6 | 24451783 | 0.742 | 24451783 | 0.483 |
| 2 | 24451815 | G | A | G | 0.983 | 24451815 | 0.983 | 24451815 | 0.983 |
| 2 | 24452162 | C | T | C | 0.983 | 24452162 | 0.967 | 24452162 | 0.966 |
| 2 | 24452243 | G | A | G | 0.983 | 24452243 | 0.975 | 24452243 | 0.975 |
| 2 | 24453789 | C | T | C | 0.983 | 24453789 | 0.967 | 24453789 | 0.975 |

populations, although CHB+JPT p value in that window is also quite low. All peak windows are in the intronic regions of this gene, and there are no functionally known variants.

***SDK1***: This gene also showed strong signals in all three populations (Figure 3.11 C). Interestingly, the peak signals of all three populations do not overlap, though the peaks of CEU and YRI are quite close. The product of this gene is a cell adhesion protein that guides axonal terminals to specific synapses in developing neurons. Studies have shown that dysregulation of this protein may play an important role in podocyte dysfunction in HIV-associated nephropathy[145,146]. It was also shown that a variant within this gene, rs645106, is associated with hypertension[147] in the Japanese population. This variant is not within any of the peaks, but is closest to the peak of the CHB+JPT population (about 100 kb downstream).

***ULK4***: This gene shows strong signals in both the CEU and YRI populations, and also low, although not significant based on our stringent threshold, p values in the CHB+JPT population (Figure 3.11 D). The CEU and YRI peak signals are more than 300 kb away from each other. The peak signal of the CEU population contains one exon of the gene. Previous studied have shown a strong association of *ULK4* with diastolic blood pressure (DBP)[148]. There are three linked high DAF non-synonymous changes within this gene that show significant GWAS signals: rs6768438, rs9816772 and rs9852991, but they are about 100kb upstream of the peak signal in CEU and even further from the YRI signal. It is likely that this gene plays important functional roles; however, very little is known about these functions. Thus it is worth further functional investigation.

## 3.6 Discussion

In this study, we have for the first time performed a genome-wide survey of positive selection in the human genome using low-coverage whole-genome sequencing data. We faced two main challenges: one was how to choose the genome-wide significance level of our tests; the other was how to localize the selection target. To solve the first challenge, we needed to decide between a higher sensitivity and better specificity from our scan. In this study, although we hoped to identify as many real positively selected regions as possible, we preferred to obtain a small list of very likely targets instead of a long list with a high proportion of false positives. Therefore, we needed to achieve a small FDR.

With this in mind, we looked at the FDR in our simulations under different p value cutoffs, and decided to choose the one with an FDR less than 5% in all three populations. Interestingly, this p value cutoff is 0.01 with Bonferroni correction, which is considered to be the most stringent significance cutoff. Although in this case our sensitivity is low, we are still able to identify interesting candidate regions, and we are able to achieve a very low FDR.

Although we could measure our specificity by calculating the FDR based on the neutral simulations, we were unable to reliably measure the specificity, i.e. the power of our test to detect positive selection. There are two main reasons for this. Firstly, unlike the neutral scenario, positive selection has different stages and strengths, and we do not know the strengths and ages of the selective sweeps that happened in the human genome. Although we could simulate several combinations of different selection coefficients and ages of sweeps, we are very unlikely to mimic the real situation. Secondly, in reality, there are many other factors that can affect the selective sweep, for example, change of environment, bottlenecks, population expansion, inbreeding, admixture, and so on. Although in our simulations, we used the best-fit demographic model to mimic the major population events, it was not a 100% replication of the real population history. Therefore, although we could measure the false negative rate of our simulation, it may not reflect the reality and may be misleading. For example, in our simulations, we had 16 scenarios of selection, among which we could only effectively detect selective sweeps with a selection coefficient of at least 0.007, and an age of at least 1,500 generations. We found that in the empirical data, we had a large number of windows with much lower p values then the lowest p value in our selection simulations, indicating that there may have been much stronger selection in our genome. Therefore, our simulations could only provide general guidance of how strong the selection has to be in order to be readily distinguished from the neutral scenario. But needless to say, this information is crucial in our study.

It is worth noting that the number of candidate regions and genes in the CHB+JPT population was much smaller than the other two populations. We believe that this was due to the lower quality data in this population. The

proportion of singletons in the CHB+JPT population is only about one third of CEU and half of YRI population (Figure 3.12). If we assume that the whole-genome frequency spectra of the European and Asian population should be similar, this lack of extremely low frequency alleles in the CHB+JPT population is largely due to the heavy filtering of uncertain variants during quality control. As our tests are looking for extreme patterns of the frequency spectra, this will affect the strengths of our signals. Although we have filtered our neutral simulations to match the frequency spectra of low-coverage Pilot data, this still could not fully eliminate the bias, as the proportion of extremely low frequency alleles will be much larger in regions under positive selection, whereas the missing alleles in the variant calling process of the empirical data should be pretty much randomly distributed. Therefore, more low frequency alleles will be missing in regions with an excess number of them. Therefore, it is understandable that the power of detection in the CHB+JPT population was much lower, and this should not be mistakenly interpreted as less selection in this population.

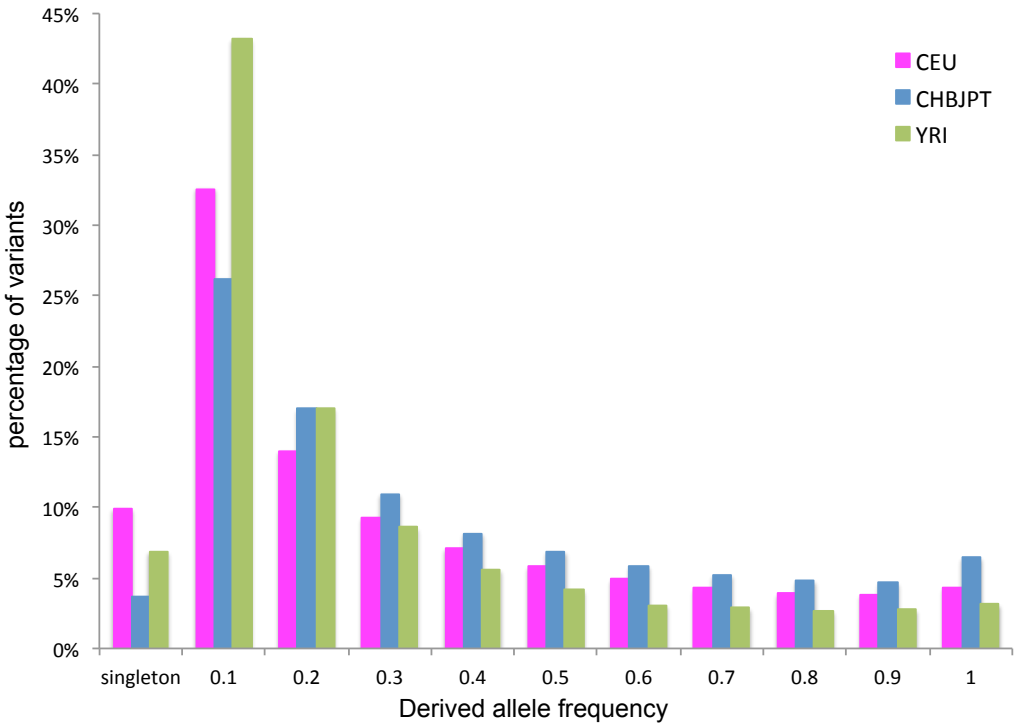


**Figure 3.12 Frequency spectra of 1000 Genomes low-coverage pilot data in each population.**

Although in our case, we used a very stringent p value significance threshold in order to obtain a confident list of selected regions, the cutoff is by no means black and white. As indicated by our simulations, relatively weak selective sweeps or sweeps that have not yet reached a late stage, may have more moderate p values. In fact, some genes that are known to have undergone a selective sweep may not have very strong signals. For example, the gene *EDAR*, which is related to hair thickness and tooth morphology, showed multiple evidence of positive selection in the East Asian population in previous studies[66,67,78,149-153]. In our scan, *EDAR* showed a peak p value of $1.4\times 10^{-7}$ ($-\log_e$ value 15.8) in the CHB+JPT population (Figure 3.13). Although this did not pass our genome-wide significance threshold, it would be considered as a significant p value if the threshold was slightly lower. Furthermore, as discussed earlier, due to the ascertainment bias in the CHB+JPT data, the level of significance of p values in this population is much lower than in the other two populations (Figure 3.13). Interestingly, the density of significant p values corresponds to the proportion of singletons in the data in each population (Figure 3.12 and Figure 3.14). On one hand, this demonstrates the importance of extremely low-

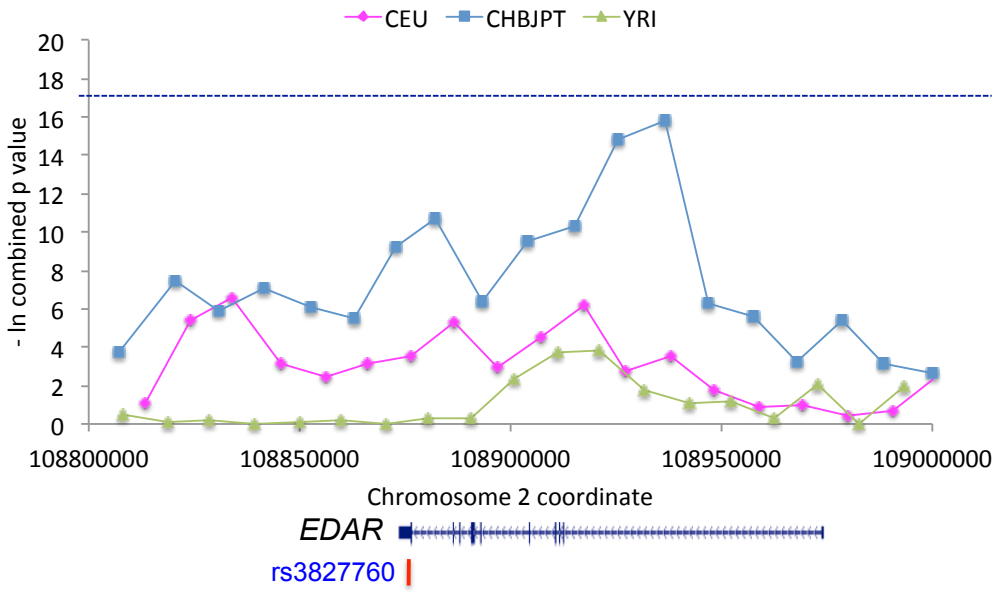


**Figure 3.13 Signals of positive selection of EDAR gene in the CHB+JPT population.** As observed previously, the peak signal lies in an intron, and not over the non-synonymous SNP rs3827760 often assumed to be the target of selection.

frequency alleles for detecting selection signals. On the other hand, it shows us that although for a genome-wide scale study like this, we may set up a stringent significance threshold to start with, we should not ignore the many other signals that are not so strong but may still indicate signals of positive selection. However, for those cases, stronger independent supporting evidence may be needed to confirm the signals of positive selection.

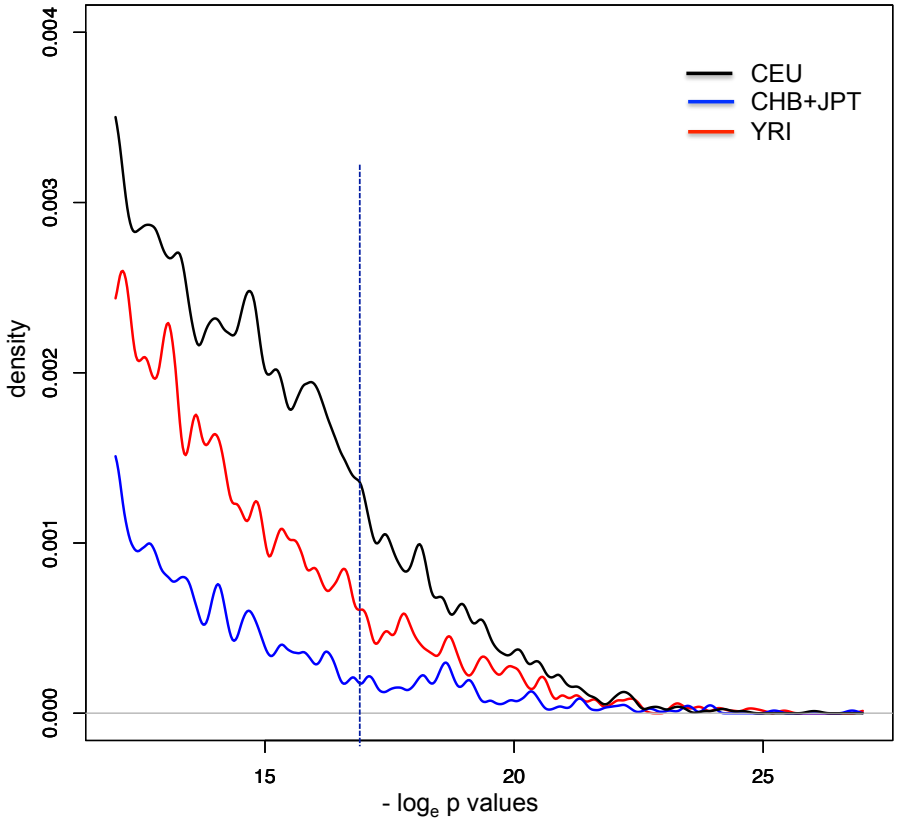


**Figure 3.14 Distributions of p values in three populations.** In order to show the difference of densities of significant p values in each population, we only showed the distributions of those (-$\log_e$ p) values bigger than 12 (equivalent to p values smaller than 6.1E-6). The blue dashed line is the significance threshold.

With a much higher density of variants in the sequencing data, we were hoping to achieve a better resolution of signals, which may lead to higher power of localization of selection targets. In our genome-wide scan, we used windows sized about 10 kb, which in general contain enough variants to have the statistical power, and at the same time are small enough for further investigation to identify the selected variant. However, although on average, it is mostly likely that the selected allele will fall into the window with the strongest signal, there is still a high chance that the selected allele is elsewhere. Our simulations

suggested that there is about 75% chance that the selected allele will be within the ~100 kb regions centered by the peak signal, though the signal pattern is made more complicated by recombination near the selected allele. Therefore, although it is still not easy to localize the selection target into a very small region or even a variant, by taking into account recombination, we were able to localize the selected region into a reasonable size for further investigations.

The identification and interpretation of biological targets of selection has for long been one of the biggest challenges in human evolutionary genetics. Two main constraints limit our abilities to do so: one is the low power of current statistical approaches to narrow down the selected genomic region, and the other is the limited understanding of functions of our genome. We have shown here that in some cases, selection targets can be narrowed down to a few tens of kb, so that functional variants can be sought and investigated further. However, due to the lack of known functional elements within many candidate regions, biological targets of selection are often hard to identify and interpret. Follow-up biological experiments can sometimes be done to investigate functions of plausible selected variants, but it is often time- and resource-consuming, and difficult to carry out on a large scale. New experimental assays to examine biological functions of variants on a large scale will be extremely beneficial for the investigation of biological targets of selection.

# 4 A search for genomic regions with the most recent coalescence times in all humans

## 4.1 Introduction

One of the most interesting questions for human evolutionary geneticists is whether or not there were genetic contributions to the emergence of modern humans around 200 KYA, and to the uniqueness of modern humans compared to other species, including archaic humans. Two hypotheses can be made. One is that all the necessary genetic changes were already present in the genomes of our immediate ancestors before the emergence of modern humans, and those mutations might have occurred at different times. In combination with environmental and social or cultural factors, they led to the emergence of modern human traits and behaviors at the times discussed in Chapter 1. The alternative hypothesis would be that some important mutations occurred shortly before modern humans emerged, and those mutations were so advantageous that they spread quickly among our ancestors, which then contributed to the traits of modern humans and thus the emergence of our species. If the first hypothesis were true, then there would be no or very few human-specific variants of genes or other functional regions in the human genome that are shared between all humans with relatively low diversity, but are not present in this form in our immediate ancestors or sister species. In contrast, if the second hypothesis were true, then there would have been some strong selective sweeps in the genomes of early humans, and those sweeps would have reached fixation in our African ancestors before fully modern humans emerged and the current populations split. This would have resulted in shared haplotypes in all humans at those selected loci, and those haplotypes would likely be human-specific, i.e. they would not be present in our sister species.

Under the second scenario, the identification of such genomic loci would provide great insights into the genetic uniqueness of modern humans. The common statistical approaches for detecting recent positive selection, however, have

almost no power to identify positive selection that started more than 100 KYA. The main reason for this is that such positive selection events are likely to have reached fixation before 100 KYA, and thus the signatures of selection on the patterns of LD or frequency spectra would have been erased by recombination or new mutations after the completion of those sweeps. There also would not be any population differentiation, as those selection events should have happened before modern human populations split. So the statistical approaches mentioned earlier are not able to detect such older selection events. Therefore, new approaches that do not rely on these patterns of variation in contemporary humans need to be applied in order to identify these regions.

Because of the diploid nature of the human genome and the action of recombination, different pieces of our genome derive from different common ancestors. According to coalescent theory, the expected time to the most recent common ancestor (TMRCA) of a genomic segment in a diploid population is $4N_e$[85,154]. For modern humans, although many studies have used genetic data to estimate effective sizes, realistic effective population sizes of both subpopulations and the global population are still unclear. Based on the Wright-Fisher model, the global ancestral population size of modern humans is $N_e$ = 10,000, and the present-day continental populations may have an effective population size of around 100,000[26], due to the recent expansion of human populations after the agricultural revolution. If we assume 20 years per generation, the expected average TMRCA of a particular non-recombined region in current global human population might be around 800,000 years. However, the TMRCA of different regions in the human genome must vary, and it is not easy to estimate the variation of TMRCAs between different genomic regions only based on the estimates of population parameters.

As has been noted, anatomically modern humans emerged around 200 KYA. This ancestral human population lived in Africa (with a temporary expansion into the Levant) until around 50-60 KYA, when a subgroup of them with fully modern characteristics migrated out of Africa and populated other parts of the world. Selective sweeps on alleles that contributed to modern human traits should have occurred around the time when modern human emerged, and should have

reached fixation before the out-of-Africa migrations. Therefore, if we trace back to the common ancestor of one of these loci, the TMRCA should be around or slightly more than 200,000 years. If we use the unit of $2N_e$ generations, and an $N_e$ of 10,000 for the human population, the TMRCA should be around or a little more than 0.5 and less than 2, as 2 should be the expected value of TMRCA for a diploid Wright-Fisher population. Therefore, the TMRCA of the selected locus should be much less than what we would expect from a neutral region, so we may distinguish these regions that had undergone a complete selective sweep during modern human evolution from neutral regions by calculating TMRCAs of human genomic regions and identifying the most recent ones.

In this study, we aimed to answer two questions: (1) are there regions in the human genome that support the second hypothesis, and, if the answer is "yes", (2) where are these regions and what functions do they have? To achieve this goal, we calculated TMRCAs of 5 kb non-overlapping windows in the human genome with relatively low diversity/divergence ratio from 54 unrelated human samples from 11 populations around the globe, using high-coverage whole-genome sequencing data. Then we compared the distributions of TMRCAs in the empirical data with simulated neutral regions. We also compared the variants of humans in regions with a TMRCA of less than $2N_e$ generations with those in a high-coverage Denisovan genome, to see whether or not these regions have the characteristics of strong classic selective sweeps. Public datasets were used, and all analyses described in this chapter were performed by the author of this thesis.

## 4.2 Materials and Methods

### 4.2.1 Data

To estimate the coalescence time of a particular genomic region, we need the complete set of single nucleotide variation in a set of unrelated samples. According to coalescent theory, in an unstructured population the probability of a sample size $n$ containing the most recent common ancestor of the whole population is $(n\text{-}1)/(n\text{+}1)$, so even with a small sample size of 10, we would still have a more than 80% chance to obtain the TMRCA of the whole population from the sample. However, due to the complex structures of human populations, in

order to obtain TMRCAs in all humans, we need samples that can represent at least all the main continental human populations. So in order to conduct a genome-wide survey of TMRCAs in humans, we needed high-coverage whole-genome sequencing data from a diverse collection of human samples. When this project started in 2010, there were 15 personal genomes sequenced at high coverage by different research groups around the world. These include a YRI and a CEU trio from the 1000 Genomes Project pilot 2[40], Venter's[155] and Watson's[156] genomes, one Chinese genome (YH)[157], two Korean genomes[158,159], two European genomes from Complete Genomics Inc.[160], and one Bantu and one Khoisan individual from southern Africa[161]. These individual genomes have diverse population backgrounds, thus formed a good sample of the global human population. We first used 13 out of these 15 individuals (excluding offspring in the two trios) to calculate coalescence times, but found that due to the diversity of platforms used in sequencing those genomes, and different algorithms applied in variation calling, the data quality was not consistent from one genome to another, and when putting these genomes together, there were a lot of genotype gaps and violation of the infinitely-many-sites model. Therefore, it was not useful to calculate coalescence times on these genomes.

In 2011, Complete Genomics Inc. (CGI hereafter) released 69 high-coverage whole genome sequences from a diverse panel of samples (http://www.completegenomics.com/sequence-data/). The consistency of sequencing platform and variants calling algorithm, together with the stringent quality control by CGI made this a much better data set to use for this study. Among these 69 samples, 54 are unrelated individuals, and these individuals are from 11 diverse populations (Table 4.1). So we decided to use these 54 genomes for coalescent time calculations and further analyses.

Low quality sites were removed and missing genotypes were filled before using these data for our analyses. Firstly, trialelic sites, telomere and centromere regions, as well as sites that are not consistent with the Mendelian inheritance in the CGI trios and the pedigree panel were excluded. Because of the highly diverse samples, we avoided using inference algorithms to infer missing genotypes, as inferences from a large number of mixed populations may be inaccurate. Instead,

we filled the majority of missing genotypes using the 1000 Genomes Project Phase 1 data in the same samples (34 samples in common) (http://www.1000genomes.org/). We then discarded sites that still had more than two missing genotype calls. For those with one or two missing genotypes, we assigned either the reference or alternative allele as the genotype based on the genotypes of other samples in the same population. After the filtering, around 95% of the SNPs were retained.

**Table 4.1 Sample information.**

| Population | No. of samples |
|---|---|
| ASW (African ancestry in Southwest USA) | 5 |
| CEU (Utah residents with Northern and Western European ancestry) | 9 |
| CHB (Han Chinese in Beijing, China) | 4 |
| GIH (Gujarati Indian in Houston, Texas, USA) | 4 |
| JPT (Japanese in Tokyo, Japan) | 4 |
| LWK (Luhya in Webuye, Kenya) | 4 |
| MKK (Maasai in Kinyawa, Kenya) | 4 |
| MXL (Mexican ancestry in Los Angeles, California) | 5 |
| PUR (Puerto Rican in Puerto Rico) | 2 |
| TSI (Toscans in Italy) | 4 |
| YRI (Yoruba in Ibadan, Nigeria) | 9 |

### 4.2.2 Divergence and diversity

Since it was not practical to calculate TMRCA across the whole genome using GENETREE, we first compared divergence and diversity. We calculated the intra-species diversity in 5-kb non-overlapping windows throughout the genome within these 54 humans by calculating the average pairwise difference per site in each window.

In order to calculate human divergence from the ancestor, we obtained the inferred ancestral state of each locus across the whole genome from Ensembl (http://www.ensembl.org/). The ancestral states are inferred from the six primates EPO (Enredo-Pecan-Ortheus) pipeline (see Ensembl website for

details). We then identified fixed derived alleles in humans based on the 54 CGI genomes and the ancestral alleles. Divergence per site on the same 5-kb non-overlapping windows was calculated as for diversity.

We further filtered the data by removing windows with less than 80% ancestral state information and/or less than 90% callable sites in the CGI data. This gave us 277,256 5kb windows (total length ~1,386Mb), which is about 46% of the genome. Then we calculated the diversity/divergence ratio for all these eligible windows across the genome.

### 4.2.3 TMRCA calculations

Firstly, we inferred haplotypes from the genotype data of the 54 samples in each window, using BEAGLE[162]. We used the five parent-offspring trios from the CGI sequence data (three CEU trios, one YRI and one PUR trio) to increase the accuracy of the phasing. We then pruned the data to fit the infinitely-many-sites model in order to build the gene tree, using the PRUNE algorithm[163]. Sites or samples that did not fit the model were removed. On average, ~13% of the SNPs were removed by PRUNE. In most windows, all samples were retained, and a maximum of two samples were pruned out. On average, 0.08 samples were removed per 5-kb window. We estimated the local mutation rate of each window by comparing the human reference sequence and the chimpanzee sequence, assuming that the split time between human and chimpanzee genomes was 7 million years ago, with 20 years per generation. We then calculated an initial estimation of theta ($4N_e\mu$, 4 times the effective population size times the local mutation rate) using the estimated mutation rate and a human effective population size of 10,000. We used the GENETREE[86,164-167] package to obtain the best theta of each 5 kb window using the above estimated theta as a seed, and then used the best estimate of theta to calculate the TMRCA using GENETREE (See Appendix G for parameters and command lines). We used 100,000 simulations in estimating the theta, but in order to increase the accuracy of the TMRCA estimation, we used 10,000,000 simulations in calculating the coalescence time. All the TMRCAs are in the unit of $2N_e$ generations.

### 4.2.4 Simulations

We simulated 1000 independent 100 kb neutral regions in 54 samples, using the *cosi* package[26] and the best-fit demographic model[26]. Due to the limited demographic models, only three main continental populations, i.e. African, European and Asian, were simulated. We categorized the 11 populations in the CGI samples into these three population groups, which gave us 22 Africans, 18 Europeans and 14 Asians. We first used *cosi* to generate a random recombination map using the distribution of recombination rates in autosomes in the deCODE genetic map[168], and then used this recombination map in the simulations. A genome-wide average mutation rate of $1.5\times10^{-8}$ and gene conversion rate of $4.5\times10^{-9}$ were used. All other parameters are the same as in previous simulations.

### 4.2.5 Comparison with two high-coverage southern African genomes and a high-coverage Denisovan genome

We picked all the 5-kb windows with a TMRCA of less than $2N_e$ generations, and combined adjacent windows into one region. Then we picked regions with at least two adjacent windows (10 kb) to form a list of 143 regions with recent TMRCAs. These regions have the lengths of 10 kb to 25 kb. We used this set of regions for comparison with other genomes.

In order to investigate whether or not these regions with recent coalescence times calculated from CGI data are likely to have undergone strong selective sweeps during the emergence of modern humans, we compared the variants in the 143 regions with those in two high-coverage southern African genomes – one Bantu and one Khoisan[161]. The Bantu sample ABT was sequenced to over 30-fold coverage using the SOLiD 3.0 platform from Applied Biosystems. The Khoisan sample KB1 was sequenced by two platforms: 10.2-fold coverage using the Roche/454 GS FLX platform, plus 12.3-fold non-redundant clone coverage with long-insert libraries, and 23.2-fold using the Illumina platform[161]. We used the variation data generated by the authors. We also compared the variants with those of a Denisovan genome, sequenced by Reich et al. (http://www.eva.mpg.de/denisova/), with approximately 30-fold coverage using the Illumina GAIIx sequencing platform[12]. First of all, we called variants

differing from the human reference genome GRCh37 from the alignment generated by the authors, using SAMtools[169]. We used a maximum read depth of 100 as a filter (~3 times of the average read depth). We then further filtered out heterozygous calls where the ratio of the second-highest:total read depth was less than 0.3:1, or the second-highest read depth was less than 2. Then we obtained all variants in the 54 CGI samples, two southern African genomes and the Denisovan genome within the 143 regions with recent TMRCAs, as well as 100 sets of random windows matching the number of windows in the recent coalescent regions. Firstly, we used the two southern African genomes to validate our human-fixed derived alleles. Only those derived alleles that were fixed in both the CGI and the southern African samples were considered as fixed derived alleles in humans. We then counted the number of the following four types of loci in each set of regions: (1) the derived allele was only seen in the Denisovan genome: a "Denisovan specific variant"; (2) the derived allele was fixed in humans but not seen in the Denisovan genome: a "human specific variant"; (3) the derived allele was seen in both humans and the Denisovan genome, with a frequency in the 54 humans higher than or equal to 50%: "high DAF shared variant"; and (4) the derived allele was seen in both humans and the Denisovan genome, with a frequency in the 54 humans less than 50%: a "low DAF shared variant". In order to test whether or not there was any enrichment, we randomly picked 100 sets of windows with calculated TMRCAs, matching the number of windows in our recent coalescent region set. Then we ranked the numbers of these four types of derived alleles in the recent coalescent regions against the 100 random sets of matched windows to see if any type of alleles was enriched in the recent coalescent windows compared to the random windows.

### 4.2.6 Phylogenetic network analysis on regions with recent TMRCAs

In order to further understand the relationship between the haplotypes in humans and the Denisovan, we performed phylogenetic network analysis on some regions with recent TMRCAs using the NETWORK software[170] ( http://www.fluxus-engineering.com/sharenet.htm ). Human haplotypes were inferred using BEAGLE as described before, and heterozygous sites in the Denisovan were assigned to the two chromosomes manually based on the

similarities with the human haplotypes. For those Denisovan variants that are not shared with humans, alleles were randomly assigned to the two haplotypes. Then these haplotypes were grouped into African (ASW, LWK, MKK, YRI and southern African), European (CEU and TSI), Asian (CHB, JPT and GIH), other human populations (MXL and PUR), and Denisovan. Phylogenetic networks were built, and each node was marked with colors representing the relevant population group(s).

## 4.3 Results

### 4.3.1 Divergence and diversity

We expect that regions in the genome with low diversity compared to divergence tend to have more recent common ancestors than regions with high diversity compared to divergence. Therefore, we first calculated intra-species diversity within the 54 humans, and the inter-species divergence of humans and chimpanzees. The local diversity of 5kb windows in the 54 samples ranged from 0% to 0.39% per nucleotide, with the median of 0.07% per nucleotide. This means that on average, in a 1-kb long region, two randomly drawn chromosomes would be expected to have 0.7-nucleotide difference. This was in line with the widely-accepted estimation that two random individual chromosomes would on average have one nucleotide difference per kb. The local divergence on the same data, based on the comparison with inferred ancestral data from six primates, ranged from 0% to 1.27%, with the median of 0.50%. The diversity/divergence ratio ranged from as small as 0.002 to as large as 200, with a median of 0.145. The distribution of diversity/divergence ratio has a long tail on the right-hand side (Figure 4.1).

### 4.3.2 TMRCA distribution on low and high diversity/divergence regions

As discussed earlier, for a diploid population, the TMRCA in a Wright-Fisher population is expected to be $4N_e$. As the TMRCAs calculated by GENETREE are in the unit of $2N_e$, we should expect an average TMRCA of 2 across the genome. To test whether or not the TMRCAs calculated by GENETREE on these 54 samples reflect our expectation, we calculated TMRCAs on windows with 1% lowest and

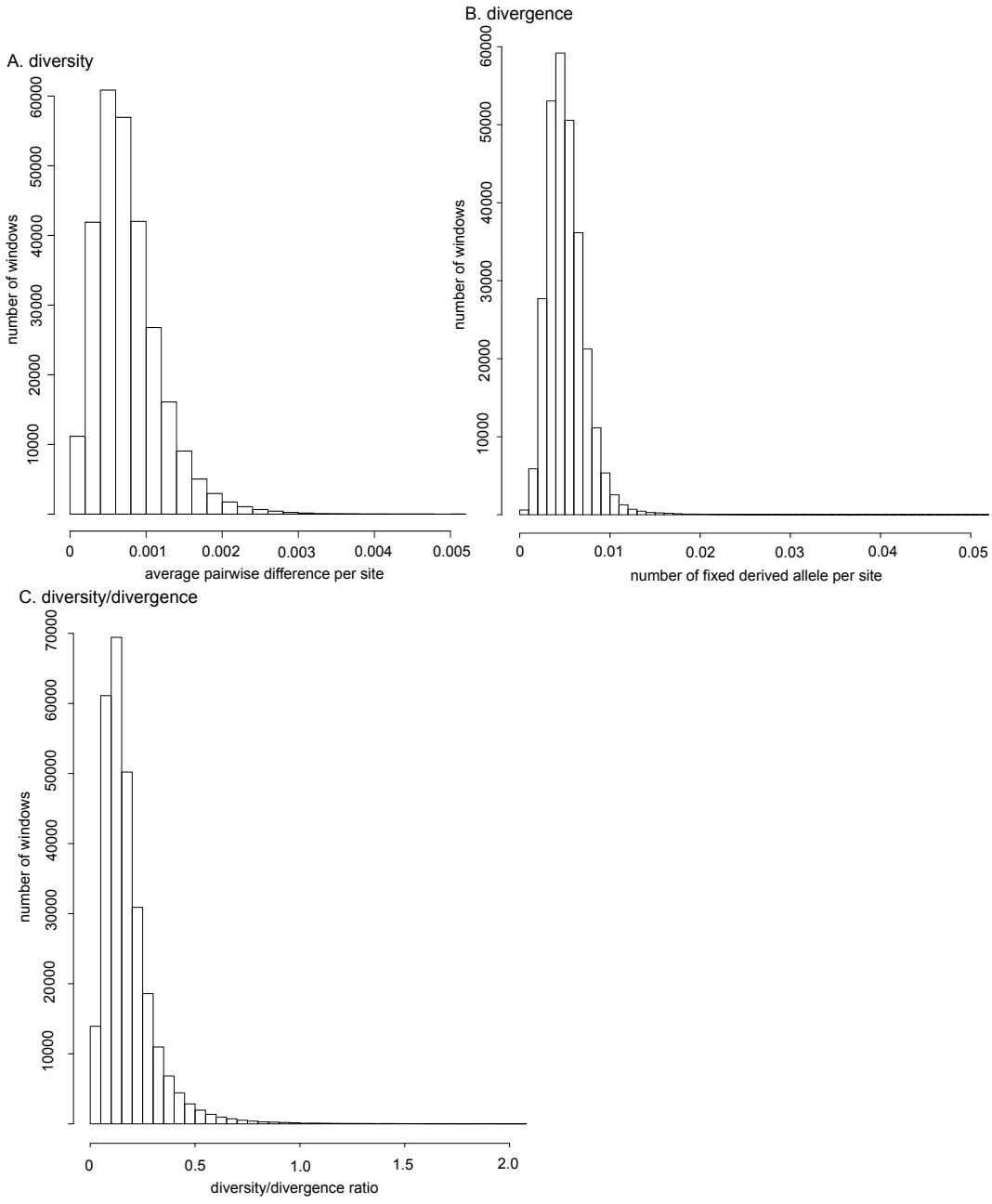


**Figure 4.1 Diversity and divergence distributions of the 5-kb windows in the CGI data.**

1% highest diversity/divergence ratio, as well as those with the 20% lowest diversity/divergence. As we would expect, the distribution of TMRCAs of the 1% lowest diversity/divergence windows is narrow and sharp, with a median of ~1.5, while that of the 1% highest diversity/divergence windows is much wider and flatter, with a median of ~6.3 (Figure 4.2). The TMRCA distribution of 20% lowest diversity/divergence windows, as we would expect, is slightly fatter and more towards the right, compared to the 1% lowest distribution (Figure 4.2). The median of these TMRCAs is ~1.6, slightly smaller than the expected genome

average of 2, which is as we would expect, since in general, lower diversity/divergence regions tend to have a smaller coalescence time.

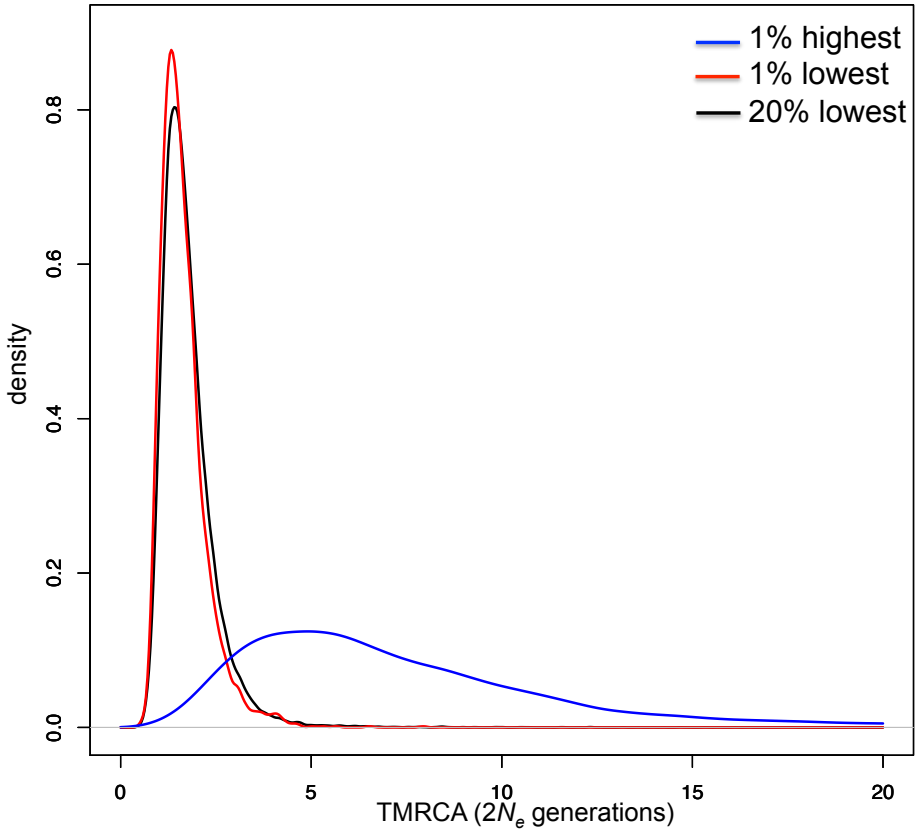


**Figure 4.2 TMRCA distributions in the CGI data.** This plot shows density distributions of TMRCAs in the 1% highest diversity/divergence windows (blue), 1% lowest diversity/divergence windows (red), and 20% lowest diversity/divergence windows (black).

### 4.3.3 Validation of TMRCA estimations by simulation

In order to further understand whether or not our TMRCAs reflect the reality, and whether they are unusual compared to neutral regions, we simulated 1,000 independent 100-kb neutral regions in 54 samples. We then compared the minor allele frequency spectra of the CGI and simulated data, and found very similar distributions (Figure 4.3). We next chunked the simulated regions into 20,000 windows of 5 kb, and calculated diversity. As we were unable to estimate divergence on simulated data due to the lack of information on fixed derived sites, we could only compare diversity of the simulated data with CGI data. We found that the distributions were very similar, except that the simulated neutral data lacked extremely low diversity windows (Figure 4.4), which is as expected. We then calculated TMRCA on the windows with 1% and 20% lowest, and 1%

highest diversity. Interestingly, the distributions of low diversity windows were very similar to the empirical data, but the high diversity windows had a much narrower range of TMRCAs, and there are no extremely high TMRCA windows in the simulated data (Figure 4.5). A Q-Q plot of the 20% lowest diversity simulated windows versus CGI windows shows quite a few outliers at the higher end, i.e. extremely large TMRCAs, that are only present in the CGI data. In contrast, there is only one outlier at the lower end; i.e. only one window's TMRCA is lower than expected from the neutral simulation (Figure 4.6). This indicates that there may not be enrichment for outliers with low TMRCAs in our genome; i.e. there are no more regions in the human genome with extremely recent TMRCAs than expected from a neutral model. However, we had windows with a TMRCA of less than $4N_e$ generations. These windows are worth further analysis to see if they are likely to have undergone selective sweeps. In contrast, we had some extreme outliers on the higher end of the TMRCA distribution. The majority of these windows, as expected, have high diversity/divergence ratio in humans. There are three plausible explanations for this. One is that many of these regions might have undergone balancing selection, where a high level of diversity or a combination of ancestral and derived alleles is beneficial to the individual or the population as a whole. Therefore, some very ancient alleles from our ancestors were maintained in current humans. The second explanation might be that there had been archaic admixture in the history of modern humans, which resulted in some gene flow between humans and their sister species, so that some of their alleles have been derived from other archaic humans. The third explanation is simply sequencing/mapping errors in the data. Although efforts have been made to produce a high-quality set of variant calls from the sequencing data, due to the complexity of the genome, some variants might have been called wrongly, especially those within highly repetitive regions, short insertion or deletions, or copy number variants. Furthermore, because of the diverse panel of samples, missing genotypes could not be inferred from the genotypes of other samples in the panel. These factors might have contributed to some artifactual high-diversity regions. In reality, these three reasons may all have played some role in causing the outliers with extremely ancient TMRCAs. More detailed examination

of these “ancient regions” is needed to figure out whether these regions are truly ancient in humans, but is beyond the scope of this thesis.

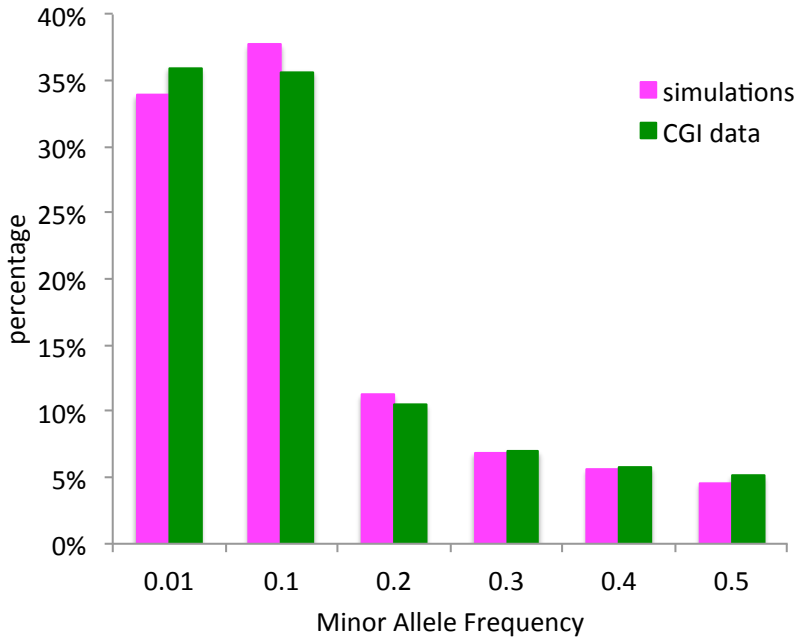


**Figure 4.3 Minor allele frequency spectra in the CGI and simulated data.**

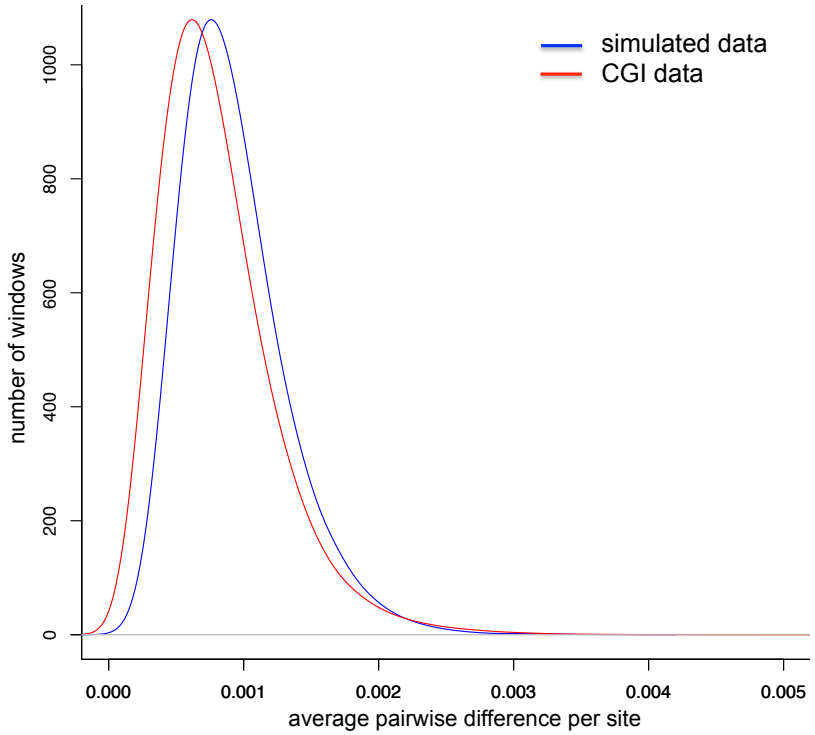


**Figure 4.4 Diversity distribution of simulated and CGI data.** This plot shows the density distributions of diversity per nucleotide in the CGI data (red) and the simulated data (blue).

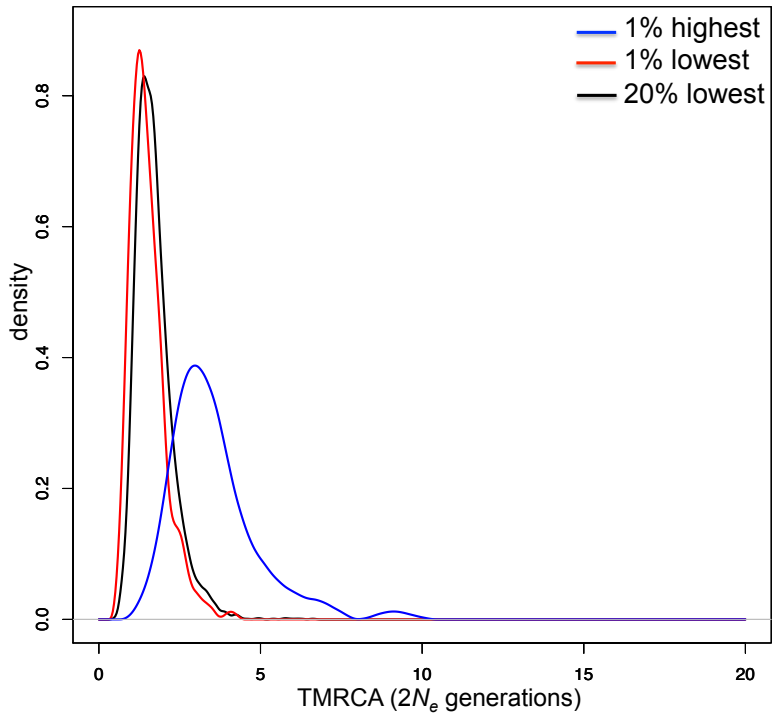


**Figure 4.5 TMRCA distributions on simulated windows with different diversity.** This plot shows density distributions of TMRCAs in the 1% highest diversity/divergence windows (blue), 1% lowest diversity/divergence windows (red), and 20% lowest diversity windows (black) in the simulated data.

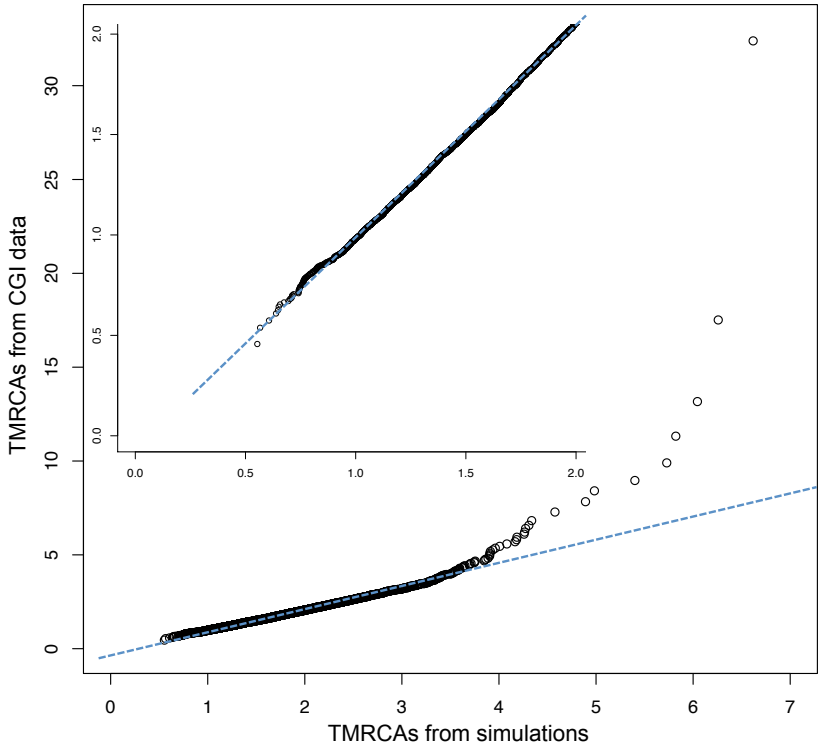


**Figure 4.6 Q-Q plot of TMRCAs in simulated data versus the CGI data.** This Q-Q plot shows the TMRCAs of simulated data (X axis) versus CGI data (Y axis), blue dashed line is the trend line. The smaller plot on the upper left corner is the magnified Q-Q plot for the part where TMRCAs are less than 2.

### 4.3.4 Comparison of variants in low-TMRCA regions with southern African and Denisovan genomes

Although the distribution of TMRCAs in our CGI data matches the neutral simulations, there are still windows with TMRCAs more recent than expectation. We identified 3259 windows with a TMRCA of less than 1 ($2N_e$ generations). If we assume a genome-wide average recombination rate of $1\times10^{-8}$ per nucleotide per generation, and 20 years per generation, for modern humans with a history of 200,000 years, the average length of a non-recombining segment in humans should be around 10 kb. Of course as the recombination rates across the genome vary a lot, the non-recombining segment lengths also vary dramatically. Nevertheless, as a rough guide, if a region has been positively selected at the same time when modern humans emerged, we would expect the recently coalesced region should not be shorter than 10 kb. Therefore, we combined adjacent windows with TMRCAs less than 1, and discarded single windows. This resulted in 143 regions sized from 10 kb to 25 kb.

We then investigated whether or not these regions are likely to have undergone selective sweeps during the emergence of modern humans. If they have, they should possess two features: all humans should share one or more derived alleles in these regions, and most of these fixed derived alleles should be human-specific. In order to test these features in the 143 regions, we first used the Bantu and Khoisan genomes to further filter for and confirm human-fixed derived alleles. The Khoisan belong to the indigenous hunter-gatherer peoples in southern Africa, and are believed to be descendants of the oldest known split among modern human populations. If the fixed derived alleles in our 54 CGI samples are also homozygous in these two genomes, we can be more confident that they are very likely to be shared by all humans.

In order to investigate whether or not the fixed derived alleles in these regions are human-specific, we should compare them with a sister species of modern humans that diverged from humans after the human-chimpanzee split but before the divergence of present-day populations. There are draft genomes of two non-human archaic hominins that can serve as the sister species to modern humans:

the Neanderthal genome sequence[17] and the Denisovan genome[12]. However, due to the low coverage (< 2x) of these sequences, we could only call a variant in the Neanderthal or Denisovan sequences if this variant is observed in humans, which therefore would not be suitable for our purpose, as we are hoping to identify shared and non-shared variants between humans and the archaic hominins in those regions. Fortunately, the authors of the first Denisovan genome[12] released an additional high-coverage (average coverage ~30x) Denisovan genome sequence data set recently, which allowed us to perform the comparison with a good level of confidence. We counted four types derived alleles, as described in section 4.2.5: (1) Denisovan-specific derived allele; (2) human-specific derived allele; (3) high DAF shared derived allele; and (4) low DAF shared derived allele. In theory, if the regions have undergone strong selective sweeps during the early times of the human lineage, we should expect high numbers of type (1) and (2) alleles, but no or very low numbers in type (3) and (4). However, there are some limitations of these counts. First of all, there might be some ascertainment bias in the data. For example, the Denisovan variants were called using the human genome as the reference, which might have introduced some bias towards shared alleles. Secondly, we only have one Denisovan genome, so even if we do not see a particular derived allele in this Denisovan genome, it does not mean that it is not present in the Denisovan population. Thirdly, although we had a diverse panel of human samples plus two other divergent human genomes, we still could not guarantee that the fixed derived alleles seen in these samples were truly fixed in all humans. Therefore, the absolute counts of these alleles might not be ideal to serve the purpose of testing the two features mentioned above. However, we could safely form a hypothesis that if these regions have undergone selective sweeps, type (1) and (2) alleles should be enriched in these regions while type (3) and (4) should be depleted. In order to test this hypothesis, we also generated 100 sets of random windows matching the number of windows in those 143 recent coalescent regions, and compared the number of each of the above four types of variants within the random regions and the 143 recent coalescent regions. We found no enrichment or depletion in any of the above categories in those 143 regions compared to random matched regions (Figure 4.7). This indicates that the regions with a TMRCA of less than $2N_e$ generations

were not as a whole likely to have undergone strong classic selective sweeps when modern humans emerged.

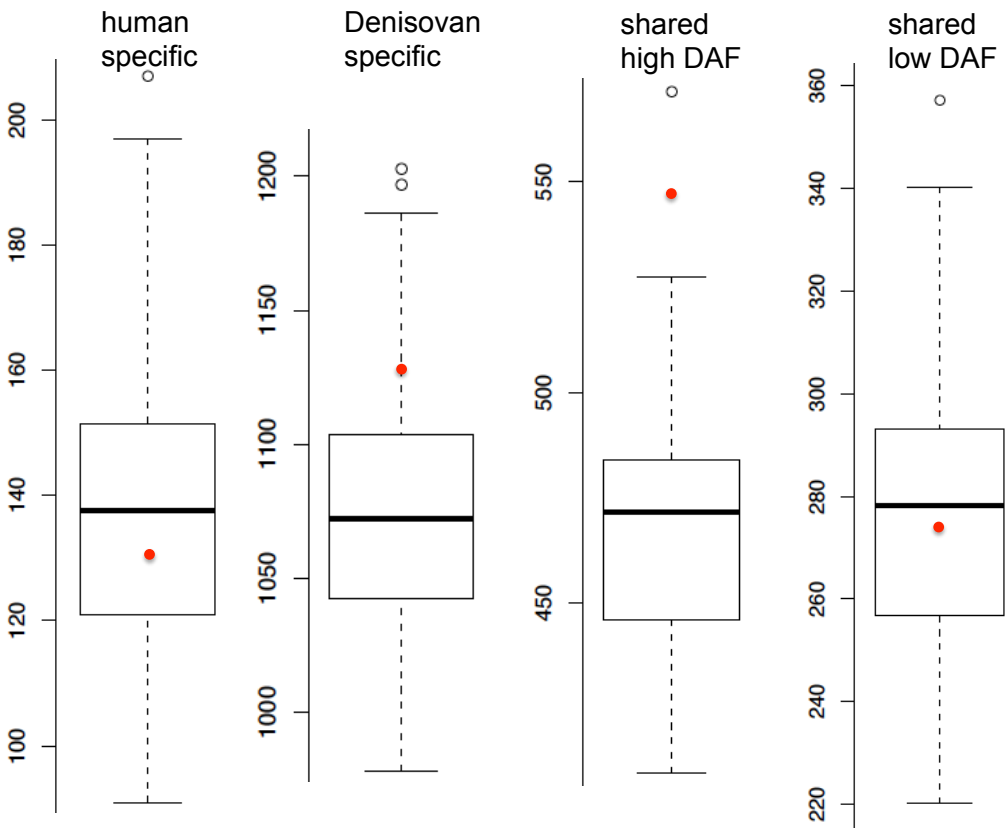


**Figure 4.7 Comparison of numbers of four types of derived alleles in humans and the Denisovan genome.** These box plots show distributions of numbers of each of the four types of derived alleles in 100 random sets of windows matching the recent-TMRCA windows. Red dots are the corresponding values of the recent-TMRCA windows.

### 4.3.5 Phylogenetic network analysis on regions with recent TMRCAs

To further understand whether or not the regions with recent TMRCAs are likely to have undergone an expansion in early modern humans, we performed phylogenetic network analysis on some of these recently coalesced regions in 54 CGI humans, two southern Africans and a Denisovan. If a particular haplotype in a genomic region had expanded to all modern humans but not in our sister species, we should expect to see that the branches of humans and the Denisovan in the gene network are well-separated. For the purpose of comparison, we looked at the phylogenetic network on the five 5-kb windows with a TMRCA of less than $N_e$ generations, a few regions from the 143 regions with a TMRCA of less than $2N_e$ generations, and a few regions with a TMRCA between $2N_e$ and $4N_e$

generations. We found that there was no population cluster in the phylogenetic networks in any of these regions, and the haplotype with highest frequency was present in all populations (Figure 4.8, Appendix H). This indicates that these regions all derived from one haplotype before the populations split. However, not all the human regions are well distinguished from the Denisovan. In Figure 4.8 A, the region has a clear pattern of recent expansion from the haplotype represented by the largest circle, and this is likely to have happened before the out-of-Africa migrations, as this ancestral haplotype is present in all populations, with the highest frequency in Africans. However, the Denisovan haplotypes did not appear much further away from the human haplotypes, and in fact, some human haplotypes have the same or a longer distance from the high-frequency haplotypes than the Denisovan. There are two possible explanations for this. One is that the haplotype that expanded in humans might already have existed before the human-Denisovan split, and the other is that there had been gene flow between humans and Denisovans, so that this haplotype in Denisovans was derived from humans. To test these hypotheses, more knowledge about the population history of Denisovans and their relationship with modern humans is needed.

Some regions with recent TMRCAs do show patterns where human and the Denisovan haplotypes are well distinguished. In Figure 4.8 B and C, the Denisovan haplotypes were much further away from the highest-frequency human haplotypes than any other humans, indicating that these human haplotypes were differentiated from their sister species. In fact, Regions with this type of pattern tend to have more Denisovan-private variants than European-Asian-private variants. The reason is obvious: if the region in humans differentiated from Denisovans before the human population split, Denisovans would have more time for accumulating new mutations than the European and Asian populations. In order to identify these regions, we compared the number of Denisovan-private variants (variants only present in the Denisovan genome) and European-Asian private variants (variants only present in the European and Asian samples) in each of the 143 regions with a TMRCA of less than $2N_e$ generations and the 5 windows with a TMRCA of less than $N_e$ generations. We

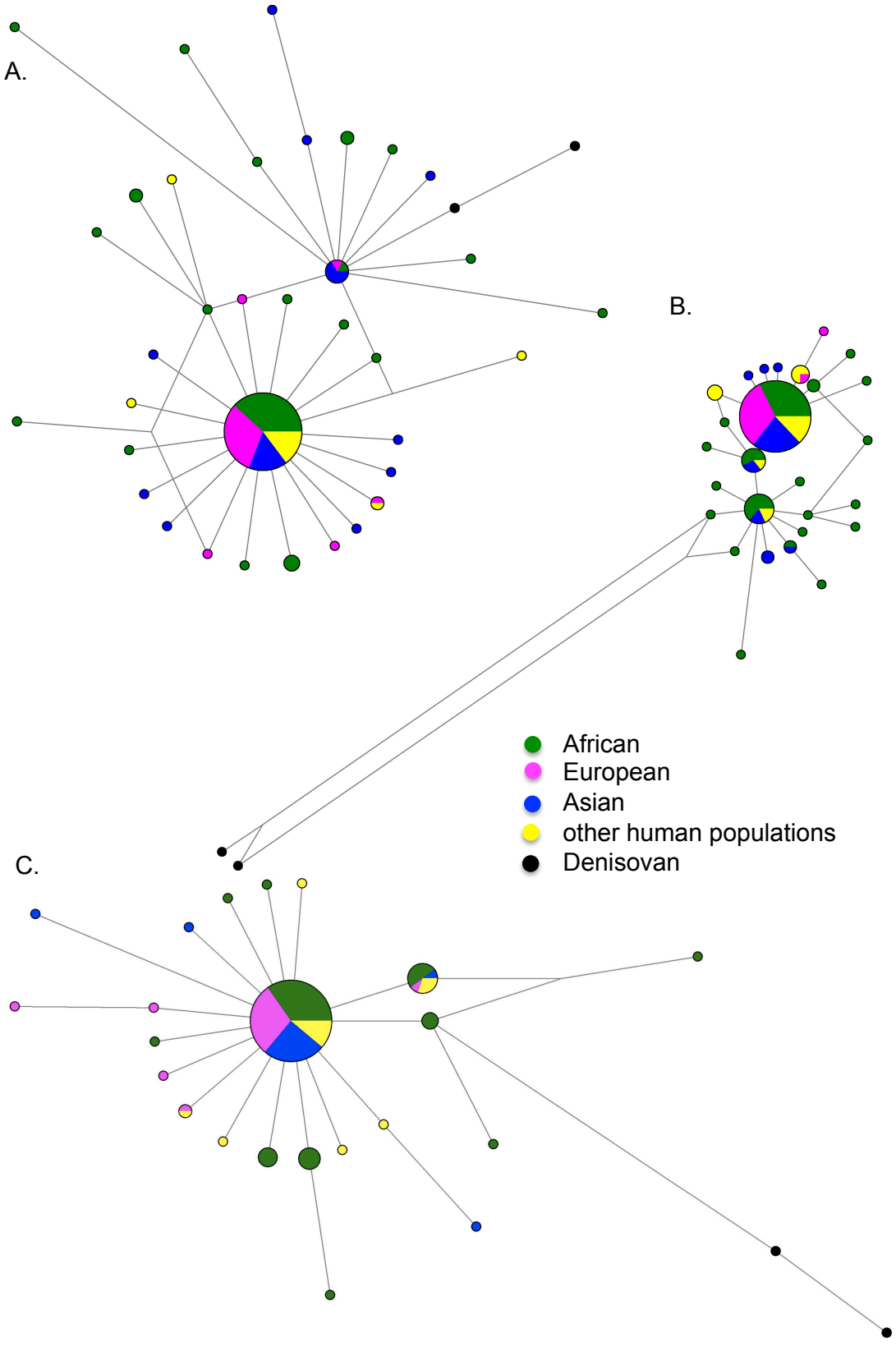


**Figure 4.8 Phylogenetic networks of three regions with recent TMRCAs.** A: region chr1:32,660,001-32,665,000; TMRCA 0.952 $N_e$ generations. B: region chr19:16,465,001-16,470,000; TMRCA 0.992 $N_e$ generations. C: region chr11:46,430,001:46,440,000; TMRCA 3.692 $N_e$ generations.

identified 22 regions with a larger number of Denisovan-private variants than European-Asian-private variants (Table 4.2). We believe that these regions are worth further investigations on whether or not they have undergone selective sweeps in the early stages of modern humans. It is worth noting that we were very conservative in comparing the private alleles in Denisovans and the European-Asian populations, since we only had one Denisovan sample but 32 European-Asian samples. With more Denisovan samples, we should expect more Denisovan-private alleles, which means that these regions may be even more differentiated from humans than we have seen here.

**Table 4.2 Regions with recent TMRCAs and more Denisovan-private alleles than Eurasian-private alleles.** Chromosome coordinates are in GRCh37. The table also shows the number of private variants in each population group, the TMRCAs of the regions in the unit of $N_e$ generations and genes in those regions.

| Chr | Start | End | Length (kb) | Denisovan private | African private | Eurasian private | TMRCA ($N_e$ generations) | Gene(s) |
|---|---|---|---|---|---|---|---|---|
| 1 | 28,465,001 | 28,480,000 | 15 | 23 | 47 | 18 | 0.996 | *PTAFR* |
| 1 | 70,195,001 | 70,210,000 | 15 | 21 | 52 | 17 | 0.978 | *LRRC7* |
| 2 | 197,675,001 | 197,690,000 | 15 | 20 | 42 | 8 | 0.961 | |
| 2 | 200,335,001 | 200,345,000 | 10 | 16 | 33 | 12 | 0.987 | *SATB2* |
| 3 | 110,875,001 | 110,885,000 | 10 | 12 | 29 | 11 | 0.967 | *PVRL3* |
| 4 | 84,130,001 | 84,140,000 | 10 | 5 | 29 | 4 | 0.949 | |
| 6 | 156,030,001 | 156,040,000 | 10 | 14 | 35 | 9 | 0.983 | |
| 6 | 157,065,001 | 157,075,000 | 10 | 10 | 31 | 9 | 0.992 | |
| 8 | 10,970,001 | 10,985,000 | 15 | 11 | 67 | 8 | 0.988 | *XKR6* |
| 8 | 43,360,001 | 43,375,000 | 15 | 29 | 67 | 21 | 0.873 | |
| 8 | 74,125,001 | 74,135,000 | 10 | 21 | 31 | 11 | 1.185 | |
| 8 | 82,120,001 | 82,130,000 | 10 | 13 | 27 | 3 | 0.871 | |
| 9 | 133,720,001 | 133,735,000 | 15 | 30 | 51 | 20 | 0.957 | *ABL1* |
| 10 | 400,001 | 425,000 | 25 | 82 | 111 | 44 | 0.921 | *DIP2C* |
| 10 | 22,105,001 | 22,120,000 | 15 | 13 | 43 | 12 | 0.963 | *DNAJC1* |
| 11 | 61,025,001 | 61,035,000 | 10 | 11 | 43 | 6 | 0.879 | *VWCE* |
| 12 | 80,370,001 | 80,385,000 | 15 | 13 | 45 | 12 | 0.926 | |
| 15 | 25,225,001 | 25,235,000 | 10 | 12 | 40 | 10 | 0.911 | *SNRPN, SNURF* |
| 17 | 74,765,001 | 74,775,000 | 10 | 8 | 41 | 6 | 0.952 | *MFSD11* |
| 19 | 15,380,001 | 15,390,000 | 10 | 10 | 27 | 9 | 0.987 | *BRD4* |
| 19 | 16,465,001 | 16,470,000 | 5 | 11 | 13 | 7 | 0.992 | *EPS15L1* |
| 20 | 54,990,001 | 55,005,000 | 15 | 41 | 50 | 13 | 0.988 | *CASS4* |

## 4.4 Discussion

This study has for the first time used whole-genome sequencing data from a diverse panel of human samples to systematically estimate coalescence times across the genome in humans, aiming to identify regions that share a very recent common ancestor among all humans, which may indicate positive selection

during the early stage of modern human history ~200 KYA. This approach is complementary to the statistical tests used in Chapters 2 and 3, as well as to other LD-based tests, and differs from them in two aspects: one is that it detects selective sweeps that were much older than those statistical tests, and the other is that it only detects complete selective sweeps, where the statistical tests have very limited power.

We first set out to answer the question of whether there are regions in the human genome that coalesce within the anatomically modern human lineage. Assuming that (1) modern human emerged around 200 KYA, (2) the human effective population size is 10,000 and (3) there are 20 years per generation, these regions should have a TMRCA around $N_e$ generations. However, as these assumptions have very limited accuracy, this threshold can only serve as a general guideline, and a range of TMRCAs around this value should be considered. In fact, a recent study suggested that generation times are about 29 years in humans and 25 years in chimpanzees, and also estimated the population-split time between Neanderthals and modern humans as 400-800 KYA[171]. If these estimations are reasonable, and if we look for regions that coalesce after human-Neanderthals split, then we should look for a TMRCA between around 0.5 and 1.5 $N_e$ generations. Among our calculated TMRCAs, very few windows had a TMRCA less than $N_e$ generations (5 out of 55,467 windows). Our simulations suggested that this number does not differ from expectations based on neutral assumptions. Comparisons of derived alleles with the Denisovan genome and the phylogenetic analysis also suggested that those regions with recent TMRCAs were not all completely human specific.

Based on these results, it seemed that we could draw the preliminary conclusion that there is no excess of "human-exclusive" regions spreading to all humans during the early stage of modern human history. However, there are several limitations to this study, which may prevent us from drawing such a conclusion. Firstly, the model used by GENETREE might be too simplistic, so the estimation of TMRCAs might not be accurate. GENETREE assumes a Wright-Fisher population, with no recombination, and an infinitely-many-sites model. Although these may provide a good approximation in most cases, in order to make an

accurate estimation of coalescence times, we may need a more realistic model. Secondly, we do not have proper independent sister species to use as outgroups of modern humans in this study. Ideally, we hope to have genomic information from some hominin species that diverged from humans not too long before the modern human emergence, and did not experience much gene flow with modern humans. Although the high-coverage Denisovan genome provided the closest approach to these requirements, it has limitations. For example, there might have been substantial gene flow between Denisovans and humans[12,18], and we only had one Denisovan genome sequence to use. Thirdly, the ancestral alleles were inferred from the primates that split from ancestors of humans several million years ago. This timescale might be too long for our purpose in this study, because multiple mutations will have occurred at some sites. It may be better if the ancestral alleles were inferred from species that are closer to humans.

Nevertheless, despite the limitations mentioned above, the results from this study serve as a first step for the genetic understanding of early modern human evolution. We have seen that strong classic selective sweeps might not have played a major role in the emergence of modern humans. It is more likely that the traits made us modern humans were the results of accumulation of mutations throughout a long period of time, and those genetic changes might have been present in our ancestors for a long time before modern human emerged. However, this does not mean that positive selection did not play a role in shaping early modern humans. Instead, this may indicate in most cases selection might have happened on existing alleles, or in a moderate manner rather than strong selective sweeps. In fact, by drawing gene networks of some regions with TMRCAs of less than $4N_e$ generations in humans and the Denisovan, we found some that showed patterns of rapid expansion of one haplotype specifically in humans. An example is the gene *AMBRA1*. A 10-kb region within this gene had a TMRCA of less than 4 $N_e$ generations, and gene network analysis of this region shows a clear pattern where all human halplotypes in this region were derived from one central haplotype from the African population, which was different from the Denisovan haplotype (Figure 4.8 C). Studies have shown that this gene involved in autophagy and may regulate the development of the

nervous system[172]. Some other genes that overlap with the regions with recent TMRCAs listed in Table 4.2 also seem to play important roles in humans. For example, the gene *PTAFR* is a receptor for platelet activating factor, a chemotactic phospholipid mediator that possesses potent inflammatory, smooth-muscle contractile and hypotensive activity (http://www.uniprot.org/uniprot/P25105 - section_comments); another gene *SATB2* may play an important role in palate formation and act as a molecular node in a transcriptional network regulating skeletal development and osteoblast differentiation (http://www.uniprot.org/uniprot/Q9UPW6 - section_comments). These examples indicate that classic positive selection might have shaped some genes or regions that contributed to modern human traits, but the number of such regions is not large. Also, due to our limited knowledge about gene functions in humans, it is often difficult to judge whether a gene is likely to have contributed to the modern human uniqueness. Of course, more studies are needed to understand these processes and to answer the question of what are the critical genetic changes that made us modern humans. Apart from estimating coalescence times of human genomic regions using more realistic models, we could systematically build phylogenetic trees of regions in humans and our sister species, in order to identify regions that are well-separated between the species. Therefore, the availability of additional high-quality genetic information from those hominin groups will be a key factor for the success of this type of study.

# 5 Discussion

## 5.1 The detection of positive selection: from genotyping to sequencing

Detecting signatures of positive selection by applying statistical approaches to genetic data has been a prolonged endeavor among evolutionary geneticists. The advancement of sequencing technology in recent years has raised the possibilities of larger-scale, higher-power and better-resolution detection of positive selection in the human genome, compared to genotype data that were previously used in such studies. Here, we discuss the benefits of sequencing data in the detection of positive selection in the human genome, as well as the challenges we are still facing.

Genome sequencing aims to detect all variation in the genome, with no bias towards certain types or frequencies of variants. Although in reality this is not achieved, sequencing does reveal many more variants with very low frequencies, which are otherwise undetectable with genotyping techniques because they are not included on standard chips. This provides higher power in detecting selective sweeps that are nearly complete or have just completed, as an excess of extremely low frequency alleles characterizes those sweeps, and genotyping may miss those variants. For example, in our genome-wide scan of positive selection using the 1000 Genomes low-coverage Pilot data, we detected an extremely strong signal in the *ITSN2* gene (Figure 3.11A). This signal ranks in the top 10 strongest signals in all three populations, yet was not discovered in any of the previous genome-wide scans using genotype data. In order to understand the data contributing to the difference between the strongest signal from the 1000 Genomes sequencing data and the HapMap genotype data, we looked at the minor allele frequency (MAF) spectra of the ~30kb peak region (chr2: 24,430,000-24,460,000) in the 1000 Genomes low-coverage Pilot and the HapMap Phase II data in the CEU population. Strikingly, although the number of samples in the HapMap data is higher (90 individuals), there is only one SNP

with a frequency of less than 1% (Figure 5.1), while in 1000 Genomes data, there are 19 such extremely-low frequency SNPs. Even if overall SNP densities are considered, lower-than-one-percent-frequency SNPs only account for 3% of the total number of SNPs in the HapMap data, while they account for 16% of the total SNPs in the 1000 Genomes data. These SNPs make great contributions to the selection signals, as an excess of extremely low frequency alleles is one of the most important features of a nearly completed or completed selective sweep that most frequency-spectrum based tests are able to detect.

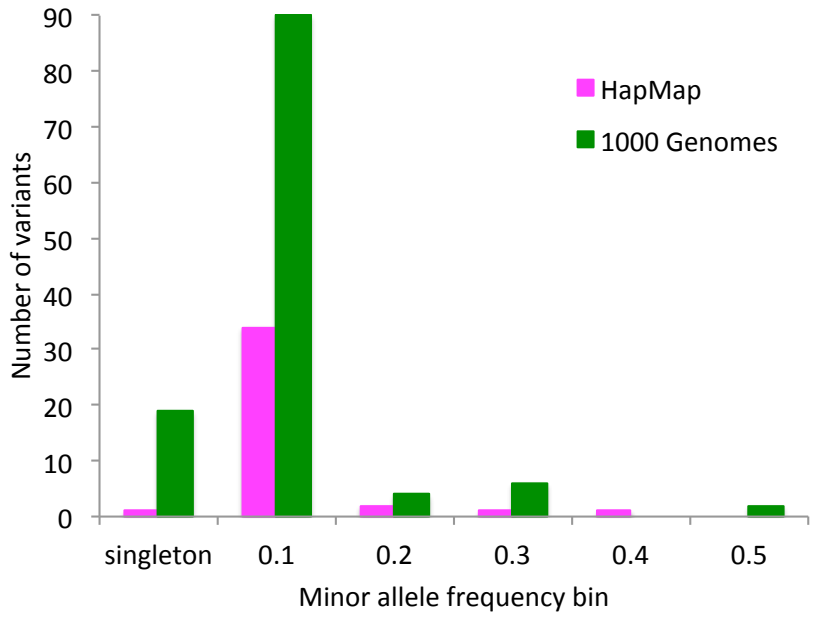


**Figure 5.1 MAF comparison of 1000 Genomes low-coverage Pilot and HapMap data in ITSN2 peak windows.**

In addition to revealing low frequency alleles, sequencing also discovers novel variants, which genotyping does not. All genotyping techniques are based on pre-designed assays, which means that the set of variants being genotyped are pre-determined. This not only makes it impossible to detect new variants, but may also inadvertently eliminate population-specific variants, which are very important features in some population-specific selective sweeps. This is especially true when the genotyping chips used are designed or based on one population (in practice mostly European) that is very different from the one being investigated.

Although sequencing provides better data for the detection of positive selection in the human genome, there are still challenges needing to be addressed in order to achieve a comprehensive understanding of which regions in the human genome have truly undergone classic selective sweeps during modern human evolution. The first challenge is the lack of realistic demographic models for many populations. Demographic factors, for example population expansion, bottlenecks and admixture, can have great impact on the patterns of signals of selection in the genome, which may result in false negative or positive detection. In order to allow for these effects, demographic histories of the populations under investigation need to be modeled in simulations, so that the p values obtained reflect real departures from neutrality under that demographic model and thus plausible signals of selection. Great efforts have been made in developing demographic models for some of the main continental populations, such as African, European and Asian as represented in HapMap, yet for specific sub-populations or populations with admixture, due to the lack of sufficient data and the complexity of population structure, very few satisfactory models have been developed to mimic their population histories. Therefore, the elimination of demographic effects without a model of population history remains a challenging task.

The second challenge faced by us is the determination of p values and their significance thresholds. In genome-wide scans of positive selection, p values are usually generated by either of two means: one is from the distribution of test values of simulated neutral data, and the other is from the distribution of test values in the empirical data. Among previous genome-wide scans of positive selection, both simulation-based p values and empirical p values were used. For example, Sabeti et al. used 10 Gb simulated data to determine the significance cutoff of their test values[66], whereas Voight et al. treated the 1% most extreme values as significant in their iHS test[67]. There are pros and cons in both approaches. Using simulated data is powerful and unbiased if realistic demographic models are used in the simulations, and adequate quality control techniques are used to make sure that the simulated data mimic the empirical data in a neutral scenario. Because the simulated data are independent of the

empirical data under investigation, they can provide an objective view of what proportion of the genome has been under positive selection, and this may differ between populations. However, as mentioned earlier, if the demographic models used do not reflect the real population history, simulations can be unrealistic, and by using them to generate p values, a large number of false positive results can be generated, which may be in fact caused by demographic effects, and some real positive selection signals may also be disguised. Using empirical distributions, in contrast, is not confounded by demographic effects, as those would have impact on the whole genome, or at least a large proportion of it, instead of specific regions. "Outliers" whose test values are higher than the vast majority of regions in the genome can be identified, so any baseline effect on the whole genome is eliminated. However, empirical p values cannot answer the very important question of what proportion of the genome has been positively selected, as the threshold of outliers is set artificially. Therefore, in reality, either of these two means can be used according to the particular study, and sometimes both simulation-based p values and empirical p values are used to complement each other. After p values are calculated, which value should be the threshold of significance is the next question that is critical in the detection process. Traditionally, the simplest way has been to set an artificial baseline threshold for the p value, which is usually 0.05 or 0.01, and then correct it from multiple comparisons by using the Bonferroni correction, the Benjamini-Hochberg procedure[173], or other similar approaches. In a global-scale study like a genome-wide scan of positive selection, a Bonferroni correction tends to be very conservative, which results in greatly reduced power of detection. However, using a looser threshold has a danger of high FDR in a large-scale study where the number of times the tests are applied is very high, since even a very small false positive rate can result in hundreds or thousands of false positive detections. Therefore, it is crucial to calculate the FDR under different thresholds, and use one that gives a satisfactory FDR for the particular study. That being said, there is always a tradeoff between specificity and sensitivity of any statistical evaluation based study, and one needs to decide which one should have more weight based on the purpose of the study. Also, p values are always relative, so there is no black-or-white cutoff. One needs to consider multiple factors and

other sources of evidence to judge whether a region or gene shows real signals of positive selection.

The third challenge is the limited range of selective sweeps that are detectable by current statistical approaches. As demonstrated by our simulations, all frequency-spectrum-based tests have high power only for classic selective sweeps that are relatively strong and have reached a late stage (selected allele frequency more than ~70%) or have just completed. For weak sweeps (for example, selection coefficient is 0.001), or early stage sweeps, or sweeps that have completed a long time ago, the power of detection is very low. This is also true for many LD-based tests, for example, EHH[65], as the principles behind these statistical tests are similar: they look for patterns of the genetic variation that reflect the footprint of a selective sweep, which only exists under certain conditions. When selective sweeps are weak, it takes a long time for the selected allele to reach a high frequency, so it is more likely that new mutations, recombination or gene conversion will break down the patterns of the selective sweep in the genomic sequence. Therefore, it is very difficult to distinguish them from neutral regions, the variation patterns of which are determined by genetic drift and demographic effects. Likewise, if the sweep is at its early stage, the patterns are likely to be undistinguishable from the neutral scenario; and if the sweep had completed hundreds of generations ago, new mutations and recombination may have erased the patterns. Some other tests, for example, XP-EHH[66] and iHS[67], utilize the population differentiation of allele frequencies or long haplotypes, to detect selective sweeps that are population specific. These tests are more robust in detecting selective sweeps with different stages and strengths, but are not able to detect sweeps that are not population specific. Therefore, by using the current statistical tests, we are likely to miss selective sweeps that are out of the detectable range, and the development of methodologies to detect those sweeps remains a challenge.

Finally, using low-coverage sequencing data in the detection of positive selection can also be a challenge. Large sequencing projects, for example, the 1000 Genomes Project, sequence a large number of individuals in many populations around the globe. Due to their primary aim of discovering SNPs in the most cost-

effective way, these whole genome sequences mostly have low coverage, for example, 2-4x. This is often insufficient to call a variant at a certain locus in a single individual, especially if it is heterozygous. The common way to deal with this issue is to split variant discovery and genotyping into separate steps. First, evidence for a non-reference variant in the pooled data from all individuals is sought. Then the most likely genotype at each variable locus is inferred by referring to other samples in the same population and the LD of nearby sites. Although this approach has been proven to be able to impute fairly accurate genotype calls effectively, the error rates are still high in heterozygous sites of low-frequency alleles. Therefore, although low-coverage sequencing data are a good starting point for genome-wide investigations of positive selection, it may be helpful to subsequently re-sequence some candidate regions at much deeper coverage, in order to obtain the full set of variants in the region to eliminate any bias and to enhance the chance of identifying the selection targets.

## 5.2 The localization of selection targets

The detection of signatures of selective sweeps is just the first step in the exploration of positive selection in the human genome. After finding the signals, we need to identify which loci or alleles were favored by natural selection. This is, in most cases, not an easy task, as most candidate positively selected regions are tens or hundreds kb in length, and sometimes, especially when LD-based tests are applied, even several Mb long. Three types of approaches were commonly used to localize selection targets. One is by identifying the strongest signal from the statistical tests, a second is by looking for derived alleles with a high frequency in the selected population but not in the non-selected population, and a third is to look for derived variants that have clear functional impact. These approaches are usually complementary to each other, so are often used together whenever possible, to localize selection targets, and finding the strongest statistical signal is often the first step. Using frequency-spectrum based tests on sequencing data improves localization power in at least two aspects. One is the higher resolution of the signals. Due to the much higher density of variants in sequencing data compared to genotype data, statistical tests can be applied on a smaller genomic region, so that the density of the signals is higher. This

obviously helps to identify a peak signal covering a smaller region so that looking for the target variant is easier. The other advantage is the completeness of discovery of the variants in the region. As discussed before, sequencing data contain almost all variants in a given genomic region, while genotype data only contain a small proportion of the variants. Furthermore, most of the missing variants in genotype data have low minor allele frequencies (MAF), and in near-complete selective sweeps, the selected variant usually has a low MAF. Thus there is a high chance that the genotype data do not include the variant under positive selection, which makes it even more difficult to localize the target. Furthermore, by combining signals from multiple tests, real strong signals can be amplified, while moderate signals, or signals in only one test, which are more likely to be false, can be diluted or eliminated. In our studies, we combined three independent frequency-spectrum based neutrality tests, therefore increased the power of localizing the signals. Similarly, a previous study combined multiple LD-based tests and DAF differentiation scores to generate a compound score, called composite of multiple signals (CMS)[68]. Their results also showed significant enhancement in the power of localizing the selected variants in the candidate regions.

As demonstrated by our simulations, although peak signals are on average enriched at the locus containing the selected allele, there is still a high chance that they are located far away from the selection target. This is most likely due to recombination happening during the course of selective sweep, or other mutational effects that break down or blur the patterns. If recombination hotspots exist close to the selected locus, peak signals can be further away from the selection target and their strength can be reduced. Therefore, knowing the location and intensity of recombination hotspots is critical when trying to localize selection targets. Although the localization power is significantly reduced when recombination hotspots are close to the selection target, by having this information, one can extend the length of the candidate target region under investigation, so that the real selection target will not so readily be missed.

Another way to help localize selection target is to focus on loci with high derived allele frequencies (DAF), as the selected allele is most likely to be a derived allele

with a relatively high frequency in a detectable classic selective sweep. This has two potential challenges. One is that there are often quite a few such high-frequency derived alleles due to hitchhiking effects, and it is almost impossible to figure out which one is the selection target without other information. The other is that sometimes the information about ancestral status of some loci is unavailable, inaccurate or lost due to recurrent mutation, and in these cases, derived alleles cannot be identified reliably. When selection signals are only present in one population but not the other, we can also compare DAFs in these two populations, and those that have a high DAF in the selected population but not the other are likely to be at or near the selection target. This approach of course requires adequate sample sizes from both populations and that the variation calling methods do not skew the allele frequencies.

Functional information is potentially very helpful for narrowing down the potential candidates for selection targets. For example, if there is a high-frequency allele that changes an amino acid in a target region, it is quite likely that this allele has been positively selected, especially if we know that this amino acid change has a functional impact. However, this scenario is unfortunately very rare. In most cases, we have no or very little information about functions of variants, especially if the variants are not in or close to protein-coding regions. Although researchers have been gradually improving annotation of the human genome, for example with projects like ENCODE[56], we still only know very little about the functional elements, and this remains a challenge for us when trying to identify the selection targets. On the other hand, even in cases where there are genes or known functional elements at or near the selection signals, due to the fact that the candidate regions are often quite long, and the hitchhiking effect of selective sweeps, there may be many variants nearby the selection target that show very similar features as the selected variant, which makes it challenging to distinguish the real selected variant from the hitchhiked variants.

## 5.3 Biological interpretation of alleles under positive selection

The aim of identifying positively selected regions and localizing selection targets in the human genome is to understand which functional changes have undergone

positive selection. As mentioned earlier, this is often a two-way process. On one hand, known functional variants or fixed derived mutations from ancestors, especially those likely to affect the individual's fitness, are good candidates for further investigation of whether or not these changes have undergone positive selection. On the other hand, signals of positive selection in functionally unknown regions of the genome may indicate the functional importance of those regions, and are thus worth further investigation of their biological functions.

Some traits or biological functions are considered more likely to be positively selected than others. These traits are often involved in reproduction, metabolism, disease resistance, environment-related morphological features (e.g. skin color, hair thickness), and so on. These can be categorized into three types: (1) biological functions that are directly involved in reproduction, for example, sperm mobility[174]; (2) traits related to adaptations to the climate, natural environment and life style, for example, pigmentation of skin and hair[66,99]; and (3) resistance to debilitating or life-threatening diseases, for example, malaria[175,176]. Genes within each of the three categories have been identified as positively selected recently in modern human evolutionary history, yet there are more to be discovered. In most of these cases, positive selection signals were revealed after the functional impact of the variants within those genes were discovered, or the functions of the genes where the variants lie were known.

Although the identification of functional targets of selection is challenging, as discussed in Section 5.2, there are many bioinformatic and experimental approaches that can reduce the number of candidate variants or even discover the real target. The advancement of technologies and accumulation of new findings has been constantly contributing to these approaches in at least three ways. Firstly, more and higher quality data have made it possible to obtain a nearly complete set of variants in a large number of samples. This can be beneficial in two ways: one is to prevent biases of the variants data that may lead to false results, and the other is to provide more population-genetic information on the variants, which can help identify the variants that show unique patterns. Secondly, more advanced modeling techniques and statistical algorithms can help narrow down the number of candidate variants, which of course makes it

easier to further identify the real target. Thirdly, a constantly improved understanding of functions of the human genome is providing valuable information to assist the identification of possible selection targets. Furthermore, experimental functional studies on model organisms and humans are yielding fruitful results that significantly improve our understanding of functions of our genes. For example, the Knockout Mouse Project (KOMP), initiated by the National Institutes of Health (NIH) in the US, aims to generate a comprehensive and public resource comprised of mice containing a null mutation in every gene in the mouse genome[177]. Similarly, the Zebrafish Mutation Project (ZMP, http://www.sanger.ac.uk/Projects/D_rerio/zmp/) at the Wellcome Trust Sanger Institute aims to create a knockout allele in every protein-coding gene in the zebrafish genome. These resources will certainly add knowledge to the understanding of human gene functions, and lead to the systematic studies of human gene functions and phenotypes. Researchers have raised the concept of "Human Phenome Project"[178], proposing comprehensive databases of human phenotypic data. Many research groups around the world are carrying out GWAS on various human traits and diseases, which are constantly contributing to our understanding of the functions of human genomic variants. A combination of these bioinformatic tools, large-scale experimental projects and databases is leading to progress in understanding positive selection in modern humans.

## 5.4 Impact of the studies in this thesis

Next Generation Sequencing (NGS) technologies have provided geneticists with seemingly unlimited possibilities for exploring our genomes in a large-scale and comprehensive manner. Being in one of the greatest genomics institutions and one of the largest sequencing centers in the world has provided me with the access to cutting-edge technologies, high-quality large data sets, and high-impact research projects, for example, the 1000 Genomes Project. The three projects during my PhD study were all based on NGS data, and for the first time used these exciting data sets to explore positive selection in the human genome in a holistic and comprehensive manner. There are three major impacts that my PhD research has made to the field of human evolutionary genetics, which are discussed below.

The project discussed in Chapter 2 provided, for the first time, an understanding of how large-scale sequencing data may benefit the detection and localization of positive selection. All previous large-scale studies of positive selection were based on genotype data. As discussed earlier, due to the fact that genotyping techniques only detect a subset of "known" variants, genotype data may miss a large proportion of low-frequency variants, which will severely reduce the power to detect and localize selection signals. By resequencing at a very high coverage two regions that showed strong signals of positive selection from a genome-wide scan on genotype data, we demonstrated that using frequency-spectrum based tests on sequencing data can not only detect the signals, but also effectively increase the power to localize the signal, for example, by ten-fold in both regions we investigated. This study provided the first insight into how we can maximize the benefits of sequencing data in studies of positive selection in the genome.

In Chapter 3, several sets of simulations using various scenarios were presented, aiming to understand how recombination affects signals of positive selection, and the sensitivity and specificity of detecting selection signals, as well as localizing positive selection using sequencing data. This study demonstrated the effects of recombination hotspots on the localization of selection signals. It benefits the research community by showing the importance of considering the recombination rates of the region in question when trying to localize selection signals, and also by providing general guidelines on how well a selection target can be localized by the frequency-spectrum approach.

Our genome-wide scan using 1000 Genomes low coverage Pilot sequencing data provided a list of candidate regions in the human genome that may have undergone positive selection in the course of modern human evolution. This is the first map of positive selection in the human genome generated from whole-genome sequencing data. As Chapter 2 and the simulations in Chapter 3 demonstrated, this map has a higher resolution in terms of the positions of selection targets, and provides higher power in detecting selective sweeps that may not be detected from genotype data. This new generation map of positive selection in the human genome will benefit the research community in at least

two ways. On the one hand, it provides a valuable resource for further evolutionary studies of specific types of human genes, regulatory elements or functions that have been evolving under recent positive selection in the human lineage. On the other hand, it provides guidance on studies of human genomic functions and discoveries of new functional elements in the human genome. If a genomic region shows strong signals of positive selection, it is very likely that the region is or has been functionally important, even if we do not yet know what functional roles the region plays. Functional studies usually involve extensive wet lab experiments that are costly and time-consuming. Therefore, some prior knowledge about which regions in the genome are more likely to be functional is critical when choosing candidates for experimental functional studies. The list of candidate positively selected regions from our scan is a good list to choose from, for example, the regions ranked at the top of the extremely low p values (see Appendix D for candidate regions and p values) should be worth further functional investigation.

The coalescence project described in Chapter 4 is a pioneering investigation of whether we can identify genomic regions with recent coalescence times using sequencing data and find ones that are uniquely shared by all humans and not by Neanderthals and Denisovans. Such regions may have played critical roles in making humans as what we are today, and may have been favored by positive selection in the critical early stage of modern human evolution when modern behavior was evolving. However, these regions cannot be detected by standard neutrality tests, as the signatures of positive selection will have been erased by new mutations and recombination over time. Although our results showed that such recently-coalesced regions are not abundant in modern humans, we were able to identify regions with recent TMRCAs in humans that were differentiated from the Denisovan genome, which potentially may have played important roles in shaping modern humans. Although more in-depth investigations are needed to further understand the recently coalesced regions in the human genome, this is the first time that this type of techniques has been used on a genome-wide scale, and will certainly shed new lights on our understanding of early modern human evolution.

## 5.5 Future directions

Human evolutionary genetics has entered an exciting era with numerous opportunities to better understand how humans have been evolving during the last hundreds of thousands of years. Thanks to the advancement of new technologies, whole-genome or targeted sequencing data of individuals from many populations across the world have become available and more are coming out all the time. These data help researchers to better understand human population histories, recombination and mutation patterns and population differences. They also provide more power for researchers to investigate selection in the human genome, as discussed earlier. However, current methodologies for detecting positive selection do not take into account all these new factors. Therefore, more comprehensive algorithms or statistical approaches need to be developed, taking the new knowledge and sequencing data into account, in order to maximize the benefit of sequencing data, and achieve detection of positive selection with higher power and lower false positive or false negative rate.

While my PhD research focused on hard sweeps, which are the most straightforward form of positive selection to detect, the new data sets available should make it possible to detect more complex sweeps, for example, soft sweeps. Extensive simulations and modeling are necessary to figure out the most effective way to detect soft sweeps. In fact, these can be developed based on the simulations and knowledge gained from the studies of hard sweeps, and this will be an important area in the near future. In fact, researchers have been studying potential selection on standing variants related to some human polygenic traits that showed population differentiation. For example, a recent study showed evidence of positive selection on standing alleles associated with increased height in Northern Europeans compared to Southern Europeans[179], by systematically comparing allele frequencies of those variants in these two populations.

As discussed in earlier chapters, understanding the functions of positively selected regions is the most important and exciting, yet most challenging step in

studies of positive selection in humans. Most of the regions showing signals of positive selection have no obvious candidate functional elements, and it remains a big challenge for us to demystify their functions. Traditional experimental studies of human cells or model organisms to investigate functions of genes are probably the most reliable approaches. These studies, however, usually take months or years to investigate a single locus and are difficult to scale up. Therefore, they may not be the most efficient way for large-scale functional studies, especially for regulatory elements, functions of which may be indirect, subtle and not easy to observe. Array and RNA sequencing techniques have enabled large-scale studies of gene expression in different tissues or organs, which provides power to large-scale functional studies, although this may be just the first step of the investigation of gene functions by detecting eQTLs. Various computational approaches for functional studies have also been developed. These approaches usually use available experimental data sets to identify general features of certain functional elements, and then construct algorithms to identify novel ones from the genome. These approaches are of course less reliable than experimental studies, but they are better-suited to large-scale studies of regulatory elements, which otherwise are difficult to design experiments for. To maximize the effectiveness and efficiency of functional investigation of candidate selected genes, we need to combine computational and experimental approaches. Generally speaking, computational methods can serve as a preliminary filtering tool to help choose the right candidates or the right direction for the design of wet-lab experiments. In fact, this process can be very dynamic, as results from experimental studies can feed back to the computational part of investigation, which will again guide the next steps of experiments. Therefore, we need to link computational and experimental studies more closely in order to maximize the effectiveness and efficiency of functional studies.

Many current or previous studies of candidate genes focused on a single gene or a few functionally related genes. However, to understand a complete biological process, it helps to identify genes involved in a certain pathway or network. It is possible that a biological process, rather than a specific gene, has been positively

selected. In this situation, many of the genes involved in the pathway or network may have been selected, and each of them may only have played part of the role and the selection strength on each gene may be relatively weak. Therefore, it is worth grouping genes into their pathways and networks to understand the functional targets of positive selection. In fact, researchers have studied positive or other forms of natural selection in some gene networks in humans[180,181]. For example, by investigating genetic adaptations of the human antibacterial innate immunity network, Casals et al. found different patterns of selection on genes at different positions of the network, and that functional classes involved in autoinflammatory and autoimmune diseases are enriched with evidence of balancing selection[180]. As more and more studies have revealed pathways and networks of genes in lots of biological processes, and such databases have been built up and enriched[182], a next step is to utilize this knowledge to understand more about the functional targets of positive selection.

In the last four years, the field has moved from the first tentative attempts to sequence whole human genomes to established whole-genome sequencing platforms and global-scale sequencing projects. Sequencing data are no longer the limiting factor for studies of positive selection. In the next few years, more and better-quality whole-genome sequences from more populations and even more sister species of humans, along with more advanced computational models, will enable more exciting discoveries on positive selection in humans, and provide more insights into the understanding of modern human evolution.

# References


1 Kono, R. T. Molar enamel thickness and distribution patterns in extant great apes and humans: new insights based on a 3-dimensional whole crown perspective. *Anthropol Sci* **112**, 121-146 (2004).

2 Richmond, B. G. & Strait, D. S. Evidence that humans evolved from a knuckle-walking ancestor. *Nature* **404**, 382-385 (2000).

3 Thorpe, S. K. S., Holder, R. L. & Crompton, R. H. Origin of human bipedalism as an adaptation for locomotion on flexible branches. *Science* **16**, 1328-1331 (2007).

4 Yunis, J. J. & Sanchez, O. G-Banding and Chromosome Structure. *Chromosoma* **44**, 15-23 (1973).

5 Sarich, V. M. & Wilson, A. C. Immunological time-scale for Hominid evolution. *Science* **158**, 1200-1203 (1967).

6 Sibley, C. G. & Ahlquist, J. E. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *J Mol Evol.* **20**, 2-15 (1984).

7 Chen, F.-C. & Li, W.-H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68**, 444-456 (2001).

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

9 Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169-175 (2012).

10 Fischer, A. *et al.* Bonobos Fall within the Genomic Variation of Chimpanzees. *PLoS ONE* **6**, e21605 (2011).

11 Hublin, J.-J. The origin of Neandertals. *Proc Natl Acad Sci USA* **106**, 16022–16027 (2009).

12 Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010).

13 Caramelli, D. *et al.* A highly divergent mtDNA sequence in a Neandertal individual from Italy. *Curr Biol* **16**, R630–R632 (2006).

14 Green, R. E. *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416-426 (2008).

15 Briggs, A. W. *et al.* Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *Science* **325**, 318-321 (2009).

16 Krause, J. *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894-897 (2011).

17 Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710-722 (2010).

18 Reich, D. *et al.* Denisova admixture and the first modern human dispersals into southeast Asia and Oceania. *Am J Hum Genet* **89**, 516-528 (2011).

19 Eriksson, A. & Manica, A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc Natl Acad Sci USA* (2012).

20 Wood, B. & Harrison, T. The evolutionary context of the first hominins. *Nature* **470**, 347-352 (2011).
21 Leakey, M. D. & Hay, R. L. Pliocene footprints in the Laetolil Beds at Laetoli, northern Tanzania. *Nature* **278**, 317 - 323 (1979).
22 Walker, A. & Leakey, R. E. F. *The Nariokatome Homo erectus skeleton* (Harvard University Press, Cambridge, 1993).
23 Ammerman, A. J. & Cavalli-Sforza, L. L. *The Neolithic Transition and the Genetics of Populations in Europe*. (Princeton University Press, Princeton, NJ, USA, 1984).
24 Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708-713 (2000).
25 Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. & Feldman, M. W. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci USA* **97**, 7360-7365 (2000).
26 Schaffner, S. F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* **15**, 1576-1583 (2005).
27 Manica, A., Amos, W., Balloux, F. & Hanihara, T. The effect of ancient population bottlenecks on human phenotypic variation. *Nature* **448**, 346-348 (2007).
28 von Cramon-Taubadel, N. & Lycett, S. J. Brief communication: human cranial variation fits iterative founder effect model with African origin. *Am J Phys Anthropol* **136**, 108-113 (2008).
29 Liu, H., Prugnolle, F., Manica, A. & Balloux, F. A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet* **79**, 230-237 (2006).
30 Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104.
31 Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).
32 Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**, 712-714 (2011).
33 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).
34 Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007 (2002).
35 Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-732 (2005).
36 Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712 (2009).
37 Stephens, M. & Donnelly, P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* **73**, 1162-1169 (2003).
38 Rapley, R. & Harbron, S. *Molecular Analysis and Genome Discovery*. (John Wiley & Sons Ltd, 2004).
39 The International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).

40 The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
41 Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).
42 Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci USA* **108**, E1128-E1136 (2011).
43 Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65 (2011).
44 Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucl Acids Res* **36**, e105 (2008).
45 Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**, R32 (2009).
46 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858 (2008).
47 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
48 Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-1729 (2001).
49 Warren, R. L. & Holt, R. A. Targeted assembly of short sequence reads. *PLoS One* **6**, e19816 (2011).
50 Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* **11**, 473-483 (2010).
51 Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trend Genet* **16**, 369-372 (2000).
52 Frazer, K. A. *et al.* Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res* **14**, 367-372 (2004).
53 King, D. *et al.* Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* **15**, 1051-1060 (2005).
54 Wang, Q. F. *et al.* Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. *Genome Biol* **8**, R1 (2007).
55 Prabhakar, S. *et al.* Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* **16**, 855-863 (2006).
56 The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640 (2004).
57 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
58 Wright, S. Classification of the factors of evolution. *Cold Spring Harb Symp Quant Biol* **20**, 16-24 (1955).
59 Darwin, C. *On The Origin of Species*. (Oxford University Press, 1859).

60 Michon, P. *et al.* Duffy-null promoter heterozygosity reduces DARC expression and abrogates adhesion of the P. vivax ligand required for blood-stage infection. *FEBS Lett* **495**, 111-114 (2001).
61 Hamblin, M. T. & Di Rienzo, A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* **66**, 1669-1679 (2000).
62 Hamblin, M. T., Thompson, E. E. & Di Rienzo, A. Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet* **70**, 369-383 (2002).
63 Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-1157 (2005).
64 Lewontin, R. C. The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics* **49**, 49-67 (1964).
65 Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837 (2002).
66 Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-918 (2007).
67 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72 (2006).
68 Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883-886 (2010).
69 Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072-1079 (2005).
70 Wang, E. T., Kodama, G., Baldi, P. & Moyzis, R. K. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc Natl Acad Sci USA* **103**, 135-140 (2006).
71 Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).
72 Fay, J. C. & Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405-1413 (2000).
73 Kim, Y. & Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765-777 (2002).
74 Kim, Y. & Nielsen, R. Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513-1524 (2004).
75 Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F. & Bustamante, C. D. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**, 1401-1410 (2005).
76 Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res* **15**, 1566-1575 (2005).
77 Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W. & Akey, J. M. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res* **16**, 980-989 (2006).
78 Williamson, S. H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS Genet* **3**, e90 (2007).
79 Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97-159 (1931).
80 Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**, 1805-1814 (2002).

81 Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M. & Hill, W. G. Measures of human population structure show heterogeneity among genomic regions. *Genome Res* **15**, 1468-1476 (2005).
82 Beaumont, M. A. & Balding, D. J. Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **13**, 969-980 (2004).
83 The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).
84 Oleksyk, T. K. *et al.* Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE* **3**, e1712 (2008).
85 Kingman, J. F. C. The coalescent. *Stochastic Processes and Their Applications* **13**, 235-248 (1982).
86 R. C. Griffiths, S. T. Ancestral inference in population genetics. *Statistical Science* **9**, 307-319 (1994).
87 Harding, R. M. *et al.* Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* **60**, 772-789 (1997).
88 Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338 (2002).
89 Spencer, C. C. & Coop, G. SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* **20**, 3673-3675 (2004).
90 Mailund, T. *et al.* CoaSim: a flexible environment for simulating genetic data under coalescent models. *BMC bioinformatics* **6**, 252 (2005).
91 Marjoram, P. & Wall, J. D. Fast "coalescent" simulation. *BMC Genetics* **7**, 16 (2006).
92 Peng, B. & Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**, 3686-3687 (2005).
93 Hoggart, C. J. *et al.* Sequence-level population simulations over large genomic regions. *Genetics* **177**, 1725-1731 (2007).
94 Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* **19**, 826-837 (2009).
95 Harding, R. M. *et al.* Evidence for variable selective pressures at MC1R. *Am J Hum Genet* **66**, 1351-1361 (2000).
96 Soejima, M., Tachida, H., Ishida, T., Sano, A. & Koda, Y. Evidence for recent positive selection at the human *AIM1* locus in a European population. *Mol Biol Evol* **23**, 179-188 (2006).
97 Aoki, K. Sexual selection as a cause of human skin colour variation: Darwin's hypothesis revisited. *Ann Hum Biol* **29**, 589-608 (2002).
98 Juzeniene, A., Setlow, R., Porojnicu, A., Steindal, A. H. & Moan, J. Development of different human skin colors: a review highlighting photobiological and photobiophysical aspects. *J Photochem Photobiol B* **96**, 93-100 (2009).
99 Lamason, R. L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310**, 1782-1786 (2005).
100 Akey, J. M. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res* **19**, 711-722 (2009).
101 Akey, J. M. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* **19**, 711-722 (2009).

102 Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920-924 (2011).
103 Hu, M. *et al.* Exploration of signals of positive selection derived from genotype-based human genome scans using re-sequencing data. *Hum Genet* **131**, 665-674 (2012).
104 Fisher, R. A. *Statistical Methods for Research Workers*. 12th edn, (Oliver and Boyd, 1954).
105 Xue, Y. *et al.* Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am J Hum Genet* **78**, 659-670 (2006).
106 Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**, 111-118 (2010).
107 Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nature Methods* **5**, 1005-1010 (2008).
108 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
109 Guerra-Assuncao, J. A. & Enright, A. J. MapMi: automated mapping of microRNA loci. *BMC bioinformatics* **11**, 133 (2010).
110 Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie* **125**, 167-188 (1994).
111 Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112 (2009).
112 Xue, Y. *et al.* Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. *Genetics* **183**, 1065-1077 (2009).
113 Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883-886 (2010).
114 Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* **4**, e1000214 (2008).
115 Piriyapongsa, J. & Jordan, I. K. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PloS ONE* **2**, e203 (2007).
116 King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).
117 Quach, H. *et al.* Signatures of purifying and local positive selection in human miRNAs. *Am J Hum Genet* **84**, 316-327 (2009).
118 Neilson, L. I. *et al.* cDNA cloning and characterization of a human sperm antigen (SPAG6) with homology to the product of the *Chlamydomonas PF16* locus. *Genomics* **60**, 272-280 (1999).
119 Sapiro, R. *et al.* Male infertility, impaired sperm motility, and hydrocephalus in mice deficient in sperm-associated antigen 6. *Mol Cell Biol* **22**, 6298-6305 (2002).
120 Maine, G. N. & Burstein, E. COMMD proteins: COMMing to the scene. *Cell Mol Life Sci* **64**, 1997-2005 (2007).
121 Shakhova, O., Leung, C. & Marino, S. *Bmi1* in development and tumorigenesis of the central nervous system. *J. Mol. Med.* **83**, 596-600 (2005).
122 Schuringa, J. J. & Vellenga, E. Role of the polycomb group gene BMI1 in normal and leukemic hematopoietic stem and progenitor cells. *Curr Opin Hematol* **17**, 294-299 (2010).

123 Ginjala, V. *et al.* BMI1 is recruited to DNA breaks and contributes to DNA damage induced H2A ubiquitination and repair. *Mol Cell Biol* **31**, 1972-1982 (2011).
124 van der Lugt, N. M. *et al.* Posterior transformation, neurological abnormalities, and severe hematopoietic defects in mice with a targeted deletion of the *bmi-1* proto-oncogene. *Genes & Development* **8**, 757-769 (1994).
125 Zhang, J. & Sarge, K. D. Identification of a polymorphism in the RING finger of human Bmi-1 that causes its degradation by the ubiquitin-proteasome system. *FEBS Lett* **583**, 960-964 (2009).
126 Yngvadottir, B., Macarthur, D. G., Jin, H. & Tyler-Smith, C. The promise and reality of personal genomics. *Genome Biol* **10**, 237 (2009).
127 Hellenthal, G. & Stephens, M. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics* **23**, 520-521 (2007).
128 Abdi, H. in *Encyclopedia of Measurement and Statistics* (ed Neil Salkind) (Sage, 2007).
129 Gonzalez-Perez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* **88**, 440-449 (2011).
130 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4**, 1073-1082 (2009).
131 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248-249 (2010).
132 Mann, H. B. & whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann Math Statist* **18**, 50-60 (1947).
133 The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **submitted** (2012).
134 Dennis, G., Jr. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3 (2003).
135 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* **106**, 9362-9367 (2009).
136 Yngvadottir, B. *et al.* A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. *Am J Hum Genet* **84**, 224-234 (2009).
137 Harvey, K. F., Dinudom, A., Cook, D. I. & Kumar, S. The Nedd4-like protein KIAA0439 is a potential regulator of the epithelial sodium channel. *J Biol Chem* **276**, 8597-8601 (2001).
138 Raikwar, N. S. & Thomas, C. P. Nedd4-2 isoforms ubiquitinate individual epithelial sodium channel subunits and reduce surface expression and function of the epithelial sodium channel. *Am J Physiol Renal Physiol.* **294**, 1157-1165 (2008).
139 Dahlberg, J., Nilsson, L.-O., Wowern, F. v. & Melander, O. Polymorphism in NEDD4L Is Associated with Increased Salt Sensitivity, Reduced Levels of P-renin and Increased Levels of Nt-proANP. *PLoS ONE* **2**, e432 (2007).

140 Fava, C. *et al.* 24-h ambulatory blood pressure is linked to chromosome 18q21–22 and genetic variation of NEDD4L associates with cross-sectional and longitudinal blood pressure in Swedes. *Kidney Int* **70**, 562-569 (2006).
141 Russo, C. J. *et al.* Association of NEDD4L ubiquitin ligase with essential hypertension. *Hypertension* **46**, 488-491 (2005).
142 Pucharcos, C., Estivill, X. & Luna, S. d. l. Intersectin 2, a new multimodular protein involved in clathrin-mediated endocytosis. *FEBS Lett* **478**, 43-51 (2000).
143 McGavin, M. K. H. *et al.* The Intersectin 2 Adaptor Links Wiskott Aldrich Syndrome Protein (WASp)-mediated Actin Polymerization to T Cell Antigen Receptor Endocytosis. *J Exp Med* **194**, 1777-1787 (2001).
144 Winther, M., Berezin, V. & Walmod, P. S. NCAM2/OCAM/RNCAM: cell adhesion molecule with a role in neuronal compartmentalization. *Int J Biochem Cell Biol* **44**, 441-446 (2012).
145 Kaufman, L., Hayashi, K., Ross, M. J., Ross, M. D. & Klotman, P. E. Sidekick-1 is upregulated in glomeruli in HIV-associated nephropathy. *J Am Soc Nephrol* **15**, 1721-1730 (2004).
146 Kaufman, L. *et al.* The homophilic adhesion molecule sidekick-1 contributes to augmented podocyte aggregation in HIV-associated nephropathy. *FASEB J* **21**, 1367-1375 (2007).
147 Oguri, M. *et al.* Assessment of a polymorphism of SDK1 with hypertension in Japanese Individuals. *Am J Hypertens* **23**, 70-77 (2010).
148 Levy, D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nature Genet* **41**, 677-687 (2009).
149 Tang, K., Thornton, K. R. & Stoneking, M. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLoS Biol* **5**, e171 (2007).
150 Carlson, C. S. *et al.* Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res* **15**, 1553-1565 (2005).
151 Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W. & Akey, J. M. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res* **16**, 980-989 (2006).
152 Akey, J. M. *et al.* Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* **2**, e286 (2004).
153 Bryk, J. *et al.* Positive selection in East Asians for an *EDAR* allele that enhances NF-kappaB activation. *PLoS ONE* **3**, e2209 (2008).
154 Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* Vol. 7 (ed Douglas Futuyma and Janis Antonovics) 1-44 (Oxford University Press, 1991).
155 Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254 (2007).
156 Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876 (2007).
157 Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60-66 (2008).
158 Ahn, S.-M. *et al.* The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res* **19**, 1622-1629 (2009).

159 Kim, J.-I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011-1016 (2009).
160 Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2009).
161 Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943-947 (2010).
162 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-1097 (2007).
163 Lenhard, J. Kritische Untersuchung einer Methode zur Sch¨atzung Phylogenetischer Gr¨oßen. *PhD thesis, Mathematics Department, Johann Wolfgang Goethe University* (1997).
164 R. C. Griffiths, S. T. Simulating probability distributions in the coalescent. *Theor Popn Biol* **46**, 131-159 (1994).
165 R. C. Griffiths, S. T. Sampling theory for neutral alleles in a varying environment. *Phil Trans R Soc Lond B* **344**, 403-410 (1994).
166 R. C. Griffiths, S. T. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math Biosci* **127**, 77-98 (1995).
167 Harding, R. M. *et al.* Archaic African *and* Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* **60**, 772-789 (1997).
168 Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-1103 (2010).
169 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
170 Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**, 37-48 (1999).
171 Langergraber, K. E. *et al.* Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci USA* (2012).
172 Fimia, G. M. *et al.* Ambra1 regulates autophagy and development of the nervous system. *Nature* **447**, 1121-1125 (2007).
173 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc, Series B (Methodological)* **57**, 289-300 (1995).
174 Podlaha, O. & Zhang, J. Positive selection on protein-length in the evolution of a primate sperm ion channel. *Proc Natl Acad Sci USA* **100**, 12241-12246 (2003).
175 Tishkoff, S. A. *et al.* Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**, 455-462 (2001).
176 Ayodo, G. *et al.* Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am J Hum Genet* **81**, 234-242 (2007).
177 Skarnes, W. C. *et al.* A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337-342 (2011).
178 Freimer, N. & Sabatti, C. The human phenome project. *Nat Genet* **34**, 15-21 (2003).
179 Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet* (2012).

180 Casals, F. *et al.* Genetic adaptation of the antibacterial human innate immunity network. *BMC Evol Biol* **11**, 202 (2011).
181 Dall'olio, G. M. *et al.* Distribution of events of positive selection and population differentiation in a metabolic pathway: the case of asparagine N-glycosylation. *BMC Evol Biol* **12**, 98 (2012).
182 Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**, D561-568 (2011).
183 Xue, Y. *et al.* Adaptive evolution of UGT2B17 copy-number variation. *Am J Hum Genet* **83**, 337-346 (2008).

# Appendix A

## Parameters and commands for Chapter 2 simulations

***cosi* parameters:**

#random_seed 1001

# in bp.

length 1000000

# per bp per generation

mutation_rate 1.9e-8

recomb_file ../cosi_files/rec_r1

# population info

pop_define 1 european

pop_define 4 asian

pop_define 5 african

#european

pop_size 1 7700

sample_size 1 10

#asian

pop_size 4 7700

sample_size 4 1540

#african

pop_size 5 24000

sample_size 5 4800

pop_event split "asian and european split" 1 4 0

pop_event bottleneck "OoA bottleneck" 1 1499 .085

pop_event split "out of Africa" 5 1 1500

pop_event change_size "african pop size" 5 15000 12500

***mpop* commands:**

# mpop commands for Asian population (selection coefficient 0.007)

./mpop -i mpop_input_r1_Asian -o mpop_out_r1_Asian_1 -N 1540 -S -m 0.09 -r 0.46 -s 0.01 -h 0.5 -g 1;

./mpop -i mpop_out_r1_Asian_1 -o mpop_out_r1_Asian_2 -N 145 -g 1;

./mpop -i mpop_out_r1_Asian_2 -o mpop_out_r1_Asian_3 -N 1540 -g 318;

./mpop -i mpop_out_r1_Asian_3 -o mpop_out_r1_Asian_4 -N 20000 -g 80;

./mpop -i mpop_out_r1_Asian_4 -o mpop_out_r1_Asian -N 50 -g 0;

# mpop commands for African population (neutral)

./mpop -i mpop_input_r1_African -o mpop_out_r1_African_1 -N 4800 -m 0.09 -r 0.46 -g 360;

./mpop -i mpop_out_r1_African_1 -o mpop_out_r1_African_2 -N 20000 -g 40;

./mpop -i mpop_out_r1_African_2 -o mpop_out_r1_African -N 50 -g 0;

# Appendix B

## Chapter 2 targeted resequencing of two regions: PCR primers and protocols

### PCR enrichments*:

| Primer Name | Primer sequences | PCR product coordinates F=start, R=end | PCR product size (bp) |
|---|---|---|---|
| Region1_1F | TCTATCTCCTCCCTTACCCTTTG | chr4:158702285 | |
| Region1_1R | GATTCTTTCAGTGTTGATCTGGG | chr4:158708600 | 6316 |
| Region1_2F | TAATGCACCTTTGTTCTTGGTCT | chr4:158708036 | |
| Region1_2R | CTTGTAGTCCCATCATCTCCTTG | chr4:158715605 | 7570 |
| Region1_3F | ACCTCTCCCTACTCCCAGAGTC | chr4:158715109 | |
| Region1_3R | ATCTGCCATGAATACAGAAAGGA | chr4:158722818 | 7710 |
| Region1_4F | CTCCATGACTTTAGAGGCTACGA | chr4:158722140 | |
| Region1_4R | GAAGTAGGGTTGGAGAGGGTCTA | chr4:158730090 | 7951 |
| Region1_5F | TCCTCCTATCTTGTCTCTTGCTG | chr4:158729433 | |
| Region1_5R | GAGAAAGAAATTGTGTTGCATCC | chr4:158735391 | 5959 |
| Region1_6F | AGCCAGCCACACTTACTATGAAC | chr4:158734888 | |
| Region1_6R | GCAACTTCCCTCTAATATGCCTT | chr4:158741661 | 6774 |
| Region1_7F | AAATGGACTGTGCTTTCAAAGAG | chr4:158740885 | |
| Region1_7R | GTATTTGTCCTTCTGTGCCTGAC | chr4:158747831 | 6947 |
| Region1_8F | AAACGATTGACAGAGTGAAGAGC | chr4:158747100 | |
| Region1_8R | CCAGTCAGAAATATTGCAAGTCC | chr4:158753068 | 5969 |
| Region1_9F | CTGGAATTTCTTATCCTCGTCCT | chr4:158752465 | |
| Region1_9R | AGGTCTCGGATTACAGACATGAA | chr4:158758682 | 6218 |
| Region1_10F | GCAAGCTTCTCAATGGAGTTAAA | chr4:158757964 | |
| Region1_10R | TTGGGTGGAGAAGAAGTAATGAA | chr4:158767669 | 9705 |

| | | | |
|---|---|---|---|
| Region1_11F | AAGACCTGGAATCAGTAGAAGGG | chr4:158762445 | |
| Region1_11R | GGAGATTTACCAAGGCTTCACTT | chr4:158772142 | 9697 |
| Region1_12F | CTCACTATGGATATTGACGAGGC | chr4:158770930 | |
| Region1_12R | CCTTAATTTCGTTCTCCTGCTTT | chr4:158778727 | 7798 |
| Region1_13F | GGGCTCCTCACTTACCCAGTAG | chr4:158777605 | |
| Region1_13R | TGCTTCCGAAATTATTGTTCTGT | chr4:158785309 | 7704 |
| Region1_14F | ACAGCTGCCATTCAATAAATGTT | chr4:158783817 | |
| Region1_14R | TGCCAGGTAACCTAGATGAGGTA | chr4:158791494 | 7678 |
| Region1_15F | TGACTGACCATTATTGACCATGA | chr4:158790760 | |
| Region1_15R | TAGCTATGATTGATTGGGTGCTT | chr4:158797058 | 6299 |
| Region1_16F | TTGAACAGACGAATGAATGATTG | chr4:158796560 | |
| Region1_16R | TTTATGCTAATTGGCTCTGGGTA | chr4:158802768 | 6209 |
| Region1_17F | TCTTTATCTTGCCAGTTGAGCAT | chr4:158802175 | |
| Region1_17R | TATTTGTGTTCCCTTTCCTGCTA | chr4:158808423 | 6249 |
| Region1_18F | GTGAGAATTCATCTCAAAGCCAC | chr4:158807818 | |
| Region1_18R | GGAAGCTATTTACAGTTTGCCCT | chr4:158815385 | 7568 |
| Region1_19F | CAGTAAGCCCAAATGTTAAGGTG | chr4:158814909 | |
| Region1_19R | ACCTGACTTTATTTCCCTCTTCG | chr4:158822313 | 7405 |
| Region1_20F | GGATGCTGATCAATACCTGATGT | chr4:158821535 | |
| Region1_20R | CTACTTACGGCAACTCACAGCTT | chr4:158829166 | 7632 |
| Region1_21F | AGGAATGCTCAGTTCTTGTTCTG | chr4:158827934 | |
| Region1_21R | TTATTTCTGAGGGCTCTGTTCTG | chr4:158836728 | 8794 |
| Region1_22F | CATGGAAACTGAATAACCTGCTC | chr4:158834952 | |
| Region1_22R | ACAAGGATTCTCATTTGAGTGGA | chr4:158840977 | 6026 |
| Region1_23F | GGAAGTTGAAAGATGAATAGAACAAA | chr4:158840377 | |
| Region1_23R | ACGGTCAATATTCTCTCCTCACA | chr4:158847471 | 7095 |
| Region1_24F | ATCATGAGCCAAGTAAGCACAAT | chr4:158846985 | |
| Region1_24R | GGCACCTATGTGAAATCTGACTC | chr4:158853923 | 6939 |
| Region1_25F | ATGCCTTGCTTTCATAACTCTTG | chr4:158853372 | |

| | | | |
|---|---|---|---|
| Region1_25R | CGGAAAGTCTAATTTGAACAACG | chr4:158860928 | 7557 |
| Region1_26F | TCAAACTGAGTCTCTGGGAATGT | chr4:158859085 | |
| Region1_26R | TGGCTGGTAACTCATTAGGTCAT | chr4:158868003 | 8918 |
| Region1_27F | CACACAATTTATCCAACATCCCT | chr4:158866183 | |
| Region1_27R | TTACATTGATTGGATGCAGTGAG | chr4:158874110 | 7928 |
| Region1_28F | CTGAGGAATACTGCCGTATCAAG | chr4:158872529 | |
| Region1_28R | ACCAATCCCAGTCCTTTATGAAT | chr4:158881348 | 8819 |
| Region1_29F | GCAAAGCTAATTCGATACACCTG | chr4:158880521 | |
| Region1_29R | TCAAGATCAAATGCAGTCAGAGA | chr4:158887599 | 7079 |
| Region1_30F | CAAAGGTAATTGTGAGGTGAAGG | chr4:158886844 | |
| Region1_30R | TTGGGAGTTGAAGCTGGTATAAA | chr4:158894432 | 7589 |
| Region1_31F | TTCCTCTCTGTAAATGTGGCAAT | chr4:158893844 | |
| Region1_31R | AGTTTGAACAAAGCAGCAGGTAG | chr4:158901107 | 7264 |
| Region1_32F | GCTTGTCTATGCTTCACGAAGTT | chr4:158900212 | |
| Region1_32R | TTCTATCGCAATACTCCCTTTCA | chr4:158907520 | 7308 |
| Region1_33F | CACCAGGCTACAGTTTCTTCATC | chr4:158906161 | |
| Region1_33R | CATTGCTCCACATTCTCATTACA | chr4:158913779 | 7619 |
| Region1_34F | CTGAAGTGTGTAGAATGGTGCTG | chr4:158913234 | |
| Region1_34R | TTGAATCCACAAGGTGAAGCTAT | chr4:158920243 | 7010 |
| Region1_35F | AAGGATCATTTCTCTGCCCTAAC | chr4:158919497 | |
| Region1_35R | TTATTAGTGGTGCTTTCAGGGAA | chr4:158927302 | 7806 |
| Region1_36F | CAGTGGGTACTCTATGTTGAGGC | chr4:158926740 | |
| Region1_36R | CCTCTTCATGGTACAGATTCCTG | chr4:158934735 | 7996 |
| Region1_37F | AGGTCCAACTATAGGAGGAGTGG | chr4:158934050 | |
| Region1_37R | AATCACAAGTCAAGGGAGATTCA | chr4:158940035 | 5986 |
| Region1_38F | TGAGCAGTGTGAGAGTGGACTAA | chr4:158939317 | |
| Region1_38R | GTGGGATGGACACATATTCTGTT | chr4:158945415 | 6099 |
| Region1_39F | AGAGCTCCCTTCTCTCTGACATT | chr4:158944781 | |
| Region1_39R | TTCTGTGAGATTCCAACCCTTTA | chr4:158951872 | 7092 |

| | | | |
|---|---|---|---|
| Region1_40F | TGAGAATTTAGGTGAGGCTGTGT | chr4:158951262 | |
| Region1_40R | TCTTTCCTTCTCTCAGCCCTACT | chr4:158958029 | 6768 |
| Region1_41F | TTTGACAGAAGGGAAGTAAACCA | chr4:158956621 | |
| Region1_41R | GAGCTTGTCTTCATGCTCTGAAT | chr4:158966261 | 9640 |
| Region1_42F | AGAAGGAACTCTCCAGCTGATCT | chr4:158964566 | |
| Region1_42R | TTTGGCATAAACCACTCCTCTAA | chr4:158972100 | 7535 |
| Region1_43F | GCCCATCCATGTATGTTCTGTAT | chr4:158971591 | |
| Region1_43R | CACCCTGAAAGCATTCTTAATTG | chr4:158979627 | 8037 |
| Region1_44F | CATCCACCAAGGTTATAGCTCAG | chr4:158979058 | |
| Region1_44R | ATGGAGAAGAATGGACAAACTCA | chr4:158986215 | 7158 |
| Region1_45F | CATAGTGCTTCAAGATGTCCTCC | chr4:158985263 | |
| Region1_45R | TAAAGACAGCCTACAGAATGGGA | chr4:158994297 | 9034 |
| Region1_46F | CCCACTGTTCACCTTACAGACTC | chr4:158992842 | |
| Region1_46R | TGCCAAGATAATTGTTAGAGGGA | chr4:158999198 | 6356 |
| Region1_47F | GGACAATGACACTATGCTTCACA | chr4:158998528 | |
| Region1_47R | ACATCCTCCTAGCACTAACTCCC | chr4:159006314 | 7786 |
| Region1_48F | AAATCCAACATTAGAGCGACAAA | chr4:159004387 | |
| Region1_48R | ATGCGACAGAAAGAGAATCAGAG | chr4:159010873 | 6487 |
| Region1_49F | CACTTGCTCATGAACTAAAGCCT | chr4:159010262 | |
| Region1_49R | GATCCTCAAATGGTGAGTCTGTC | chr4:159016211 | 5950 |

*PCR protocol: Xue et al. [183]

In total, 49 pairs of PCR primers were designed for chr4:158Mb, 42 for chr10:22Mb and 4 pairs for the Y chromosome to amplify 5-11 kb PCR products with overlap of >500 bp, using a Perl script (http://droog.gs.washington.edu/PCR-Overlap.html). Two previous pairs for *CASP12* [105] were also used. The three base pairs at the 3' end of all primers were

confirmed not to overlap with any SNP in dbSNP127 (http://www.ncbi.nlm.nih.gov/projects/SNP/). The primer sequences and PCR conditions are listed in above table. Forty-four out of 49 fragments from chr4:158Mb, 37 out of 42 from chr10:22Mb and all from the Y chromosome and *CASP12* were successfully amplified in initial tests. These fragments were subsequently amplified in 28 CHB and 2 YRI samples from the HapMap collection. Three CHB provided poor quality data for chr4:158Mb, and four for chr10:22Mb, and were excluded from all subsequent analyses. Amplification was tested by agarose gel electrophoresis followed by ethidium bromide staining, and approximate quantification was performed from the band intensity. Thirty nine out of 49 (~80%) long PCR primer pairs worked well for 22 or more samples for chr4:158Mb, and 32/42 (~75%) for 20 or more samples for chr10:22Mb. The PCR products from each individual sample were pooled, approximately equalizing the molar yield for the Illumina sequencing paired end library construction.

# Appendix C

## Parameters and commands for Chapter 3 simulations

***cosi* parameters:**

#random_seed 1001  # Specifies a particular random number seed

# in bp.

length 300000

# per bp per generation

mutation_rate 1.0e-8

recomb_file  rec_file

# population info

pop_define 1 european

pop_define 4 asian

pop_define 5 african

#european

pop_size 1 100000

sample_size 1 120

#asian

pop_size 4 100000

sample_size 4 120

#african

pop_size 5 100000

sample_size 5 120

pop_event migration_rate "afr->eur migration" 5 1 0 0.000032

pop_event migration_rate "eur->afr migration" 1 5 0 0.000032

pop_event migration_rate "afr->as migration" 5 4 0 0.000008

pop_event migration_rate "as->afr migration" 4 5 0 0.000008

#pop_event admix "african american admix" 3 1 5. .2

#pop_event split "african to aa" 5 3 7.0

pop_event change_size "agriculture - african" 5 200 24000

pop_event change_size "agriculture - european" 1 350 7700

pop_event change_size "agriculture - asian" 4 400 7700

pop_event bottleneck "african bottleneck" 5 1997 .008

pop_event bottleneck "asian bottleneck" 4 1998 .067

pop_event bottleneck "european bottleneck" 1 1999 .02

pop_event sweep "European selection" 1 0 0.01 0.5 0.9

pop_event split "asian and european split" 1 4 2000

pop_event migration_rate "afr->eur migration" 5 1 1996 0

pop_event migration_rate "eur->afr migration" 1 5 1995 0

pop_event migration_rate "afr->as migration" 5 4 1994 0

pop_event migration_rate "as->afr migration" 4 5 1993 0

pop_event bottleneck "OoA bottleneck" 1 3499 .085

pop_event split "out of Africa" 5 1 3500

pop_event change_size "african pop size" 5 17000 12500

***mpop* commands:**

Neutral:

CHBJPT:

./mpop -i ms_out_Eurasian -o mpop_tmp1 -N 3080 -m 0.015 -r rec -g 19;

./mpop -i mpop_tmp1 -o mpop_tmp2 -N 290 -g 1;

./mpop -i mpop_tmp2 -o mpop_tmp3-N 3080 -g 300;

./mpop -i mpop_tmp3 -o mpop_tmp4 -N 40000 -g 80;

./mpop -i mpop_tmp4 -o mpop_out -N 120 -g 0;

CEU:

./mpop -i ms_out_Eurasian -o mpop_tmp1 -N 3080 -m 0.015 -r rec -g 5;

./mpop -i mpop_tmp1 -o mpop_tmp2 -N 287 -g 1;

./mpop -i mpop_tmp2 -o mpop_tmp3-N 3080 -g 324;

./mpop -i mpop_tmp3 -o mpop_tmp4 -N 40000 -g 70;

./mpop -i mpop_tmp4 -o mpop_out -N 120 -g 0;

YRI:

./mpop -i ms_out_African -o mpop_tmp1 -N 9600 -m 0.015 -r rec -g 360;

./mpop -i mpop_tmp1 -o mpop_tmp2 -N 40000 -g 40;

./mpop -i mpop_tmp2 -o mpop_out -N 120 -g 0;

Selection coefficient = 0.01, age of sweep = 2000 generations:

CHBJPT:

./mpop -i ms_out_Eurasian -o mpop_tmp1 -N 3080 -m 0.015 -r rec -S -s 0.05 -h 0.5 -g 19;

./mpop -i mpop_tmp1 -o mpop_tmp2 -N 290 -g 1;

./mpop -i mpop_tmp2 -o mpop_tmp3-N 3080 -g 300;

./mpop -i mpop_tmp3 -o mpop_tmp4 -N 40000 -g 80;

./mpop -i mpop_tmp4 -o mpop_out -N 120 -g 0;

CEU:

./mpop -i ms_out_Eurasian -o mpop_tmp1 -N 3080 -m 0.015 -r rec -S -s 0.05 -h 0.5 -g 5;

./mpop -i mpop_tmp1 -o mpop_tmp2 -N 287 -g 1;

./mpop -i mpop_tmp2 -o mpop_tmp3-N 3080 -g 324;

./mpop -i mpop_tmp3 -o mpop_tmp4 -N 40000 -g 70;

./mpop -i mpop_tmp4 -o mpop_out -N 120 -g 0;

YRI:

./mpop -i ms_out_African -o mpop_tmp1 -N 9600 -m 0.015 -r rec -S -s 0.05 -h 0.5 -g 360;

./mpop -i mpop_tmp1 -o mpop_tmp2 -N 40000 -g 40;

./mpop -i mpop_tmp2 -o mpop_out -N 120 -g 0;

# Appendix D

## Candidate regions and genes in each population

Coordinates are in NCBI36.

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| CEU | | | | | | | |
| 1 | 797099 | 906801 | 847099 | 856801 | 2.29E-08 | ENSG00000187634 | *SAMD11* |
| 1 | 10068896 | 10178302 | 10118896 | 10128302 | 1.96E-08 | ENSG00000130939 | *UBE4B* |
| 1 | 13821389 | 13931321 | 13871389 | 13881321 | 5.17E-09 | ENSG00000116731 | *PRDM2* |
| 1 | 27831079 | 27940924 | 27881079 | 27890924 | 5.47E-09 | ENSG00000126709 | *IFI6* |
| 1 | 35355105 | 35464358 | 35405105 | 35414358 | 1.36E-08 | ENSG00000116560 | *SFPQ* |
| 1 | 53228676 | 53338605 | 53278676 | 53288605 | 2.00E-09 | ENSG00000116171 | *SCP2* |
| 1 | 53788087 | 53897937 | 53838087 | 53847937 | 6.93E-09 | ENSG00000174332 | *GLIS1* |
| 1 | 66419956 | 66529521 | 66469956 | 66479521 | 1.20E-08 | ENSG00000184588 | *PDE4B* |
| 1 | 76884112 | 76994105 | 76934112 | 76944105 | 6.54E-09 | n.a. | n.a. |
| 1 | 86596670 | 86706639 | 86646670 | 86656639 | 3.51E-09 | ENSG00000122417 | *ODF2L* |
| 1 | 86596670 | 86706639 | 86646670 | 86656639 | 3.51E-09 | ENSG00000137975 | *CLCA2* |
| 1 | 101748396 | 101970549 | 101798396 | 101808232 | 2.33E-09 | n.a. | n.a. |
| 1 | 102471628 | 102611505 | 102541801 | 102551773 | 4.81E-09 | n.a. | n.a. |
| 1 | 103207706 | 103327893 | 103268059 | 103277893 | 3.59E-09 | ENSG00000060718 | *COL11A1* |
| 1 | 104387351 | 104573771 | 104459973 | 104468954 | 9.67E-10 | n.a. | n.a. |
| 1 | 105857859 | 105967516 | 105907859 | 105917516 | 4.12E-10 | n.a. | n.a. |
| 1 | 106474162 | 106584135 | 106524162 | 106534135 | 1.36E-08 | n.a. | n.a. |
| 1 | 117217385 | 117327315 | 117267385 | 117277315 | 4.30E-09 | ENSG00000134247 | *PTGFRN* |
| 1 | 118134064 | 118264087 | 118184064 | 118193499 | 9.06E-09 | ENSG00000196505 | *GDAP2* |
| 1 | 151028501 | 151159115 | 151099455 | 151109115 | 1.31E-10 | ENSG00000163206 | *SMCP* |
| 1 | 154773501 | 154893188 | 154834024 | 154843188 | 4.42E-09 | ENSG00000183856 | *IQGAP3* |
| 1 | 161195217 | 161304900 | 161245217 | 161254900 | 1.98E-10 | n.a. | n.a. |
| 1 | 162325660 | 162434777 | 162375660 | 162384777 | 5.87E-09 | n.a. | n.a. |
| 1 | 163721017 | 163830863 | 163771017 | 163780863 | 3.10E-09 | ENSG00000162763 | *LRRC52* |
| 1 | 166615355 | 166725170 | 166665355 | 166675170 | 2.09E-08 | n.a. | n.a. |
| 1 | 181833404 | 181942458 | 181883404 | 181892458 | 1.87E-08 | ENSG00000143344 | *RGL1* |
| 1 | 181833404 | 181942458 | 181883404 | 181892458 | 1.87E-08 | ENSG00000173627 | *APOBEC4* |
| 1 | 183788556 | 183908552 | 183848976 | 183858552 | 2.17E-10 | n.a. | n.a. |
| 1 | 187028219 | 187158523 | 187088277 | 187097830 | 5.68E-09 | n.a. | n.a. |
| 1 | 187987903 | 188148535 | 188037903 | 188047844 | 2.52E-08 | n.a. | n.a. |
| 1 | 190884961 | 191004182 | 190945032 | 190954182 | 1.77E-08 | ENSG00000127074 | *RGS13* |
| 1 | 206899055 | 207008768 | 206949055 | 206958768 | 2.54E-09 | n.a. | n.a. |
| 1 | 212081680 | 212191240 | 212131680 | 212141240 | 3.35E-08 | n.a. | n.a. |
| 1 | 212827867 | 212937843 | 212877867 | 212887843 | 1.42E-08 | ENSG00000117724 | *CENPF* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 1 | 219734611 | 219854641 | 219784611 | 219794242 | 5.86E-09 | n.a. | n.a. |
| 1 | 232425142 | 232534909 | 232475142 | 232484909 | 4.62E-09 | ENSG00000183780 | *SLC35F3* |
| 1 | 234895573 | 235004711 | 234945573 | 234954711 | 1.53E-08 | ENSG00000077522 | *ACTN2* |
| 1 | 236273131 | 236403322 | 236323131 | 236333113 | 2.28E-10 | n.a. | n.a. |
| 2 | 6148356 | 6268236 | 6198356 | 6207722 | 1.10E-09 | n.a. | n.a. |
| 2 | 7533894 | 7643810 | 7583894 | 7593810 | 1.23E-08 | n.a. | n.a. |
| 2 | 21596445 | 21841341 | 21646445 | 21655066 | 5.01E-12 | n.a. | n.a. |
| 2 | 24385849 | 24504568 | 24435849 | 24445846 | 2.03E-10 | ENSG00000198399 | *ITSN2* |
| 2 | 39917077 | 40027005 | 39967077 | 39977005 | 3.09E-08 | n.a. | n.a. |
| 2 | 69025791 | 69134554 | 69075791 | 69084554 | 2.67E-10 | ENSG00000169605 | *GKN1* |
| 2 | 69025791 | 69134554 | 69075791 | 69084554 | 2.67E-10 | ENSG00000169604 | *ANTXR1* |
| 2 | 83128877 | 83238859 | 83178877 | 83188859 | 2.47E-08 | n.a. | n.a. |
| 2 | 107231463 | 107340791 | 107281463 | 107290791 | 8.59E-10 | n.a. | n.a. |
| 2 | 121376537 | 121486418 | 121426537 | 121436418 | 8.08E-09 | ENSG00000074047 | *GLI2* |
| 2 | 150814344 | 150924100 | 150864344 | 150874100 | 2.50E-08 | n.a. | n.a. |
| 2 | 167686338 | 167796283 | 167736338 | 167746283 | 2.62E-08 | ENSG00000163092 | *XIRP2* |
| 2 | 182245152 | 182439132 | 182295152 | 182305057 | 2.13E-11 | ENSG00000162992 | *NEUROD1* |
| 2 | 195721849 | 195831681 | 195771849 | 195781681 | 1.86E-08 | n.a. | n.a. |
| 2 | 224178700 | 224288695 | 224228700 | 224238695 | 3.11E-11 | n.a. | n.a. |
| 2 | 237029301 | 237139132 | 237079301 | 237089132 | 2.46E-08 | ENSG00000132321 | *IQCA1* |
| 3 | 355682 | 773168 | 426407 | 436102 | 7.65E-11 | ENSG00000134121 | *CHL1* |
| 3 | 3835933 | 3945844 | 3885933 | 3895844 | 8.43E-09 | ENSG00000144455 | *SUMF1* |
| 3 | 4050353 | 4170446 | 4110452 | 4120446 | 3.55E-09 | n.a. | n.a. |
| 3 | 7018406 | 7128393 | 7068406 | 7078393 | 3.53E-08 | ENSG00000196277 | *GRM7* |
| 3 | 7296272 | 7406271 | 7346272 | 7356271 | 5.74E-09 | n.a. | n.a. |
| 3 | 8482559 | 8592440 | 8532559 | 8542440 | 3.17E-08 | ENSG00000071282 | *LMCD1* |
| 3 | 11925199 | 12035027 | 11975199 | 11985027 | 4.46E-09 | ENSG00000157152 | *SYN2* |
| 3 | 14825904 | 14935524 | 14875904 | 14885524 | 6.04E-09 | ENSG00000154783 | *FGD5* |
| 3 | 15972893 | 16082532 | 16022893 | 16032532 | 4.36E-09 | n.a. | n.a. |
| 3 | 29593202 | 29773402 | 29643202 | 29652970 | 7.48E-09 | ENSG00000144642 | *RBMS3* |
| 3 | 40617631 | 40894999 | 40667631 | 40677443 | 2.99E-09 | n.a. | n.a. |
| 3 | 41645755 | 41755115 | 41695755 | 41705115 | 3.59E-08 | ENSG00000168038 | *ULK4* |
| 3 | 58659078 | 58768392 | 58709078 | 58718392 | 1.16E-08 | ENSG00000163689 | *C3orf67* |
| 3 | 59781034 | 59890749 | 59831034 | 59840749 | 7.59E-09 | ENSG00000189283 | *FHIT* |
| 3 | 66556933 | 66665955 | 66606933 | 66615955 | 7.96E-09 | ENSG00000144749 | *LRIG1* |
| 3 | 89800572 | 89910474 | 89850572 | 89860474 | 3.23E-08 | n.a. | n.a. |
| 3 | 97896644 | 98017341 | 97957359 | 97967341 | 8.19E-09 | ENSG00000080224 | *EPHA6* |
| 3 | 99217131 | 99327009 | 99267131 | 99277009 | 2.23E-08 | ENSG00000196578 | *OR5AC2* |
| 3 | 104252756 | 104362666 | 104302756 | 104312666 | 2.53E-08 | n.a. | n.a. |
| 3 | 107434455 | 107544376 | 107484455 | 107494376 | 1.75E-09 | n.a. | n.a. |
| 3 | 111775748 | 111882789 | 111825748 | 111832789 | 2.83E-08 | n.a. | n.a. |
| 3 | 112565463 | 112675318 | 112615463 | 112625318 | 3.98E-08 | n.a. | n.a. |
| 3 | 113254707 | 113364412 | 113304707 | 113314412 | 2.29E-08 | ENSG00000114529 | *C3orf52* |
| 3 | 124864552 | 124973930 | 124914552 | 124923930 | 2.86E-08 | ENSG00000065534 | *MYLK* |
| 3 | 136816358 | 136926078 | 136866358 | 136876078 | 3.16E-08 | n.a. | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 3 | 144995803 | 145115822 | 145055942 | 145065822 | 6.06E-10 | ENSG00000181804 | *SLC9A9* |
| 3 | 147407375 | 147517084 | 147457375 | 147467084 | 2.45E-08 | ENSG00000114698 | *PLSCR4* |
| 3 | 157455823 | 157565750 | 157505823 | 157515750 | 2.91E-08 | ENSG00000169282 | *KCNAB1* |
| 3 | 174493364 | 174603316 | 174543364 | 174553316 | 1.25E-08 | n.a. | n.a. |
| 3 | 175277765 | 175387764 | 175327765 | 175337764 | 2.31E-08 | ENSG00000169760 | *NLGN1* |
| 3 | 177707190 | 177816865 | 177757190 | 177766865 | 2.94E-09 | n.a. | n.a. |
| 3 | 178886726 | 178996686 | 178936726 | 178946686 | 1.96E-09 | n.a. | n.a. |
| 3 | 187149342 | 187258965 | 187199342 | 187208965 | 1.40E-08 | ENSG00000171656 | *ETV5* |
| 3 | 190059450 | 190189991 | 190109450 | 190119392 | 6.72E-09 | ENSG00000145012 | *LPP* |
| 3 | 191940911 | 192050897 | 191990911 | 192000897 | 3.86E-08 | n.a. | n.a. |
| 3 | 193235593 | 193345464 | 193285593 | 193295464 | 3.08E-09 | ENSG00000114279 | *FGF12* |
| 3 | 194163940 | 194364328 | 194304531 | 194314328 | 1.32E-09 | n.a. | n.a. |
| 3 | 195632688 | 195742208 | 195682688 | 195692208 | 1.82E-08 | ENSG00000133657 | *ATP13A3* |
| 3 | 196051892 | 196161845 | 196101892 | 196111845 | 7.44E-09 | n.a. | n.a. |
| 4 | 3505256 | 3615203 | 3555256 | 3565203 | 1.45E-08 | ENSG00000216560 | n.a. |
| 4 | 4938663 | 5048572 | 4988663 | 4998572 | 1.42E-08 | n.a. | n.a. |
| 4 | 5261778 | 5505182 | 5352718 | 5362388 | 1.65E-09 | ENSG00000152953 | *STK32B* |
| 4 | 14278140 | 14387761 | 14328140 | 14337761 | 3.16E-09 | n.a. | n.a. |
| 4 | 15278537 | 15388271 | 15328537 | 15338271 | 3.03E-08 | ENSG00000109743 | *BST1* |
| 4 | 24338604 | 24448532 | 24388604 | 24398532 | 7.18E-09 | ENSG00000109610 | *SOD3* |
| 4 | 32806311 | 32936928 | 32877427 | 32886928 | 4.69E-10 | n.a. | n.a. |
| 4 | 33216406 | 33326273 | 33266406 | 33276273 | 8.53E-09 | n.a. | n.a. |
| 4 | 34349592 | 34479595 | 34409609 | 34419485 | 2.04E-09 | n.a. | n.a. |
| 4 | 38403778 | 38513647 | 38453778 | 38463647 | 2.70E-08 | ENSG00000174123 | *TLR10* |
| 4 | 42316679 | 42426548 | 42366679 | 42376548 | 1.65E-09 | ENSG00000124406 | *ATP8A1* |
| 4 | 43008451 | 43128526 | 43058451 | 43068071 | 2.04E-09 | n.a. | n.a. |
| 4 | 55697521 | 55891731 | 55757643 | 55766818 | 9.58E-11 | n.a. | n.a. |
| 4 | 57067392 | 57176318 | 57117392 | 57126318 | 2.37E-08 | ENSG00000196503 | *ARL9* |
| 4 | 60560440 | 60670150 | 60610440 | 60620150 | 1.95E-09 | n.a. | n.a. |
| 4 | 64226667 | 64336107 | 64276667 | 64286107 | 4.77E-09 | n.a. | n.a. |
| 4 | 64772668 | 64882457 | 64822668 | 64832457 | 3.29E-08 | ENSG00000205678 | n.a. |
| 4 | 67269221 | 67378380 | 67319221 | 67328380 | 3.71E-09 | n.a. | n.a. |
| 4 | 71736097 | 71845931 | 71786097 | 71795931 | 1.40E-08 | ENSG00000132467 | *UTP3* |
| 4 | 71736097 | 71845931 | 71786097 | 71795931 | 1.40E-08 | ENSG00000018189 | *RUFY3* |
| 4 | 74361956 | 74512708 | 74411956 | 74421185 | 5.95E-09 | n.a. | n.a. |
| 4 | 75543994 | 75653814 | 75593994 | 75603814 | 3.95E-08 | n.a. | n.a. |
| 4 | 75931273 | 76041154 | 75981273 | 75991154 | 1.31E-08 | ENSG00000174808 | *BTC* |
| 4 | 79977676 | 80087610 | 80027676 | 80037610 | 2.69E-09 | ENSG00000138756 | *BMP2K* |
| 4 | 79977676 | 80087610 | 80027676 | 80037610 | 2.69E-09 | ENSG00000163291 | *PAQR3* |
| 4 | 84281633 | 84391554 | 84331633 | 84341554 | 1.42E-08 | n.a. | n.a. |
| 4 | 85482704 | 85591519 | 85532704 | 85541519 | 8.85E-09 | n.a. | n.a. |
| 4 | 89092930 | 89202832 | 89142930 | 89152832 | 3.97E-08 | ENSG00000118762 | *PKD2* |
| 4 | 93758997 | 93868355 | 93808997 | 93818355 | 9.60E-09 | ENSG00000152208 | *GRID2* |
| 4 | 94797451 | 94907355 | 94847451 | 94857355 | 5.56E-09 | n.a. | n.a. |
| 4 | 96065573 | 96175204 | 96115573 | 96125204 | 1.92E-08 | ENSG00000138696 | *BMPR1B* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 4 | 96856764 | 96966688 | 96906764 | 96916688 | 3.90E-09 | n.a. | n.a. |
| 4 | 134308254 | 134437993 | 134368280 | 134377531 | 1.21E-09 | ENSG00000138650 | *PCDH10* |
| 4 | 148389178 | 148499112 | 148439178 | 148449112 | 2.21E-08 | n.a. | n.a. |
| 4 | 156963127 | 157082944 | 157013127 | 157022938 | 2.51E-09 | ENSG00000023843 | *ACCN5* |
| 4 | 163364280 | 163474168 | 163414280 | 163424168 | 2.18E-08 | n.a. | n.a. |
| 4 | 163663033 | 163772877 | 163713033 | 163722877 | 8.85E-09 | n.a. | n.a. |
| 4 | 165662952 | 165772885 | 165712952 | 165722885 | 1.12E-08 | n.a. | n.a. |
| 4 | 167365959 | 167485922 | 167415959 | 167425306 | 8.65E-10 | n.a. | n.a. |
| 4 | 168404471 | 168514387 | 168454471 | 168464387 | 2.77E-08 | n.a. | n.a. |
| 4 | 171478904 | 171745218 | 171528904 | 171538771 | 3.81E-09 | n.a. | n.a. |
| 4 | 172691703 | 172811785 | 172741703 | 172751630 | 2.54E-08 | n.a. | n.a. |
| 4 | 176393686 | 176746676 | 176443686 | 176453372 | 2.98E-09 | n.a. | n.a. |
| 4 | 178045518 | 178155489 | 178095518 | 178105489 | 5.98E-10 | n.a. | n.a. |
| 4 | 179180564 | 179290500 | 179230564 | 179240500 | 1.21E-08 | n.a. | n.a. |
| 4 | 180031210 | 180141029 | 180081210 | 180091029 | 8.48E-09 | n.a. | n.a. |
| 4 | 180379745 | 180530927 | 180439785 | 180449666 | 1.18E-09 | n.a. | n.a. |
| 4 | 182020705 | 182140717 | 182081056 | 182090717 | 2.01E-08 | n.a. | n.a. |
| 4 | 186586040 | 186695933 | 186636040 | 186645933 | 5.95E-09 | ENSG00000168491 | *CCDC110* |
| 4 | 186586040 | 186695933 | 186636040 | 186645933 | 5.95E-09 | ENSG00000154553 | *PDLIM3* |
| 5 | 888770 | 998682 | 938770 | 948682 | 3.66E-08 | ENSG00000188818 | *ZDHHC11* |
| 5 | 4195348 | 4326216 | 4266360 | 4276216 | 1.21E-10 | n.a. | n.a. |
| 5 | 5251501 | 5361050 | 5301501 | 5311050 | 3.20E-08 | ENSG00000145536 | *ADAMTS16* |
| 5 | 6149837 | 6259708 | 6199837 | 6209708 | 3.74E-08 | n.a. | n.a. |
| 5 | 11405272 | 11513912 | 11455272 | 11463912 | 3.55E-08 | ENSG00000169862 | *CTNND2* |
| 5 | 11754601 | 11916207 | 11856492 | 11866207 | 3.11E-11 | n.a. | n.a. |
| 5 | 13602516 | 13712488 | 13652516 | 13662488 | 3.90E-08 | n.a. | n.a. |
| 5 | 15068086 | 15177819 | 15118086 | 15127819 | 2.60E-08 | n.a. | n.a. |
| 5 | 16291882 | 16411560 | 16341882 | 16351531 | 3.96E-09 | n.a. | n.a. |
| 5 | 18046735 | 18156677 | 18096735 | 18106677 | 2.35E-10 | n.a. | n.a. |
| 5 | 20108845 | 20217295 | 20158845 | 20167295 | 1.74E-08 | n.a. | n.a. |
| 5 | 23609527 | 23719495 | 23659527 | 23669495 | 1.16E-08 | n.a. | n.a. |
| 5 | 26369074 | 26478948 | 26419074 | 26428948 | 2.04E-09 | n.a. | n.a. |
| 5 | 29692634 | 29802426 | 29742634 | 29752426 | 5.32E-09 | n.a. | n.a. |
| 5 | 30484474 | 30594443 | 30534474 | 30544443 | 1.32E-08 | n.a. | n.a. |
| 5 | 31453154 | 31562933 | 31503154 | 31512933 | 2.65E-08 | ENSG00000113360 | *RNASEN* |
| 5 | 31453154 | 31562933 | 31503154 | 31512933 | 2.65E-08 | ENSG00000082213 | *C5orf22* |
| 5 | 33000096 | 33109984 | 33050096 | 33059984 | 2.14E-08 | n.a. | n.a. |
| 5 | 34402524 | 34512413 | 34452524 | 34462413 | 4.78E-10 | n.a. | n.a. |
| 5 | 38021327 | 38131205 | 38071327 | 38081205 | 3.11E-08 | n.a. | n.a. |
| 5 | 53958992 | 54068710 | 54008992 | 54018710 | 2.74E-08 | n.a. | n.a. |
| 5 | 54902588 | 55012059 | 54952588 | 54962059 | 3.19E-08 | ENSG00000177058 | *SLC38A9* |
| 5 | 55766947 | 55875700 | 55816947 | 55825700 | 5.86E-09 | n.a. | n.a. |
| 5 | 67955416 | 68072958 | 68005416 | 68015396 | 2.63E-10 | n.a. | n.a. |
| 5 | 75151384 | 75261335 | 75201384 | 75211335 | 2.75E-08 | n.a. | n.a. |
| 5 | 97070120 | 97220944 | 97162942 | 97170944 | 1.06E-08 | n.a. | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 5 | 99506633 | 99616542 | 99556633 | 99566542 | 3.89E-08 | n.a. | n.a. |
| 5 | 100710045 | 100840352 | 100780449 | 100790352 | 1.98E-10 | n.a. | n.a. |
| 5 | 100927427 | 101047459 | 100977427 | 100987371 | 3.92E-10 | n.a. | n.a. |
| 5 | 102002354 | 102112090 | 102052354 | 102062090 | 2.39E-09 | n.a. | n.a. |
| 5 | 109214330 | 109324059 | 109264330 | 109274059 | 9.91E-10 | ENSG00000112893 | *MAN2A1* |
| 5 | 111946180 | 112055873 | 111996180 | 112005873 | 6.96E-09 | n.a. | n.a. |
| 5 | 115908166 | 116017795 | 115958166 | 115967795 | 2.92E-08 | ENSG00000092421 | *SEMA6A* |
| 5 | 116405732 | 116515721 | 116455732 | 116465721 | 3.02E-08 | n.a. | n.a. |
| 5 | 117899353 | 118009270 | 117949353 | 117959270 | 2.00E-08 | n.a. | n.a. |
| 5 | 121071350 | 121234503 | 121121350 | 121131298 | 4.34E-11 | n.a. | n.a. |
| 5 | 126209968 | 126462023 | 126259968 | 126266104 | 7.17E-10 | ENSG00000173926 | *Mar-03* |
| 5 | 144884095 | 144992878 | 144934095 | 144942878 | 2.16E-09 | n.a. | n.a. |
| 5 | 147361224 | 147471200 | 147411224 | 147421200 | 3.71E-09 | ENSG00000133710 | *SPINK5* |
| 5 | 147566850 | 147676554 | 147616850 | 147626554 | 1.41E-08 | ENSG00000178172 | *SPINK6* |
| 5 | 159112961 | 159232971 | 159173066 | 159182971 | 5.64E-10 | n.a. | n.a. |
| 5 | 168674391 | 168784360 | 168724391 | 168734360 | 2.04E-09 | n.a. | n.a. |
| 5 | 175034965 | 175144944 | 175084965 | 175094944 | 1.95E-08 | ENSG00000113749 | *HRH2* |
| 6 | 2031068 | 2140862 | 2081068 | 2090862 | 1.17E-09 | ENSG00000112699 | *GMDS* |
| 6 | 3808344 | 3917849 | 3858344 | 3867849 | 2.50E-08 | n.a. | n.a. |
| 6 | 10675596 | 10785508 | 10725596 | 10735508 | 1.72E-08 | ENSG00000111846 | *GCNT2* |
| 6 | 29470141 | 29580071 | 29520141 | 29530071 | 6.31E-09 | ENSG00000112462 | *OR12D3* |
| 6 | 30022655 | 30132654 | 30072655 | 30082654 | 3.94E-08 | ENSG00000204623 | *C6orf12* |
| 6 | 30022655 | 30132654 | 30072655 | 30082654 | 3.94E-08 | ENSG00000204622 | *HLA-J* |
| 6 | 30755426 | 30865118 | 30805426 | 30815118 | 1.83E-08 | ENSG00000196230 | *TUBBP2* |
| 6 | 30755426 | 30865118 | 30805426 | 30815118 | 1.83E-08 | ENSG00000137312 | *FLOT1* |
| 6 | 34095971 | 34205789 | 34145971 | 34155789 | 5.31E-09 | ENSG00000124493 | *GRM4* |
| 6 | 40298670 | 40418882 | 40348670 | 40358454 | 1.21E-09 | n.a. | n.a. |
| 6 | 40493514 | 40603300 | 40543514 | 40553300 | 1.27E-08 | ENSG00000156564 | *LRFN2* |
| 6 | 43507550 | 43617547 | 43557550 | 43567547 | 1.37E-08 | ENSG00000171462 | *DLK2* |
| 6 | 51550467 | 51670722 | 51600467 | 51610364 | 1.41E-09 | ENSG00000170927 | *PKHD1* |
| 6 | 56563975 | 56673746 | 56613975 | 56623746 | 2.87E-08 | ENSG00000151914 | *DST* |
| 6 | 57648043 | 57757862 | 57698043 | 57707862 | 1.82E-08 | n.a. | n.a. |
| 6 | 58818960 | 58928887 | 58868960 | 58878887 | 1.20E-08 | n.a. | n.a. |
| 6 | 71801383 | 71911333 | 71851383 | 71861333 | 2.61E-09 | n.a. | n.a. |
| 6 | 72283845 | 72393767 | 72333845 | 72343767 | 5.93E-09 | n.a. | n.a. |
| 6 | 73805330 | 73966332 | 73865381 | 73875098 | 2.63E-10 | ENSG00000185760 | *KCNQ5* |
| 6 | 75527921 | 75637322 | 75577921 | 75587322 | 7.22E-09 | n.a. | n.a. |
| 6 | 78105977 | 78215941 | 78155977 | 78165941 | 2.97E-08 | n.a. | n.a. |
| 6 | 85292936 | 85401929 | 85342936 | 85351929 | 1.33E-08 | n.a. | n.a. |
| 6 | 95457138 | 95567100 | 95507138 | 95517100 | 3.18E-08 | n.a. | n.a. |
| 6 | 97524281 | 97634118 | 97574281 | 97584118 | 1.29E-08 | ENSG00000186231 | *KLHL32* |
| 6 | 102420908 | 102530890 | 102470908 | 102480890 | 2.63E-08 | ENSG00000164418 | *GRIK2* |
| 6 | 103306890 | 103427375 | 103367406 | 103377375 | 7.08E-11 | n.a. | n.a. |
| 6 | 121576846 | 121697335 | 121637341 | 121647335 | 1.25E-08 | ENSG00000146350 | *C6orf170* |
| 6 | 124072425 | 124234953 | 124175144 | 124184953 | 8.87E-09 | ENSG00000188580 | *NKAIN2* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 6 | 125760855 | 126018604 | 125810855 | 125820802 | 6.44E-09 | n.a. | n.a. |
| 6 | 128435397 | 128544982 | 128485397 | 128494982 | 5.86E-09 | ENSG00000152894 | *PTPRK* |
| 6 | 131672420 | 131782151 | 131722420 | 131732151 | 4.60E-09 | n.a. | n.a. |
| 6 | 132479486 | 132589184 | 132529486 | 132539184 | 1.24E-08 | n.a. | n.a. |
| 6 | 137332672 | 137442017 | 137382672 | 137392017 | 1.63E-09 | ENSG00000016402 | *IL20RA* |
| 6 | 145265416 | 145375412 | 145315416 | 145325412 | 3.22E-09 | n.a. | n.a. |
| 6 | 150643133 | 150762907 | 150703492 | 150712907 | 5.46E-09 | ENSG00000009765 | *IYD* |
| 7 | 1211748 | 1321585 | 1261748 | 1271585 | 1.07E-08 | ENSG00000164853 | *UNCX* |
| 7 | 3349469 | 3459397 | 3399469 | 3409397 | 1.32E-08 | ENSG00000146555 | *SDK1* |
| 7 | 3688578 | 3797977 | 3738578 | 3747977 | 1.34E-08 | n.a. | n.a. |
| 7 | 10067005 | 10176870 | 10117005 | 10126870 | 3.94E-09 | n.a. | n.a. |
| 7 | 18395394 | 18504697 | 18445394 | 18454697 | 4.17E-09 | ENSG00000048052 | *HDAC9* |
| 7 | 19093001 | 19202990 | 19143001 | 19152990 | 3.81E-08 | ENSG00000146618 | *FERD3L* |
| 7 | 19437502 | 19557610 | 19497666 | 19507610 | 3.73E-09 | n.a. | n.a. |
| 7 | 27995905 | 28105595 | 28045905 | 28055595 | 1.19E-09 | ENSG00000153814 | *JAZF1* |
| 7 | 30172426 | 30292686 | 30222426 | 30232347 | 2.03E-08 | n.a. | n.a. |
| 7 | 30797248 | 30906903 | 30847248 | 30856903 | 1.66E-08 | ENSG00000106121 | *C7orf67* |
| 7 | 32286745 | 32396675 | 32336745 | 32346675 | 3.58E-08 | ENSG00000154678 | *PDE1C* |
| 7 | 38567244 | 38675831 | 38617244 | 38625831 | 2.64E-08 | ENSG00000078053 | *AMPH* |
| 7 | 38887364 | 38997130 | 38937364 | 38947130 | 1.47E-08 | ENSG00000006715 | *VPS41* |
| 7 | 42336317 | 42446204 | 42386317 | 42396204 | 1.06E-08 | n.a. | n.a. |
| 7 | 46043377 | 46153174 | 46093377 | 46103174 | 9.18E-09 | n.a. | n.a. |
| 7 | 48479572 | 48589536 | 48529572 | 48539536 | 1.22E-08 | ENSG00000179869 | *ABCA13* |
| 7 | 78616613 | 78726128 | 78666613 | 78676128 | 4.04E-09 | ENSG00000187391 | *MAGI2* |
| 7 | 79049197 | 79158088 | 79099197 | 79108088 | 5.45E-10 | n.a. | n.a. |
| 7 | 80886668 | 80996120 | 80936668 | 80946120 | 1.63E-09 | n.a. | n.a. |
| 7 | 86760734 | 86870643 | 86810734 | 86820643 | 1.21E-08 | ENSG00000182165 | *TP53TG1* |
| 7 | 86760734 | 86870643 | 86810734 | 86820643 | 1.21E-08 | ENSG00000005471 | *ABCB4* |
| 7 | 93306422 | 93416374 | 93356422 | 93366374 | 1.33E-08 | ENSG00000105825 | *TFPI2* |
| 7 | 96169763 | 96279575 | 96219763 | 96229575 | 2.24E-08 | ENSG00000127922 | *SHFM1* |
| 7 | 107961491 | 108071157 | 108011491 | 108021157 | 3.42E-08 | ENSG00000128590 | *DNAJB9* |
| 7 | 109257448 | 109367225 | 109307448 | 109317225 | 1.86E-09 | n.a. | n.a. |
| 7 | 112927739 | 113037473 | 112977739 | 112987473 | 8.09E-09 | n.a. | n.a. |
| 7 | 120256851 | 120397509 | 120306851 | 120315576 | 4.15E-09 | ENSG00000106025 | *TSPAN12* |
| 7 | 134888399 | 134998230 | 134938399 | 134948230 | 1.49E-08 | ENSG00000155561 | *NUP205* |
| 7 | 135322320 | 135432160 | 135372320 | 135382160 | 2.74E-09 | n.a. | n.a. |
| 7 | 136841043 | 137051779 | 136992020 | 137001779 | 8.31E-09 | ENSG00000157680 | *DGKI* |
| 7 | 139165373 | 139275331 | 139215373 | 139225331 | 4.21E-09 | ENSG00000059377 | *TBXAS1* |
| 7 | 145185621 | 145295288 | 145235621 | 145245288 | 3.29E-09 | n.a. | n.a. |
| 7 | 145531771 | 145683118 | 145623501 | 145633118 | 3.63E-09 | ENSG00000174469 | *CNTNAP2* |
| 7 | 145978272 | 146088244 | 146028272 | 146038244 | 1.51E-08 | n.a. | n.a. |
| 7 | 152157001 | 152378499 | 152207001 | 152216725 | 5.74E-09 | ENSG00000133627 | *ACTR3B* |
| 7 | 155550844 | 155660839 | 155600844 | 155610839 | 8.66E-09 | n.a. | n.a. |
| 8 | 208415 | 318356 | 258415 | 268356 | 3.69E-08 | ENSG00000182366 | *FAM87A* |
| 8 | 1705411 | 1815223 | 1755411 | 1765223 | 2.23E-08 | ENSG00000104728 | *ARHGEF10* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 8 | 2012608 | 2122585 | 2062608 | 2072585 | 2.86E-09 | ENSG00000036448 | *MYOM2* |
| 8 | 2805335 | 2914910 | 2855335 | 2864910 | 2.20E-08 | ENSG00000183117 | *CSMD1* |
| 8 | 5089077 | 5199074 | 5139077 | 5149074 | 2.12E-09 | n.a. | n.a. |
| 8 | 9715838 | 9824929 | 9765838 | 9774929 | 1.48E-08 | n.a. | n.a. |
| 8 | 15971622 | 16465601 | 16021622 | 16031379 | 3.61E-09 | ENSG00000038945 | *MSR1* |
| 8 | 16558329 | 16668184 | 16608329 | 16618184 | 3.95E-08 | n.a. | n.a. |
| 8 | 17745658 | 17855653 | 17795658 | 17805653 | 1.08E-08 | ENSG00000104760 | *FGL1* |
| 8 | 18537864 | 18657976 | 18587864 | 18597648 | 4.24E-09 | ENSG00000156011 | *PSD3* |
| 8 | 19774737 | 19882676 | 19824737 | 19832676 | 2.68E-08 | ENSG00000175445 | *LPL* |
| 8 | 21035976 | 21145846 | 21085976 | 21095846 | 4.78E-09 | n.a. | n.a. |
| 8 | 23872153 | 24053220 | 23993884 | 24003220 | 1.19E-09 | n.a. | n.a. |
| 8 | 30040446 | 30254387 | 30195378 | 30204387 | 7.39E-09 | ENSG00000104671 | *DCTN6* |
| 8 | 34429457 | 34538406 | 34479457 | 34488406 | 6.82E-09 | n.a. | n.a. |
| 8 | 35652255 | 35869146 | 35723988 | 35733305 | 8.39E-10 | ENSG00000156687 | *UNC5D* |
| 8 | 39384731 | 39515235 | 39455323 | 39465235 | 1.43E-09 | ENSG00000197475 | n.a. |
| 8 | 43282130 | 43401494 | 43332130 | 43342076 | 1.55E-09 | ENSG00000188877 | *POTEA* |
| 8 | 51648670 | 51758579 | 51698670 | 51708579 | 2.51E-09 | ENSG00000147481 | *SNTG1* |
| 8 | 53198222 | 53308206 | 53248222 | 53258206 | 3.95E-08 | ENSG00000147488 | *ST18* |
| 8 | 56092434 | 56202352 | 56142434 | 56152352 | 2.19E-08 | ENSG00000206579 | *XKR4* |
| 8 | 58405055 | 58514501 | 58455055 | 58464501 | 1.15E-08 | n.a. | n.a. |
| 8 | 75090808 | 75200595 | 75140808 | 75150595 | 1.98E-09 | ENSG00000154589 | *LY96* |
| 8 | 82752068 | 82861800 | 82802068 | 82811800 | 1.98E-08 | ENSG00000164695 | *CHMP4C* |
| 8 | 90213031 | 90322251 | 90263031 | 90272251 | 6.97E-10 | n.a. | n.a. |
| 8 | 101654683 | 101764371 | 101704683 | 101714371 | 1.33E-08 | ENSG00000174226 | *SNX31* |
| 8 | 102425851 | 102535707 | 102475851 | 102485707 | 2.71E-09 | n.a. | n.a. |
| 8 | 102665235 | 102774334 | 102715235 | 102724334 | 7.18E-09 | ENSG00000083307 | *GRHL2* |
| 8 | 108836877 | 108946357 | 108886877 | 108896357 | 3.38E-08 | n.a. | n.a. |
| 8 | 123762254 | 123871894 | 123812254 | 123821894 | 1.41E-08 | ENSG00000178764 | *ZHX2* |
| 8 | 124847797 | 124957595 | 124897797 | 124907595 | 4.01E-09 | ENSG00000176853 | *FAM91A1* |
| 8 | 125030117 | 125139840 | 125080117 | 125089840 | 1.20E-08 | ENSG00000214814 | *FER1L6* |
| 8 | 125030117 | 125139840 | 125080117 | 125089840 | 1.20E-08 | ENSG00000181171 | *C8orf54* |
| 9 | 828249 | 937949 | 878249 | 887949 | 2.84E-08 | ENSG00000137090 | *DMRT1* |
| 9 | 9068470 | 9178227 | 9118470 | 9128227 | 8.33E-10 | ENSG00000212829 | *RPS26P3* |
| 9 | 12422510 | 12532330 | 12472510 | 12482330 | 2.69E-08 | n.a. | n.a. |
| 9 | 15034374 | 15144031 | 15084374 | 15094031 | 3.09E-08 | n.a. | n.a. |
| 9 | 16060688 | 16170630 | 16110688 | 16120630 | 3.10E-08 | n.a. | n.a. |
| 9 | 16493888 | 16654533 | 16596474 | 16604533 | 6.02E-10 | ENSG00000173068 | *BNC2* |
| 9 | 16834638 | 16944635 | 16884638 | 16894635 | 2.22E-09 | n.a. | n.a. |
| 9 | 23423526 | 23533213 | 23473526 | 23483213 | 1.16E-09 | n.a. | n.a. |
| 9 | 25141263 | 25251193 | 25191263 | 25201193 | 1.85E-09 | n.a. | n.a. |
| 9 | 26520067 | 26629986 | 26570067 | 26579986 | 8.75E-09 | n.a. | n.a. |
| 9 | 29746957 | 29856465 | 29796957 | 29806465 | 8.91E-09 | n.a. | n.a. |
| 9 | 33156428 | 33264833 | 33206428 | 33214833 | 1.40E-08 | ENSG00000122711 | *SPINK4* |
| 9 | 75515677 | 75625498 | 75565677 | 75575498 | 2.15E-08 | n.a. | n.a. |
| 9 | 78889455 | 78999267 | 78939455 | 78949267 | 2.49E-08 | ENSG00000197969 | *VPS13A* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 9 | 86242046 | 86351675 | 86292046 | 86301675 | 9.67E-10 | n.a. | n.a. |
| 9 | 92846673 | 92956032 | 92896673 | 92906032 | 1.73E-08 | n.a. | n.a. |
| 9 | 101043505 | 101152971 | 101093505 | 101102971 | 2.79E-08 | n.a. | n.a. |
| 9 | 106948244 | 107058151 | 106998244 | 107008151 | 3.24E-08 | ENSG00000070214 | *SLC44A1* |
| 9 | 114690945 | 114810936 | 114740945 | 114750746 | 2.42E-09 | ENSG00000119457 | *SLC46A2* |
| 9 | 124528802 | 124638591 | 124578802 | 124588591 | 2.33E-08 | ENSG00000148215 | *OR5C1* |
| 9 | 125932298 | 126041700 | 125982298 | 125991700 | 6.06E-09 | n.a. | n.a. |
| 10 | 457451 | 567339 | 507451 | 517339 | 2.24E-08 | ENSG00000151240 | *DIP2C* |
| 10 | 2964799 | 3074760 | 3014799 | 3024760 | 8.02E-09 | n.a. | n.a. |
| 10 | 9500977 | 9610968 | 9550977 | 9560968 | 4.16E-10 | n.a. | n.a. |
| 10 | 10338229 | 10448024 | 10388229 | 10398024 | 3.87E-09 | n.a. | n.a. |
| 10 | 28448769 | 28558609 | 28498769 | 28508609 | 2.56E-08 | ENSG00000150054 | *MPP7* |
| 10 | 52636659 | 52746506 | 52686659 | 52696506 | 1.91E-08 | ENSG00000185532 | *PRKG1* |
| 10 | 55947070 | 56056778 | 55997070 | 56006778 | 2.35E-08 | ENSG00000150275 | *PCDH15* |
| 10 | 57976303 | 58086266 | 58026303 | 58036266 | 1.81E-09 | n.a. | n.a. |
| 10 | 58447738 | 58556956 | 58497738 | 58506956 | 1.41E-08 | n.a. | n.a. |
| 10 | 59350285 | 59460215 | 59400285 | 59410215 | 3.51E-08 | n.a. | n.a. |
| 10 | 67839770 | 67960010 | 67900085 | 67910010 | 8.10E-10 | ENSG00000183230 | *CTNNA3* |
| 10 | 68095045 | 68205021 | 68145045 | 68155021 | 2.66E-08 | n.a. | n.a. |
| 10 | 72364993 | 72474888 | 72414993 | 72424888 | 5.61E-09 | n.a. | n.a. |
| 10 | 77030367 | 77139863 | 77080367 | 77089863 | 3.91E-08 | ENSG00000148655 | *C10orf11* |
| 10 | 80845369 | 80955098 | 80895369 | 80905098 | 2.08E-08 | ENSG00000165424 | *ZCCHC24* |
| 10 | 91042186 | 91152142 | 91092186 | 91102142 | 1.73E-08 | ENSG00000107798 | *LIPA* |
| 10 | 91042186 | 91152142 | 91092186 | 91102142 | 1.73E-08 | ENSG00000119917 | *IFIT3* |
| 10 | 107959593 | 108069331 | 108009593 | 108019331 | 3.67E-09 | n.a. | n.a. |
| 10 | 118102470 | 118212240 | 118152470 | 118162240 | 3.25E-08 | ENSG00000203837 | *PNLIPRP3* |
| 10 | 119927194 | 120036247 | 119977194 | 119986247 | 4.24E-09 | ENSG00000165669 | *C10orf84* |
| 10 | 121821212 | 121940070 | 121881671 | 121890070 | 1.49E-08 | n.a. | n.a. |
| 10 | 122491162 | 122601064 | 122541162 | 122551064 | 2.24E-08 | ENSG00000120008 | *BRWD2* |
| 10 | 122845219 | 122965152 | 122905267 | 122915152 | 2.88E-10 | n.a. | n.a. |
| 10 | 129358894 | 129478427 | 129418898 | 129428427 | 5.64E-09 | ENSG00000186766 | *FOXI2* |
| 10 | 130720206 | 130830092 | 130770206 | 130780092 | 2.66E-09 | n.a. | n.a. |
| 10 | 132063026 | 132193172 | 132113026 | 132122982 | 1.98E-09 | n.a. | n.a. |
| 11 | 19599663 | 19709441 | 19649663 | 19659441 | 1.44E-08 | ENSG00000166833 | *NAV2* |
| 11 | 20333900 | 20443811 | 20383900 | 20393811 | 1.32E-08 | ENSG00000185238 | *PRMT3* |
| 11 | 21498292 | 21608073 | 21548292 | 21558073 | 1.71E-08 | ENSG00000165973 | *NELL1* |
| 11 | 38081876 | 38397628 | 38131876 | 38141394 | 7.65E-11 | n.a. | n.a. |
| 11 | 42656749 | 42766347 | 42706749 | 42716347 | 2.98E-08 | n.a. | n.a. |
| 11 | 43967603 | 44077272 | 44017603 | 44027272 | 1.46E-08 | ENSG00000205126 | *ACCSL* |
| 11 | 58330409 | 58440287 | 58380409 | 58390287 | 2.16E-10 | ENSG00000156689 | *GLYATL2* |
| 11 | 59269972 | 59379867 | 59319972 | 59329867 | 3.84E-08 | ENSG00000166900 | *STX3* |
| 11 | 59269972 | 59379867 | 59319972 | 59329867 | 3.84E-08 | ENSG00000166902 | *MRPL16* |
| 11 | 60117880 | 60237203 | 60167880 | 60177422 | 4.66E-09 | ENSG00000181995 | *C11orf64* |
| 11 | 62313644 | 62423505 | 62363644 | 62373505 | 3.53E-08 | ENSG00000133316 | *WDR74* |
| 11 | 79189185 | 79298636 | 79239185 | 79248636 | 1.36E-08 | n.a. | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 11 | 87039769 | 87149407 | 87089769 | 87099407 | 2.92E-08 | n.a. | n.a. |
| 11 | 94189954 | 94340418 | 94281500 | 94290418 | 8.28E-10 | ENSG00000166025 | *AMOTL1* |
| 11 | 94189954 | 94340418 | 94281500 | 94290418 | 8.28E-10 | ENSG00000150316 | *CWC15* |
| 11 | 98546472 | 98656077 | 98596472 | 98606077 | 3.93E-08 | n.a. | n.a. |
| 11 | 104211100 | 104320241 | 104261100 | 104270241 | 5.27E-09 | ENSG00000204403 | *CASP12* |
| 11 | 108856049 | 108965664 | 108906049 | 108915664 | 1.72E-08 | n.a. | n.a. |
| 11 | 115314545 | 115424243 | 115364545 | 115374243 | 2.90E-09 | n.a. | n.a. |
| 11 | 117032884 | 117142563 | 117082884 | 117092563 | 8.07E-10 | ENSG00000177103 | *DSCAML1* |
| 11 | 133741819 | 133851594 | 133791819 | 133801594 | 7.41E-09 | ENSG00000149328 | *GLB1L2* |
| 11 | 133741819 | 133851594 | 133791819 | 133801594 | 7.41E-09 | ENSG00000109956 | *B3GAT1* |
| 12 | 8514544 | 8749273 | 8689395 | 8699273 | 1.00E-08 | ENSG00000197614 | *MFAP5* |
| 12 | 10216569 | 10357713 | 10297834 | 10307713 | 7.14E-09 | ENSG00000139112 | *GABARAPL1* |
| 12 | 10940252 | 11050194 | 10990252 | 11000194 | 1.98E-10 | ENSG00000212127 | *TAS2R14* |
| 12 | 15367046 | 15551364 | 15491436 | 15501364 | 7.99E-09 | ENSG00000151490 | *PTPRO* |
| 12 | 15642821 | 15752509 | 15692821 | 15702509 | 1.65E-08 | ENSG00000151491 | *EPS8* |
| 12 | 27468053 | 27578003 | 27518053 | 27528003 | 2.82E-08 | ENSG00000165935 | *C12orf70* |
| 12 | 27829119 | 27938883 | 27879119 | 27888883 | 9.22E-09 | ENSG00000087448 | *KLHDC5* |
| 12 | 37373017 | 37482558 | 37423017 | 37432558 | 3.81E-08 | ENSG00000139117 | *CPNE8* |
| 12 | 39042747 | 39152452 | 39092747 | 39102452 | 8.62E-09 | ENSG00000188906 | *LRRK2* |
| 12 | 42947052 | 43056574 | 42997052 | 43006574 | 5.99E-09 | ENSG00000139173 | *TMEM117* |
| 12 | 43903077 | 44012159 | 43953077 | 43962159 | 1.31E-08 | ENSG00000177119 | *ANO6* |
| 12 | 47976799 | 48086591 | 48026799 | 48036591 | 1.28E-08 | ENSG00000178401 | *DNAJC22* |
| 12 | 51471942 | 51591311 | 51532039 | 51541311 | 1.19E-10 | ENSG00000170423 | *KRT78* |
| 12 | 53447620 | 53557566 | 53497620 | 53507566 | 9.93E-09 | ENSG00000172551 | *MUCL1* |
| 12 | 53672209 | 53782077 | 53722209 | 53732077 | 1.00E-08 | ENSG00000123307 | *NEUROD4* |
| 12 | 57487096 | 57596550 | 57537096 | 57546550 | 5.07E-09 | ENSG00000139263 | *LRIG3* |
| 12 | 82082786 | 82192689 | 82132786 | 82142689 | 1.25E-08 | n.a. | n.a. |
| 12 | 90069420 | 90188955 | 90119420 | 90129118 | 3.52E-09 | ENSG00000011465 | *DCN* |
| 12 | 92512995 | 92622599 | 92562995 | 92572599 | 1.65E-08 | ENSG00000220515 | n.a. |
| 12 | 104744718 | 104854597 | 104794718 | 104804597 | 3.23E-08 | n.a. | n.a. |
| 12 | 111651772 | 111761376 | 111701772 | 111711376 | 5.77E-09 | ENSG00000089169 | *RPH3A* |
| 12 | 124796117 | 124905655 | 124846117 | 124855655 | 9.75E-09 | n.a. | n.a. |
| 12 | 125697760 | 125807745 | 125747760 | 125757745 | 1.76E-09 | ENSG00000189238 | n.a. |
| 12 | 127336467 | 127446321 | 127386467 | 127396321 | 6.75E-09 | n.a. | n.a. |
| 12 | 129970602 | 130080381 | 130020602 | 130030381 | 6.14E-10 | ENSG00000111452 | *GPR133* |
| 13 | 18826967 | 18936302 | 18876967 | 18886302 | 1.23E-08 | ENSG00000132958 | *TPTE2* |
| 13 | 21811159 | 21920958 | 21861159 | 21870958 | 6.31E-09 | n.a. | n.a. |
| 13 | 33045825 | 33155700 | 33095825 | 33105700 | 9.43E-09 | n.a. | n.a. |
| 13 | 37735362 | 37897881 | 37817996 | 37825888 | 5.96E-09 | ENSG00000120686 | *UFM1* |
| 13 | 38709451 | 38819343 | 38759451 | 38769343 | 1.47E-08 | ENSG00000183722 | *LHFP* |
| 13 | 41080309 | 41190243 | 41130309 | 41140243 | 1.32E-08 | ENSG00000102763 | *KIAA0564* |
| 13 | 59019597 | 59129025 | 59069597 | 59079025 | 4.14E-10 | n.a. | n.a. |
| 13 | 67205241 | 67315076 | 67255241 | 67265076 | 1.65E-09 | n.a. | n.a. |
| 13 | 69250850 | 69381120 | 69310997 | 69320981 | 6.02E-10 | ENSG00000150361 | *KLHL1* |
| 13 | 69935300 | 70045211 | 69985300 | 69995211 | 2.64E-08 | n.a. | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 13 | 78100336 | 78210043 | 78150336 | 78160043 | 1.43E-08 | ENSG00000152193 | *RNF219* |
| 13 | 88515298 | 88696158 | 88575469 | 88585278 | 1.82E-09 | n.a. | n.a. |
| 13 | 90985406 | 91095220 | 91035406 | 91045220 | 1.96E-08 | ENSG00000179399 | *GPC5* |
| 13 | 92591140 | 92700988 | 92641140 | 92650988 | 3.90E-08 | ENSG00000183098 | *GPC6* |
| 13 | 102790532 | 102900338 | 102840532 | 102850338 | 3.11E-08 | n.a. | n.a. |
| 13 | 103029936 | 103264439 | 103204988 | 103214439 | 7.73E-10 | n.a. | n.a. |
| 13 | 104117902 | 104227897 | 104167902 | 104177897 | 1.65E-08 | n.a. | n.a. |
| 13 | 107473958 | 107583809 | 107523958 | 107533809 | 8.43E-09 | n.a. | n.a. |
| 14 | 33332299 | 33475474 | 33415622 | 33425474 | 3.46E-09 | ENSG00000129521 | *EGLN3* |
| 14 | 39502985 | 39633018 | 39563054 | 39573038 | 9.36E-10 | n.a. | n.a. |
| 14 | 44235814 | 44503667 | 44285814 | 44295703 | 1.64E-08 | n.a. | n.a. |
| 14 | 51451495 | 51561385 | 51501495 | 51511385 | 1.73E-08 | ENSG00000186469 | *GNG2* |
| 14 | 56863240 | 56973172 | 56913240 | 56923172 | 3.43E-08 | ENSG00000139977 | *NAT12* |
| 14 | 66785355 | 66895034 | 66835355 | 66845034 | 3.86E-08 | ENSG00000072415 | *MPP5* |
| 14 | 78298536 | 78408276 | 78348536 | 78358276 | 3.02E-08 | ENSG00000021645 | *NRXN3* |
| 14 | 79894340 | 80004336 | 79944340 | 79954336 | 1.58E-08 | n.a. | n.a. |
| 14 | 80497989 | 80607253 | 80547989 | 80557253 | 1.85E-08 | ENSG00000165409 | *TSHR* |
| 14 | 83493714 | 83603697 | 83543714 | 83553697 | 4.67E-09 | n.a. | n.a. |
| 14 | 89445739 | 89565316 | 89495739 | 89505678 | 1.48E-09 | ENSG00000140025 | *C14orf143* |
| 14 | 92505675 | 92615092 | 92555675 | 92565092 | 3.81E-09 | ENSG00000100605 | *ITPK1* |
| 14 | 97037280 | 97146999 | 97087280 | 97096999 | 1.67E-10 | n.a. | n.a. |
| 14 | 102737704 | 102847561 | 102787704 | 102797561 | 1.14E-08 | n.a. | n.a. |
| 14 | 105544863 | 105652736 | 105594863 | 105602736 | 1.37E-08 | n.a. | n.a. |
| 14 | 105896339 | 106004171 | 105946339 | 105954171 | 2.05E-08 | ENSG00000214398 | n.a. |
| 15 | 24522549 | 24632371 | 24572549 | 24582371 | 9.68E-09 | ENSG00000166206 | *GABRB3* |
| 15 | 25983676 | 26093570 | 26033676 | 26043570 | 1.32E-09 | ENSG00000104044 | *OCA2* |
| 15 | 31915235 | 32024854 | 31965235 | 31974854 | 9.65E-09 | ENSG00000198838 | *RYR3* |
| 15 | 34045559 | 34155535 | 34095559 | 34105535 | 4.91E-10 | n.a. | n.a. |
| 15 | 42143313 | 42263882 | 42193313 | 42203285 | 1.17E-08 | ENSG00000171877 | *FRMD5* |
| 15 | 46107395 | 46216678 | 46157395 | 46166678 | 5.27E-09 | ENSG00000188467 | *SLC24A5* |
| 15 | 50280355 | 50389827 | 50330355 | 50339827 | 1.60E-08 | ENSG00000128833 | *MYO5C* |
| 15 | 58233152 | 58342375 | 58283152 | 58292375 | 2.46E-08 | n.a. | n.a. |
| 15 | 66850000 | 66959768 | 66900000 | 66909768 | 6.04E-09 | ENSG00000140350 | *ANP32A* |
| 15 | 91574654 | 91684241 | 91624654 | 91634241 | 1.92E-09 | n.a. | n.a. |
| 15 | 98331160 | 98451106 | 98381160 | 98390857 | 8.01E-09 | ENSG00000140470 | *ADAMTS17* |
| 16 | 6192831 | 6302510 | 6242831 | 6252510 | 3.99E-08 | n.a. | n.a. |
| 16 | 7465646 | 7575546 | 7515646 | 7525546 | 1.13E-08 | ENSG00000078328 | n.a. |
| 16 | 8237634 | 8357193 | 8297823 | 8307193 | 4.72E-09 | n.a. | n.a. |
| 16 | 8431090 | 8695537 | 8635588 | 8645537 | 2.59E-09 | ENSG00000067365 | *C16orf68* |
| 16 | 46456451 | 46587768 | 46527985 | 46537768 | 1.67E-08 | n.a. | n.a. |
| 16 | 50453993 | 50563859 | 50503993 | 50513859 | 2.11E-08 | n.a. | n.a. |
| 16 | 52889717 | 52999318 | 52939717 | 52949318 | 2.56E-08 | n.a. | n.a. |
| 16 | 56644417 | 56753769 | 56694417 | 56703769 | 1.66E-08 | ENSG00000070761 | *C16orf80* |
| 16 | 76961320 | 77071263 | 77011320 | 77021263 | 3.50E-09 | ENSG00000186153 | *WWOX* |
| 16 | 81740232 | 81860227 | 81790232 | 81800178 | 5.64E-10 | ENSG00000140945 | *CDH13* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 16 | 83598485 | 83708477 | 83648485 | 83658477 | 2.73E-08 | ENSG00000153786 | *ZDHHC7* |
| 17 | 21088013 | 21198002 | 21138013 | 21148002 | 1.31E-08 | ENSG00000034152 | *MAP2K3* |
| 17 | 37028879 | 37138251 | 37078879 | 37088251 | 2.57E-09 | ENSG00000173812 | *EIF1* |
| 17 | 51318754 | 51428539 | 51368754 | 51378539 | 7.16E-10 | n.a. | n.a. |
| 17 | 56142640 | 56252518 | 56192640 | 56202518 | 6.68E-09 | ENSG00000141376 | *BCAS3* |
| 17 | 60365115 | 60474458 | 60415115 | 60424458 | 5.28E-09 | ENSG00000120063 | *GNA13* |
| 17 | 60778059 | 60907702 | 60848310 | 60857702 | 1.98E-10 | n.a. | n.a. |
| 17 | 63844587 | 63954520 | 63894587 | 63904520 | 9.55E-09 | ENSG00000141337 | *ARSG* |
| 17 | 72195918 | 72325982 | 72266413 | 72275982 | 8.53E-09 | ENSG00000182534 | *MXRA7* |
| 18 | 28618964 | 28791790 | 28732104 | 28741790 | 2.52E-10 | ENSG00000166960 | *C18orf34* |
| 18 | 36029899 | 36139593 | 36079899 | 36089593 | 1.01E-08 | n.a. | n.a. |
| 18 | 53189468 | 53299216 | 53239468 | 53249216 | 3.43E-08 | ENSG00000119547 | *ONECUT2* |
| 18 | 59529215 | 59639135 | 59579215 | 59589135 | 1.01E-08 | ENSG00000166396 | *SERPINB7* |
| 18 | 64138301 | 64248289 | 64188301 | 64198289 | 1.25E-08 | n.a. | n.a. |
| 18 | 64754914 | 64958486 | 64804914 | 64814826 | 5.60E-10 | ENSG00000150636 | *CCDC102B* |
| 18 | 65711578 | 65821486 | 65761578 | 65771486 | 1.81E-09 | ENSG00000150637 | *CD226* |
| 18 | 68603842 | 68713818 | 68653842 | 68663818 | 3.16E-08 | ENSG00000166342 | *NETO1* |
| 18 | 71520972 | 71640987 | 71581089 | 71590987 | 4.17E-09 | n.a. | n.a. |
| 19 | 11669484 | 11778917 | 11719484 | 11728917 | 2.16E-08 | ENSG00000197933 | *ZNF823* |
| 19 | 11898124 | 12007462 | 11948124 | 11957462 | 3.98E-08 | ENSG00000197054 | *ZNF763* |
| 19 | 33747932 | 33869312 | 33797932 | 33807586 | 1.88E-09 | n.a. | n.a. |
| 19 | 36767849 | 36884574 | 36827973 | 36834574 | 1.42E-09 | n.a. | n.a. |
| 19 | 39135597 | 39255304 | 39185597 | 39195195 | 1.92E-08 | ENSG00000186008 | n.a. |
| 19 | 41007576 | 41117442 | 41057576 | 41067442 | 8.14E-10 | ENSG00000167595 | *C19orf55* |
| 19 | 45223926 | 45374319 | 45314937 | 45324319 | 4.80E-09 | ENSG00000197782 | *ZNF780A* |
| 19 | 48579080 | 48688321 | 48629080 | 48638321 | 1.21E-09 | ENSG00000131126 | *TEX101* |
| 19 | 48579080 | 48688321 | 48629080 | 48638321 | 1.21E-09 | ENSG00000124466 | *LYPD3* |
| 19 | 51237331 | 51347303 | 51287331 | 51297303 | 1.49E-08 | ENSG00000204866 | *IGFL2* |
| 19 | 52471796 | 52580956 | 52521796 | 52530956 | 1.97E-08 | ENSG00000197405 | *C5AR1* |
| 19 | 52471796 | 52580956 | 52521796 | 52530956 | 1.97E-08 | ENSG00000134830 | *GPR77* |
| 19 | 56427836 | 56536923 | 56477836 | 56486923 | 3.91E-08 | n.a. | n.a. |
| 19 | 57575452 | 57716035 | 57636087 | 57646041 | 8.83E-10 | ENSG00000167555 | *ZNF534* |
| 19 | 59434571 | 59544505 | 59484571 | 59494505 | 9.68E-09 | ENSG00000204577 | *LILRA6* |
| 19 | 59736464 | 59846379 | 59786464 | 59796379 | 4.88E-09 | ENSG00000187095 | *LILRA2* |
| 20 | 951502 | 1061318 | 1001502 | 1011318 | 1.04E-08 | ENSG00000125818 | *PSMF1* |
| 20 | 6241475 | 6351403 | 6291475 | 6301403 | 5.52E-09 | n.a. | n.a. |
| 20 | 19170791 | 19280743 | 19220791 | 19230743 | 1.89E-08 | ENSG00000185052 | *SLC24A3* |
| 20 | 20622014 | 20731953 | 20672014 | 20681953 | 3.55E-08 | ENSG00000188559 | *C20orf74* |
| 20 | 24173207 | 24293349 | 24223207 | 24233177 | 4.23E-10 | n.a. | n.a. |
| 20 | 41716665 | 41826636 | 41766665 | 41776636 | 2.99E-08 | ENSG00000101057 | *MYBL2* |
| 20 | 43279190 | 43399244 | 43329190 | 43338818 | 1.23E-08 | ENSG00000124107 | *SLPI* |
| 20 | 44708708 | 44818465 | 44758708 | 44768465 | 3.91E-09 | ENSG00000172315 | *TP53RK* |
| 20 | 44708708 | 44818465 | 44758708 | 44768465 | 3.91E-09 | ENSG00000197496 | *SLC2A10* |
| 20 | 49288547 | 49398425 | 49338547 | 49348425 | 2.89E-08 | n.a. | n.a. |
| 20 | 52190430 | 52422425 | 52362487 | 52372425 | 1.90E-08 | n.a. | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 20 | 57837075 | 57947032 | 57887075 | 57897032 | 2.15E-08 | ENSG00000196074 | *SYCP2* |
| 20 | 58623432 | 58743393 | 58683447 | 58693393 | 1.24E-09 | n.a. | n.a. |
| 21 | 21295000 | 21404989 | 21345000 | 21354989 | 6.10E-10 | ENSG00000154654 | *NCAM2* |
| 21 | 22932648 | 23042420 | 22982648 | 22992420 | 1.98E-09 | n.a. | n.a. |
| 21 | 27938947 | 28048735 | 27988947 | 27998735 | 1.36E-09 | n.a. | n.a. |
| 21 | 29727785 | 29837153 | 29777785 | 29787153 | 1.23E-09 | ENSG00000171189 | *GRIK1* |
| 21 | 30380160 | 30489836 | 30430160 | 30439836 | 3.09E-10 | ENSG00000156282 | *CLDN17* |
| 21 | 35817141 | 35926909 | 35867141 | 35876909 | 2.92E-09 | n.a. | n.a. |
| 21 | 37893505 | 38003502 | 37943505 | 37953502 | 2.25E-09 | ENSG00000157542 | *KCNJ6* |
| 22 | 24480901 | 24590850 | 24530901 | 24540850 | 9.59E-09 | ENSG00000133454 | *MYO18B* |
| 22 | 32730938 | 32840882 | 32780938 | 32790882 | 2.61E-08 | n.a. | n.a. |
| CHB+JPT | | | | | | | |
| 1 | 37025194 | 37133411 | 37075194 | 37083411 | 1.44E-10 | ENSG00000163873 | *GRIK3* |
| 1 | 63942953 | 64052760 | 63992953 | 64002760 | 8.65E-09 | ENSG00000185483 | *ROR1* |
| 1 | 64184683 | 64293968 | 64234683 | 64243968 | 3.12E-09 | n.a. | n.a. |
| 1 | 75416996 | 75526492 | 75466996 | 75476492 | 1.35E-09 | ENSG00000137968 | *SLC44A5* |
| 1 | 103469185 | 103578851 | 103519185 | 103528851 | 8.47E-09 | n.a. | n.a. |
| 1 | 109758910 | 109868628 | 109808910 | 109818628 | 3.05E-10 | ENSG00000143028 | *SYPL2* |
| 1 | 179416676 | 179526507 | 179466676 | 179476507 | 2.53E-10 | ENSG00000179452 | n.a. |
| 1 | 181805606 | 181915529 | 181855606 | 181865529 | 3.83E-08 | ENSG00000162704 | *ARPC5* |
| 1 | 181805606 | 181915529 | 181855606 | 181865529 | 3.83E-08 | ENSG00000143344 | *RGL1* |
| 1 | 194034592 | 194144585 | 194084592 | 194094585 | 6.22E-09 | n.a. | n.a. |
| 1 | 206138351 | 206248286 | 206188351 | 206198286 | 2.39E-10 | ENSG00000174059 | *CD34* |
| 1 | 245170726 | 245280081 | 245220726 | 245230081 | 7.94E-09 | ENSG00000197472 | *ZNF695* |
| 2 | 4161943 | 4271354 | 4211943 | 4221354 | 9.47E-09 | n.a. | n.a. |
| 2 | 7544018 | 7653585 | 7594018 | 7603585 | 6.78E-09 | n.a. | n.a. |
| 2 | 15042617 | 15152501 | 15092617 | 15102501 | 3.62E-09 | n.a. | n.a. |
| 2 | 24383855 | 24503789 | 24433855 | 24443679 | 1.57E-09 | ENSG00000198399 | *ITSN2* |
| 2 | 37629764 | 37739660 | 37679764 | 37689660 | 9.21E-09 | ENSG00000163171 | *CDC42EP3* |
| 2 | 81642954 | 81752433 | 81692954 | 81702433 | 9.19E-09 | n.a. | n.a. |
| 2 | 86420650 | 86530548 | 86470650 | 86480548 | 2.24E-08 | ENSG00000115548 | *JMJD1A* |
| 2 | 107132184 | 107252362 | 107182184 | 107191113 | 7.87E-10 | n.a. | n.a. |
| 2 | 151224799 | 151334526 | 151274799 | 151284526 | 4.69E-09 | n.a. | n.a. |
| 2 | 177424441 | 177533521 | 177474441 | 177483521 | 4.74E-09 | n.a. | n.a. |
| 2 | 188528149 | 188638106 | 188578149 | 188588106 | 1.73E-08 | n.a. | n.a. |
| 2 | 194687010 | 194817154 | 194737010 | 194746818 | 4.07E-11 | n.a. | n.a. |
| 2 | 202052305 | 202162058 | 202102305 | 202112058 | 3.38E-08 | ENSG00000155754 | *ALS2CR11* |
| 2 | 215921581 | 216031383 | 215971581 | 215981383 | 1.74E-08 | ENSG00000115414 | *FN1* |
| 2 | 223625575 | 223755822 | 223695838 | 223705822 | 4.62E-10 | n.a. | n.a. |
| 3 | 17757230 | 17867059 | 17807230 | 17817059 | 2.06E-08 | ENSG00000131374 | *TBC1D5* |
| 3 | 18971639 | 19080822 | 19021639 | 19030822 | 5.05E-09 | n.a. | n.a. |
| 3 | 97781565 | 97890978 | 97831565 | 97840978 | 1.94E-08 | n.a. | n.a. |
| 3 | 99331808 | 99441778 | 99381808 | 99391778 | 1.07E-08 | ENSG00000198068 | *OR5H15* |
| 3 | 109057897 | 109167716 | 109107897 | 109117716 | 2.94E-09 | n.a. | n.a. |
| 3 | 110294094 | 110403858 | 110344094 | 110353858 | 3.76E-08 | ENSG00000114487 | *MORC1* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 3 | 136796133 | 136906087 | 136846133 | 136856087 | 1.69E-08 | n.a. | n.a. |
| 3 | 146472561 | 146582522 | 146522561 | 146532522 | 2.04E-10 | n.a. | n.a. |
| 3 | 155702265 | 155812163 | 155752265 | 155762163 | 5.66E-09 | n.a. | n.a. |
| 3 | 160810006 | 160919764 | 160860006 | 160869764 | 5.02E-09 | ENSG00000151967 | *SCHIP1* |
| 3 | 164187145 | 164297136 | 164237145 | 164247136 | 8.85E-09 | n.a. | n.a. |
| 4 | 6288142 | 6408528 | 6338142 | 6348095 | 1.32E-09 | ENSG00000109501 | *WFS1* |
| 4 | 17250145 | 17360122 | 17300145 | 17310122 | 4.34E-09 | ENSG00000047662 | *FAM184B* |
| 4 | 19648308 | 19758257 | 19698308 | 19708257 | 1.02E-08 | n.a. | n.a. |
| 4 | 32612876 | 32722011 | 32662876 | 32672011 | 2.20E-08 | n.a. | n.a. |
| 4 | 41648592 | 41835140 | 41719050 | 41728350 | 5.46E-10 | ENSG00000014824 | *SLC30A9* |
| 4 | 70013085 | 70122666 | 70063085 | 70072666 | 8.97E-09 | ENSG00000213759 | *UGT2B11* |
| 4 | 86616416 | 86725866 | 86666416 | 86675866 | 8.25E-09 | ENSG00000138639 | *ARHGAP24* |
| 4 | 135610281 | 135720147 | 135660281 | 135670147 | 3.01E-10 | n.a. | n.a. |
| 4 | 159516223 | 159637555 | 159577788 | 159587555 | 1.27E-10 | n.a. | n.a. |
| 4 | 167024921 | 167134656 | 167074921 | 167084656 | 9.55E-09 | ENSG00000038295 | *TLL1* |
| 4 | 170916232 | 171026110 | 170966232 | 170976110 | 3.74E-08 | n.a. | n.a. |
| 4 | 176524029 | 176633194 | 176574029 | 176583194 | 1.15E-08 | n.a. | n.a. |
| 4 | 178028544 | 178158603 | 178078544 | 178088424 | 1.57E-09 | n.a. | n.a. |
| 4 | 178841769 | 178951541 | 178891769 | 178901541 | 2.14E-08 | n.a. | n.a. |
| 5 | 8941824 | 9051609 | 8991824 | 9001609 | 1.67E-08 | n.a. | n.a. |
| 5 | 97214622 | 97324613 | 97264622 | 97274613 | 7.33E-09 | n.a. | n.a. |
| 5 | 117563179 | 117693097 | 117623357 | 117632864 | 4.50E-09 | n.a. | n.a. |
| 5 | 124457396 | 124567289 | 124507396 | 124517289 | 1.22E-08 | n.a. | n.a. |
| 5 | 127902655 | 128012328 | 127952655 | 127962328 | 2.80E-08 | n.a. | n.a. |
| 5 | 137003365 | 137113016 | 137053365 | 137063016 | 8.45E-09 | ENSG00000146021 | *KLHL3* |
| 5 | 141304610 | 141413293 | 141354610 | 141363293 | 2.17E-09 | ENSG00000113552 | *GNPDA1* |
| 6 | 25155429 | 25265396 | 25205429 | 25215396 | 1.27E-08 | ENSG00000168405 | *CMAH* |
| 6 | 63776582 | 63886531 | 63826582 | 63836531 | 1.15E-08 | n.a. | n.a. |
| 6 | 67166270 | 67276239 | 67216270 | 67226239 | 4.07E-11 | n.a. | n.a. |
| 6 | 112815920 | 112925256 | 112865920 | 112875256 | 3.33E-08 | n.a. | n.a. |
| 6 | 129338050 | 129447104 | 129388050 | 129397104 | 2.23E-08 | ENSG00000196569 | *LAMA2* |
| 7 | 3992147 | 4101869 | 4042147 | 4051869 | 1.63E-09 | ENSG00000146555 | *SDK1* |
| 7 | 19431848 | 19541741 | 19481848 | 19491741 | 2.70E-08 | n.a. | n.a. |
| 7 | 49162814 | 49271468 | 49212814 | 49221468 | 1.34E-08 | n.a. | n.a. |
| 7 | 54604266 | 54714221 | 54654266 | 54664221 | 7.81E-09 | ENSG00000170419 | *VSTM2A* |
| 7 | 101468685 | 101578345 | 101518685 | 101528345 | 2.17E-08 | ENSG00000160967 | *CUX1* |
| 7 | 110928160 | 111037066 | 110978160 | 110987066 | 3.93E-11 | ENSG00000184903 | *IMMP2L* |
| 7 | 119178681 | 119288390 | 119228681 | 119238390 | 3.92E-08 | n.a. | n.a. |
| 7 | 131201425 | 131311164 | 131251425 | 131261164 | 3.69E-10 | n.a. | n.a. |
| 7 | 155367424 | 155477353 | 155417424 | 155427353 | 1.63E-08 | ENSG00000204876 | n.a. |
| 8 | 10846529 | 10956442 | 10896529 | 10906442 | 3.98E-08 | ENSG00000171044 | *XKR6* |
| 8 | 11780143 | 11890010 | 11830143 | 11840010 | 2.21E-10 | ENSG00000205882 | *DEFB134* |
| 8 | 50218674 | 50338697 | 50278737 | 50288697 | 1.07E-08 | n.a. | n.a. |
| 8 | 56934254 | 57044165 | 56984254 | 56994165 | 2.47E-08 | ENSG00000147507 | *LYN* |
| 8 | 120373469 | 120483155 | 120423469 | 120433155 | 2.05E-09 | n.a. | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 8 | 120935575 | 121045403 | 120985575 | 120995403 | 7.94E-09 | ENSG00000155792 | *DEPDC6* |
| 9 | 10600129 | 10709873 | 10650129 | 10659873 | 1.97E-09 | n.a. | n.a. |
| 9 | 11045707 | 11165140 | 11095707 | 11105500 | 2.33E-12 | n.a. | n.a. |
| 9 | 16551692 | 16660889 | 16601692 | 16610889 | 1.24E-08 | ENSG00000173068 | *BNC2* |
| 9 | 75961384 | 76071044 | 76011384 | 76021044 | 5.75E-09 | n.a. | n.a. |
| 9 | 123769116 | 123878892 | 123819116 | 123828892 | 2.53E-09 | ENSG00000175764 | *TTLL11* |
| 9 | 131607709 | 131717544 | 131657709 | 131667544 | 1.48E-08 | ENSG00000136878 | *USP20* |
| 10 | 3902232 | 4012090 | 3952232 | 3962090 | 1.11E-11 | n.a. | n.a. |
| 10 | 55577106 | 55687093 | 55627106 | 55637093 | 1.69E-09 | ENSG00000150275 | *PCDH15* |
| 10 | 55782701 | 55892539 | 55832701 | 55842539 | 7.64E-09 | n.a. | n.a. |
| 10 | 59322139 | 59575727 | 59402926 | 59412905 | 6.56E-11 | n.a. | n.a. |
| 10 | 82071433 | 82253516 | 82121433 | 82129901 | 1.34E-09 | ENSG00000133665 | *DYDC2* |
| 10 | 87117583 | 87226925 | 87167583 | 87176925 | 1.84E-08 | n.a. | n.a. |
| 10 | 127061186 | 127170208 | 127111186 | 127120208 | 3.95E-08 | n.a. | n.a. |
| 11 | 4718977 | 4828330 | 4768977 | 4778330 | 1.32E-08 | ENSG00000167346 | *MMP26* |
| 11 | 23620797 | 23730561 | 23670797 | 23680561 | 5.83E-09 | n.a. | n.a. |
| 11 | 25087541 | 25208180 | 25137541 | 25146942 | 7.38E-09 | n.a. | n.a. |
| 11 | 25971643 | 26090783 | 26021643 | 26031382 | 5.98E-10 | n.a. | n.a. |
| 11 | 37908714 | 38018331 | 37958714 | 37968331 | 8.80E-09 | n.a. | n.a. |
| 11 | 39263416 | 39372951 | 39313416 | 39322951 | 3.73E-09 | n.a. | n.a. |
| 11 | 39692796 | 39802176 | 39742796 | 39752176 | 7.02E-09 | n.a. | n.a. |
| 11 | 87138674 | 87248425 | 87188674 | 87198425 | 1.43E-08 | n.a. | n.a. |
| 11 | 96698477 | 96807835 | 96748477 | 96757835 | 5.68E-09 | n.a. | n.a. |
| 11 | 97528280 | 97638256 | 97578280 | 97588256 | 7.40E-09 | n.a. | n.a. |
| 12 | 32672358 | 32782327 | 32722358 | 32732327 | 2.36E-08 | ENSG00000087470 | *DNM1L* |
| 12 | 53231280 | 53340940 | 53281280 | 53290940 | 5.61E-10 | ENSG00000135447 | *PPP1R1A* |
| 12 | 53231280 | 53340940 | 53281280 | 53290940 | 5.61E-10 | ENSG00000135413 | *LACRT* |
| 12 | 84618003 | 84727973 | 84668003 | 84677973 | 5.94E-11 | ENSG00000198774 | *RASSF9* |
| 12 | 97169923 | 97278998 | 97219923 | 97228998 | 1.40E-08 | n.a. | n.a. |
| 12 | 129969331 | 130079205 | 130019331 | 130029205 | 3.46E-08 | ENSG00000111452 | *GPR133* |
| 13 | 45225436 | 45333565 | 45275436 | 45283565 | 7.00E-09 | ENSG00000215475 | *SIAH3* |
| 13 | 60298813 | 60408388 | 60348813 | 60358388 | 3.47E-08 | n.a. | n.a. |
| 13 | 104456061 | 104566048 | 104506061 | 104516048 | 3.36E-09 | n.a. | n.a. |
| 13 | 105018750 | 105128603 | 105068750 | 105078603 | 7.58E-09 | n.a. | n.a. |
| 15 | 40336454 | 40446426 | 40386454 | 40396426 | 2.39E-09 | ENSG00000214013 | *GANC* |
| 15 | 61892963 | 62002122 | 61942963 | 61952122 | 2.60E-08 | ENSG00000103657 | *HERC1* |
| 15 | 61892963 | 62002122 | 61942963 | 61952122 | 2.60E-08 | ENSG00000035664 | *DAPK2* |
| 15 | 86513328 | 86623194 | 86563328 | 86573194 | 1.31E-09 | ENSG00000140538 | *NTRK3* |
| 15 | 98336682 | 98446142 | 98386682 | 98396142 | 8.46E-09 | ENSG00000140470 | *ADAMTS17* |
| 16 | 5482839 | 5592836 | 5532839 | 5542836 | 1.45E-08 | n.a. | n.a. |
| 16 | 79629749 | 79739291 | 79679749 | 79689291 | 1.22E-08 | ENSG00000140905 | *GCSH* |
| 16 | 81746275 | 81854494 | 81796275 | 81804494 | 6.43E-11 | ENSG00000140945 | *CDH13* |
| 17 | 8891763 | 9001519 | 8941763 | 8951519 | 1.57E-09 | ENSG00000065320 | *NTN1* |
| 17 | 36944831 | 37054809 | 36994831 | 37004809 | 3.60E-08 | ENSG00000186847 | *KRT14* |
| 17 | 53182090 | 53291839 | 53232090 | 53241839 | 6.13E-10 | ENSG00000181610 | *MRPS23* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 17 | 56131340 | 56240825 | 56181340 | 56190825 | 9.07E-11 | ENSG00000141376 | *BCAS3* |
| 17 | 65545487 | 65655093 | 65595487 | 65605093 | 1.40E-09 | ENSG00000153822 | *KCNJ16* |
| 18 | 11005451 | 11115425 | 11055451 | 11065425 | 9.09E-10 | n.a. | n.a. |
| 18 | 53823966 | 53933503 | 53873966 | 53883503 | 5.42E-10 | ENSG00000049759 | *NEDD4L* |
| 18 | 59506677 | 59616415 | 59556677 | 59566415 | 4.61E-09 | ENSG00000166396 | *SERPINB7* |
| 18 | 61582394 | 61702170 | 61632394 | 61642103 | 7.88E-09 | ENSG00000081138 | *CDH7* |
| 18 | 63684655 | 63794519 | 63734655 | 63744519 | 8.11E-09 | n.a. | n.a. |
| 18 | 70573244 | 70683006 | 70623244 | 70633006 | 2.16E-08 | ENSG00000215421 | *ZNF407* |
| 19 | 6087442 | 6197373 | 6137442 | 6147373 | 1.45E-09 | ENSG00000130377 | *ACSBG2* |
| 19 | 36749293 | 36858184 | 36799293 | 36808184 | 3.01E-08 | n.a. | n.a. |
| 20 | 11938626 | 12048402 | 11988626 | 11998402 | 4.46E-09 | n.a. | n.a. |
| 20 | 22404140 | 22513528 | 22454140 | 22463528 | 2.92E-08 | ENSG00000125798 | *FOXA2* |
| 20 | 31120574 | 31230422 | 31170574 | 31180422 | 9.59E-09 | ENSG00000186191 | *C20orf186* |
| 21 | 20989362 | 21099246 | 21039362 | 21049246 | 3.26E-09 | n.a. | n.a. |
| 21 | 21707908 | 21816420 | 21757908 | 21766420 | 5.02E-09 | ENSG00000154654 | *NCAM2* |
| YRI | | | | | | | |
| 1 | 4870735 | 4980651 | 4920735 | 4930651 | 8.07E-09 | n.a. | n.a. |
| 1 | 13823309 | 13933241 | 13873309 | 13883241 | 4.48E-10 | ENSG00000116731 | *PRDM2* |
| 1 | 20413206 | 20522555 | 20463206 | 20472555 | 1.10E-08 | ENSG00000158816 | *VWA5B1* |
| 1 | 36934932 | 37097131 | 37037187 | 37047131 | 1.91E-09 | ENSG00000163873 | *GRIK3* |
| 1 | 42809335 | 42918788 | 42859335 | 42868788 | 2.70E-08 | ENSG00000186409 | n.a. |
| 1 | 59811568 | 59921231 | 59861568 | 59871231 | 6.47E-09 | ENSG00000172456 | *FGGY* |
| 1 | 73159742 | 73269292 | 73209742 | 73219292 | 8.16E-09 | n.a. | n.a. |
| 1 | 79829438 | 79949721 | 79889943 | 79899721 | 3.87E-09 | n.a. | n.a. |
| 1 | 86419276 | 86528838 | 86469276 | 86478838 | 3.96E-08 | n.a. | n.a. |
| 1 | 99855589 | 100046207 | 99966739 | 99976721 | 8.25E-11 | ENSG00000156869 | *FRRS1* |
| 1 | 106377980 | 106487733 | 106427980 | 106437733 | 4.08E-10 | n.a. | n.a. |
| 1 | 155242061 | 155351428 | 155292061 | 155301428 | 3.18E-09 | ENSG00000132694 | *ARHGEF11* |
| 1 | 173448741 | 173558687 | 173498741 | 173508687 | 1.74E-11 | ENSG00000116147 | *TNR* |
| 1 | 184419087 | 184529045 | 184469087 | 184479045 | 1.94E-09 | ENSG00000143341 | *HMCN1* |
| 1 | 191738540 | 191848404 | 191788540 | 191798404 | 2.65E-08 | n.a. | n.a. |
| 1 | 213667620 | 213777582 | 213717620 | 213727582 | 3.91E-08 | n.a. | n.a. |
| 1 | 221394736 | 221504627 | 221444736 | 221454627 | 1.05E-08 | ENSG00000143502 | *SUSD4* |
| 1 | 246384277 | 246494129 | 246434277 | 246444129 | 2.86E-09 | ENSG00000177233 | *OR2M3* |
| 2 | 24382278 | 24502243 | 24432278 | 24441844 | 1.94E-11 | ENSG00000198399 | *ITSN2* |
| 2 | 54476778 | 54586707 | 54526778 | 54536707 | 2.70E-08 | ENSG00000115306 | *SPTBN1* |
| 2 | 84191980 | 84311960 | 84241980 | 84251876 | 3.48E-09 | n.a. | n.a. |
| 2 | 114848255 | 114958065 | 114898255 | 114908065 | 7.01E-09 | ENSG00000175497 | *DPP10* |
| 2 | 132735607 | 132855558 | 132785607 | 132795561 | 2.37E-09 | n.a. | n.a. |
| 2 | 141894443 | 142004338 | 141944443 | 141954338 | 7.28E-09 | ENSG00000168702 | *LRP1B* |
| 2 | 182246665 | 182369058 | 182296665 | 182306170 | 1.81E-10 | ENSG00000162992 | *NEUROD1* |
| 2 | 205719400 | 205828370 | 205769400 | 205778370 | 1.14E-09 | ENSG00000116117 | *PARD3B* |
| 3 | 536594 | 646563 | 586594 | 596563 | 4.57E-09 | n.a. | n.a. |
| 3 | 2373053 | 2482381 | 2423053 | 2432381 | 2.66E-08 | ENSG00000144619 | *CNTN4* |
| 3 | 5320406 | 5430190 | 5370406 | 5380190 | 2.13E-08 | n.a. | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 3 | 23081221 | 23202040 | 23142243 | 23152040 | 1.08E-09 | n.a. | n.a. |
| 3 | 36364625 | 36494650 | 36414625 | 36424144 | 8.25E-11 | ENSG00000144681 | *STAC* |
| 3 | 41349401 | 41459183 | 41399401 | 41409183 | 2.20E-09 | ENSG00000168038 | *ULK4* |
| 3 | 43693436 | 43801778 | 43743436 | 43751778 | 2.54E-08 | ENSG00000011198 | *ABHD5* |
| 3 | 64426658 | 64536323 | 64476658 | 64486323 | 1.77E-09 | ENSG00000163638 | *ADAMTS9* |
| 3 | 75304187 | 75414076 | 75354187 | 75364076 | 8.48E-09 | n.a. | n.a. |
| 3 | 78245678 | 78355243 | 78295678 | 78305243 | 7.75E-09 | n.a. | n.a. |
| 3 | 79104378 | 79213990 | 79154378 | 79163990 | 2.04E-09 | ENSG00000169855 | *ROBO1* |
| 3 | 95290048 | 95399666 | 95340048 | 95349666 | 7.15E-09 | ENSG00000178694 | *NSUN3* |
| 3 | 95751850 | 95861845 | 95801850 | 95811845 | 1.67E-08 | n.a. | n.a. |
| 3 | 99247597 | 99357585 | 99297597 | 99307585 | 8.66E-09 | ENSG00000196578 | *OR5AC2* |
| 3 | 104242065 | 104352042 | 104292065 | 104302042 | 1.93E-10 | n.a. | n.a. |
| 3 | 105261477 | 105371078 | 105311477 | 105321078 | 2.11E-08 | ENSG00000214405 | n.a. |
| 3 | 108157665 | 108267642 | 108207665 | 108217642 | 2.86E-10 | n.a. | n.a. |
| 3 | 141867812 | 141977715 | 141917812 | 141927715 | 1.92E-08 | ENSG00000155890 | *TRIM42* |
| 3 | 146485421 | 146594742 | 146535421 | 146544742 | 1.81E-10 | n.a. | n.a. |
| 3 | 149306627 | 149416518 | 149356627 | 149366518 | 2.48E-08 | n.a. | n.a. |
| 3 | 167698224 | 167807954 | 167748224 | 167757954 | 2.84E-08 | n.a. | n.a. |
| 3 | 179975850 | 180085645 | 180025850 | 180035645 | 4.38E-09 | ENSG00000197584 | *KCNMB2* |
| 3 | 184434340 | 184544050 | 184484340 | 184494050 | 1.91E-10 | ENSG00000053524 | *MCF2L2* |
| 3 | 189887078 | 189996944 | 189937078 | 189946944 | 2.79E-08 | ENSG00000145012 | *LPP* |
| 3 | 190943750 | 191053505 | 190993750 | 191003505 | 1.54E-08 | ENSG00000073282 | *TP63* |
| 3 | 193191467 | 193301396 | 193241467 | 193251396 | 1.40E-08 | n.a. | n.a. |
| 3 | 197331035 | 197441026 | 197381035 | 197391026 | 1.81E-08 | ENSG00000163958 | *ZDHHC19* |
| 4 | 983248 | 1093172 | 1033248 | 1043172 | 1.97E-08 | ENSG00000178222 | *RNF212* |
| 4 | 3500900 | 3621054 | 3550900 | 3560375 | 1.03E-08 | ENSG00000163956 | *LRPAP1* |
| 4 | 8971648 | 9081406 | 9021648 | 9031406 | 2.09E-08 | ENSG00000186146 | *DEFB131* |
| 4 | 11616229 | 11726216 | 11666229 | 11676216 | 3.61E-08 | n.a. | n.a. |
| 4 | 14278121 | 14387775 | 14328121 | 14337775 | 1.98E-08 | n.a. | n.a. |
| 4 | 17782634 | 17912613 | 17832634 | 17842088 | 8.48E-09 | n.a. | n.a. |
| 4 | 21263121 | 21372734 | 21313121 | 21322734 | 3.79E-08 | ENSG00000185774 | *KCNIP4* |
| 4 | 35748496 | 35858336 | 35798496 | 35808336 | 2.26E-10 | ENSG00000047365 | *ARAP2* |
| 4 | 42709663 | 42819646 | 42759663 | 42769646 | 3.38E-09 | ENSG00000215203 | *GRXCR1* |
| 4 | 43806692 | 43916412 | 43856692 | 43866412 | 4.00E-09 | ENSG00000183783 | *KCTD8* |
| 4 | 55388408 | 55497816 | 55438408 | 55447816 | 3.62E-09 | n.a. | n.a. |
| 4 | 57065485 | 57175425 | 57115485 | 57125425 | 3.65E-08 | ENSG00000196503 | *ARL9* |
| 4 | 58080145 | 58190090 | 58130145 | 58140090 | 1.06E-09 | n.a. | n.a. |
| 4 | 63343946 | 63453869 | 63393946 | 63403869 | 1.29E-08 | n.a. | n.a. |
| 4 | 64015330 | 64125212 | 64065330 | 64075212 | 8.39E-09 | n.a. | n.a. |
| 4 | 88950767 | 89060764 | 89000767 | 89010764 | 1.54E-08 | ENSG00000152595 | *MEPE* |
| 4 | 88950767 | 89060764 | 89000767 | 89010764 | 1.54E-08 | ENSG00000183199 | *HSP90AB3P* |
| 4 | 93811847 | 93921502 | 93861847 | 93871502 | 1.91E-09 | ENSG00000152208 | *GRID2* |
| 4 | 97557869 | 97790761 | 97730837 | 97740761 | 5.04E-09 | n.a. | n.a. |
| 4 | 100210496 | 100330411 | 100270745 | 100280411 | 3.38E-11 | ENSG00000198099 | *ADH4* |
| 4 | 108083711 | 108193607 | 108133711 | 108143607 | 1.36E-08 | ENSG00000155011 | *DKK2* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 4 | 119894987 | 120075813 | 119944987 | 119954789 | 4.92E-11 | ENSG00000150961 | *SEC24D* |
| 4 | 131910075 | 132020008 | 131960075 | 131970008 | 5.68E-09 | n.a. | n.a. |
| 4 | 138091938 | 138201765 | 138141938 | 138151765 | 1.84E-08 | n.a. | n.a. |
| 4 | 143671542 | 143781515 | 143721542 | 143731515 | 2.12E-09 | n.a. | n.a. |
| 4 | 148291906 | 148475054 | 148415209 | 148425054 | 4.79E-11 | n.a. | n.a. |
| 4 | 153392228 | 153502206 | 153442228 | 153452206 | 2.83E-10 | ENSG00000109670 | *FBXW7* |
| 4 | 157265631 | 157375276 | 157315631 | 157325276 | 1.37E-09 | n.a. | n.a. |
| 4 | 159758019 | 159867896 | 159808019 | 159817896 | 7.97E-09 | ENSG00000205208 | *C4orf46* |
| 4 | 159758019 | 159867896 | 159808019 | 159817896 | 7.97E-09 | ENSG00000171503 | *ETFDH* |
| 4 | 162172027 | 162281987 | 162222027 | 162231987 | 2.79E-08 | n.a. | n.a. |
| 4 | 163615182 | 163725014 | 163665182 | 163675014 | 4.34E-09 | n.a. | n.a. |
| 4 | 176361854 | 176471559 | 176411854 | 176421559 | 3.73E-08 | n.a. | n.a. |
| 4 | 176790390 | 176900273 | 176840390 | 176850273 | 3.38E-11 | ENSG00000150625 | *GPM6A* |
| 4 | 184204236 | 184313605 | 184254236 | 184263605 | 7.42E-09 | ENSG00000151718 | *WWC2* |
| 4 | 189969174 | 190079118 | 190019174 | 190029118 | 3.34E-09 | n.a. | n.a. |
| 4 | 190658887 | 190819031 | 190739050 | 190749027 | 1.14E-09 | n.a. | n.a. |
| 5 | 4255299 | 4365223 | 4305299 | 4315223 | 1.33E-09 | n.a. | n.a. |
| 5 | 18324455 | 18434442 | 18374455 | 18384442 | 2.15E-08 | n.a. | n.a. |
| 5 | 21505816 | 21655905 | 21565840 | 21575831 | 9.29E-09 | ENSG00000198014 | n.a. |
| 5 | 23155746 | 23265626 | 23205746 | 23215626 | 2.28E-09 | n.a. | n.a. |
| 5 | 65647250 | 65780812 | 65720959 | 65730812 | 2.16E-09 | ENSG00000205619 | n.a. |
| 5 | 71514708 | 71624621 | 71564708 | 71574621 | 2.00E-08 | ENSG00000113048 | *MRPS27* |
| 5 | 83807680 | 83917381 | 83857680 | 83867381 | 1.52E-08 | n.a. | n.a. |
| 5 | 94619514 | 94729254 | 94669514 | 94679254 | 2.25E-08 | ENSG00000175471 | *MCTP1* |
| 5 | 103953314 | 104062812 | 104003314 | 104012812 | 2.55E-09 | n.a. | n.a. |
| 5 | 104967588 | 105076529 | 105017588 | 105026529 | 1.95E-08 | n.a. | n.a. |
| 5 | 117597895 | 117707254 | 117647895 | 117657254 | 1.27E-09 | n.a. | n.a. |
| 5 | 117763488 | 118019158 | 117959270 | 117969158 | 1.14E-08 | n.a. | n.a. |
| 5 | 128649555 | 128759390 | 128699555 | 128709390 | 5.72E-09 | n.a. | n.a. |
| 5 | 141305842 | 141415580 | 141355842 | 141365580 | 1.95E-08 | ENSG00000113552 | *GNPDA1* |
| 5 | 151213949 | 151323931 | 151263949 | 151273931 | 5.66E-10 | ENSG00000145888 | *GLRA1* |
| 5 | 152869545 | 152989663 | 152919545 | 152929422 | 5.98E-10 | ENSG00000155511 | *GRIA1* |
| 5 | 160127170 | 160257093 | 160187277 | 160197053 | 1.61E-10 | ENSG00000118322 | *ATP10B* |
| 5 | 163518471 | 163628135 | 163568471 | 163578135 | 1.41E-08 | n.a. | n.a. |
| 5 | 166345073 | 166454989 | 166395073 | 166404989 | 5.64E-11 | n.a. | n.a. |
| 5 | 174377023 | 174486770 | 174427023 | 174436770 | 3.47E-08 | n.a. | n.a. |
| 6 | 29434364 | 29554196 | 29484364 | 29494303 | 2.02E-09 | ENSG00000112462 | *OR12D3* |
| 6 | 30433695 | 30543670 | 30483695 | 30493670 | 9.33E-09 | n.a. | n.a. |
| 6 | 30844975 | 30954911 | 30894975 | 30904911 | 3.09E-08 | ENSG00000214894 | n.a. |
| 6 | 31345585 | 31455426 | 31395585 | 31405426 | 3.38E-09 | ENSG00000204525 | *HLA-C* |
| 6 | 32669640 | 32859076 | 32719640 | 32729569 | 2.71E-10 | ENSG00000179344 | *HLA-DQB1* |
| 6 | 38950190 | 39060170 | 39000190 | 39010170 | 2.67E-09 | ENSG00000124721 | *DNAH8* |
| 6 | 48315912 | 48580494 | 48520533 | 48530494 | 1.55E-08 | n.a. | n.a. |
| 6 | 57420721 | 57767987 | 57698266 | 57707862 | 7.48E-09 | n.a. | n.a. |
| 6 | 63776582 | 63886526 | 63826582 | 63836526 | 2.72E-09 | n.a. | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 6 | 64056119 | 64166056 | 64106119 | 64116056 | 1.25E-09 | ENSG00000146166 | *LGSN* |
| 6 | 66513711 | 66623618 | 66563711 | 66573618 | 4.55E-09 | n.a. | n.a. |
| 6 | 75710972 | 75832458 | 75772925 | 75782458 | 2.11E-08 | n.a. | n.a. |
| 6 | 75932671 | 76042343 | 75982671 | 75992343 | 3.55E-08 | ENSG00000111799 | *COL12A1* |
| 6 | 75932671 | 76042343 | 75982671 | 75992343 | 3.55E-08 | ENSG00000112695 | *COX7A2* |
| 6 | 77147061 | 77257059 | 77197061 | 77207059 | 3.09E-08 | n.a. | n.a. |
| 6 | 82353052 | 82462699 | 82403052 | 82412699 | 2.95E-08 | n.a. | n.a. |
| 6 | 85803437 | 85913316 | 85853437 | 85863316 | 3.48E-08 | n.a. | n.a. |
| 6 | 95133876 | 95243862 | 95183876 | 95193862 | 3.61E-08 | n.a. | n.a. |
| 6 | 109231959 | 109371825 | 109281959 | 109291853 | 2.40E-09 | ENSG00000118690 | *ARMC2* |
| 6 | 120421792 | 120531600 | 120471792 | 120481600 | 1.08E-08 | n.a. | n.a. |
| 6 | 122083715 | 122193546 | 122133715 | 122143546 | 2.94E-08 | n.a. | n.a. |
| 6 | 124400554 | 124510298 | 124450554 | 124460298 | 2.84E-08 | ENSG00000188580 | *NKAIN2* |
| 6 | 140346414 | 140456180 | 140396414 | 140406180 | 1.23E-08 | n.a. | n.a. |
| 6 | 151154546 | 151263329 | 151204546 | 151213329 | 1.77E-08 | ENSG00000120278 | *PLEKHG1* |
| 6 | 154853049 | 154962780 | 154903049 | 154912780 | 7.90E-09 | ENSG00000153721 | *CNKSR3* |
| 6 | 157990558 | 158100241 | 158040558 | 158050241 | 6.16E-09 | ENSG00000175048 | *ZDHHC14* |
| 6 | 158409833 | 158519226 | 158459833 | 158469226 | 2.23E-08 | ENSG00000122335 | *SERAC1* |
| 6 | 160592439 | 160702211 | 160642439 | 160652211 | 2.49E-08 | ENSG00000112499 | *SLC22A2* |
| 6 | 168799357 | 168909344 | 168849357 | 168859344 | 2.95E-09 | ENSG00000112562 | *SMOC2* |
| 6 | 169488931 | 169598810 | 169538931 | 169548810 | 2.33E-08 | n.a. | n.a. |
| 6 | 170511524 | 170621330 | 170561524 | 170571330 | 1.99E-08 | ENSG00000112584 | *FAM120B* |
| 7 | 3735339 | 3845287 | 3785339 | 3795287 | 1.74E-09 | ENSG00000146555 | *SDK1* |
| 7 | 4163560 | 4273457 | 4213560 | 4223457 | 2.02E-08 | n.a. | n.a. |
| 7 | 17913563 | 18106534 | 17963563 | 17973313 | 2.51E-09 | ENSG00000071189 | *SNX13* |
| 7 | 18439265 | 18549221 | 18489265 | 18499221 | 1.05E-09 | ENSG00000048052 | *HDAC9* |
| 7 | 19430315 | 19540260 | 19480315 | 19490260 | 1.05E-08 | n.a. | n.a. |
| 7 | 29893633 | 30003330 | 29943633 | 29953330 | 1.58E-08 | ENSG00000136193 | *SCRN1* |
| 7 | 35376998 | 35497388 | 35426998 | 35436796 | 1.10E-09 | n.a. | n.a. |
| 7 | 41972250 | 42081762 | 42022250 | 42031762 | 1.51E-08 | ENSG00000106571 | *GLI3* |
| 7 | 42335633 | 42445581 | 42385633 | 42395581 | 8.62E-10 | n.a. | n.a. |
| 7 | 53396948 | 53506758 | 53446948 | 53456758 | 5.19E-09 | n.a. | n.a. |
| 7 | 54410200 | 54529392 | 54470275 | 54479392 | 8.08E-10 | n.a. | n.a. |
| 7 | 55320648 | 55430641 | 55370648 | 55380641 | 1.45E-11 | ENSG00000132434 | *LANCL2* |
| 7 | 64855038 | 64964875 | 64905038 | 64914875 | 2.04E-08 | ENSG00000169921 | n.a. |
| 7 | 91012764 | 91121597 | 91062764 | 91071597 | 2.57E-09 | n.a. | n.a. |
| 7 | 92435168 | 92545093 | 92485168 | 92495093 | 1.31E-09 | n.a. | n.a. |
| 7 | 118474018 | 118614637 | 118554731 | 118564637 | 5.68E-09 | n.a. | n.a. |
| 7 | 125428929 | 125538905 | 125478929 | 125488905 | 2.87E-09 | n.a. | n.a. |
| 7 | 133069253 | 133229109 | 133170018 | 133179109 | 2.45E-10 | ENSG00000131558 | *EXOC4* |
| 7 | 134848292 | 134967619 | 134908327 | 134917619 | 5.37E-09 | ENSG00000155561 | *NUP205* |
| 7 | 139169143 | 139278799 | 139219143 | 139228799 | 4.18E-10 | ENSG00000059377 | *TBXAS1* |
| 7 | 140963597 | 141073140 | 141013597 | 141023140 | 6.74E-11 | ENSG00000127359 | *KIAA1147* |
| 7 | 146943839 | 147053376 | 146993839 | 147003376 | 7.56E-10 | ENSG00000174469 | *CNTNAP2* |
| 7 | 158732284 | 158852296 | 158792304 | 158802296 | 1.68E-09 | n.a. | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 8 | 1702312 | 1812271 | 1752312 | 1762271 | 1.62E-08 | ENSG00000104728 | *ARHGEF10* |
| 8 | 3654876 | 3764742 | 3704876 | 3714742 | 5.43E-10 | n.a. | n.a. |
| 8 | 3906844 | 4016736 | 3956844 | 3966736 | 2.18E-09 | n.a. | n.a. |
| 8 | 4594679 | 4704634 | 4644679 | 4654634 | 9.24E-09 | n.a. | n.a. |
| 8 | 5710355 | 5820327 | 5760355 | 5770327 | 1.14E-08 | n.a. | n.a. |
| 8 | 23016243 | 23126155 | 23066243 | 23076155 | 1.10E-09 | ENSG00000173530 | *TNFRSF10D* |
| 8 | 32897840 | 33007836 | 32947840 | 32957836 | 3.44E-08 | n.a. | n.a. |
| 8 | 79200198 | 79310061 | 79250198 | 79260061 | 1.11E-08 | n.a. | n.a. |
| 8 | 82332049 | 82441679 | 82382049 | 82391679 | 1.90E-08 | ENSG00000164687 | *FABP5L2* |
| 8 | 85398122 | 85518354 | 85458531 | 85468354 | 9.92E-10 | n.a. | n.a. |
| 8 | 92489226 | 92598877 | 92539226 | 92548877 | 1.76E-08 | n.a. | n.a. |
| 8 | 120051706 | 120161675 | 120101706 | 120111675 | 2.73E-08 | ENSG00000184374 | *COLEC10* |
| 8 | 127158213 | 127268063 | 127208213 | 127218063 | 2.24E-08 | n.a. | n.a. |
| 8 | 130241742 | 130351579 | 130291742 | 130301579 | 4.31E-09 | n.a. | n.a. |
| 8 | 130816197 | 130926195 | 130866197 | 130876195 | 1.04E-08 | ENSG00000147697 | *GSDMC* |
| 8 | 132362634 | 132472596 | 132412634 | 132422596 | 1.63E-08 | n.a. | n.a. |
| 8 | 134440359 | 134550339 | 134490359 | 134500339 | 1.67E-09 | ENSG00000008513 | *ST3GAL1* |
| 8 | 135926238 | 136036101 | 135976238 | 135986101 | 1.93E-08 | n.a. | n.a. |
| 8 | 137092623 | 137212194 | 137142623 | 137152586 | 6.95E-10 | n.a. | n.a. |
| 8 | 138496674 | 138606420 | 138546674 | 138556420 | 3.61E-08 | n.a. | n.a. |
| 9 | 9680605 | 9790514 | 9730605 | 9740514 | 7.64E-09 | n.a. | n.a. |
| 9 | 10388782 | 10498677 | 10438782 | 10448677 | 1.67E-09 | n.a. | n.a. |
| 9 | 12037077 | 12147065 | 12087077 | 12097065 | 2.49E-08 | n.a. | n.a. |
| 9 | 12652157 | 12762146 | 12702157 | 12712146 | 1.88E-08 | ENSG00000107165 | *TYRP1* |
| 9 | 15955008 | 16064891 | 16005008 | 16014891 | 2.29E-08 | ENSG00000164989 | *C9orf93* |
| 9 | 24653105 | 24763039 | 24703105 | 24713039 | 2.10E-09 | n.a. | n.a. |
| 9 | 31644278 | 31754115 | 31694278 | 31704115 | 2.44E-08 | n.a. | n.a. |
| 9 | 44684944 | 44794637 | 44734944 | 44744637 | 2.62E-08 | n.a. | n.a. |
| 9 | 91690411 | 91800172 | 91740411 | 91750172 | 3.98E-08 | n.a. | n.a. |
| 9 | 100412521 | 100522373 | 100462521 | 100472373 | 4.75E-09 | ENSG00000136928 | *GABBR2* |
| 9 | 107308722 | 107417306 | 107358722 | 107367306 | 3.38E-09 | ENSG00000106701 | *FSD1L* |
| 9 | 107308722 | 107417306 | 107358722 | 107367306 | 3.38E-09 | ENSG00000106692 | *FKTN* |
| 9 | 138250273 | 138356368 | 138300273 | 138306368 | 7.29E-09 | ENSG00000165661 | *QSOX2* |
| 9 | 138822720 | 138930998 | 138872720 | 138880998 | 1.76E-08 | ENSG00000177943 | *MAMDC4* |
| 9 | 138822720 | 138930998 | 138872720 | 138880998 | 1.76E-08 | ENSG00000107223 | *EDF1* |
| 10 | 25169318 | 25279195 | 25219318 | 25229195 | 8.03E-09 | ENSG00000099256 | *PRTFDC1* |
| 10 | 47057619 | 47167276 | 47107619 | 47117276 | 5.69E-09 | ENSG00000198250 | *ANTXRL* |
| 10 | 55093068 | 55202990 | 55143068 | 55152990 | 4.29E-10 | n.a. | n.a. |
| 10 | 57604041 | 57723632 | 57664215 | 57673632 | 4.23E-09 | n.a. | n.a. |
| 10 | 58913807 | 59023667 | 58963807 | 58973667 | 3.30E-08 | n.a. | n.a. |
| 10 | 68147491 | 68256569 | 68197491 | 68206569 | 3.31E-08 | ENSG00000183230 | *CTNNA3* |
| 10 | 92049635 | 92159497 | 92099635 | 92109497 | 1.31E-08 | n.a. | n.a. |
| 10 | 107134271 | 107244178 | 107184271 | 107194178 | 2.75E-08 | n.a. | n.a. |
| 10 | 109003054 | 109112833 | 109053054 | 109062833 | 1.19E-08 | n.a. | n.a. |
| 10 | 118507959 | 118617886 | 118557959 | 118567886 | 2.26E-08 | ENSG00000188316 | *C10orf134* |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 10 | 134207902 | 134412074 | 134278316 | 134287774 | 3.45E-10 | ENSG00000068383 | *INPP5A* |
| 11 | 5007434 | 5117204 | 5057434 | 5067204 | 1.41E-08 | ENSG00000176787 | *OR52E2* |
| 11 | 6015825 | 6125787 | 6065825 | 6075787 | 4.43E-10 | n.a. | n.a. |
| 11 | 6753243 | 6873391 | 6813393 | 6823391 | 1.79E-08 | n.a. | n.a. |
| 11 | 34336403 | 34446297 | 34386403 | 34396297 | 1.06E-08 | ENSG00000121691 | *CAT* |
| 11 | 48255938 | 48387052 | 48305938 | 48315892 | 3.94E-09 | ENSG00000176547 | *OR4C3* |
| 11 | 55119145 | 55229067 | 55169145 | 55179067 | 3.75E-08 | ENSG00000181927 | *OR4P4* |
| 11 | 55119145 | 55229067 | 55169145 | 55179067 | 3.75E-08 | ENSG00000174982 | *OR4S2* |
| 11 | 58322153 | 58432053 | 58372153 | 58382053 | 8.66E-09 | ENSG00000156689 | *GLYATL2* |
| 11 | 71924296 | 72034209 | 71974296 | 71984209 | 2.46E-10 | ENSG00000186642 | *PDE2A* |
| 11 | 93804371 | 93913922 | 93854371 | 93863922 | 3.48E-08 | ENSG00000020922 | *MRE11A* |
| 11 | 93804371 | 93913922 | 93854371 | 93863922 | 3.48E-08 | ENSG00000168876 | *ANKRD49* |
| 11 | 97528804 | 97638753 | 97578804 | 97588753 | 1.40E-08 | n.a. | n.a. |
| 11 | 98832809 | 98942728 | 98882809 | 98892728 | 1.77E-08 | n.a. | n.a. |
| 11 | 126138569 | 126248207 | 126188569 | 126198207 | 6.92E-09 | n.a. | n.a. |
| 12 | 2190860 | 2300493 | 2240860 | 2250493 | 8.07E-09 | ENSG00000151067 | *CACNA1C* |
| 12 | 10945041 | 11167569 | 10995041 | 11004915 | 1.82E-08 | ENSG00000212127 | *TAS2R14* |
| 12 | 17722238 | 17832178 | 17772238 | 17782178 | 9.18E-09 | n.a. | n.a. |
| 12 | 41400616 | 41510579 | 41450616 | 41460579 | 1.10E-09 | n.a. | n.a. |
| 12 | 43420652 | 43530560 | 43470652 | 43480560 | 1.45E-09 | ENSG00000184613 | *NELL2* |
| 12 | 57301424 | 57411025 | 57351424 | 57361025 | 3.59E-09 | n.a. | n.a. |
| 12 | 72491600 | 72611625 | 72541600 | 72551447 | 1.14E-09 | n.a. | n.a. |
| 12 | 86107277 | 86217091 | 86157277 | 86167091 | 7.17E-10 | n.a. | n.a. |
| 12 | 113947438 | 114057220 | 113997438 | 114007220 | 1.48E-08 | n.a. | n.a. |
| 12 | 124356881 | 124466320 | 124406881 | 124416320 | 3.05E-08 | ENSG00000139364 | *TMEM132B* |
| 12 | 125356974 | 125466886 | 125406974 | 125416886 | 6.52E-09 | n.a. | n.a. |
| 13 | 18795614 | 18905535 | 18845614 | 18855535 | 9.58E-09 | n.a. | n.a. |
| 13 | 25176123 | 25285418 | 25226123 | 25235418 | 1.74E-08 | ENSG00000132932 | *ATP8A2* |
| 13 | 28986360 | 29096326 | 29036360 | 29046326 | 2.39E-08 | ENSG00000139514 | *SLC7A1* |
| 13 | 42430804 | 42540355 | 42480804 | 42490355 | 3.69E-10 | ENSG00000133106 | *EPSTI1* |
| 13 | 45940089 | 46050084 | 45990089 | 46000084 | 2.24E-09 | ENSG00000136141 | *LRCH1* |
| 13 | 51855900 | 51965708 | 51905900 | 51915708 | 1.22E-08 | ENSG00000136100 | *VPS36* |
| 13 | 62485115 | 62595091 | 62535115 | 62545091 | 8.65E-09 | n.a. | n.a. |
| 13 | 67284250 | 67393806 | 67334250 | 67343806 | 7.20E-09 | n.a. | n.a. |
| 13 | 100825274 | 100945266 | 100885429 | 100895266 | 5.51E-09 | ENSG00000198542 | *ITGBL1* |
| 14 | 21811403 | 21921291 | 21861403 | 21871291 | 5.73E-09 | n.a. | n.a. |
| 14 | 22482624 | 22592369 | 22532624 | 22542369 | 3.14E-08 | ENSG00000100802 | *C14orf93* |
| 14 | 37655799 | 37765761 | 37705799 | 37715761 | 3.52E-08 | ENSG00000139874 | *SSTR1* |
| 14 | 54006122 | 54115972 | 54056122 | 54065972 | 1.38E-08 | ENSG00000100532 | *CGRRF1* |
| 14 | 68858623 | 68978676 | 68918680 | 68928676 | 2.00E-10 | ENSG00000100632 | *ERH* |
| 14 | 69034751 | 69144401 | 69084751 | 69094401 | 8.14E-10 | ENSG00000175985 | n.a. |
| 14 | 72302462 | 72412083 | 72352462 | 72362083 | 7.67E-09 | ENSG00000205683 | *DPF3* |
| 14 | 76462869 | 76572825 | 76512869 | 76522825 | 1.92E-08 | ENSG00000119669 | *C14orf4* |
| 14 | 87461459 | 87571399 | 87511459 | 87521399 | 8.04E-10 | ENSG00000054983 | *GALC* |
| 14 | 105942432 | 106052419 | 105992432 | 106002419 | 2.73E-08 | ENSG00000187156 | n.a. |

| Chr | Region start | Region end | Peak start | Peak end | Peak p value | Gene EnsemblID | Gene name |
|---|---|---|---|---|---|---|---|
| 15 | 21711179 | 21821057 | 21761179 | 21771057 | 1.83E-08 | n.a. | n.a. |
| 15 | 25872564 | 25982406 | 25922564 | 25932406 | 2.22E-09 | ENSG00000104044 | *OCA2* |
| 15 | 57479321 | 57589164 | 57529321 | 57539164 | 2.77E-08 | ENSG00000157470 | *FAM81A* |
| 15 | 83658565 | 83768373 | 83708565 | 83718373 | 9.70E-09 | ENSG00000170776 | *AKAP13* |
| 15 | 93371423 | 93481402 | 93421423 | 93431402 | 4.12E-09 | n.a. | n.a. |
| 15 | 93607440 | 93716802 | 93657440 | 93666802 | 4.09E-09 | n.a. | n.a. |
| 15 | 95867099 | 95976469 | 95917099 | 95926469 | 3.00E-09 | n.a. | n.a. |
| 16 | 7005291 | 7115269 | 7055291 | 7065269 | 1.73E-08 | n.a. | n.a. |
| 16 | 8082635 | 8192555 | 8132635 | 8142555 | 2.50E-08 | n.a. | n.a. |
| 16 | 8576022 | 8706045 | 8646123 | 8656045 | 7.08E-11 | ENSG00000067365 | *C16orf68* |
| 16 | 9090535 | 9200491 | 9140535 | 9150491 | 7.41E-09 | ENSG00000182831 | *C16orf72* |
| 16 | 10229060 | 10337712 | 10279060 | 10287712 | 2.45E-08 | n.a. | n.a. |
| 16 | 22817413 | 22988761 | 22877911 | 22887894 | 1.78E-12 | ENSG00000122254 | *HS3ST2* |
| 16 | 46790944 | 46900813 | 46840944 | 46850813 | 1.12E-09 | ENSG00000102910 | *LONP2* |
| 16 | 59901557 | 60011534 | 59951557 | 59961534 | 2.33E-08 | n.a. | n.a. |
| 16 | 74887489 | 74997217 | 74937489 | 74947217 | 1.87E-08 | ENSG00000152910 | *CNTNAP4* |
| 16 | 75724465 | 75834452 | 75774465 | 75784452 | 2.50E-09 | ENSG00000103111 | *MON1B* |
| 16 | 80663239 | 80773163 | 80713239 | 80723163 | 6.43E-09 | ENSG00000135698 | *MPHOSPH6* |
| 17 | 21091602 | 21213130 | 21141602 | 21151092 | 1.24E-08 | ENSG00000034152 | *MAP2K3* |
| 17 | 23649233 | 23768988 | 23709249 | 23718988 | 1.96E-08 | ENSG00000160629 | *TMEM199* |
| 17 | 23649233 | 23768988 | 23709249 | 23718988 | 1.96E-08 | ENSG00000109072 | *VTN* |
| 17 | 48778049 | 48887499 | 48828049 | 48837499 | 5.42E-09 | n.a. | n.a. |
| 18 | 11654262 | 11764260 | 11704262 | 11714260 | 1.88E-09 | ENSG00000141404 | *GNAL* |
| 18 | 48518185 | 48628069 | 48568185 | 48578069 | 1.41E-08 | ENSG00000187323 | *DCC* |
| 18 | 49678230 | 49788193 | 49728230 | 49738193 | 6.28E-09 | n.a. | n.a. |
| 18 | 71336451 | 71446441 | 71386451 | 71396441 | 3.88E-09 | n.a. | n.a. |
| 18 | 74017896 | 74127878 | 74067896 | 74077878 | 3.10E-08 | n.a. | n.a. |
| 19 | 1518037 | 1627456 | 1568037 | 1577456 | 7.77E-09 | ENSG00000181588 | *MEX3D* |
| 19 | 21556254 | 21676350 | 21606254 | 21615861 | 8.48E-09 | ENSG00000213976 | n.a. |
| 19 | 33428608 | 33548642 | 33488767 | 33498642 | 9.10E-09 | n.a. | n.a. |
| 19 | 33814772 | 33924767 | 33864772 | 33874767 | 3.29E-09 | ENSG00000205243 | n.a. |
| 19 | 34459241 | 34569238 | 34509241 | 34519238 | 1.11E-08 | n.a. | n.a. |
| 19 | 36871058 | 36981031 | 36921058 | 36931031 | 1.66E-08 | n.a. | n.a. |
| 19 | 43457904 | 43567429 | 43507904 | 43517429 | 5.59E-10 | ENSG00000099337 | *KCNK6* |
| 19 | 52474978 | 52583615 | 52524978 | 52533615 | 4.57E-09 | ENSG00000134830 | *GPR77* |
| 19 | 56816295 | 56926116 | 56866295 | 56876116 | 1.42E-08 | n.a. | n.a. |
| 20 | 18733277 | 18839699 | 18783277 | 18789699 | 4.15E-09 | ENSG00000149443 | *C20orf78* |
| 20 | 24163282 | 24293922 | 24233949 | 24243922 | 9.10E-12 | n.a. | n.a. |
| 20 | 51860627 | 51970421 | 51910627 | 51920421 | 7.26E-09 | n.a. | n.a. |
| 20 | 58592629 | 58702351 | 58642629 | 58652351 | 2.22E-10 | n.a. | n.a. |
| 21 | 21293678 | 21403136 | 21343678 | 21353136 | 2.53E-08 | ENSG00000154654 | *NCAM2* |
| 21 | 27493692 | 27603684 | 27543692 | 27553684 | 2.83E-08 | n.a. | n.a. |
| 21 | 29759048 | 29878999 | 29819050 | 29828999 | 6.23E-09 | ENSG00000171189 | *GRIK1* |
| 21 | 35819529 | 35929507 | 35869529 | 35879507 | 4.01E-09 | n.a. | n.a. |

# Appendix E

The table below lists regions in Akey's review that overlap with our candidate regions, and the number of previous genome-wide scans that identified the region as positively selected. Coordinates are in NCBI36.

| Chr | Start | End | No. of scans |
|---|---|---|---|
| 1 | 35091347 | 36450032 | 7 |
| 1 | 52834684 | 53397600 | 5 |
| 1 | 63896341 | 64300000 | 4 |
| 1 | 73110567 | 73763239 | 2 |
| 1 | 75412213 | 75792745 | 4 |
| 1 | 102950280 | 103520567 | 5 |
| 1 | 186911638 | 187544804 | 2 |
| 1 | 187734833 | 188413044 | 2 |
| 1 | 212836722 | 213574853 | 2 |
| 1 | 245200000 | 245400000 | 2 |
| 2 | 5937460 | 6624193 | 2 |
| 2 | 21520822 | 21691853 | 2 |
| 2 | 39873524 | 40313548 | 2 |
| 2 | 69000000 | 69100000 | 2 |
| 2 | 83200000 | 83491853 | 3 |
| 2 | 84300000 | 85001443 | 7 |
| 2 | 86420824 | 86700685 | 3 |
| 2 | 107252946 | 107707862 | 4 |
| 2 | 121388162 | 121500000 | 2 |
| 2 | 167488441 | 167829843 | 3 |
| 2 | 177021882 | 178332739 | 8 |
| 2 | 194388441 | 194872739 | 4 |
| 2 | 195445042 | 197337972 | 5 |
| 2 | 215700000 | 216047276 | 4 |
| 2 | 237100000 | 237806446 | 4 |
| 3 | 17174903 | 17918047 | 7 |
| 3 | 43171008 | 43826735 | 3 |
| 3 | 59659140 | 59859140 | 2 |
| 3 | 66549620 | 66649620 | 2 |
| 3 | 108648481 | 109250364 | 5 |
| 3 | 110200000 | 110400000 | 3 |
| 3 | 124800000 | 124908517 | 3 |
| 3 | 144980752 | 145392790 | 4 |
| 3 | 189890333 | 190421069 | 3 |
| 3 | 195600000 | 195700000 | 2 |

| Chr | Start | End | No. of scans |
|---|---|---|---|
| 4 | 1000000 | 1100000 | 2 |
| 4 | 3500000 | 3600000 | 2 |
| 4 | 14117624 | 14591044 | 5 |
| 4 | 32715092 | 33416543 | 6 |
| 4 | 33453829 | 34600000 | 4 |
| 4 | 41100000 | 42012240 | 9 |
| 4 | 60433581 | 60833044 | 2 |
| 4 | 71712335 | 71739976 | 2 |
| 4 | 85052048 | 85711845 | 2 |
| 4 | 93711845 | 93911845 | 2 |
| 4 | 96559478 | 96900000 | 2 |
| 4 | 99861845 | 100861845 | 4 |
| 4 | 135100000 | 135637586 | 3 |
| 4 | 147933837 | 148461845 | 4 |
| 4 | 159500000 | 160383102 | 5 |
| 4 | 165478100 | 165764287 | 2 |
| 4 | 170733084 | 171062639 | 4 |
| 4 | 171504228 | 172245075 | 4 |
| 4 | 172390869 | 173690711 | 2 |
| 4 | 176402530 | 177137450 | 5 |
| 4 | 177900000 | 178108882 | 2 |
| 5 | 11428374 | 11964581 | 3 |
| 5 | 21591332 | 22138852 | 4 |
| 5 | 54919872 | 55028135 | 2 |
| 5 | 75118302 | 75321300 | 2 |
| 5 | 100585818 | 101100000 | 3 |
| 5 | 109051683 | 109351683 | 2 |
| 5 | 116900000 | 118039500 | 6 |
| 5 | 124199397 | 124569221 | 3 |
| 5 | 141324037 | 141367520 | 2 |
| 6 | 48261123 | 48400000 | 2 |
| 6 | 56600000 | 56700000 | 2 |
| 6 | 67209702 | 67610632 | 2 |
| 6 | 95545225 | 95800000 | 3 |
| 6 | 102261123 | 102461123 | 2 |
| 6 | 125797887 | 126100000 | 4 |
| 6 | 129300000 | 129400000 | 2 |
| 6 | 132524241 | 132708765 | 2 |
| 6 | 144769146 | 145480251 | 2 |
| 6 | 158164208 | 158660281 | 4 |
| 7 | 3616258 | 3831778 | 2 |
| 7 | 19000000 | 19100000 | 2 |
| 7 | 28092602 | 28200000 | 2 |
| 7 | 30117938 | 30486920 | 3 |
| 7 | 100190728 | 102393972 | 4 |

| Chr | Start | End | No. of scans |
|---|---|---|---|
| 7 | 118426018 | 118562339 | 2 |
| 7 | 119136728 | 119800000 | 3 |
| 7 | 135077048 | 135518370 | 3 |
| 7 | 136703651 | 137238735 | 3 |
| 7 | 145499216 | 145879474 | 4 |
| 8 | 9245944 | 9900000 | 3 |
| 8 | 10630705 | 11614773 | 7 |
| 8 | 16087027 | 16507429 | 3 |
| 8 | 20836402 | 21282410 | 3 |
| 8 | 34000000 | 35059528 | 3 |
| 8 | 35611003 | 36378014 | 3 |
| 8 | 50212593 | 50500000 | 3 |
| 8 | 50580000 | 52150000 | 8 |
| 8 | 56962654 | 57180754 | 2 |
| 8 | 82065718 | 82400000 | 4 |
| 9 | 12500000 | 12800000 | 3 |
| 9 | 15900000 | 16100000 | 3 |
| 9 | 24300000 | 24679974 | 3 |
| 9 | 25922398 | 27000000 | 3 |
| 10 | 2950000 | 3100000 | 3 |
| 10 | 55489628 | 55857080 | 6 |
| 10 | 58559648 | 59725403 | 6 |
| 10 | 92000000 | 92100000 | 2 |
| 10 | 107024983 | 107512933 | 4 |
| 10 | 118125403 | 118276595 | 4 |
| 11 | 23600000 | 23730424 | 2 |
| 11 | 24900000 | 25092267 | 2 |
| 11 | 37368196 | 38750000 | 7 |
| 11 | 39594336 | 40051695 | 3 |
| 11 | 48292267 | 48392267 | 2 |
| 12 | 10900000 | 11100000 | 2 |
| 12 | 42500000 | 43185479 | 6 |
| 12 | 48000000 | 48300000 | 2 |
| 12 | 84396635 | 85067701 | 2 |
| 13 | 18774831 | 19612163 | 6 |
| 13 | 21669520 | 22133385 | 2 |
| 13 | 33055423 | 33455792 | 2 |
| 13 | 37603687 | 38490126 | 5 |
| 13 | 51800000 | 51909915 | 2 |
| 13 | 62067478 | 62850000 | 4 |
| 13 | 67150000 | 67350000 | 2 |
| 13 | 102845482 | 103213227 | 2 |
| 13 | 103955437 | 104200000 | 2 |
| 13 | 104398149 | 104598149 | 2 |
| 13 | 104950261 | 105324693 | 2 |

| Chr | Start | End | No. of scans |
|---|---|---|---|
| 14 | 44280000 | 44770503 | 6 |
| 14 | 56700000 | 57000000 | 3 |
| 14 | 66779712 | 66900000 | 2 |
| 14 | 87235602 | 87763024 | 2 |
| 14 | 89492048 | 89579438 | 2 |
| 14 | 105800000 | 105900000 | 2 |
| 15 | 25800000 | 26378746 | 6 |
| 15 | 42200000 | 42300000 | 2 |
| 15 | 45937993 | 46804624 | 6 |
| 15 | 49844415 | 50441408 | 3 |
| 15 | 61145232 | 62319917 | 9 |
| 15 | 86550993 | 87140285 | 3 |
| 16 | 22840948 | 23040948 | 2 |
| 16 | 45959870 | 47212009 | 5 |
| 16 | 76927006 | 77155299 | 4 |
| 17 | 55211782 | 56901284 | 6 |
| 17 | 60400000 | 60500000 | 2 |
| 18 | 28600000 | 29361325 | 7 |
| 18 | 61600000 | 61700000 | 2 |
| 18 | 64675139 | 64941495 | 3 |
| 18 | 65689235 | 66040000 | 5 |
| 19 | 11875627 | 12000000 | 2 |
| 19 | 43400000 | 43600000 | 2 |
| 19 | 45194869 | 45300000 | 2 |
| 20 | 19831968 | 20720000 | 3 |
| 20 | 31200000 | 31300000 | 2 |
| 20 | 57849002 | 58049002 | 2 |
| 21 | 29719326 | 30020705 | 2 |

# Appendix F

This table shows genes within each enriched functional cluster in the CEU and YRI populations. Genes are shown in Ensembl ID.

| Functional cluster | No. of genes | Bonferroni p-value | Genes |
|---|---|---|---|
| CEU | | | |
| Cell adhesion | 27 | 0.001 | ENSG00000138650, ENSG00000164853, ENSG00000183230, ENSG00000140945, ENSG00000138696, ENSG00000152894, ENSG00000169760, ENSG00000156282, ENSG00000174469, ENSG00000177103, ENSG00000165973, ENSG00000118762, ENSG00000169862, ENSG00000077522, ENSG00000169604, ENSG00000060718, ENSG00000112699, ENSG00000134121, ENSG00000021645, ENSG00000146555, ENSG00000151914, ENSG00000170927, ENSG00000150275, ENSG00000154654, ENSG00000137975, ENSG00000150637, ENSG00000145012. |
| Signal | 74 | 0.002 | ENSG00000138650, ENSG00000152208, ENSG00000144749, ENSG00000104974, ENSG00000183117, ENSG00000162763, ENSG00000138696, ENSG00000152894, ENSG00000174808, ENSG00000174469, ENSG00000156687, ENSG00000109743, ENSG00000126709, ENSG00000016402, ENSG00000154589, ENSG00000124159, ENSG00000204866, ENSG00000149328, ENSG00000169604, ENSG00000102763, ENSG00000092421, ENSG00000169605, ENSG00000134121, ENSG00000188467, ENSG00000021645, ENSG00000214510, ENSG00000203837, ENSG00000105825, ENSG00000151490, ENSG00000166342, ENSG00000131126, ENSG00000164418, ENSG00000150275, ENSG00000124493, ENSG00000179399, ENSG00000172551, ENSG00000137975, ENSG00000165409, ENSG00000166206, ENSG00000111452, ENSG00000011465, ENSG00000150637, ENSG00000133710, ENSG00000140945, ENSG00000169760, ENSG00000175445, ENSG00000156564, ENSG00000107798, ENSG00000177103, ENSG00000145536, ENSG00000165973, ENSG00000144455, ENSG00000171189, ENSG00000109610, ENSG00000060718, ENSG00000142549, ENSG00000196277, ENSG00000183722, ENSG00000183098, ENSG00000122711, ENSG00000146555, ENSG00000170927, ENSG00000134247, ENSG00000197614, ENSG00000104760, ENSG00000141337, ENSG00000187095, ENSG00000154654, ENSG00000080224, ENSG00000174123, ENSG00000124107, ENSG00000140470, ENSG00000009765, ENSG00000139263. |
| Ig-like C2-type 3 domain | 12 | 0.001 | ENSG00000154654, ENSG00000134121, ENSG00000144749, ENSG00000104974, ENSG00000146555, ENSG00000065534, ENSG00000036448, ENSG00000177103, ENSG00000134247, ENSG00000139263, ENSG00000142549, ENSG00000187095 |

| Functional cluster | No. of genes | Bonferroni p-value | Genes |
|---|---|---|---|
| YRI | | | |
| N-linked glycosylation site | 60 | 0.0007 | ENSG00000152208, ENSG00000139364, ENSG00000163638, ENSG00000174469, ENSG00000112562, ENSG00000163956, ENSG00000112499, ENSG00000143341, ENSG00000204525, ENSG00000112462, ENSG00000109072, ENSG00000187323, ENSG00000155011, ENSG00000150625, ENSG00000165661, ENSG00000152910, ENSG00000169855, ENSG00000173530, ENSG00000151067, ENSG00000158816, ENSG00000155511, ENSG00000179344, ENSG00000152595, ENSG00000107165, ENSG00000145888, ENSG00000008513, ENSG00000116147, ENSG00000099337, ENSG00000099338, ENSG00000106692, ENSG00000122254, ENSG00000168702, ENSG00000198542, ENSG00000177233, ENSG00000171189, ENSG00000143502, ENSG00000197584, ENSG00000184374, ENSG00000139514, ENSG00000156869, ENSG00000175497, ENSG00000184613, ENSG00000144619, ENSG00000134830, ENSG00000146555, ENSG00000212127, ENSG00000176787, ENSG00000139874, ENSG00000196578, ENSG00000054983, ENSG00000166363, ENSG00000154654, ENSG00000177943, ENSG00000174982, ENSG00000136928, ENSG00000176555, ENSG00000163873, ENSG00000197865, ENSG00000176547, ENSG00000104044. |
| RhoGEF domain | 6 | 0.01 | ENSG00000170776, ENSG00000104728, ENSG00000198399, ENSG00000120278, ENSG00000132694, ENSG00000053524. |
| glutamate receptor activity | 5 | 0.04 | ENSG00000152208, ENSG00000155511, ENSG00000171189, ENSG00000136928, ENSG00000163873. |

# Appendix G

### Command lines for GENETREE

Generate tree structure:
./seq2tr seq2tr_input seq2tr_output;

Estimate the best theta:
./genetree seq2tr_output seed_theta 100000 6666 -f surf_output -g seed_theta/10 seed_theta*10 500 -m mg_3pop -y 100 -2 -x 1000 > estimate_theta_out;

Estimate TMRCA:
./genetree  seq2tr_output estimated_theta 10000000 6666 -m mg_3pop -y 100 -x 1000 > estimate_TMRCA_output;

# Appendix H

Phylogenetic networks of two regions with recent TMRCAs. A: chr1:28,465,001-28,480,000; TMRCA 1.992 $N_e$ generations. B: chr1:28,920,001-28,940,000; TMRCA 1.962 $N_e$ generations.