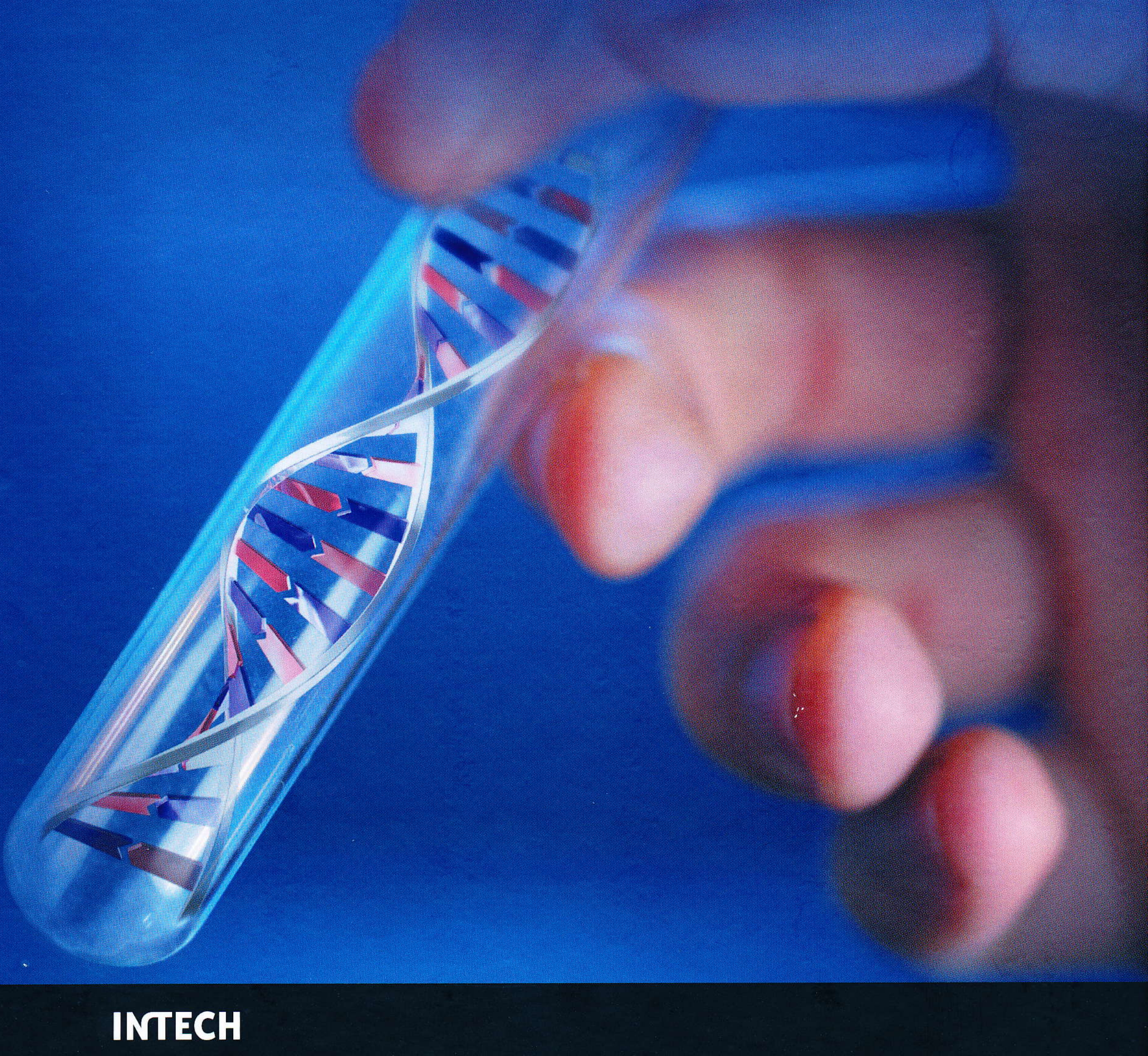
STUDIES IN

POPULATION GENETICS

Edited by **M. Carmen Fusté**



INTECH

Contents

Preface VII

Chapter 1 **Piecing the *punicus* Puzzle 1**
Byron Baron

Chapter 2 **The Evolution of Plant Mating System: Is It Time for a Synthesis? 17**
Cheptou Pierre-Olivier

Chapter 3 **Population Genetics of the"*Aeromonas hydrophila* Species Complex" 39**
M^{a} Carmen Fusté, Maribel Farfán, David Miñana-Galbis, Vicenta Albarral, Ariadna Sanglas and José Gaspar Lorén

Chapter 4 **Minisatellite DNA Markers in Population Studies 55**
Svetlana Limborska, Andrey Khrunin and Dmitry Verbenko

Chapter 5 **Polymorphism 85**
Oliver Mayo

Chapter 6 **Speciation in Brazilian AtlanticForest Mosquitoes: A Mini-Review of the *Anopheles cruzii* Species Complex 105**
Luísa D.P. Rona, Carlos J. Carvalho-Pinto and Alexandre A. Peixoto

Chapter 7 **The Next Step in Understanding Population Dynamics: Comprehensive Numerical Simulation 117**
John C. Sanford and Chase W. Nelson

Chapter 8 **Population Genetics in the Genomic Era 137**
Shuhua Xu and Wenfei Jin

Population Genetics in the Genomic Era

Shuhua Xu[*] and Wenfei Jin
Chinese Academy of Sciences Key Laboratory of Computational Biology,
Chinese Academy of Sciences and Max Planck Society (CAS-MPG)
Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences, Shanghai,
China

1. Introduction

Over the past decades, scientific research on population genetics has been facilitated greatly by advances in DNA genotyping and sequencing technologies, during which time it has transformed from a theory-driven field with little empirical data into a data-driven discipline with a deluge of data. The emergence of population genomics demarcated a transition from single-locus-based studies to genome-wide analyses of genetic variations, which benefits the identification of disease/trait associated genes and targets of natural selection. In particular, with the development of next-generation sequencing (NGS), a considerable number of individual genomes are becoming available. Nonetheless, the availability of genome-wide data not only is a big challenge for computational capability, which drives the statistics and methods to be more efficient, but also leads to some transitions on strategies and methodologies. For example, the traditional strategy for identifying targets of positive selection was based on candidate gene strategy that priori assumed a gene under selection and compared it with the null hypothesis. Genomic approaches, however, usually involve constructing an empirical distribution of a summary statistics across all loci, which quantifies a characteristic of the genetic variation, with the extreme tail of the distribution being defined as putative targets.

The analysis of genome-wide data has greatly improved our knowledge on mechanism of mutations and recombination, human origins and history, adaptation to local environment and variants underlying disease. However, some of these discoveries conflicted with the traditional views or models. For example, it is traditionally believed that beneficial mutation usually arises and increases in frequency to fixation, which was referred to as classic selective sweep model that almost all statistics detecting positive selection relied on, while recent genome-wide analysis showed that it was rarely the case in recent human evolution, which demands new models and statistics to depict and detect the signatures of positive selection. Meanwhile, great progress could be made by integrating these new discoveries to improve the mathematical models such as coalescent theory. In a nutshell, the advent of genomic data has influenced and will continue its influence on every corner of genetics, thus significantly accelerating the development of population genetics.

[*] Corresponding Author

Besides, genomic approach has also greatly facilitated the identification of diseases/traits associated genes. Genome-wide association studies (GWAS), genotyping millions of SNPs on thousands of individuals, has become a standard method in disease gene discovery in the past several years. However, the common variants identified by GWAS only account for a small fraction of the heritability thus fail to explain the majority of phenotypic variance in population. Therefore, as an alternative to the common disease common variants hypothesis (CDCV), several new hypotheses have been proposed.

In this chapter, we will focus only on those topics concerned with population genomics, i.e. methods, statistics and analysis based on high-density genome-wide data (either genotyping data or sequencing data), so the research category can be different from that of traditional population genetic studies relying on single locus or sparse loci.

2. Variation, recombination, haplotypes and inference of population parameters

2.1 Overview of genome-wide high-density data

Based on the current technologies and features, the genome-wide data can be roughly classified into genotyping data and sequencing data.

DNA Genotyping is the process of determining the status of DNA using biological assays and comparing it with known sequences. It is used either to track the alleles an individual inherited from his/her parents, or to reveal differentiations between individuals and populations. With most SNPs discovered in a small set of samples, the genotyping data have high proportion of SNPs with intermediate allele frequency. This ascertainment bias is likely to affect all statistics based on allele frequencies[1].

DNA Sequencing, on the other hand, includes several methods and technologies to determine the order of the nucleotides in DNA. The high demand for sequencing data has promoted the development of low-cost high-throughput sequencing technologies (also referred to as next-generation sequencing) that parallelize the sequencing process by producing thousands or millions of sequences at once[2]. These low-cost high-throughput technologies, including 454 Life Sciences (Roche) sequencing, Illumina Solexa sequencing, and Applied Biosystems SOLiD sequencing, will finally make the individual genomes affordable and accessible, initiating individual genomic era.

2.2 Genetic variations in human genome

Genetic variations refer to any genetic differences among individuals within one population or species, which provide the genetic basis of evolution. Since the nucleotide differentiation between individuals is estimated to be about 0.1%[3], meaning that there are about 3-million nucleotide differences between two unrelated individuals. The genetic variations in human genome can be classified into single nucleotide polymorphism (SNP), short insertion and deletion (indel), copy number variation (CNV), variable number tandem repeat (VNTR: including microsatellite and minisatellite), haplotype (including haplogroup), epigenetic and so on[4]. Among all these, SNP is a type of variation with one nucleotide differentiation in sequence, which is generally caused by single mutations, and it is estimated that there are about 30 million SNPs existing in human genomes, which makes them the most common genetic variations in human genomes.

There were various scientific endeavors for identifying the genetic variations after the completion of human genome project (HGP). For example, the International Haplotype Map Project (http://hapmap.ncbi.nlm.nih.gov) has provided the allele frequency of about 4 million SNPs in at least one population; the 1000 Genome Project is another international collaboration trying to provide the accurate haplotype information on all forms of human DNA polymorphism in multiple human populations, and the completion of its pilot phase has provided the location, allele frequency and local haplotype structure of ~15 million SNPs, 1 million indel and 20,000 structural variations, most of which being novel[5]. The sequencing data also showed that each individual cherished ~3 million variant SNPs, ~350,000 indel, 250-300 loss-of-function variants in annotated genes and 50-100 variants previously implicated in inherited disorders[5].

2.3 Linkage disequilibrium pattern and haplotype

In meiosis, one allele often transmits together with the alleles around it, which leads to association or correlation between loci close to each other, such phenomenon is called linkage disequilibrium (LD). The distribution of LD in human becomes a topic of great interest due to its fundamental role in gene mapping, recombination and human history[6]. Assuming two alleles A and B at two loci with frequencies π_A and π_B in the population, respectively, we expect the frequency of the AB haplotype to be $\pi_A\pi_B$ if the two loci are independent. If the frequency of AB haplotype in a population does not fit the following: $\pi_{AB} = \pi_A\pi_B$, the two loci are in LD. A wide variety of statistics have been proposed to measure LD, the simplest of which is $D = \pi_{AB}\text{-}\pi_A\ \pi_B$. Although D originates from the definition, its values are affected seriously by allele frequencies. Normalization of D is the most common way to address the dependence of D on marginal allele frequencies. Lewontin's *D'*[7, 8] , one of the normalized *D*, has the desirable property, and |D'| = 1 if there are only three gametic types or if two SNPs are in complete LD. Calculation of D' is:

$$D' = \begin{cases} \dfrac{D}{\min(\pi_A\pi_b, \pi_a\pi_B)} & D > 0 \\ \dfrac{D}{\min(\pi_A\pi_B, \pi_a\pi_b)} & D < 0 \end{cases}$$

where πa and πb are the allele frequencies of the alleles a and b, which are the counterpart of A and B in the same loci, respectively. The obvious drawback of D' is that its sampling properties are poorly understood when |D'| < 1. In addition, estimation of D' is strongly inflated in small samples, especially for rare variants. Currently, the most popular measure of LD between biallelic loci is r^2 (also referred to as Δ^2), whose values reflect the amount of information provided about each other:

$$r^2 = \frac{D^2}{\pi_A\pi_a\pi_B\pi_b}$$

In the case of $r^2 = 1$, which is known as perfect LD, it means the observation at one marker provides complete information about the other.

The completion of HapMap project has provided the fine-scale LD pattern of the human genome[9, 10]. First, it is found that LD varies remarkably on scales of 1-100kb, which are always discontinued and compose block-like structures. Second, haplotype diversity arises solely through mutation in the genome when recombine is absent. Thus SNPs arising on the same branch of the genealogy are in complete LD ($|D'| = 1$), while those happened on different branches have limited or no correlation. Third, although different populations have different haplotype frequencies, both common and rare haplotypes are usually shared among populations. Fourth, some SNPs located in recombination hotspots have very weak LD with neighboring SNPs, which is not well represented in tag SNPs. Finally, LD correlates with many genomic features such as recombination rate, mutation rate, G+C content, sequencing variation, repeat composition and chromosome length. A worldwide survey of haplotype and LD in the human genome showed that human history, to some extent, is reflected by the geographic distribution of haplotype, which looses diversity as distance increases from Africa[11].

2.4 Recombination in human population and its implication

By shaping the landscape of individual genomes per generation, recombination plays an important role in reproduction. However, it was not considered in the initial models of population genetics, which assumed that all loci are independent. In recent decades, LD pattern and haplotype pattern created by recombination have been realized harbored much information about recent population history, which has been extended to estimate population parameters and population history[12]. Despite the limited numbers of meioses in pedigree, the genetic map was reconstituted by counting crossover, which revealed the sex difference and recombination variation at megabase[13]. Sperm-typing studies on dozens of regions have demonstrated that most recombination events concentrated in very short regions of 1-2 kb, which are referred to as recombination hotspots with intensity from 4×10^{-4} cM to 0.14 cM[14].

Coalescent theory is a stochastic process that describes the distribution of underlying genealogic tree of individuals from idealized population[15]. When no recombination happened, the inheritance relationships of a group of samples can be represented by a genealogical tree, on which all samples are traced back to a single ancestral copy known as the most recent common ancestor (MRCA)[16], and differences of these samples on the tree are due to mutation events. However, recombination breaks up the chromosomes each generation and reduces LD, which increases haplotype diversity. Thus different regions on the genome may produce different trees due to recombination[17, 18]. If two loci are close to each other and recombination rarely occurred between them, the two trees are likely to be the same. However, as the genetic distance between the two loci increases, the correlation between the two trees decreases[17]. Therefore, a simple genealogic tree may not fully represent the ancestry of the sample of recombined chromosome. Instead, the complex graph[19], which includes a series of coalescence and recombination events, was proposed to represent the recombined chromosome, allowing us to recover the marginal genealogy at any given position[20]. The coalescent that integrated recombination was often referred to as the ancestral recombination graph (ARG)[18, 19], which benefits us a lot in understanding the effect of recombination. By considering where coalescent, recombination and mutation happened on the tree, we are able to infer the impact of recombination on patterns of genetic variation. On the other hand, if the procedure that generates the graph can be modeled, it is also possible for us to estimate the population parameters such as recombination rates[20].

2.5 Estimating recombination rates based on population genetic methods

On one hand, traditional pedigree studies generally do not provide fine-scale recombination rates due to the limited numbers of meioses in a few generations[13, 21]. On the other hand, sperm-typing analysis is extremely laborious and expensive despite its high-resolution[22]. Therefore, with the availability of a deluge of SNP data, statistical inferences of recombination rates using population genetic methods become the major strategy to obtain the landscape of fine-scale recombination. The simplest and most direct way to identify the historical recombination events is to analyze the closest pair-wise SNPs, which, however, does not model the recombination process and recombination graph. For examples, we assume two bi-allelic loci with ancestral and derived allele A/B and a/b, respectively, and all potential haplotypes constituted by the two loci are AB, Ab, aB, ab. For infinite sites mutation model, recombination must have occurred in history when all of these haplotypes were found[20]. Performing four-gamete test (FGT) on all pairs of loci in a region can identify the intervals at which recombination occurred. Assuming all recombination originated from the same recombination event, R_{m} conservatively estimated the minimum number of recombination that had occurred in history[20], which would underestimate the real recombination events[23].

To overcome the shortcoming of counting recombination events directly, it is necessary to model the underlying recombination process. The coalescent that integrated recombination was often referred to as the ancestral recombination graph (ARG)[18, 19]. As a traditional coalescent, ARC model assumes neutral evolution, constant population size with random mating and uniformity of recombination rates across the genome, as well as straightforward extension on complex demographic events. The key parameter in determining patterns of LD is the product of the per-generation recombination (c), and the effective population size (N_e): ρ=4$N_e c$, where ρ is the population recombination rate. Studies such as sperm typing or pedigree analysis, counting the recombination events in one generation, can be used to calculate c, which depends on genomic features such as local DNA motif; while ρ, which depends on demographic history (effect on N_e), and therefore differs substantially among populations.

The numerous methods and statistics based on ARC can be classified into summary statistics, full-likelihood approaches and approximate-likelihood approaches[20, 24]. Summary statistics, though very easy to calculate, provides limited information and is therefore rarely used at present. Full-likelihood approaches try to incorporate all information contained in the data and always integrate many variable dimension genealogies. In this way, Markov Chain Monte Carlo (MCMC), Bayesian MCMC and important sampling (IS) have been used to estimate the likelihood surface for model parameters[20, 25]. For examples, inspired by the observed patterns of recombination in sperm-typing studies, Wang and Rannala[24] developed a Bayesian full-likelihood method using MCMC to estimate background recombination rates and hotspots. However, it seems impossible to apply them on moderate dataset since they are notorious for computational intensity.

Up to now, various approximate-likelihood approaches have been developed to investigate the large population genetic data available[22, 26, 27]. However, these methods either ignore rare and low frequency genetic loci which harbor little information of recombination, or consider only a small number of sites at a time, calculating the likelihood of each subset separately and combining them to obtain the approximate likelihood. In the simplest case, maximum-likelihood estimator of the recombination parameter for each linked two-loci pair was calculated independently[27], and a composite-likelihood estimator was constructed by multiplying all pair-wise likelihood. This approach is not only able to handle both genotyping data and sequencing data efficiently, but also straightforward to incorporating complex mutation and population models with high accuracy. McVean *et al.*[22] extended the two-locus composite-likelihood approach by allowing different recombination rates between each pair of makers, and took a Bayesian implementation using prior distribution to avoid over-fitting, thus estimating recombination rates from a fine-scale of kilo-bases up to that of mega-bases.

Based on the observed ancestry switch points in admixed population, Hinch et al.[29] and Wegmann et al.[30] constructed a high-resolution recombination map of African American in 2000, respectively. It was novel to use ancestry-based approach to identify recombination events and recombination hotspots. Both studies showed that the recombination maps of admixed population are consistent with those of non-admixed populations at mega-bases levels. However, the recombination maps of the African American differ significantly from those of the European at fine-scale. In addition, Hinch *et al.*[29] also identified about 2,500 active recombination hotspots in African Americans but not in European. The 17bp DNA motif enriched in African specific hotspots is well matched to a predicted *PRDM9* binding alleles common in Africans. And individuals who carry the motif tend to have a higher risk of the disease-causing genomic rearrangements. Wegmann *et al.*[30] showed that recombination rates at regions with known large chromosomal structural variants, especially inversions, are likely to be highly population specific compared with those at other regions.

2.6 Inference of effective population size from genetic data

The effective population size (N_e) is defined as the number of breeding individuals in an idealized population that shows the same allele frequency spectrum as the population under consideration. N_e is an important parameter in population genetics that helps to explain population demographic history and genetic structure of complex traits. It is also an important parameter in conservation biology, ecology and evolutionary biology[31]. Based on heterozygosity, LD, temporal changed allele frequency and pattern of genetic variation within or between populations, various strategies and methods have been proposed to estimate current, past and ancient N_e[32].

In recent years, the most quickly developed methodology for estimating N_e was based on LD. Hayes *et al.*[33] proposed chromosome segment homozygosity (CSH), a new statistics, to estimate N_e in different time scales using haplotype or haplotype frequencies data. CSH is the probability of two homologous chromosomal segments coming from the same ancestor without recombination interference. When population size changes linearly over time, the expectation of CSH will be $1/(4N_tc + 1)$, where N_t is the effective population size at t=$1/2c$ generations ago, and c is the length of the chromosome segment in morgans. Thus, CSHs for chromosomal segments of different length can be used to estimate the Ne at different time scales. In other words, CSH over a long distance reflects recent N_e, whereas that over a short distance reflects N_e far more back. When the statistics was applied on human haplotype data[33], N_e was estimated to be ~5,000 at about 2,000 generations ago, and ~15,000 at about 182 generations ago. The results reflect an exponential growth of human population in the past.

The distribution of the time since the most recent common ancestor (TMRCA) between two alleles cherishes much information about population demographic history[34]. In order to take advantage of the genome-wide sequencing data, Li and Durbin[34, 35] proposed a pairwise sequentially Markovian coalescent (PSMC) model that scaled mutation, recombination and piecewise ancestral population size to reveal the detailed population history. In the PSMC-HMM model[34], the observation is a binary sequence of '0', '1' and '.'. The emission probability from state t is $e(1|t) = e^{-\theta}t$, $e(0|t) = 1\text{-}e^{-\theta}t$, and $e(.|t) = 1$; the transition probability from s to t is:

$$p(t\,|\,s) = (1 - e^{-\rho t})q(t\,|\,s) + e^{-\rho s}\delta(t - s)$$

where θ and ρ is the scaled mutation rate and recombination rate, respectively; $\delta(.)$ is the Dirac delta function and

$$q(t\,|\,s) = \frac{1}{\lambda(t)}\int_0^{\min\{s,t\}} \frac{1}{s} \times e^{-\int_u^t \frac{dv}{\lambda(v)}} du$$

is the transition probability condition on there being a recombination event, where $\lambda(t) = N_e(t)/N_0$ is the relative population size at t. Li and Durbin[34] applied the PSMC model on seven high accurate individual genomes from Africa, Europe and East Asia, whose results showed that all populations shared similar N_e between 150 and 1,500 thousand year (kyr) ago. The N_e of African is different from that of non-African populations around 100-120 kyr ago (at 110 kyr ago, N_e of African = 15,313 ± 559; N_e of non-African = 12,829 ± 485). The estimated N_e of European and East Asian populations before 11 kyr ago were almost the same as both experienced a serious bottleneck between 150 kyr and 20-40 kyr ago, during which time their N_e declined from 13,500 to 1,200, and increased sharply afterwards.

3. Human origins, population structure and population history

3.1 Human origins and its early history

Although population genetics focus on variation changes within species, emergence of new species due to long population divergence is also an interesting topic. Recent theoretical studies have focused on the "Isolation with Migration" model of population divergence, which integrated many parameters using methods of population genetics[36]. It is known that genus *Homo* diverged from Australopithecines about 2.3 to 2.5 million years ago in Africa[37]. However, no species except modern human (*Homo sapiens*) in the genus *Homo* has survived in the long evolutionary history. Traditionally, there are two major competing models on the origin of our anatomically modern human: Recent African origin and multiregional evolution. The debates focus on whether modern human originated solely in

Africa. The former proposes that all modern human originated in Africa and dispersed into other parts of the world; while the latter holds that local *archaic hominin* evolved into modern human separately. With evidences from both mtDNA and Y chromosome, recent African origin model has been widely accepted and become the mainstream since the end of last century[38-40]. According to this model, *archaic hominin* evolved into anatomically modern humans solely in Africa about 200,000 years ago. Then a branch of modern humans left Africa 125,000 to 60,000 years ago, and replaced earlier *archaic hominin* such as Neanderthals and *Homo erectus* in other parts of the world.

However, mtNDA or Y chromosome only represents a genetic locus and is suffered from serious genetic drift. Analyses of two extinct *archaic hominin* genomes have enriched our understanding of human origins[41, 42]. The first sequenced *archaic hominin* was Neanderthal, the closest evolutionary cousin of present modern human, who used to live in large parts of Western Eurasia before extinction 30,000 years ago. Analysis of Neanderthal genome essentially supported the recent African origin of modern human. However, Neanderthal shared more genetic variants with present modern humans in Eurasia than in sub-Saharan Africa, which suggested that gene flow from Neanderthals into the ancestry of non-African occurred before the divergence of Eurasian groups[41].It is estimated that 1%-4% of the DNA in modern Eurasia is contributed by Neanderthal. Another sequenced *archaic hominin* was referred to as 'Denisovans' due to their bones was found in Denisova Cave in southern Siberia, who shared a common origin with Neanderthals[42]. Denisovans was not involved in the putative gene flow from Neanderthals into Eurasians; however, data analysis suggested that it has contributed 4–6% of DNA to present Melanesian. Gene flows from Denisovans to New Guineans, Australians, and Mamanwa were also found, but not to mainland East Asians, western Indonesians, Jehai, and Onge[43].

In 2011, genome of an aboriginal Australian, whose DNA was extracted from a 100-year-old lock of hair, was sequenced[44]. Genomic evidences showed that aboriginal Australians were descendents of population colonizing Australian about 62,000 to 75,000 years ago, which is different from that of modern East Asians possibly 25,000 to 38,000 years ago. Thus, present aboriginal Australians should be descendents of the earliest Australian colonist, possibly representing one of the oldest continuous populations outside Africa. In a nutshell, recent studies based on genome-wide data essentially support that all modern human originated in Africa and dispersed into other parts of the world by at least two waves. However, gene flows from some archaic hominin contributed to the gene pool of modern non-African.

3.2 Inference of population history from allele frequency spectrum

Analysis of allele frequency spectrum is one of the most commonly used strategies to infer population history. For example, a large proportion of rare alleles indicates recent population expansion, as mutations occurred since population expansions do not have enough time to spread. Based on expected allele frequency spectrum, various methods, such as those provided by Nielson[45] and Williamson *et al.*[46], have been developed to infer the demographic history. Several studies used multiple summary statistics to compare empirical data with that of simulations under varying demographic history. For example, Voight *et al.*[47] used level of polymorphism, allele frequency spectrum and LD to fit population bottleneck of non-African population. After correcting the ascertainment in HapMap data, Keinan *et al.*[48] found that both ancestry of East Asian and that of European experienced population bottlenecks out of Africa, with the former being affected more seriously.

3.3 Population structure and human history

An ideal population is a single entity in which all members randomly mates. However, because of geographic barriers and limited tendency for individuals to spread, natural populations rarely interbreed as in theoretical model, which leads to population structure. Elucidating human population structure can not only improve our knowledge on human population history, but also reduce false-positive results caused by population stratification in association studies. The worldwide population samples were essential for studying the global pattern of human population structure. Promoted by Cavalli-Sforza et al. in the 1990s, the Human Genome Diversity Project (HGDP) provided more than 1000 samples in 53 indigenous populations from the world[49]. The traditional methods for studying population structure could be classified into phylogeography and summary statistics[50]. These methods analyzed the genetic variation based on predefined populations/ethnics classification according to culture or geographic locations, which may not reflect the true genetic relationships. The clustering methods such as STRUCTURE[51, 52] can infer the population genetic structure directly without the prior information about the origins of individuals. STRUCTURE implements a model-based clustering method that integrates Markov chain Monte Carlo (MCMC) to infer population structure with multi-locus genotype data. Rosenberg et al.[53] applied STRUCTURE on 1,056 HGDP individuals genotyping at 377 microsatellites and found that individuals from the same predefined populations always shared similar membership coefficient in inferred clusters. When the number of clusters was set to five, the genetic clusters correspond to the five geographic regions (Africa, East Asia, America, Oceania and Europe) very well [53]. In order to take advantage of the genome-wide high-density data, many computationally efficient software and algorithms have been developed[54, 55]. Based on a new method, it is found that seven genetic clusters correlate well with the seven major geographic regions, namely African, Middle East, Europe, Central/South Asia, East Asia, America and Oceania[56].

Multivariate techniques have been used to condense information of numerous loci into one or a few synthetic variables, which are especially powerful in analyzing the genome-wide high-density data. Principle component analysis (PCA) has been introduced to population genetics by Cavalli-Sforza and his colleagues[57] >30 years ago. Interests in PCA were renewed after Patterson *et al.*[58] had implemented it on individual genotypic data and plotted individuals on the graph. McVean[59] showed that for SNP data, the projection of samples onto the PCs could be obtained directly from the average coalescent times between pairs of haploid genomes. These results provided a framework for interpreting PCA projections in terms of underlying processes, including migration, geographical isolation, and admixture. McVean also demonstrated a link between PCA and Wright's F_{ST}[59]. Reich et al.[60] suggested that PCA was very useful in population genetics and highlighted three applications: detecting population substructure, correcting stratification in association studies and making qualified inferences about human history. For example, the first PC map based on European populations showed a southeast-to-northwest cline and was interpreted as the reflection of Neolithic farming spreading from the Levant to Europe about 6,000-9,000 years ago[57]. And the hypothesis about the expansion of Neolithic farming has been supported by many genetic and archaeological data[60, 61]. However, according to the study by Novembre and Stephens[62], PCs correlating with geography do not necessarily reflect major population migrations but isolation by distance, in which gene exchanges are only among neighboring populations. For example, based on the dataset of 3,000 individuals genotyped at over half a million SNPs, Novmbre et al.[63] found that the inferred principle components essentially reconstructed the geographic map of European, thus they suggested that individual genome could be used to infer the geographic origin of the individuals.

3.4 Integrating haplotype information to infer population relationships and history

Recently, methods and statistics that integrate LD and haplotype information to infer population relationships and history have been developed, which may become the mainstream of population genetics in the future. The studies integrating haplotype information essentially have greatly improved our knowledge on the human history and population relationships. For example, the LD information has been integrated into STRUCTURE as linkage model to estimate the ancestry along chromosome[51]. The average length of the migrant chromosome tracts were used to infer the change of recent gene flow in different populations[64]. In particular, based on a copying model adapted from Li and Stephens[65], the worldwide LD pattern of human was used to infer human origins and dispersals[66]. The study also found some new points on the human history such as the most northerly East-Asian population (Yakut) having received genetic contribution from the ancestors of north European. The copying model was further used to estimate parameters of population split, which illustrated LD pattern carrying historical information beyond recent migration[67].

Haplotype-sharing, which accounts for LD and haplotype information, has been proposed to infer the human history[68, 69]. For example, haplotype analysis has been used in the study on human genetic diversity in Asia[69]. With diversity decreasing from south to north, haplotype diversity was found strongly correlated with latitude (R^2 = 0.91, P < 0.0001), which was constant with a loss of diversity as populations moved to higher latitudes. Besides, more than 90% haplotypes in East Asian population could be found in Southeast and Central-South Asian populations, of which about 50% were found in Southeast only, and 5% in Central-South Asian only. Phylogenic analyses of private haplotypes indicate greater similarity between East Asian and Southeast Asian, suggesting that Southeast Asia was a major geographic source of East Asian population. Another example, although Uyghurs have been proposed as a genetic donor of the East Asian[66], haplotype-sharing analysis of Uyghur showed that more than 95% of Uyghur haplotype could be found in either European or East Asian population, which contradicts the expectation of null hypothesis. Simulation studies further indicated that the proportion of Uyghur private haplotype observed in the empirical data is only expected in alternative models assuming Uyghur is an admixed population[68].

Furthermore, Xu and Jin[12] proposed chromosome-wide haplotype sharing (CHS) as a measure of genetic similarity between human populations, which was an indirect approach to integrate recombination information. They showed that recombination and genetic differences between human populations are strongly correlated in both empirical and simulated data, indicating that recombination events in different human populations are evolutionarily related. They further demonstrated that CHS could be used to reconstruct reliable phylogenies of human populations and the majority of the variation in CHS matrix could be attributed to recombination[12]. However, for distantly related populations, the utility of CHS to reconstruct correct phylogeny is limited, suggesting that the linear correlation of CHS and population divergence could have been disturbed by recurrent recombination events over a large time scale. The CHS they proposed is a practical approach without involving computationally challenging and time-consuming estimation of recombination parameter. The advantage of CHS is rooted in its integration of both drift and recombination information, thus providing additional resolution especially for populations separated recently.

4. Natural selection and human adaptation to local environments

4.1 Genome-wide detection of natural selection

One of the most exciting prospects of genome-wide high-density polymorphic data is its implication in detecting natural section. Before the advent of genomic era, detecting natural selection was exclusively through candidate gene studies, in which a gene was priori hypothesized subjected to natural selection and the value of statistic on it was compared with that under neutrality[70]. However, besides the inefficiency, there are three significant limitations of candidate gene studies. Firstly, a priori hypothesis under which gene has been subjected to natural selection requires priori understanding of the gene functions or genotype-phenotype relationships. Genetic basis for most traits/phenotypes remaining mysterious limited our ability to intelligently nominate candidate genes for investigation[71]. Secondly, since demographic events such as population expansion can leave footprints on the genome similar to that of natural selection, the statistical power of candidate study for identifying genes under natural selection is confounded[72]. Thirdly, candidate gene study is especially inefficient in detecting selection in regulatory elements that are far from coding regions.

The availability of genomic data has offered a new paradigm for detecting signature of natural selection, which was referred to as genomic approach. In addition to the high throughput, it also shows several statistical advantages[70, 71]. First, it searches the whole genome without a priori hypothesis which gene is under selection, yielding a set of unbiased results. Second, since demographic history affects genomic loci equally while natural selection only affect a few loci, it provides a framework to distinguish natural selection signals from demographic history in principle. Third, genomic approach does not need a priori knowledge about the gene functions and genotype-phenotype relationships. With all these being said, genomic approach is extremely efficient and powerful compared with the traditional candidate approach. The most commonly used genomic approach is the "outlier of summary statistic"[71], in which a large number of sites are examined, calculating a statistic across all loci that quantifies a specific features of genetic variations, constructing empirical distribution, and defining selection candidate based on the extreme tail of the empirical distribution. Although simulations under neutrality have been conducted either to guide the definition of thresholds or to evaluate the efficiency of the outlier method, criteria for defining outlier in most studies is often arbitrary, e.g., loci falling in 99th percentile of the empirical distribution[71].

Up to now, hundreds of genome-wide studies have been conducted to detect natural selection in human, and the initial maps of positive selection in human have been produced.

Although these maps are incomplete, error-prone, and of low-resolution, it is no doubt that this progress has fundamentally changed the field of human population genetics and evolutionary studies.

4.2 Genetic basis and statistics for detecting selection using genome-wide data

As we know, selective sweep acts on the beneficial allele and makes its frequency increase quickly, and ultimately affects a large region around it due to LD. Therefore, the features of genetic variation subjected to natural selection in a region would be different from those evolve neutrally. Various methods have been developed, taking advantage of the footprints left by selection, such as changed allele frequency spectrum, increased derived allele frequencies, polymorphism deviating from interspecies divergence, population differentiation and exteneded haplotype homozygosity. However, we only introduce the recently developed methods and those whose statistical power greatly benefit from the genome-wide high-density data.

1. Extended haplotype homozygosity (EHH)

Under neutral model, it takes a long time for a new mutation to drift to high frequency in a population, during which time LD around this variant will decay substantially due to recombination[73, 74]. However, positive selection leads to rapid increase of the frequency of beneficial allele, occurring in such a short time that recombination does not have time to break the selected haplotype. Thus, an allele having extremely long LD compared with its counterparts should be seen as a signal of recent positive selection. Based on this principle, many different statistics such as EHH and integrated haplotype score (iHS) were developed[74-77]. EHH[74] is defined as the probability that two randomly chosen chromosomes carrying a tested core haplotype are homozygous at all SNPs. While iHS compares the EHH decay around an ancestral and derived allele[76]. Compared with other approaches, the most obvious advantage of these approaches is that they are relatively robust in choosing genetic markers or ascertainment bias[74]. Although it is pretty powerful to detect recent positive selection, it has little power in revealing natural selection happened about tens of thousands of years ago, because most chromosomal segments will be split into small pieces less than 100kb by recombination after 30, 000 years.

2. Population differentiation

Population differentiation is largely determined by population demographic history and genetic drift[78], which almost affect each locus similarly. However, variations of local environment impose different selection pressures on some genomic regions, leading to high population differentiation at these regions[79]. The first genome-wide study on positive selection taking advantage of population differentiation was based on the locus-specific F_{ST}[79]. Compared with others, this method is much more powerful in detecting ancient selective sweep that happened tens of thousands of years ago[80]. Another strategy based on population differentiation was the cross-population extended haplotype homozygosity (XP-EHH)[77], which is much powerful in detecting recent selection less than 1,000 generations. Recently, Chen *et al.*[81] developed a new method for detecting selective sweeps that involves jointly modeling the multi-locus allele frequency differentiation between two populations.

These methods were much robust to both ascertainment bias and recombination rate heterogeneity. However, these approaches cannot be directly applied on recently admixed population due to the confused population structure. Jin et al.[82] developed a new strategy to detect natural selection in admixed population such as African American, in which they reconstructed an ancestral African population (AAF) from African components of ancestry in African American and compared the population differentiation between indigenous African and AAF. Many targets of selection identified by this approach were associated with African-Americans specific high-risk diseases such as prostate cancer and hypertension, suggesting an important role these disease-related genes might have played in adapting to new environments[82].

3. Biased ancestry contribution in admixed populations

Under neutral evolution, the locus-specific ancestral contribution for all loci were similar, if not equal, as genetic drift on all loci were simultaneous[83]. However, a beneficial allele from an ancestry may provide higher survival or reproduction capability for the admixed individuals, leading to the increase of ancestral contribution from population carrying the beneficial allele. Thus, some genomic regions in admixed population might show excess of a particular ancestry, possibly attributable to selection pressures after the population admixture[84]. By far, several studies have used the genome-wide data to scan for signatures of selection in admixed populations[82].

4. Composite strategies and new challenge

Although hundreds of regions subjected to positive selection have been identified, most of the underlying genes and causal mutations are still unknown. A recent study by Grossman *et al.*[85] analyzed five statistics (F_{ST}, XP-EHH, iHS, ΔiHH, ΔDAF) and found their values were highly correlated only around the causal variants. Based on this observation, composite of multiple signals (CMS), a composite likelihood test, was proposed to distinguish causal variants. For each statistic *i*, the probability *P* of a score s_i was estimated whether selected or not. Assuming a uniform prior probability of selection π, the CMS score is the approximate posterior probability that the variant is selected:

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

Application of CMS to HapMap data has localized population-specific selective signals to 55kb, and identified known or novel causal variants[85]. It is novel to use composite strategy to identify causal variants, although several studies have proposed statistics considering multiple factors or combining several summary statistics.

Most aforementioned statistics and strategies detecting natural selection were developed according to the classic selective model in which a new beneficial mutation arises and increases in frequency to fix in a population[70, 71]. However, using the 1000 Genome Project data, Hernandez *et al.*[86] showed that classic selective sweeps were rare in recent human evolution, which indicates that many statistics may loose power in detecting selections that have not been modeled and considered. Thus it is a big challenge to depict, model and detect the signatures of natural selection in recent human evolutionary history.

4.3 Human adaptation to local environments

There are many completely distinguished environments on the earth, varying in temperature, moisture, light, and so on. Most animals and plants can only survive in a specific environment, even chimpanzee, our closest relative, has very limited distribution in Africa. Modern human, however, have successively colonized almost every corner of the earth within only about 100,000 years since the first branch of modern human left Africa[87]. Although the current patterns of human genetic and phenotypic diversities can, to some extent, be explained by human migration and demographic history[56], natural selection and adaptive evolution have also played very important roles in different populations adapting to local environments, as revealed by modern genetics[77, 88].

Natural selection acts on the phenotypes only when their differentiations correspond with different survival or reproductive rates, which makes the advantageous phenotypes and their underlying genetic variants become more common in a population. Over time, this process results in adaptive evolution, allowing organisms to handle the challenges from the environment to ensure their survival and reproduction. For example, animal is initially able to accommodate to the environmental challenges by simply changing their behaviors such as daily activities, and if the behavioral flexibility does not work, a range of physiological mechanisms accompanied by regulations of gene expression, such as accumulation of body fat, will act to relieve the environmental pressures. However, if the aforementioned mechanisms fail to buffer against the environmental challenges, the survival and reproductive rates began to vary from individual to individual. In this case, individuals harboring advantageous allele are more likely to have higher survival rates or have more descendents, thus frequencies of the advantage alleles increased accordingly, giving an indication of natural selection and adaptive evolution.

As a key mechanism of modern evolution, natural selection, supported by various scientific evidences, has been widely accepted at present. Although our species was not discussed in *On the Origin of Species* published in 1859 considering the religious Europe, the theory of evolution has gradually been applied to understanding human variation in the following years. Identifying targets of positive selection in human based on candidate gene studies has been frustratingly slow for just a decade. Recently, with the availability of large-scale genotyping and sequence data, we have experienced an explosive increase of studies on genome-wide scanning for signals of selection[71, 89]. Identification of genomic regions showing evidence of natural selection has helped us to find genes adapting populations to pathogens, climate, diet and possible cognitive challenges. These discoveries have greatly enriched our understanding of human origins and history, and hold large potential for identifying genes with important biological functions, thus in turn, will elucidate the genetic basis of some human diseases[71, 77]. These studies together provide many new insights into the natural selection process and mechanisms, which will ultimately improve the modern evolution theory.

4.4 Human adaptation to high-latitude climates

The early migration out of Africa exposed ancestral populations to colder environments with less sunlight, which eventually left the most obvious footprints on human populations[90]. Human morphology such as body mass, body mass index (BMI), nasal size, hair texture and density, lip size and thinness, relative sitting height (RSH) and surface area/body mass ratio have been reported shaped by climate adaptation[91]. Among all these characteristics, skin pigmentation is perhaps the most conspicuous one, with darker-skinned populations concentrated in the tropics, and lighter-skinned populations in higher latitudes[92, 93]. In fact, skin pigmentation is primarily determined by the type, amount and distribution of melanin. The global distribution of melanin can be explained by the balance selections interacting between elusion of ultraviolet radiation and photosynthesis of vitamin D_3[92, 93]. In tropic regions, the dark melanin protects people from sunburn by scattering and absorbing ultraviolet radiation, as well as limiting photo-degradation of nutrients such as folate. In this case, any mutation that impacts normal dark melanin production is deleterious, therefore, genes involved in dark melanin production such as *MC1R* are subjected to strong purifying selection in African[94]. Contrarily, the situation is very different for non-African including European, East Asian, and Southeast Asian, where *MC1R* is highly polymorphic, containing many nonsynonymous variants. In higher latitude, with less sunlight available, depigmentation may be favored since ultraviolet penetration is necessary for vitamin D_3 synthesis. Therefore, those genes that are associated with light skin such as *SLC45A2* and *OCA2* have been subjected to recent positive selection[95]. In order to cope with the ultraviolet radiation in temperate zone, human have developed a complex tanning system that includes immediate pigment darkening and delayed tanning reaction.

Although polymorphisms in *ASIP* and *OCA2* may play a shared role in shaping light skin around the world[95], it seems that the evolutions of light skin color in East Asia and western Eurasia are independent of each other, with many different genes underlying the environmental adaptation[77, 88, 95]. Based on extended haplotype homozygosity (EHH) or extremely high population differentiation using high density SNPs data, it is revealed that *SLC45A2* and *SLC24A5* have been subjected to strong positive selection in western Eurasia, while *EADR* and *ED2R* in East Asia and America[77]. Especially, the global distributions of both *CLC24A5* A111T polymorphism and *EDAR* V370A polymorphism correlate very well with the ethnic skin characteristics.

After long evolution since human out of Africa, the general populations have adapted to local environments to some extent. However, due to recent human migrations and colonization, many people lived in geographic region with completely different environmental conditions compared with their ancestry, which led to many health problems. For examples, fair skinned individuals living in low latitude are at much higher risk of skin cancer, while dark skinned individuals living in high latitude are at higher risk of vitamin D deficiency. Especially, European Australian are at about 10 times higher risk of several types of skin cancers than Australian aboriginals[92], while African American have the highest risk of vitamin D deficiency[96].

Meanwhile, our ancestry in tropic Africa also evolved a series of mechanisms to accommodate to the local environment (heat-adaptation), which include cooling through sweating[90]. As we know, sweating is accompanied by salt loss at the same time. Therefore, the low dietary salt in local environment ultimately leads to the selection for salt retention. After blood volume was depleted as a result of water loss, individuals with stronger arterial tone and cardiac contractile force are more likely to survive. However, this advantage may turn out to be negative when population migrated to temperate climate and will ultimately lead to hypertension. This has been supported by the genetic evidences from worldwide populations, including the functional alleles of seven genes, which are associated with distribution of blood pressure showing a latitudinal cline[97]. It is shown that *GNB3* 825T (one allele) accounts for a remarkable 64% of worldwide variation in blood pressure. Peoples from tropic climate migrating to temperate climate recently show high susceptibility to hypertension, which is represented by African-Americans[98, 99].

4.5 Human adaptation to high-altitude

Although latitude cline is the main pattern of human phenotype variations, another impressive example is the high-altitude adaptation. The environmental challenges for human survival and reproduction in high altitude include deceased ambient oxygen tension, increased ultraviolet radiation, extreme diurnal ranging in temperature, arid climate and so on. In particular, high-altitude hypoxia, caused by the decreased barometric pressure in high altitude, cannot be overcome simply by behavioral and cultural modification. However, there are approximately 140 million individuals living permanently at high-altitude (above 2500 meters) in North, Central and South America, East Africa, and Asia[100, 101]. Populations living at the high altitude, especially those living at Tibetan Plateau and Andes, have evolved unique physiological characteristics compared with each other and with low altitude populations[102]. It is known that Andean population, as well as high-altitude sojourners, demonstrated higher hemoglobin concentration than low-altitude populations. In contrast, Tibetan populations exhibit lower hemoglobin concentrations than expected[102]. The physiological characteristics of the high altitude residents also include blunted ventilatory response to acute hypoxia, protection from altitude-associated fetal growth restriction and so on[100, 103].

Although the physiology of these populations has been well described for hundreds of years, the genetic basis of these traits has not been revealed until recently. Several studies, using different technologies and strategies, have identified the genes that adapted Tibetans to the local environment based on genome-wide data[80, 101, 104, 105]. The candidate genes identified by these studies are essentially consistent with each other. The strongest ones are *EGLN1* (also known as *HIFPH2*, located in 1q42.2) and *EPAS1* (also known as *HIF2A*, located in 2p21), both of which are involved in HIF pathway and response to hypoxia[80, 104, 105]. Xu *et al.*[80] analyzed the local linkage disequilibrium (LD) of the two HIF genes and ultimately found the Tibetans dominant haplotype. Based on the significant overrepresentation of the carrier of dominant haplotype of *EPAS1* and *EGLN1* in Tibetans, they proposed a "dominant haplotype carrier" model to explain the roles of the two genes in adapting to high altitude[80].

4.6 Human adaptation to shifted diet

Diet is one of the most important factors in species evolution and has been highlighted in *On the Origin of Species*. Our ancestry had adapted their genome to the food from local environment in evolutionary history. However, facilitated by the development of stone tools, the master of fire and the recent domestication of plants and animals[106, 107], human evolution is characterized by significant dietary shifts. Especially, the diet and lifestyle conditions have been changed fundamentally since the introduction of agriculture and the industrial revolution. The conflicts between the genetically determined biological features and contemporary diet and lifestyle may lead to the diseases of civilization[108]. A common view is that post-Neolithic human adapted, through a 'thrifty genotypes', to a hunter-gather lifestyle of feast and famine[109, 110]. However, studies based on the recent available genome-wide data showed that the real case might be complex[111].

Lactase persistence of human is one of the best-studied examples of dietary adaptation. Lactase-phlorizin hydrolase (*LPH*) is predominantly expressed in small intestine, where it hydrolyzes lactose into glucose and galactose that can be easily absorbed[112]. In human, the capability to digest lactose, the main carbohydrate in milk, declines rapidly after weaning because of the decreasing levels of *LPH*. However, using lactose as a source of food and nutrient in adulthood provides some survival advantages for populations mainly on dairy food. It is found that distribution of lactase persistence correlates well with that of populations with a history of cattle domestication and milk drinking[113]. The lactase persistence is inherited as a dominant mendelian trait, and is thought caused by change of *cis*-acting element of *LCT*, the gene encoding LPH[114]. A linkage disequilibrium (LD) and haplotype analysis of Finnish pedigrees has identified a causative regulatory variant (C/T-13910) ~14kb upstream of *LCT* gene[112], with an estimated age of about 2,000-20,000 years[115]. A region of extensive LD spanning >1M has been observed in European chromosome with T-13910 allele, which is consistent with recent positive selection[10, 115, 116]. However, lactase persistent populations elsewhere such as African do not carry this variant[117]. Association studies on Tanzanians, Kenyans and Sudanese identified another three variants (G/C-14010, T/G-13915 and C/G-13907) that could lead to lactase persistence. These SNPs originated on different haplotype backgrounds from European C/T-13910 and from each other, which indicated the independent origin of lactase persistence[117]. Genotyping across a 3-Mb region demonstrated haplotype homozygosity extending >2Mb in chromosomes carrying C-14010, which is consistent with a selective sweep about 7,000 years ago. The different origins of lactase persistence also provide a perfect example of convergent evolution due to strong selective pressure as the shared dietary culture.

However, starch consumption is the prominent characteristic of agricultural societies, especially among the populations living on planting. Interestingly, *AMY1* (human salivary amylase gene) has been reported subjected to recent positive selection in populations with planting tradition, contrasting to *LCT* in population with stockbreeding tradition[106, 117], which might reflect the influence of different kinds of agricultures and cultures. Copy number of the salivary amylase gene (*AMY1*) are corrected positively with salivary amylase protein level and individuals from population with high-starch diets[106]. Thus individual with more copies of *AMY1* is presumably able to get more out of their starchy diet, thus providing survival advantage when food is limited. It is also suggested that higher *AMY1* copy number and protein levels might also buffer against the fitness-reducing effects of intestinal disease[106].

4.7 Adaptation to pathogens and its influence on defense genes

Infectious diseases have always been a massive burden in human evolutionary history. The human life expectancy was <25 years until the control of infections by improved hygiene, vaccines and antibiotics following the advent of Pasteur's microbial theory of disease[118]. Being a major cause of human mortality, pathogens also imposed strong selective pressure on the human genome. However, the relationship between pathogens and natural selection had not been established in a long time span, until John B. S. Haldane found the link by analyzing thalassaemia patients infecting malaria[119], one of the best examples showing how infectious disease shapes human genome and how natural selection is working. The pathogen adaptation process is more complex than any of the other kind of selection pressure as the evolutionary dynamics of host-pathogen interactions lead to constant selection for adaptation and counter-adaptation in the competing species. Through the contending with pathogens, human have to improve their immune defense mechanism to combat microbial infection.

Although human and chimpanzee split only about 3 million years ago, prevalence and severity of infectious disease such as HIV, *Plasmodium malaria*, hepatitis B, hepatitis C and influenza A between humans and non-human primates are different[120, 121]. In human populations, many genes involving in infectious diseases have also been revealed subjected to natural selection. Especially, many studies have demonstrated the correlation between genetics variability in human population worldwide and pathogen richness in the corresponding geographic regions[122, 123]. Genome-wide scans for positive selection have detected >5,000 genomic regions that present at least one genomic signals of positive selection[124]. When we focused on natural selection that occurred more recently (detected by integrated haplotype score and LD decay test), defense genes were found over-represented[76, 125]. These observations may indicate that our immune system has particularly been challenged during the recent phases of human evolution, which might propose a strong burden of infectious disease that are associated with the advent of agriculture at the beginning of the Neolithic period 10,000 years ago[126]. In this situation, genome-wide association studies (GWAS) has become a powerful tool in detecting loci associated with the susceptibility or severity of infectious diseases. These susceptibility genes identified in this way are targets, transports or some other components in the pathogen infectious pathway. Thus, careful analysis of the pathogens associated genes will finally illuminate the infectious process and the targets of selection.

The model in which human adapts to pathogens is very complex and dependents on a lot of factors, including the type of microorganism, the different temporal and spatial presence of pathogens during evolution, their varying pathogenicity, the nature of the host-pathogen interaction, and the rate at which pathogens evolve[124]. A study on Toll-like receptor (*TLR*) gene family has concluded that viruses have exerted stronger selective pressures than other pathogens by constraining amino acid diversity at viral recognition TLRs[127]. Although the immune-related genes played a role in protecting the host from infection, mutations that inactivate these genes are likely to represent a selective advantage for the host when a pathogen uses the host's immune receptors as a mechanism of cell entry and survival. Some of these genes have lost their function because of the strong selection pressure, which also provides insights into the degree of redundancy in our immune system[128]. Loss-of-function mutations in *CCR5*, *DARC*, *CASP12*, *SERPINA2* and *SIGLEC12* are such examples. However, the selection may be very complex sometimes considering the changing pathogens. For example, CCR5-Δ32 allele is a deletion mutation of *CCR5* gene that impairs the function of its coding protein and has a specific impact on the function of T cells. It has been subjected to positive selection in Europe and can block the entrance of HIV-1[129]. However, since HIV has not emerged in Europe until recently, the selection signals on *CCR5*-Δ32 might be caused by Black Death or/and smallpox[130]

There are many well-studied examples about the natural selection imposed by pathogens, among which the selection imposed by malaria may be the best. Malaria has been, and still is, one of the major causes of child mortality in tropical regions[131]. Because of the strong selective pressure, malaria has become the driven force of most common Mendelian diseases including sickle-cell anemia, α-thalassemia, β-thalassemia, glucose-6-phosphatase (*G6PD*) deficiency and so on. However, these erythrocyte variants are probably only the tip of the iceberg considering all the genes associated with susceptibility and resistance to malaria, many of which are involved in immune system and inflammatory genes[132]. All these evidences suggested that malaria was the strongest known force of evolution in recent human history. The observations that different malaria-resistance alleles arose in different regions suggested independent evolutionary history of these genes[132].

5. Complex diseases/traits: Genetic basis and identification of the underlying variants

5.1 Overview of complex diseases/traits

Complex traits (or multifactorial traits) refer to phenotypes that vary in degree and can be attributed to the effects of multiple genes in combination with environmental factors. Generally, complex traits contribute to what we see as continuous characteristics in organisms, such as height, skin color, and body mass, whose inheritances do not follow Mendel's law. Most human diseases, such as diabetes, hypertension and various cancers, can be thought as some special complex traits, and are also associated with multiple genes and environmental factors. In fact, most studies only focus on the complex diseases due to their great health implications. Current debates concerning the genetic basis of complex diseases focus on two hypotheses: Common disease common variants (CDCV) and common disease rare variants (CDRV)[133, 134]. The CDCV hypothesis argues that the major genetic susceptibility to the complex/common disease are variations with appreciable frequency in the population, but relatively low penetrance[135, 136], while the alternative CDRV hypothesis argues that multiple rare variations with high penetrance are the major contributors[137, 138].

Technology advance on DNA microarray has lured many scientists genotyping high-density SNPs on thousands of individuals. Up to now, GWAS have reproducibly identified thousands of genes associated with complex diseases/traits[139, 140], providing many new insights into the functional and biological networks of these genes. However, among these heritable components of complex disease, only a small fraction has been explained. A potential source of the majority of missing heritability is the contribution of rare variants. Although next generation sequencing has the potential to discover the entire spectrum of sequence variation in well-phenotyped individuals, it remains a challenge to develop efficient methods to integrate the rare variants and eliminate the effects of sequence error and missing data.

5.2 Models for allelic spectrum of complex diseases

Allelic spectrum is the total variations that contribute to a disease, including common variants (frequency >1%), rare variants (frequency <1%), high penetrance and low penetrance. The allelic spectrum of complex disease has important influence on both research and clinical practice. In brief, it determines the strategies and methodologies for disease gene discovery. Although various methods and statistics have been developed by simply assuming high or intermediate allele frequency, allelic spectrums of the discovered susceptibility are seemingly to be complex. Since human population experienced complicated demographic history and a series of local adaptations, it is still obscure how human history affects the allelic spectrum of complex disease.

Despite the fact that several studies have promoted the formulation of the common disease common variants (CDCV) hypothesis, it turns out to be a complete hypothesis after the publication of the allelic spectrum of human disease by Reich and Lander[141], in which they used empirical data to qualify the allelic spectrum of rare diseases and studied the changes of allele frequency under rapid population expansion. As a result, they found that CDCV is not incompatible with the reported susceptible variants and diseases. They predicted that the overall frequencies of disease alleles are not low and it is possible to use variants with allele frequency above a threshold to detect the susceptible loci. The ancestral-susceptibility model provides an evolutionary framework to explain the susceptibility alleles to be ancestral alleles (mostly common alleles)[135]. According to this model, the ancestral alleles reflect ancient human populations adaptation, whereas the derived alleles were deleterious. However, with the shift of environment and lifestyle, the ancestral alleles increase the risk of common diseases in modern populations.

Many studies have challenged CDCV hypothesis and supported the alternative hypothesis common disease rare variants (CDRV). For example, Pritchard[137] argued that population processes such as mutation, genetic drift and purifying selection are against the deleterious alleles, which reduce the frequency of disease causal alleles. In this context, common variants, with an appreciable frequency, tend to be older and are unlikely to be subjected to long-lasting purifying selection in the whole evolutionary history. However, rare variants are either new mutations or being selected against owing to their deleterious nature. Furthermore, studies have found individuals with extreme values of quantitative phenotypes significantly cherished more rare missenses variants in candidate genes and pathways[138]. The evolutionary explanation for CDRV hypothesis is selection-mutation balance model, in which most missense mutations are deleterious[142].

Based on previous observations, a decanalization model was proposed to explain the origin of complex diseases[143]. It is known that stabilizing selection drives species to an optimum that takes an intermediate value among all possible values of phenotype. In the case of complex diseases, canalized traits will influence fewer individuals than those that lack a mechanism to reduce susceptibility. Importantly, persistent stabilizing selection not only removes the risk allele, but also reduces the additive genetic effects of alleles that are present in the gene pool or arisen by mutations[144]. In decanalization model, increased variants broaden the normal distribution, which leads to some individuals excess liability threshold, rather than horizontal shift of entire distribution. Since changes of epistasis interaction must have happened in decanalized case, it is expected to observe interactions among risk alleles. As risk alleles are not selected against in normal individuals, the canalized system can facilitate genetic drift. Therefore, the risk alleles underlying complex disease may be very similar to the normal allele frequency spectrum, and gene-gene interaction and genetic pathway must be considered in the identification of risk allele.

In fact, there are supporting evidences for each of these hypotheses[133, 142, 143, 145]. Although there may be some artifacts and most causal alleles are unknown[146], hundreds of GWAS have identified thousands of high frequency susceptible alleles[139], indicating that CDCV does have its place. As for CDRV, rare mutations in *ANGPTL4* caused reduce of triglycerides[147], while rare independent mutations in renal salt handling genes (*SLC12A3*, *SLC12A1* and *KCNJ1*) contributed to blood pressure[148]. Since next generation sequencing has become affordable recently, reports on rare variants are accumulated quickly. Type 2 diabetes (T2D), immune disorders and psychological disorders have been used as examples of decanalization model[143].

5.3 Genome-wide association studies (GWAS)

The availability of high-throughput genotyping technologies, as well as the major efforts to identify genome-wide genetic diversity, has made the genome-wide association studies (GWAS) become possible. Especially, HapMap provided nearly 4 million SNPs and characterized the genome-wide LD pattern and haplotype map[9, 10]. The debates on marker selection that determines the genomic coverage at early stage have boiled down to choose a limited range of commodity chips. Although higher density will increase genomic coverage, it does not equate increasing much power in most cases. With limited funding, the overall power might be maximized by genotyping more samples using less dense and less costly array[149]. Hundreds of genome-wide association studies (GWAS) have been conducted to identify common variations that are statistically associated with particular diseases[9, 139]. The first wave of large-scale GWAS has improved our knowledge on genetic basis of many complex traits/diseases[150]. For example, we have witnessed rapid expansion in numbers of susceptible loci for some diseases/traits, such as type 1 diabetes, type 2 diabetes, prostate cancer, inflammatory bowel disease, breast cancer, height, fat mass and lipid[139, 149]. These findings have provided many valuable clues to the allelic architecture of complex traits.

Most GWAS have featured case-control designs, which has raised issues about the selection of suitable cases and controls. Optimal selection of both cases and controls is important due to the fact that they will seriously affect statistical power of GWAS. Case selection has mainly focused on improvement of statistical power by enriching specific disease-predisposing alleles including minimizing phenotypic heterogeneity. Optimal selection of control samples remains more controversial, although the accumulating empirical data indicate that many commonly expressed concerns have been overstated[149]. One economic approach is to use common-control to study a series of diseases/traits such as what Wellcome Trust Case Control Consortium (WTCCC) has done[150]. Another more economic strategy is to use the genetic matched controls in public available control resources such as Illumina iControlDB (www.illumina.com). As for the sample size, the consensus view is clear: the more the better. Increasingly, GWAS are being extended from case-control designs to population-based cohorts which offer longitudinal measures of a wide range of quantitative traits and integrate the environmental factors for systematic analysis[149].

Another potential challenge for GWAS are the presence of undetected population stratification that can mimic the signals of association, thus leading to false positive and missing statistical power[151-153]. Although the influence of population structure caused by continental differentiation is serious, these outliers can be easily removed[150]. Therefore, the residual work on population substructure should focus on identifying the cryptic population stratification in a region or an ethnic group. As for the population stratification in European such as the north-south cline, differences between Jewish and non-Jewish have been revealed[154, 155]. Even in one single ethnic population such as Han Chinese, the obvious population stratification such as north-south cline is observed[156, 157]. Several methods and statistics have been developed to detect and adjust the population stratifications, even for continental structure such as African American[151, 158].

Conclusively, GWAS are a very powerful tool in investigating the genetic basis of complex diseases/traits. It is undoubtedly that it represents an important advance compared with 'candidate gene' studies in which limited variants and samples yielded many non-replicated results. Based on a threshold of p-values < 1.0 x 10^{-5} and studies with >100,000 SNPs in the initial stage, overall 5,053 SNPs have been reported to be associated with hundreds of complex traits by September 24, 2011 (www.genome.gov) [139]. The deluge of GWAS also provided the opportunity to evaluate the potential impact of genetic variants on complex diseases by systematically cataloging and summarizing the characteristics of the identified trait/disease associated SNPs (TASs). Unsurprisingly, since GWAS were primarily powered for common variants, risk allele frequencies were well above 5% (interquartile range 21%-53%) in the populations analyzed as well as in the HapMap populations (CEU: 21-54%; YRI: 13-65%; CHB+JPT:13-58%)[139]. Based on the position and function of 465 unique TASs, 43% SNPs were located in intergenic regions, 45% were intronic, 9% were nonsynonymous, 2% were in 5'UTR or 3'UTR, and 2% were synonymous[139]. The odds ratios (ORs) of discrete traits ranged from 1.04 to 29.4 (median 1.33, interquartile range 1.20 –1.61). Evolutionary analysis of the diseases associated genes showed they have been subjected to stronger positive selection compared with that of background[159].

Although GWAS are a great success, most of the variants having been identified so far only account for a small increment in risk and explain a small fraction of estimated heritability. For example, human height is a classic complex traits with an estimated heritability of about 80%, however, with tens of thousands of individuals having been studied, more than 40 associated loci identified by GWAS explain only about 5% of phenotypic variance [160]. These phenomena have led to the heated discussion on where the missing heritability of the complex disease can be found[161]. Many explanations have been proposed, including much more variants of smaller effect that have not been identified, that rare variants are not examined in the commercial chips, epistasis undetected, epigenetic and structural variants poorly captured. From this point of view, GWAS are just a beginning in systematically identifying the disease-associated loci.

5.4 Association studies for next generation sequencing

Next-generation sequencing has the potential to discover the entire spectrum of sequence variations and has been proved to be successful in the study of Mendelian disorders[162, 163]. When association studies on expression quantitative trait loci (eQTL) were conducted using all SNPs in low-coverage pilot of the 1000 genome project, a large number of more significant eQTLs was observed compared with traditional chips[5]. Thus it is no doubt that next generation sequencing will greatly benefit the studies of complex traits/diseases in the future. However, the applications of it on complex disease studies, which generally require sequencing hundreds or thousands of individuals, remain to be a challenge due to the high costs and limits of sequencing capacity.

In order to take advantage of the next generation sequencing, three strategies have been proposed: imputation, genotyping and low-coverage sequencing[5, 163]. First, imputation of previously genotyped samples using the recent sequenced reference panel is the most economic way, albeit less accurate. Analysis of ~400 samples imputed based on 1000 genome project data has revealed that the imputed data have more power than the original genotyped data[5]. Second, commercial chips integrating the new discovered SNPs will essentially improve the statistical power in identifying the disease-associated sites. Third, low-coverage sequencing of many individuals can be used to detect polymorphic sites and infer genotypes when many individuals are sequenced (2-6 x coverage)[163]. However, low-coverage data include very high sequencing error and missing data compared with high-coverage data, which is a big challenge for genetic variants discovery and statistics of association study.

Since next generation sequence detects millions of rare variants, these data have three features: high proportion of rare variants, high error and high missing data[134]. Thus traditional statistics, testing the association of common alleles one by one, are not suitable for large amount of allelic heterogeneity presenting in sequencing data[164]. In recent years, various statistics have been proposed to analysis the coming data with new features[134, 165-167]. For examples, Li and Leal[165] developed a combined multivariate and collapsing (CMC) method taking advantage of both collapsing and multiple-marker tests, and demonstrated that CMC was both powerful and robust using sequencing data. Price et al.[167] proposed a method for detecting association of multiple rare variants based on regression of phenotypic values on individuals' genotype scores, integrating computational predications of the functional influences of missenses. Luo et al.[134] used a genome continuum model and functional principal components as a general principle to develop functional principal component analysis (FPCA) statistic for sequencing data.

5.5 Admixed population and admixture mapping

Strictly speaking, almost all human populations showed some admixture features to some extent. However, we usually only refer to the populations with recent ancestry from two or more continents as admixed populations, most of which arise from the colonization of America and trans-Atlantic slave trade. A substantial proportion of the populations in the New World are recent admixed populations, such as African Americans, Mestizos, Puerto Ricans and other Latino/Hispanic populations. In fact, admixed population also distributed in other parts of the world, such as Uyghur in Central Asia[68, 168, 169], and populations that are of African-Indian origin in South Asia[170, 171]. Although genetic differences between populations only represent a small fraction of the total genetic variation, some diseases have different prevalence in populations owing to local adaptation or genetic drift[172, 173]. In admixed population, high population differentiation allele between parental populations may be risk for a disease with varying prevalence. Therefore, local ancestry differentiation can be used for disease gene discovery, namely admixture mapping[174].

The statistical power of admixture mapping comes from the fact that population admixture creates LD between loci with different allele frequencies in ancestral parental populations[174-176]. Since population admixture creates extended LD and chromosomal segments of distinct ancestry even extending several cMs in recent admixed population, only thousands of (about 1,500-5,000) high ancestry informative markers (AIMs) will be enough for a genome-wide admixture mapping[177]. Factors influencing the statistical power of admixture mapping, such as admixture dynamics and demographic history, have been investigated in various studies[172, 176, 178, 179]. Admixture mapping using AIMs is very important for holding the statistical power and reducing costs[180]. As a kind of specialized GWAS, design of admixture mapping can be either case-control or case-only, and in the later case, the local ancestry of disease cases is compared with the local ancestry elsewhere in the genome.

Since the selected AIMs have high population differentiation and are unlinked in each ancestry populations, Hidden Markov model (HMM) based approaches are used to infer the local ancestry and are implemented in several software packages, including ADMIXMAP, ANCESTRYAMP and MALDsoft[181]. Using admixture mapping, many complex diseases associated loci have been identified by selected AIMs[99, 182-184]. For example, admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men, which can be replicated by later GWAS[185]. However, admixture mappings based on economic AIMs do not account for high-LD between markers, which makes it less powerful than studies inferring local ancestry based on genome-wide high-density data[181]. In recent years, various methods such as SABER[186], HAPAA[187] and HAPMIX[188] have been developed to infer locus-specific ancestry based on high density SNPs data. Especially, HAPMIX employs an explicit population genetic model to infer local ancestry based on fine-scale variation data for populations formed by two-way admixture[188]. HAPMIX permits small rates of miscopying from the ancestral haplotype, modeling unphased diploid data from the admixed population with the HMM. Our simulations showed that HAMPIX performed better compared with other methods when very recent admixed population were investigated[82].

6. Acknowledgements

This work was supported by the National Science Foundation of China (NSFC) grants 31171218 and 30971577, by the Shanghai Rising-Star Program 11QA1407600, and by the Science Foundation of the Chinese Academy of Sciences (CAS) (KSCX2-EW-Q-1-11; KSCX2-EW-R-01-05; KSCX2-EW-J-15-05). This research was supported in part by the Ministry of Science and Technology (MoST) International Cooperation Base of China. Shuhua Xu is Max-Planck Independent Research Group Leader and member of CAS Youth Innovation Promotion Association. Shuhua Xu also gratefully acknowledges the support of K.C.Wong Education Foundation, Hong Kong.

7. References

[1] Clark, A.G., et al., *Ascertainment bias in studies of human genome-wide polymorphism.* Genome Res, 2005. 15(11): p. 1496-502.

[2] Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. 26(10): p. 1135-45.

[3] Reich, D.E., et al., *Human genome sequence variation and the influence of gene history, mutation and recombination*. Nat Genet, 2002. 32(1): p. 135-42.

[4] Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. 7(2): p. 85-97.

[5] Durbin, R.M., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. 467(7319): p. 1061-73.

[6] Przeworski, M. and J.K. Pritchard, *Linkage disequilibrium in humans: Models and data.* Am J of Hum Genet, 2001. 69(1): p. 1-14.

[7] Lewontin, R.C., *Interaction of Selection + Linkage .2. Optimum Models*. Genetics, 1964. 50(4): p. 757-&.

[8] Lewontin, R.C., *Interaction of Selection + Linkage .I. General Considerations - Heterotic Models*. Genetics, 1964. 49(1): p. 49-&.

[9] Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. 449(7164): p. 851-61.

[10] Altshuler, D., et al., *A haplotype map of the human genome.* Nature, 2005. 437(7063): p. 1299-1320.

[11] Rosenberg, N.A., et al., *A worldwide survey of haplotype variation and linkage disequilibrium in the human genome.* Nat Genet, 2006. 38(11): p. 1251-1260.

[12] Xu, S. and L. Jin, *Chromosome-wide haplotype sharing: a measure integrating recombination information to reconstruct the phylogeny of human populations*. Ann Hum Genet, 2011. 75(6): p. 694-706.

[13] Kong, A., et al., *A high-resolution recombination map of the human genome*. Nat Genet, 2002. 31(3): p. 241-7.

[14] Kauppi, L., A.J. Jeffreys, and S. Keeney, *Where the crossovers are: recombination distributions in mammals.* Nat Rev Genet, 2004. 5(6): p. 413-24.

[15] Nordborg M: *Coalescent theory*. In Handbook of Statistical Genetics. Edited by Balding DJ, Bishop M, Cannings C. Chichester, UK: John Wiley & Sons Inc; 2001:p. 179-212.

[16] Wakeley, J. 2008. Coalescent Theory: *An Introduction*. Roberts & Co , Greenwood Village, Colorado.

[17] Wiuf, C. and J. Hein, *Recombination as a point process along sequences.* Theor Popul Biol, 1999. 55(3): p. 248-259.

[18] Wiuf, C. and J. Hein, *The ancestry of a sample of sequences subject to recombination.* Genetics, 1999. 151(3): p. 1217-28.

[19] Griffiths, R.C. and P. Marjoram, *Ancestral inference from samples of DNA sequences with recombination.* J Comput Biol, 1996. 3(4): p. 479-502.

[20] Stumpf, M.P. and G.A. McVean, *Estimating recombination rates from population-genetic data.* Nat Rev Genet, 2003. 4(12): p. 959-68.

[21] Coop, G., et al., *High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among human*s. Science, 2008. 319(5868): p. 1395-8.

[22] McVean, G.A.T., et al., *The fine-scale structure of recombination rate variation in the human genome*. Science, 2004. 304(5670): p. 581-584.

[23] Posada, D. and K.A. Crandall, *Evaluation of methods for detecting recombination from DNA sequences: computer simulations*. Proc Natl Acad Sci U S A, 2001. 98(24): p. 13757-62.

[24] Wang, Y. and B. Rannala, *Population genomic inference of recombination rates and hotspots.* Proc Natl Acad Sci U S A, 2009. 106(15): p. 6215-6219.

[25] Rannala, B. and Y. Wang, *Bayesian inference of fine-scale recombination rates using population genomic data.* Phil Trans R Soc B, 2008. 363(1512): p. 3921-3930.

[26] Fearnhead, P. and P. Donnelly, *Approximate likelihood methods for estimating local recombination rates*. J R Stat Soc B, 2002. 64: p. 657-680.

[27] Hudson, R.R., *Two-locus sampling distributions and their application*. Genetics, 2001. 159(4): p. 1805-1817.

[28] Donnelly, P., et al., *A fine-scale map of recombination rates and hotspots across the human genome*. Science, 2005. 310(5746): p. 321-324.

[29] Wegmann, D., et al., *Recombination rates in admixed individuals identified by ancestry-based inference*. Nat Genet, 2011. 43(9): p. 847-53.

[30] Hinch, A.G., et al., *The landscape of recombination in African Americans.* Nature, 2011. 476(7359): p. 170-5.

[31] Wang, J., *Estimation of effective population sizes from data on genetic markers.* Philos Trans R Soc Lond B Biol Sci, 2005. 360(1459): p. 1395-409.

[32] Hill, W.G., *Estimation of Effective Population-Size from Data on Linkage Disequilibrium.* Genet Res, 1981. 38(3): p. 209-216.

[33] Hayes, B.J., et al., *Novel multilocus measure of linkage disequilibrium to estimate past effective population size*. Genome Res, 2003. 13(4): p. 635-643.

[34] Li, H. and R. Durbin, *Inference of human population history from individual whole-genome sequences*. Nature, 2011. 475(7357): p. 493-6.

[35] McVean, G.A. and N.J. Cardin, *Approximating the coalescent with recombination.* Philos Trans R Soc Lond B Biol Sci, 2005. 360(1459): p. 1387-93.

[36] Hey, J., *Isolation with migration models for more than two populations*. Mol Biol Evol, 2010. 27(4): p. 905-20.

[37] McHenry, H.M., *Human Evolution*, in *Evolution: The First Four Billion Years*, M.R.J. Travis, Editor. 2009, The Belknap Press of Harvard University Press: Cambridge, Massachusetts: p. 265.

[38] Cann, R.L., M. Stoneking, and A.C. Wilson, *Mitochondrial DNA and human evolution.* Nature, 1987. 325(6099): p. 31-6.

[39] Ke, Y., et al., *African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes.* Science, 2001. 292(5519): p. 1151-3.

[40] Vigilant, L., et al., *African populations and the evolution of human mitochondrial DNA*. Science, 1991. 253(5027): p. 1503-7.

[41] Green, R.E., et al., *A draft sequence of the Neandertal genome.* Science, 2010. 328(5979): p. 710-22.

[42] Reich, D., et al., *Genetic history of an archaic hominin group from Denisova Cave in Siberia.* Nature, 2010. 468(7327): p. 1053-60.

[43] Reich, D., et al., *Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania*. Am J Hum Genet, 2011. 89(4): p. 516-28.

[44] Rasmussen, M., et al., *An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia*. Science, 2011. 333(6052): p. 94-98.

[45] Nielsen, R., *Estimation of population parameters and recombination rates from single nucleotide polymorphisms*. Genetics, 2000. 154(2): p. 931-42.

[46] Williamson, S.H., et al., *Simultaneous inference of selection and population growth from patterns of variation in the human genome.* Proc Natl Acad Sci U S A, 2005. 102(22): p. 7882-7.

[47] Voight, B.F., et al., *Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes.* Proc Natl Acad Sci U S A, 2005. 102(51): p. 18508-13.

[48] Keinan, A., et al., *Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans.* Nat Genet, 2007. 39(10): p. 1251-5.

[49] Cavalli-Sforza, L.L., *The Human Genome Diversity Project: past, present and future.* Nat Rev Genet, 2005. 6(4): p. 333-40.

[50] Hey, J. and C.A. Machado, *The study of structured populations--new hope for a difficult and divided science.* Nat Rev Genet, 2003. 4(7): p. 535-43.

[51] Falush, D., M. Stephens, and J.K. Pritchard, *Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.* Genetics, 2003. 164(4): p. 1567-87.

[52] Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. 155(2): p. 945-59.

[53] Rosenberg, N.A., et al., *Genetic structure of human populations.* Science, 2002. 298(5602): p. 2381-5.

[54] Tang, H., et al., *Estimation of individual admixture: analytical and study design considerations.* Genet epidemiol, 2005. 28(4): p. 289-301.

[55] Pool, J.E., et al., *Population genetic inference from genomic sequence variation*. Genome Res, 2010. 20(3): p. 291-300.

[56] Li, J.Z., et al., *Worldwide human relationships inferred from genome-wide patterns of variation*. Science, 2008. 319(5866): p. 1100-4.

[57] Menozzi, P., A. Piazza, and L. Cavalli-Sforza, *Synthetic maps of human gene frequencies in Europeans.* Science, 1978. 201(4358): p. 786-92.

[58] Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* Plos Genet, 2006. 2(12): p. e190.

[59] McVean, G., *A Genealogical Interpretation of Principal Components Analysis*. Plos Genet, 2009. 5(10).

[60] Reich, D., A.L. Price, and N. Patterson, *Principal component analysis of genetic data.* Nat Genet, 2008. 40(5): p. 491-2.

[61] Semino, O., et al., *Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area.* Am J Hum Genet, 2004. 74(5): p. 1023-34.

[62] Novembre, J. and M. Stephens, *Interpreting principal component analyses of spatial population genetic variation*. Nat Genet, 2008. 40(5): p. 646-9.

[63] Novembre, J., et al., *Genes mirror geography within Europe.* Nature, 2008. 456(7219): p. 274.

[64] Pool, J.E. and R. Nielsen, *Inference of historical changes in migration rate from the lengths of migrant tracts*. Genetics, 2009. 181(2): p. 711-9.

[65] Li, N. and M. Stephens, *Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data.* Genetics, 2003. 165(4): p. 2213-33.

[66] Hellenthal, G., A. Auton, and D. Falush, *Inferring human colonization history using a copying model.* PLoS Genet, 2008. 4(5): p. e1000078.

[67] Davison, D., J.K. Pritchard, and G. Coop, *An approximate likelihood for genetic data under a model with recombination and population splitting*. Theor Popul Biol, 2009. 75(4): p. 331-45.

[68] Xu, S., W. Jin, and L. Jin, *Haplotype-sharing analysis showing Uyghurs are unlikely genetic donors.* Mol Biol Evol, 2009. 26(10): p. 2197-206.

[69] HUGO Pan-Asian SNP Consortium, et al., *Mapping human genetic diversity in Asia.* Science, 2009. 326(5959): p. 1541-5.

[70] Sabeti, P.C., et al., *Positive natural selection in the human lineage.* Science, 2006. 312(5780): p. 1614-20.

[71] Akey, J.M., *Constructing genomic maps of positive selection in humans: where do we go from here*? Genome Res, 2009. 19(5): p. 711-22.

[72] Stajich, J.E. and M.W. Hahn, *Disentangling the effects of demography and selection in human history*. Mol Biol Evol, 2005. 22(1): p. 63-73.

[73] Kimura, M., *The neutral theory of molecular evolution.* (United Kingdom: Cambridge University Press, Cambridge), 2003.

[74] Sabeti, P.C., et al., *Detecting recent positive selection in the human genome from haplotype structure*. Nature, 2002. 419(6909): p. 832-837.

[75] Tang, K., K.R. Thornton, and M. Stoneking, *A new approach for using genome scans to detect recent positive selection in the human genome*. Plos Biol, 2007. 5(7): p. 1587-1602.

[76] Voight, B.F., et al., *A map of recent positive selection in the human genome.* Plos Biol, 2006. 4(3): p. 446-458.

[77] Sabeti, P.C., et al., *Genome-wide detection and characterization of positive selection in human populations*. Nature, 2007. 449(7164): p. 913-8.

[78] Weir, B.S. and C.C. Cockerham, *Estimating F-statistics for the analysis of population structure.* Evolution, 1984. 38(6): p. 1358-1370.

[79] Akey, J.M., et al., *Interrogating a high-density SNP map for signatures of natural selection*. Genome Res, 2002. 12(12): p. 1805.

[80] Xu, S., et al., *A genome-wide search for signals of high-altitude adaptation in Tibetans.* Mol Biol Evol, 2011. 28(2): p. 1003-11.

[81] Chen, H., N. Patterson, and D. Reich, *Population differentiation as a test for selective sweeps.* Genome Res, 2010. 20(3): p. 393-402.

[82] Jin, W., et al., *Genome-wide detection of natural selection in African Americans pre- and post-admixture*. Genome Res, 2011. doi:10.1101/gr.124784.111.

[83] Long, J.C., *The genetic structure of admixed populations.* Genetics, 1991. 127(2): p. 417-28.

[84] Tang, H., et al., *Recent genetic selection in the ancestral admixture of Puerto Ricans.* Am J Hum Genet, 2007. 81(3): p. 626-33.

[85] Grossman, S.R., et al., A *Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection*. Science, 2010. 327(5967): p. 883-886.

[86] Hernandez, R.D., et al., *Classic selective sweeps were rare in recent human evolution.* Science, 2011. 331(6019): p. 920-4.

[87] Tattersall, I., *Human origins: Out of Africa.* Proc Natl Acad Sci U S A, 2009. 106(38): p. 16018-16021.

[88] Pickrell, J.K., et al., *Signals of recent positive selection in a worldwide sample of human populations.* Genome Res, 2009. 19(5): p. 826-37.

[89] Novembre, J. and A. Di Rienzo, *Spatial patterns of variation due to natural selection in humans.* Nat Rev Genet, 2009. 10(11): p. 745-55.

[90] Balaresque, P.L., S.J. Ballereau, and M.A. Jobling, *Challenges in human genetic diversity: demographic history and adaptation.* Hum Mol Genet, 2007. 16 (R2): p. R134-9.

[91] Katzmarzyk, P.T. and W.R. Leonard, *Climatic influences on human body size and proportions: Ecological adaptations and secular trends.* Am J of Phys Anthropol, 1998. 106(4): p. 483-503.

[92] Parra, E.J., *Human pigmentation variation: Evolution, genetic basis, and implications for public health.* Am J of Phys Anthropol, 2007: p. 85-105.

[93] Jablonski, N.G. and G. Chaplin, *Human skin pigmentation as an adaptation to UV radiation.* Proc Natl Acad Sci U S A, 2010. 107: p. 8962-8968.

[94] Harding, R.M., et al., *Evidence for variable selective pressures at MC1R.* Am J Hum Genet, 2000. 66(4): p. 1351-61.

[95] Norton, H.L., et al., *Genetic evidence for the convergent evolution of light skin in Europeans and East Asians*. Mol Biol Evol, 2007. 24(3): p. 710-22.

[96] Calvo, M.S., S.J. Whiting, and C.N. Barton, *Vitamin D intake: a global perspective of current status.* J Nutr, 2005. 135(2): p. 310-6.

[97] Young, J.H., et al., *Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion.* PLoS Genet, 2005. 1(6): p. e82.

[98] Kurian, A.K. and K.M. *Cardarelli, Racial and ethnic differences in cardiovascular disease risk factors: a systematic review*. Ethn Dis, 2007. 17(1): p. 143-52.

[99] Zhu, X., et al., *Admixture mapping for hypertension loci with genome-scan markers.* Nat Genet, 2005. 37(2): p. 177-81.

[100] Moore, L.G., *Human genetic adaptation to high altitude.* High Alt Med Biol, 2001. 2(2): p. 257-79.

[101] Bigham, A., et al., *Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data.* Plos Genet, 2010. 6(9).

[102] Beall, C.M., et al., *Hemoglobin concentration of high-altitude Tibetans and Bolivian Aymara.* Am J of Phys Anthropol, 1998. 106(3): p. 385-400.

[103] Zhuang, J., et al., *Hypoxic ventilatory responsiveness in Tibetan compared with Han residents of 3,658 m*. J Appl Physiol, 1993. 74(1): p. 303-11.

[104] Yi, X., et al., *Sequencing of 50 human exomes reveals adaptation to high altitude.* Science, 2010. 329(5987): p. 75-8.

[105] Simonson, T.S., et al., *Genetic evidence for high-altitude adaptation in Tibet.* Science, 2010. 329(5987): p. 72-5.

[106] Perry, G.H., et al., *Diet and the evolution of human amylase gene copy number variation*. Nat Genet, 2007. 39(10): p. 1256-60.

[107] Diamond, J., *Evolution, consequences and future of plant and animal domestication.* Nature, 2002. 418(6898): p. 700-7.

[108] Cordain, L., et al., *Origins and evolution of the Western diet: health implications for the 21st century.* Am J Clin Nutr, 2005. 81(2): p. 341-354.

[109] Neel, J.V., Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet, 1962. 14: p. 353-62.

[110] Eaton, S.B., M. Konner, and M. Shostak, *Stone agers in the fast lane: chronic degenerative diseases in evolutionary perspective*. Am J Med, 1988. 84(4): p. 739-49.

[111] Helgason, A., et al., *Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution*. Nat Genet, 2007. 39(2): p. 218-25.

[112] Peltonen, L., et al., *Identification of a variant associated with adult-type hypolactasia.* Nat Genet, 2002. 30(2): p. 233-237.

[113] Swallow, D.M., *Genetics of lactase persistence and lactose intolerance.* Annu Rev Genet, 2003. 37: p. 197-219.

[114] Wang, Y.X., et al., *The Lactase Persistence/Non-Persistence Polymorphism Is Controlled by a Cis-Acting Element*. Hum Mol Genet, 1995. 4(4): p. 657-662.

[115] Bersaglieri, T., et al., *Genetic signatures of strong recent positive selection at the lactase gene.* Am J Hum Genet, 2004. 74(6): p. 1111-20.

[116] Poulter, M., et al., *The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans.* Ann of Hum Genet, 2003. 67: p. 298-311.

[117] Tishkoff, S.A., et al., *Convergent adaptation of human lactase persistence in Africa and Europe.* Nat Genet, 2007. 39(1): p. 31-40.

[118] Casanova, J.L. and L. Abel, *Inborn errors of immunity to infection: the rule rather than the exception*. J Exp Med, 2005. 202(2): p. 197-201.

[119] Haldane, J.B.S., *Disease and Evolution* (Reprinted from La Ricerca Scientifica Supplemento, Vol 19, Pg 1-11, 1949). Curr Sci, 1992. 63(9-10): p. 599-604.

[120] Varki, A., *A chimpanzee genome project is a biomedical imperative.* Genome Res, 2000. 10(8): p. 1065-1070.

[121] Varki, A. and T.K. Altheide, *Comparing the human and chimpanzee genomes: Searching for needles in a haystack*. Genome Res, 2005. 15(12): p. 1746-1758.

[122] Sironi, M., et al., *Widespread balancing selection and pathogen-driven selection at blood group antigen genes*. Genome Res, 2009. 19(2): p. 199-212.

[123] Prugnolle, F., et al., *Pathogen-driven selection and worldwide HLA class I diversity.* Curr Biol, 2005. 15(11): p. 1022-1027.

[124] Barreiro, L.B. and L. Quintana-Murci, *From evolutionary genetics to human immunology: how selection shapes host defence genes*. Nat Rev Genet, 2010. 11(1): p. 17-30.

[125] Moyzis, R.K., et al., *Global landscape of recent inferred Darwinian selection for Homo sapiens.* Proc Natl Acad Sci U S A, 2006. 103(1): p. 135-140.

[126] Wolfe, N.D., C.P. Dunavan, and J. Diamond, *Origins of major human infectious diseases.* Nature, 2007. 447(7142): p. 279-83.

[127] Barreiro, L.B., et al., *Evolutionary Dynamics of Human Toll-Like Receptors and Their Different Contributions to Host Defense.* Plos Genet, 2009. 5(7).

[128] Casanova, J.L., et al., *Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases*. Nat Immunol, 2007. 8(11): p. 1165-1171.

[129] Arenzana-Seisdedos, F. and M. Parmentier, *Genetics of resistance to HIV infection: Role of co-receptors and co-receptor ligands.* Seminars in Immunology, 2006. 18(6): p. 387-403.

[130] Galvani, A.P. and M. Slatkin, *Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele.* Proc Natl Acad Sci U S A, 2003. 100(25): p. 15276-15279.

[131] Snow, R.W., et al., T*he global distribution of clinical episodes of Plasmodium falciparum malaria.* Nature, 2005. 434(7030): p. 214-217.

[132] Kwiatkowski, D.P., *How malaria has affected the human genome and what human genetics can teach us about malaria.* Am J Hum Genet, 2005. 77(2): p. 171-192.

[133] Schork, N.J., et al., *Common vs. rare allele hypotheses for complex diseases.* Curr Opin Genet Dev, 2009. 19(3): p. 212-9.

[134] Luo, L., E. Boerwinkle, and M. Xiong, *Association studies for next-generation sequencing.* Genome Res, 2011. 21(7): p. 1099-108.

[135] Di Rienzo, A. and R.R. Hudson, *An evolutionary framework for common diseases: the ancestral-susceptibility model.* Trends Genet, 2005. 21(11): p. 596-601.

[136] Di Rienzo, A., *Population genetics models of common diseases.* Curr Opi Genet Dev, 2006. 16(6): p. 630-636.

[137] Pritchard, J.K., *Are rare variants responsible for susceptibility to complex diseases?* Am J Hum Genet, 2001. 69(1): p. 124-137.

[138] Kryukov, G.V., L.A. Pennacchio, and S.R. Sunyaev, *Most rare missense alleles are deleterious in humans: implications for complex disease and association studies*. Am J Hum Genet, 2007. 80(4): p. 727-39.

[139] Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*. Proc Natl Acad Sci U S A, 2009. 106(23): p. 9362-7.

[140] McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges*. Nat Rev Genet, 2008. 9(5): p. 356-369.

[141] Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease.* Trends Genet, 2001. 17(9): p. 502-10.

[142] Bodmer, W. and C. Bonilla, *Common and rare variants in multifactorial susceptibility to common diseases.* Nat Genet, 2008. 40(6): p. 695-701.

[143] Gibson, G., *Decanalization and the origin of complex disease.* Nat Rev Genet, 2009. 10(2): p. 134-40.

[144] Hermisson, J. and G.P. Wagner, *The population genetic theory of hidden variation and genetic robustness. Genetics*, 2004. 168(4): p. 2271-84.

[145] Polychronakos, C., *Common and rare alleles as causes of complex phenotypes.* Curr Atheroscler Rep, 2008. 10(3): p. 194-200.

[146] Dickson, S.P., et al., *Rare variants create synthetic genome-wide associations.* PLoS Biol, 2010. 8(1): p. e1000294.

[147] Romeo, S., et al., *Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL.* Nat Genet, 2007. 39(4): p. 513-6.

[148] Ji, W., et al., *Rare independent mutations in renal salt handling genes contribute to blood pressure variation*. Nat Genet, 2008. 40(5): p. 592-9.

[149] McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges*. Nat Rev Genet, 2008. 9(5): p. 356-69.

[150] *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls*. Nature, 2007. 447(7145): p. 661-78.

[151] Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies*. Nat Genet, 2006. 38(8): p. 904-9.

[152] Clayton, D.G., et al., *Population structure, differential bias and genomic control in a large-scale, case-control association study*. Nat Genet, 2005. 37(11): p. 1243-6.

[153] Marchini, J., et al., *The effects of human population structure on large genetic association studies.* Nat Genet, 2004. 36(5): p. 512-7.

[154] Seldin, M.F., et al., *European population substructure: clustering of northern and southern populations*. PLoS Genet, 2006. 2(9): p. e143.

[155] Tian, C., et al., *Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet*, 2008. 4(1): p. e4.

[156] Xu, S., et al., *Genomic dissection of population substructure of Han Chinese and its implication in association studies.* Am J Hum Genet, 2009. 85(6): p. 762-74.

[157] Chen, J., et al., *Genetic structure of the Han Chinese population revealed by genome-wide SNP variation*. Am J Hum Genet, 2009. 85(6): p. 775-85.

[158] Zheng, G., B. Freidlin, and J.L. Gastwirth, *Robust genomic control for association studies.* Am J Hum Genet, 2006. 78(2): p. 350-6.

[159] Jin, W., et al., *A systematic characterization of genes underlying both complex and Mendelian diseases.* Hum Mol Genet, 2011. doi:10.1093/hmg/DDR599.

[160] Visscher, P.M., *Sizing up human height variation.* Nat Genet, 2008. 40(5): p. 489-90.

[161] Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. 461(7265): p. 747-53.

[162] Ng, S.B., et al., *Targeted capture and massively parallel sequencing of 12 human exomes.* Nature, 2009. 461(7261): p. 272-6.

[163] Li, Y., et al., *Low-coverage sequencing: implications for design of complex trait association studies.* Genome Res, 2011. 21(6): p. 940-51.

[164] Gorlov, I.P., et al., *Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms.* Am J Hum Genet, 2008. 82(1): p. 100-12.

[165] Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.* Am J Hum Genet, 2008. 83(3): p. 311-21.

[166] Li, Y., A.E. Byrnes, and M. Li, *To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests.* Am J Hum Genet, 2010. 87(5): p. 728-35.

[167] Price, A.L., et al., *Pooled association tests for rare variants in exon-resequencing studies.* Am J Hum Genet, 2010. 86(6): p. 832-8.

[168] Xu, S., et al., *Analysis of genomic admixture in Uyghur and its implication in mapping strategy.* Am J Hum Genet, 2008. 82(4): p. 883-94.

[169] Xu, S. and L. Jin, *A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery.* Am J Hum Genet, 2008. 83(3): p. 322-36.

[170] Narang, A., et al., *Recent admixture in an Indian population of African ancestry.* Am J Hum Genet, 2011. 89(1): p. 111-20.

[171] Shah, A.M., et al., *Indian Siddis: African descendants with Indian admixture.* Am J Hum Genet, 2011. 89(1): p. 154-61.

[172] Smith, M.W. and S.J. O'Brien, *Mapping by admixture linkage disequilibrium: advances, limitations and guidelines.* Nat Rev Genet, 2005. 6(8): p. 623--632.

[173] Bamshad, M., et al., *Deconstructing the relationship between genetics and race.* Nat Rev Genet, 2004. 5(8): p. 598-609.

[174] Chakraborty, R. and K.M. Weiss, *Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci.* Proc Natl Acad Sci U S A, 1988. 85(23): p. 9119-23.

[175] Stephens, J.C., D. Briscoe, and S.J. O'Brien, *Mapping by admixture linkage disequilibrium in human populations: limits and guidelines.* Am J Hum Genet, 1994. 55(4): p. 809-24.

[176] Pfaff, C.L., et al., *Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium.* Am J Hum Genet, 2001. 68(1): p. 198-207.

[177] Tian, C., et al., *A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping.* Am J Hum Genet, 2006. 79(4): p. 640-9.

[178] Pfaff, C.L., R.A. Kittles, and M.D. Shriver, *Adjusting for population structure in admixed populations.* Genet Epidemiol, 2002. 22(2): p. 196-201.

[179] Seldin, M.F., et al., *Putative ancestral origins of chromosomal segments in individual african americans: implications for admixture mapping.* Genome Res, 2004. 14(6): p. 1076-84.

[180] Xu, S., et al., *Dissecting linkage disequilibrium in African-American genomes: roles of markers and individuals.* Mol Biol Evol, 2007. 24(9): p. 2049-58.

[181] Seldin, M.F., B. Pasaniuc, and A.L. Price, *New approaches to disease mapping in admixed populations.* Nat Rev Genet, 2011. 12(8): p. 523-8.

[182] Freedman, M.L., et al., *Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men.* Proc Natl Acad Sci U S A, 2006. 103(38): p. 14068-73.

[183] Cheng, C.-Y., et al., *Admixture mapping of 15,280 African Americans identifies obesity susceptibility loci on chromosomes 5 and X.* PLoS Genet, 2009. 5(5): p. e1000490.

[184] Reich, D., et al., *A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility.* Nat Genet, 2005. 37(10): p. 1113-8.

[185] Gudmundsson, J., et al., *Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24.* Nat Genet, 2007. 39(5): p. 631-7.

[186] Tang, H., et al., *Reconstructing genetic ancestry blocks in admixed individuals.* Am J Hum Genet, 2006. 79(1): p. 1-12.

[187] Sundquist, A., et al., *Effect of genetic divergence in identifying ancestral origin using HAPAA.* Genome Res, 2008. 18(4): p. 676-82.

[188] Price, A.L., et al., *Sensitive detection of chromosomal segments of distinct ancestry in admixed populations.* PLoS Genet, 2009. 5(6): p. e1000519.