5 Discussion

5.1 The detection of positive selection: from genotyping to sequencing

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

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

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

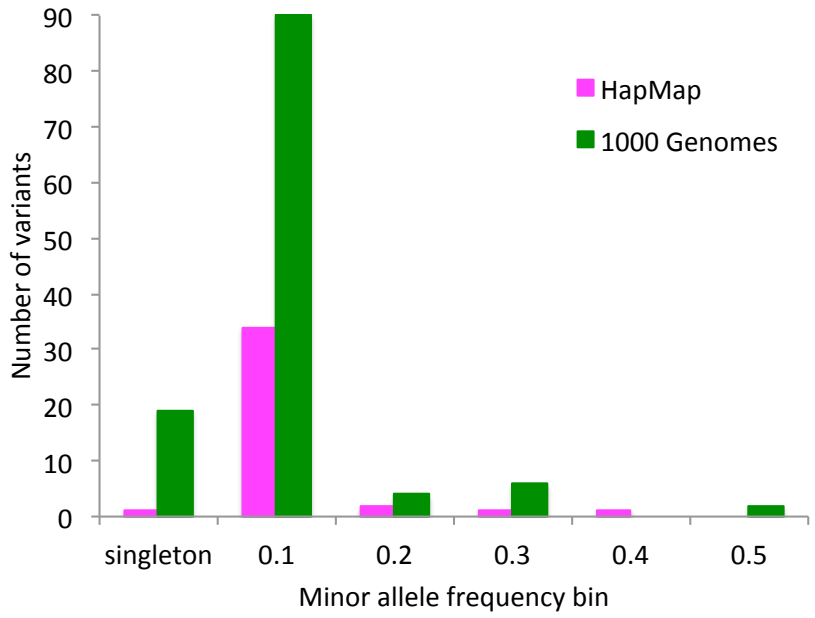

Figure 5.1 MAF comparison of 1000 Genomes low-coverage Pilot and HapMap data in ITSN2 peak windows.

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

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

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

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

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

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

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

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

5.2 The localization of selection targets

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

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

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

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

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

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

5.3 Biological interpretation of alleles under positive selection

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

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

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

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

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

5.4 Impact of the studies in this thesis

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

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

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

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

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

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

5.5 Future directions

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

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

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

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

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

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

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