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

3.1 Introduction

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

The 1000 Genomes Project is an excellent example of such resources. The Pilot 1 (low-coverage) project sequenced 179 individuals from four populations: CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), CHB+JPT (Chinese Han in Beijing, China and Japanese in Tokyo, Japan) and YRI (Yoruba in Ibadan, Nigeria), with the average coverage of 2-4x[40]. 15 million SNPs were identified in the Pilot Project along with other types of genetic polymorphism, which greatly enriched the database of human genomic variation. As demonstrated in Chapter 2, a genome-wide survey of positive selection using frequency-spectrum based methods on such sequencing data would provide deeper insights into the extent to which positive selection has shaped modern human genomic variation, as well as the biological targets that may be selected during recent modern human evolutionary history.

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

3.2 Materials and Methods

3.2.1 Simulations

We first carried out coalescent simulations using the *msHOT* package[127] to generate 1Mb long neutral haplotypes in African, European and Asian ancestral populations 2,000 generations ago, based on the best-fit demographic models[26] for the three populations. Then these simulated haplotypes were used as seed haplotypes for the forward simulations using *mpop*[94], as described in section 2.2.1. In forward simulations, one neutral scenario (1,000 independent simulations) and sixteen selective sweep scenarios were simulated in each population. Selection coefficients of 0.001, 0.004, 0.007 and 0.01, and the age of selective sweeps of 500 generations, 1,000 generations, 1,500 generations and 2,000 generations were used in the selective sweep scenarios, with 250 simulations for every combination of these two parameters. One allele with an initial frequency of 0.0006 under selection was added in the middle of the haplotypes at the starting time point of the selective sweep. The genome average mutation rate of 1.0 x 10^{-8} per nucleotide per generation was used in the simulations. In addition, to mimic the real patterns of recombination in the genome, we used the HapMap recombination map[39] to generate a recombination hotspot map, and regions of 1 Mb were drawn randomly from the genome and the recombination hotspots they contained were assigned to the simulated regions. For the purpose of comparison and understanding of the effects of recombination hotspots on the signals of selection, we also did another set of simulations with all parameters being the same, except that a strong recombination hotspot (2,000-fold greater than the background recombination rate) with 0 kb, 10 kb, 20 kb, 30 kb or 40 kb distance from the selected allele was added into the simulated haplotypes. The rest of demographic parameters were as in Schaffner et al.'s best-fit demographic model for the European population[26]. For computational efficiency, we re-scaled the parameters by a factor of 5, as described in section 2.2.1. 120 chromosomes were sampled from each simulation, to match the sample sizes of 1000 Genomes Project low-coverage Pilot data (see Appendix C for parameters and command lines).

3.2.2 Neutrality tests on simulated data

In order to mimic the real situation of 1000 Genomes low-coverage Pilot data, where rare SNPs are still under-ascertained, we filtered the simulated data by matching the proportion of SNPs in each derived allele frequency bin (bin size 0.1) of the simulated data to the 1000 Genomes low-coverage Pilot data in each population (CEU, CHB+JPT and YRI). Then three frequency-spectrum based tests, Tajima's *D*[71], Fay and Wu's *H*[72] and Nielsen's CLR[76] were applied to the simulated data in 10 kb non-overlapping windows across the simulated regions. P values of each test were calculated based on the distribution of test values of 1000 neutral simulations in each population. In order to obtain a single score representing the signals of all three tests, we calculated the correlations between the p values of every two tests in neutral simulations to see whether these tests are independent from each other. Results showed that the absolute value of the correlation of every pair of tests was less than 0.2. Therefore, we treated these tests as independent, and combined the p values of each test on the same window using Fisher's method[104].

3.2.3 Sensitivity and specificity analysis on simulated data

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

3.2.4 Neutrality tests on 1000 Genomes low-coverage Pilot data

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

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

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

3.2.5 Identification of candidate regions and genes

After the genome-wide combined p values of our neutrality tests were obtained, we needed to decide which threshold of significance to use. As we aimed to get a confident list of candidate regions, we used the stringent Bonferroni correction[128]. We divided 0.01 by the total number of windows that we applied the tests to throughout the whole genome, which yielded a threshold of ~4×10^{-8} ($-\log_e$ value 17.0). We used this as a cutoff to identify significant windows in each population. Adjacent significant windows that are less than 150 kb apart were treated as likely to originate from the same selective sweep, and combined into a single candidate region.

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

We also looked at positions of peak signals relative to the candidate protein-coding genes. We used three categories of positions: upstream of the gene, within the gene, and downstream of the gene. First of all, to determine which side of the gene is upstream or downstream, we obtained information about whether the gene is on the forward strand or reverse strand of the DNA sequence for each candidate protein-coding gene. Then we counted the number of peak windows falling into each category of position. For those peaks that cover more than one position, we used the proportion of the window in each position as the count. For example, if 40% of the peak window is in the upstream sequence, and the other 60% is in the gene, we count 0.4 into "upstream" and 0.6 into "within gene" for that candidate.

3.2.6 Comparison with previous studies and bioinformatic analyses

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

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

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

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

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

3.3 Results from simulations

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

Balancing the false positive and false negative rates in the identification of statistical significance is a crucial step in a large-scale global survey of statistical tests. As mentioned above, we chose to use the most stringent p value cutoff (Bonferroni correction, $p = 4\times10^{-8}$) to identify significant windows. This, of course, sacrifices the sensitivity of detection. An alternative measure of the p value significance threshold is the false discovery rate (FDR). Since we are applying the statistical tests a large number of times, even a very small false positive rate can result in a large FDR. To measure this, we counted the number of candidate regions under different p value thresholds, and calculated FDRs accordingly. We found that even if the false positive rate is 0.6%, the FDR is still as high as 4%. In order to get a highly confident list of candidate regions, we would like the FDR to be less than 5%. A Bonferroni-corrected threshold of 4×10^{-8} gives us 0% and 3% FDR in CEU and YRI, respectively (YRI, Table 3.2). Although in this case, we were only able to detect ~20% of the moderate-strength positive selection events, we are confident that the list of candidates we picked out is mostly real.

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

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

3.3.2 Power of localizing positive selection targets

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

3.3.3 Effects of recombination hotspots on localization of selection target

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

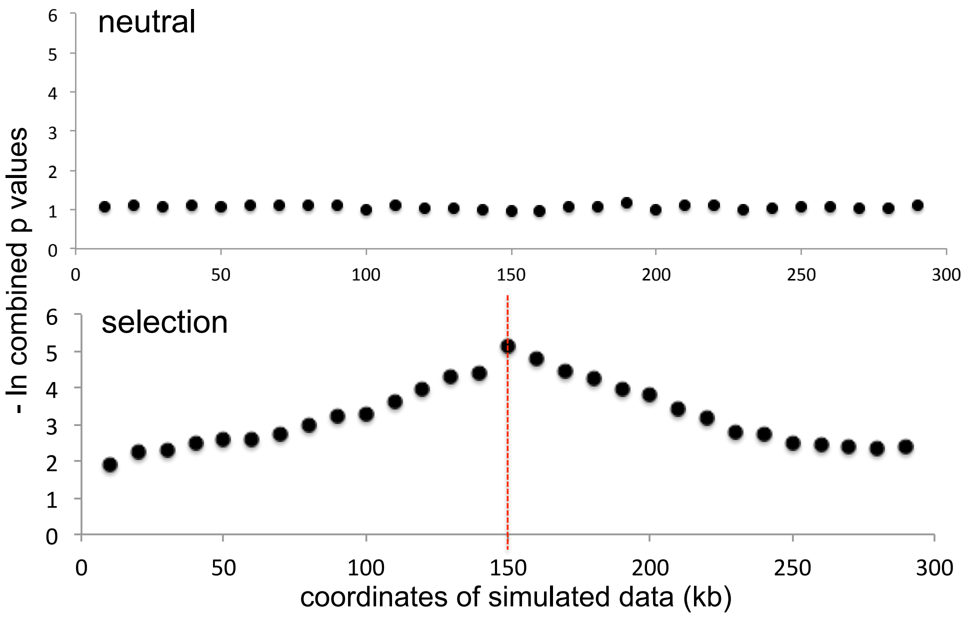


Figure 3.1 Averaged scores in neutral and positively selected simulations. The top plot shows average scores of each 10-kb window across the simulated neutral regions; the bottom plot shows the same but in simulated regions with selection. The red dashed line shows the position of the selected allele.

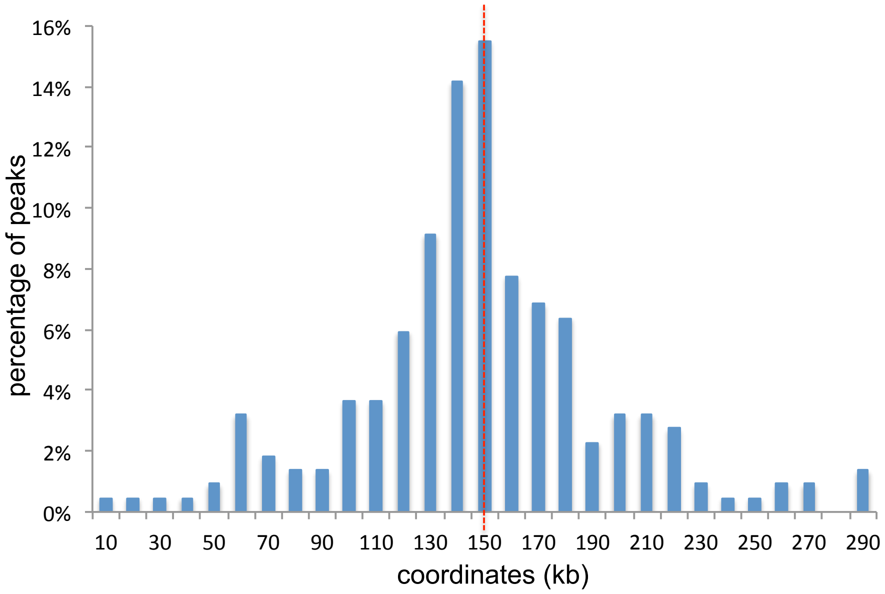


Figure 3.2 Distribution of peak signals across the simulated regions with selection. Each bar shows the percentage of peak signals falling in the particular window. The red dashed line shows the position of selected allele.

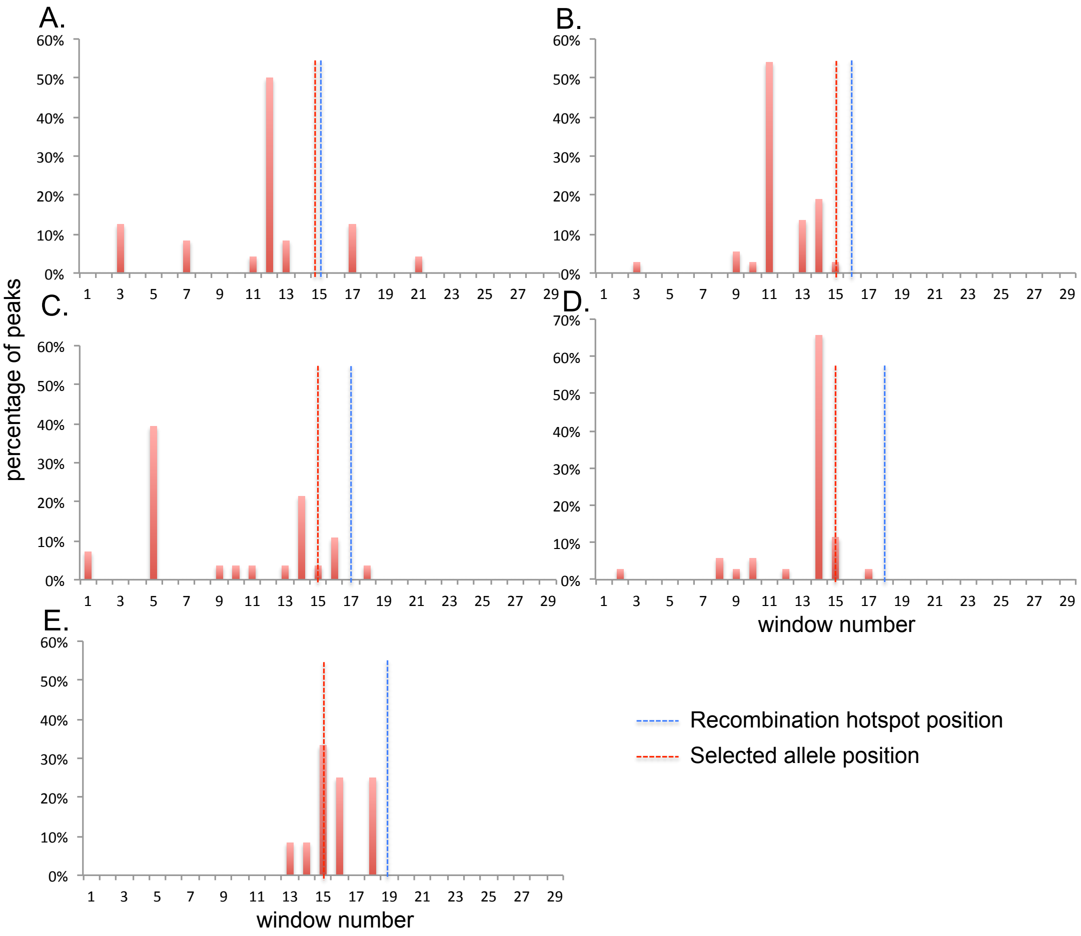


Figure 3.3 Distribution of peak signals in simulations with single strong recombination hotspots. Each plot shows the distribution of peak signals under the scenario with fixed distance between the selected allele and the recombination hotspot. The blue dashed line marks position of the recombination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

3.4 Results from 1000 Genomes Project low-coverage Pilot data

3.4.1 Genome-wide scan on 1000 Genomes low coverage data

We applied the same tests and criteria to the 1000 Genomes Project low-coverage Pilot sequencing data in ~10-kb windows across the whole genome in CEU, CHB+JPT and YRI populations. We identified 477, 137 and 290 candidate regions in the three populations, respectively. In all populations, most regions only have one significant window, but CEU have more regions with larger numbers of significant windows than the other two populations (Figure 3.4). Among these candidate regions, 65%, 59% and 64% (308, 81 and 187 regions) in each of the three populations, respectively, overlap with genes (including pseudogenes and non-coding RNAs) within the ~100 kb region around the peak signal, and among these, 258, 66 and 153 regions overlap with protein-coding genes in each population, respectively. The candidate regions are highly enriched with genes, when compared with that of randomly chosen regions across the genome ($p < 0.001$). They are also highly enriched in protein-coding genes compared to random regions ($p < 0.001$). Some candidate regions overlap with multiple genes, and as we believe that each candidate region should only have one selection target, we chose the gene(s) closest to the peak window as the candidate gene(s). We thus identified 275, 69 and 160 protein-coding genes that may have undergone positive selection in CEU, CHB+JPT and YRI populations, respectively (Table 3.3; Appendix D, candidate regions and protein-coding genes in each population). In a few cases, we identified two candidate genes in one region, either because these two genes have the same distance from the peak signal, or because these two genes overlap with each other. We then counted the number of peak signals at upstream to the candidate gene, within the candidate gene, or downstream of the candidate gene. We found that in all three populations, the biggest proportion of peaks is within the candidate genes, compared to upstream or downstream of the candidate genes (Figure 3.5).

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

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

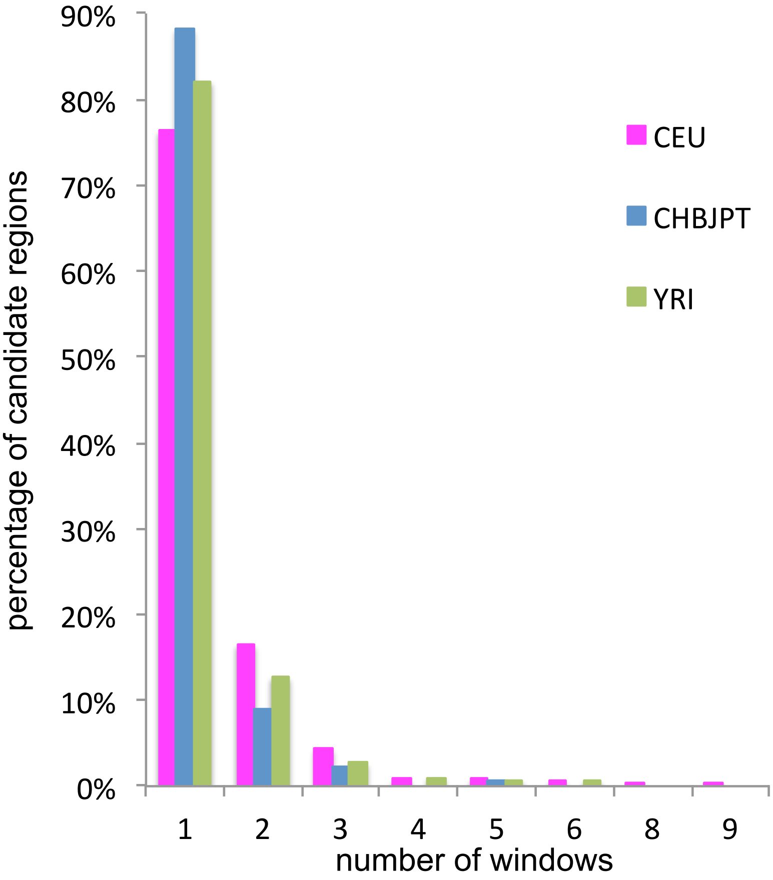


Figure 3.4 Distribution of number of significant windows per candidate region in each population.

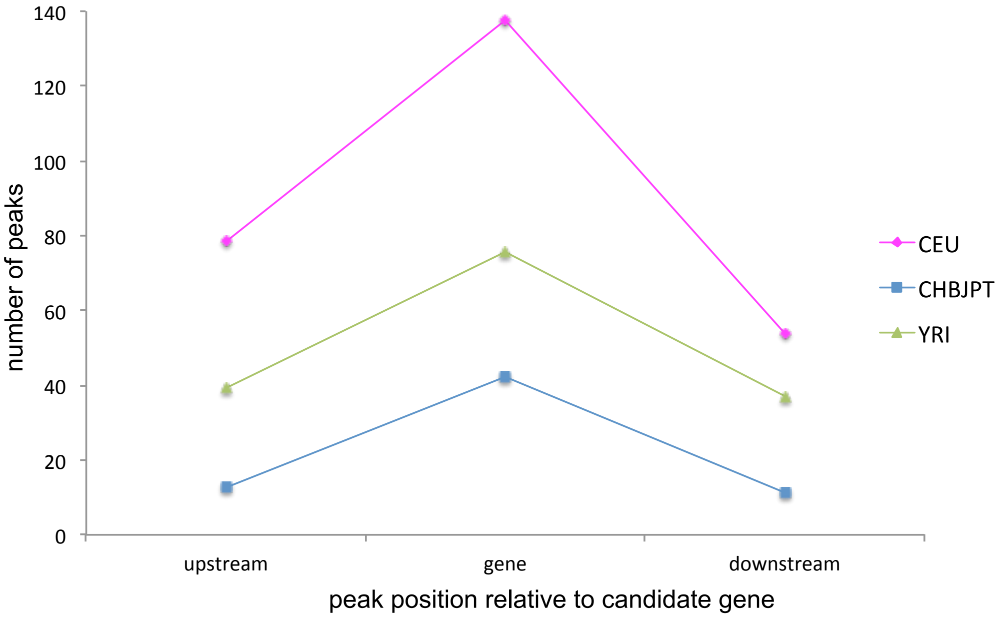


Figure 3.5 Number of peak signals at each position relative to the candidate gene in each population.

3.4.2 Comparison of candidate regions with previous studies

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

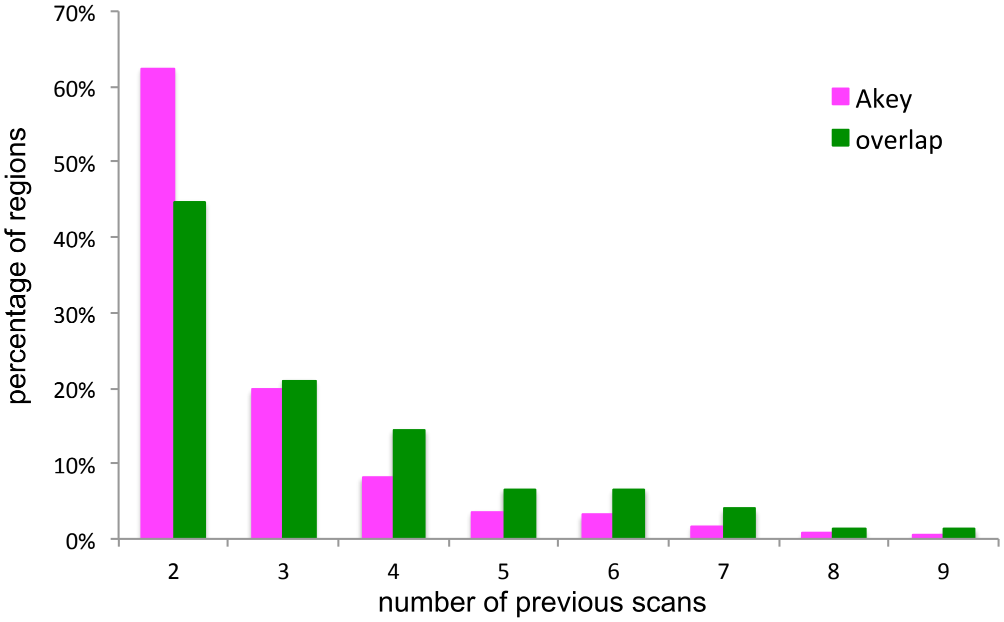


Figure 3.6 Overlap of our candidate regions with Akey's review. This plot shows the distribution of number of previous scans showing evidence of positive selection in all the candidate regions in Akey's review versus those overlap with our candidate regions.

3.4.3 Analysis of functional variants in candidate regions or genes

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

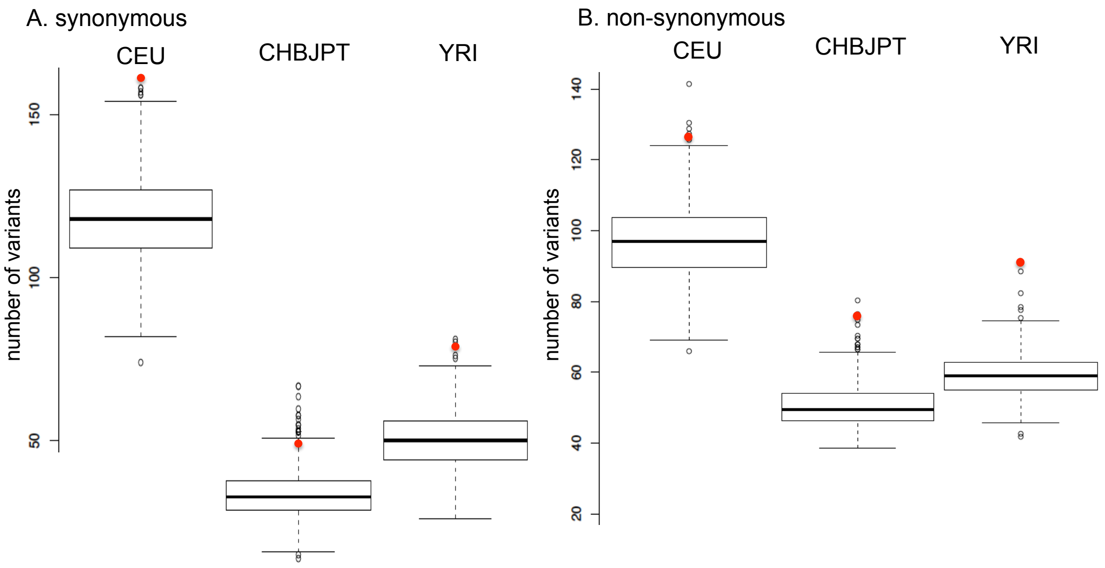


Figure 3.7 Synonymous and non-synonymous variants in candidate regions. These box plots show the distributions of the number of synonymous (A) or non-synonymous (B) changes in 1,000 sets of random genes that match the candidate genes. The upper and lower boundaries of the boxes show the 75th and 25th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

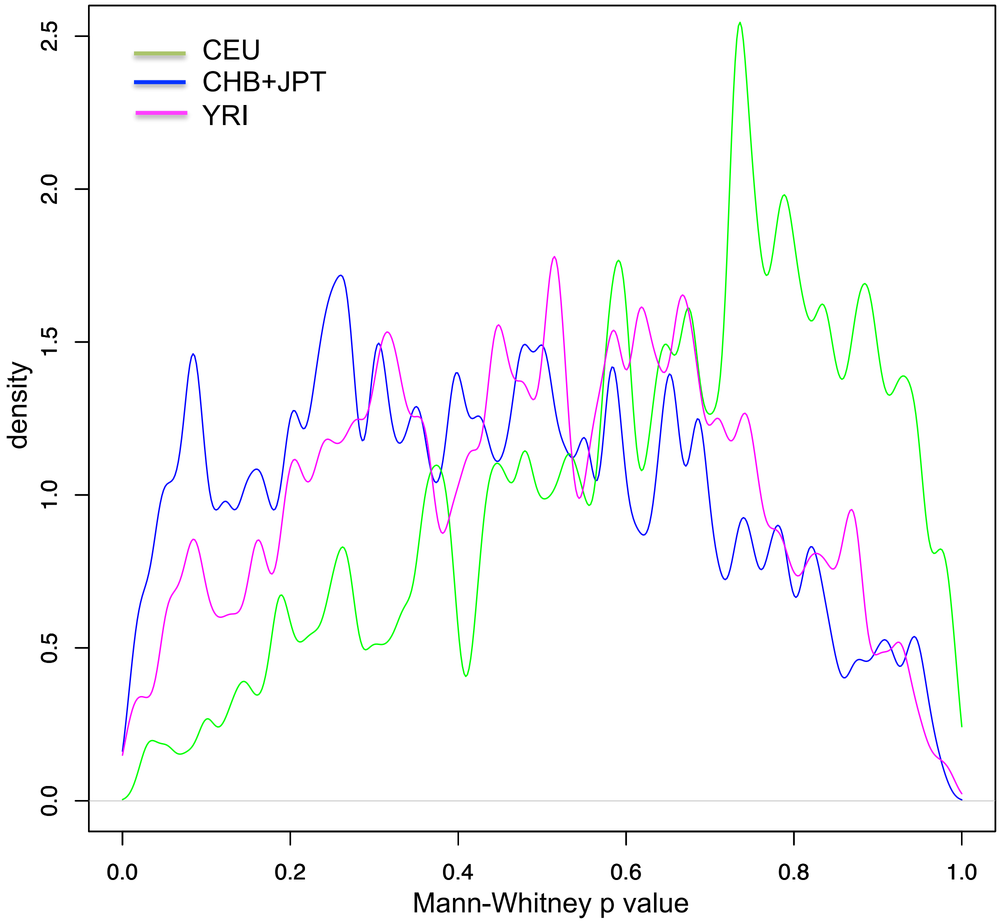
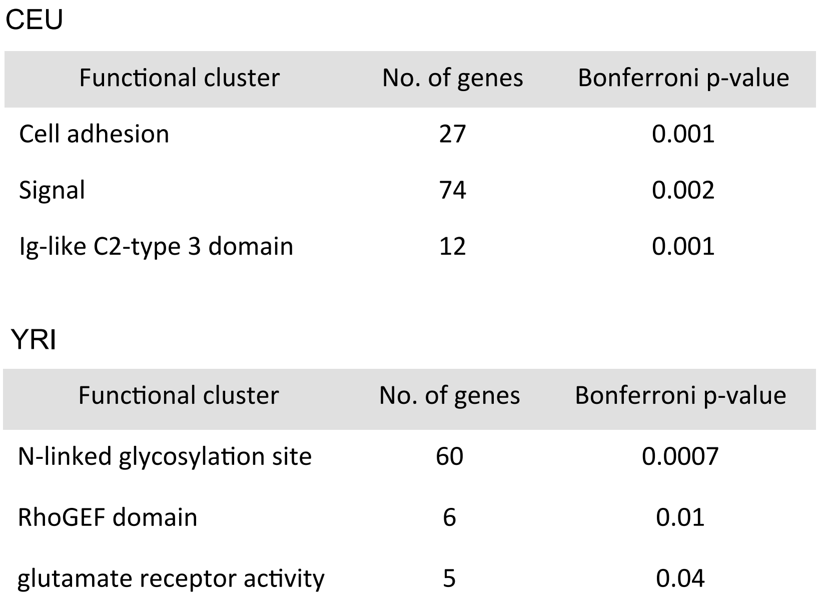


Figure 3.8 Distribution of Mann-Whitney p values on Condel scores.

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

Apart from protein-coding genes, positive selection may also act on other functional elements in the genome. In order to investigate whether there is any enrichment of non-coding functional elements, we obtained annotation of variants within UTRs, non-coding RNAs, enhancers, and TF motif gains and losses. We calculated the distributions of number of such variants with higher than or equal to 50% DAF in the 1000 Genomes low-coverage Pilot data in each population in 1,000 sets of random regions matching our candidate regions, and looked at where the corresponding number in our candidate regions fall into these distributions. We found no significant enrichment of any of the five types of non-coding functional variants in our candidate regions (Figure 3.9). There are three possible explanations for the lack of enrichment of non-coding functional variants. One is that selection on these regulatory elements might have been weaker and subtler in general, thus we were only able to identify a small proportion of them, which may not be representative of the whole set of positively selected non-coding functional elements. The second one is that our annotation of non-coding functional elements in the human genome has been very limited, in terms of both completeness and accuracy. The third one is that we did not categorize these functional elements based on their actual biological functions or processes. Positive selection may act on all types of non-coding functional elements, but favor certain types of biological function. However, due to our very limited understanding of the actual functions of those elements, we were unable to detect the enrichment.

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



CEU

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

YRI

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

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

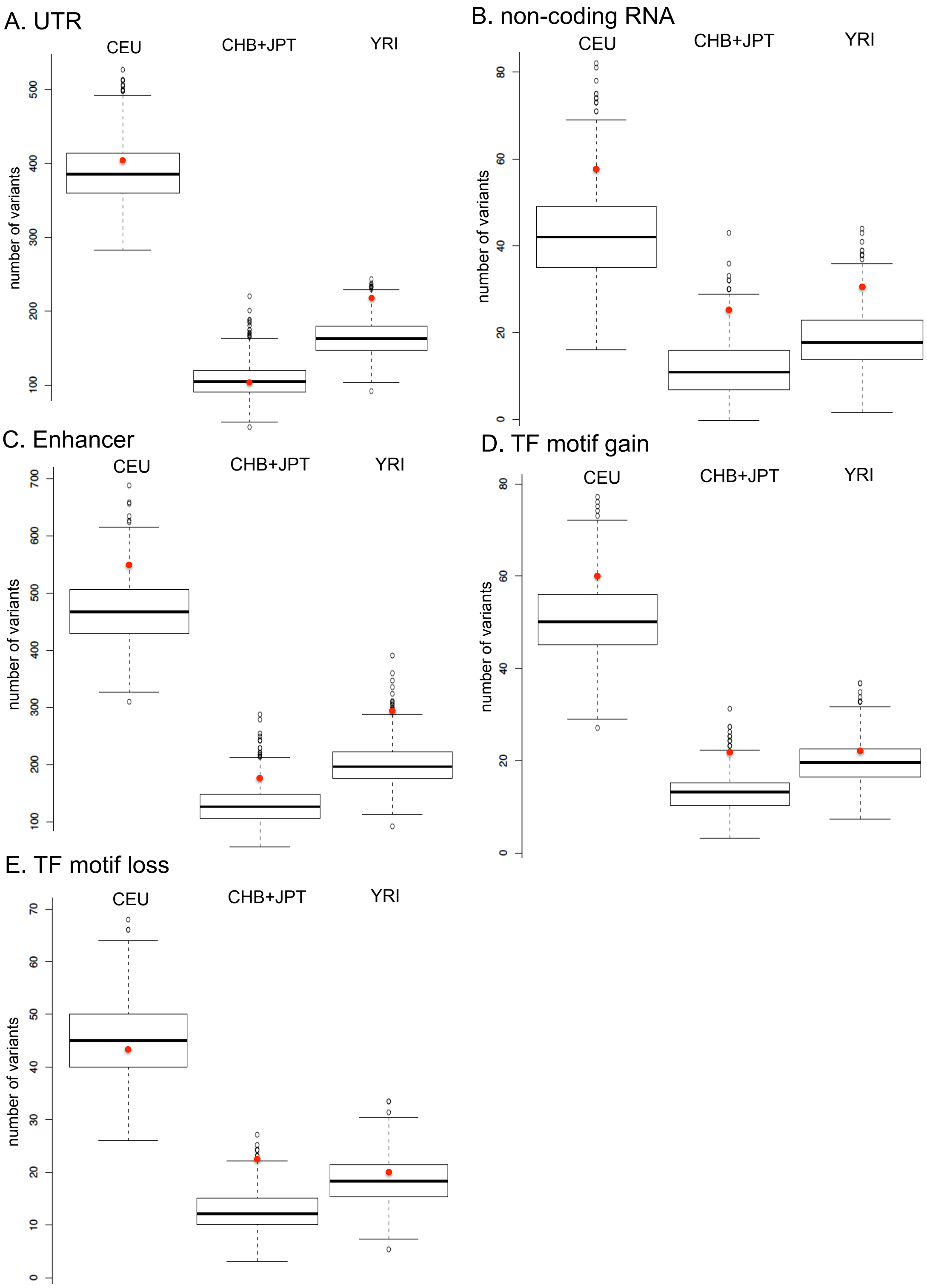
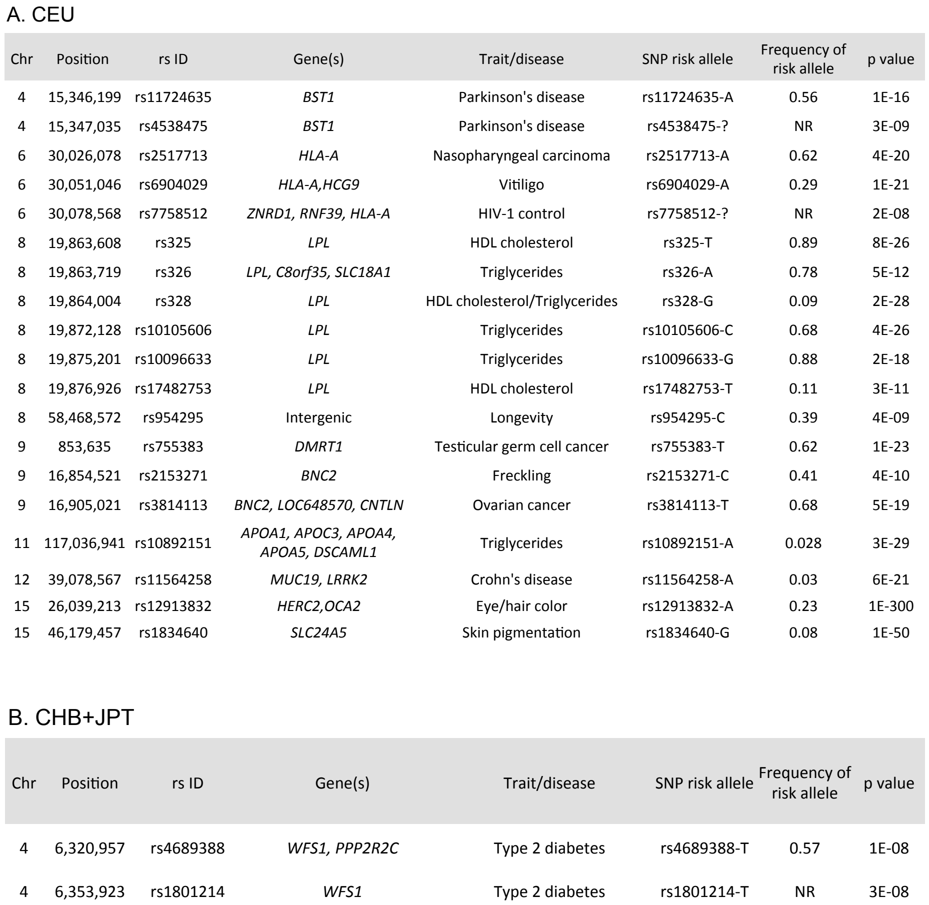


Figure 3.9 Non-coding functional variants in candidate regions. These box plots show the distributions of the number of UTR (A), non-coding RNA (B), enhancer (C), TF motif gain (D) and loss (E) variants in 1000 sets of random regions that match the candidate regions. The red dots are corresponding values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

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

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



A. CEU

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

B. CHB+JPT

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

C. YRI

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

3.5 Examples of strong candidate genes and their functions

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

3.5.1 Examples of strong positively selected genes in a particular population

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

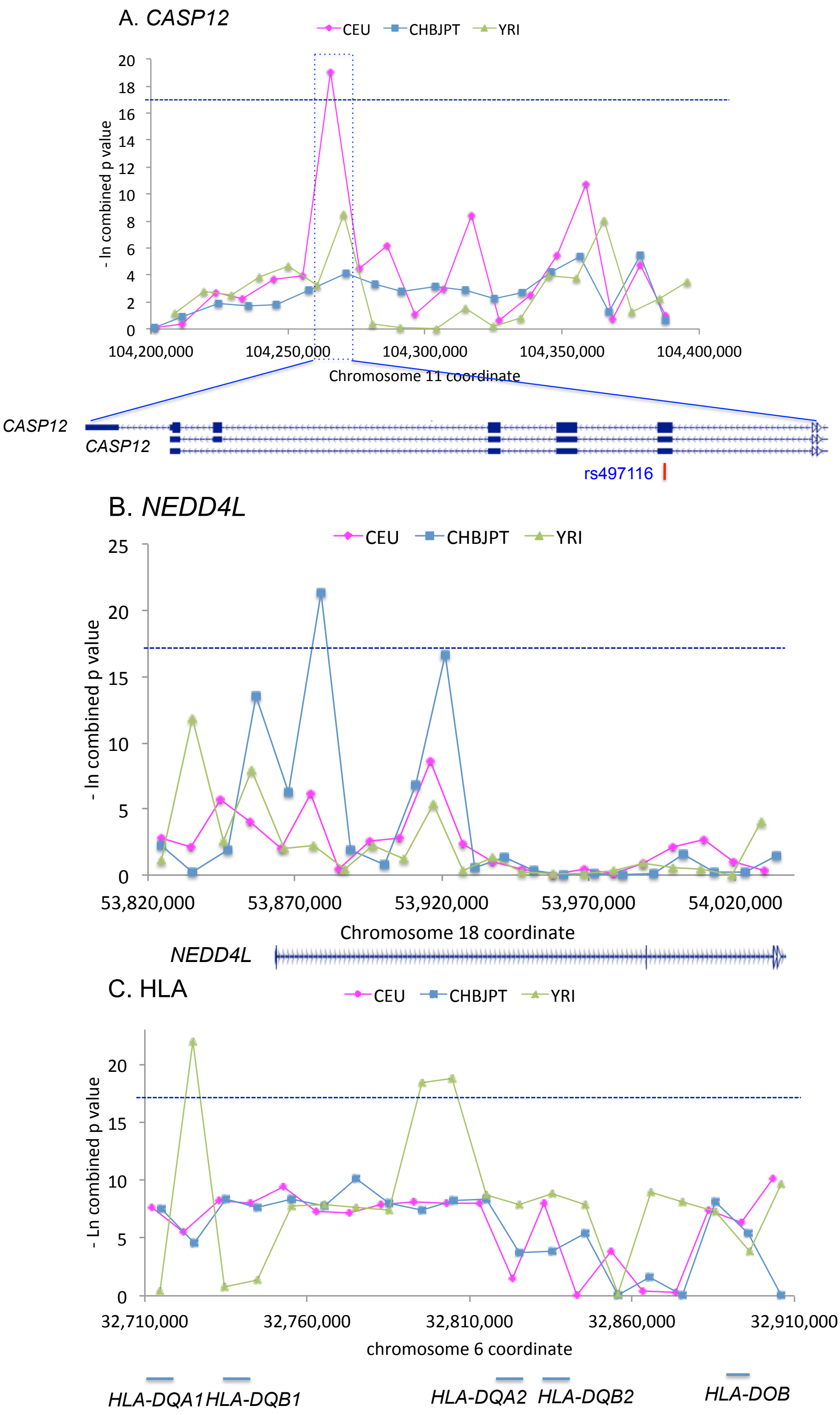


Figure 3.10 Examples of positively selected genes with signals in only one population. Blue dashed line marks the significance threshold. Candidate genes are shown and positions of putative selected SNPs are marked as red bars with the rs ID if applicable.

NEDD4L: This gene shows a very strong signal of positive selection in the CHB+JPT population, but not in the other two populations (Figure 3.10 B). The gene encodes the enzyme E3 ubiquitin-protein ligase NEDD4-like, which is believed to regulate the expression and function of the epithelial sodium channel[137,138]. It plays a very important role in salt reabsorption. Studies have shown that this gene is associated with salt sensitivity[139], blood pressure[140], and essential hypertension[141]. Interestingly, it has been reported that African-Americans are more sensitive to salt than other groups in the US, and they develop hypertension at younger ages, with more severe consequences. So it appears that Africans are more sensitive to salt than other groups. Based on these facts, it is plausible that salt-insensitivity has been positively selected outside of Africa, due to the adaptation to the new environment. The climate was hot and dry in most human habitats in Africa, and salt was rare in ancient times, so retaining salt in the body was very important for the survival of humans. However, when our ancestors moved out of Africa, the climate was cooler, and salt was easier to access especially near the sea, so retaining salt in the body was no longer advantageous, and sometimes could be harmful, as it may cause high blood pressure. Therefore, there might have been a selective force favoring less efficient salt reabsorption in out-of-Africa populations. However, if this is the case, we should expect to see signals in both European and Asian populations. There are two possible reasons that we did not see signals in the European population. One is that the selective sweep might have happened earlier in Europe than in Asia, or the strength of selective force was much higher in Europe, so that the selective sweep had already been completed for a long time, therefore the footprint of positive selection had faded. The other explanation might be that the selection strength in Europe is very low, so the sweep has not reached to a detectable stage. All in all, the strong signal of positive selection plus the interesting functional implications of this gene makes it a very good candidate for further studies on its roles in salt sensitivity and blood pressure, and its association with hypertension. It may be worth doing functional analyses on highly differentiated alleles between African and other populations within this gene to find out which variant(s) is more likely to be the selection target.

HLA gene cluster: The HLA gene cluster on Chromosome 6 showed very strong signals of positive selection in the YRI population (Figure 3.10 C). The HLA (human leukocyte antigen) system lies within the human major histocompatibility complex (MHC). This cluster contains a large number of genes related to the immune system of humans. There are different classes of HLA genes, and they play important roles in disease defense, may cause organ transplant rejections, and mediate autoimmune diseases. Many variants in this gene cluster are associated with various autoimmune or inflammatory diseases, including inflammatory bowel disease, HIV, Vitiligo, Ankylosing spondylitis, Rheumatoid arthritis and so on (Table 3.5). The positive selection signals in this locus may indicate the strong selective force of disease defense and immune functions in the African population.

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

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

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

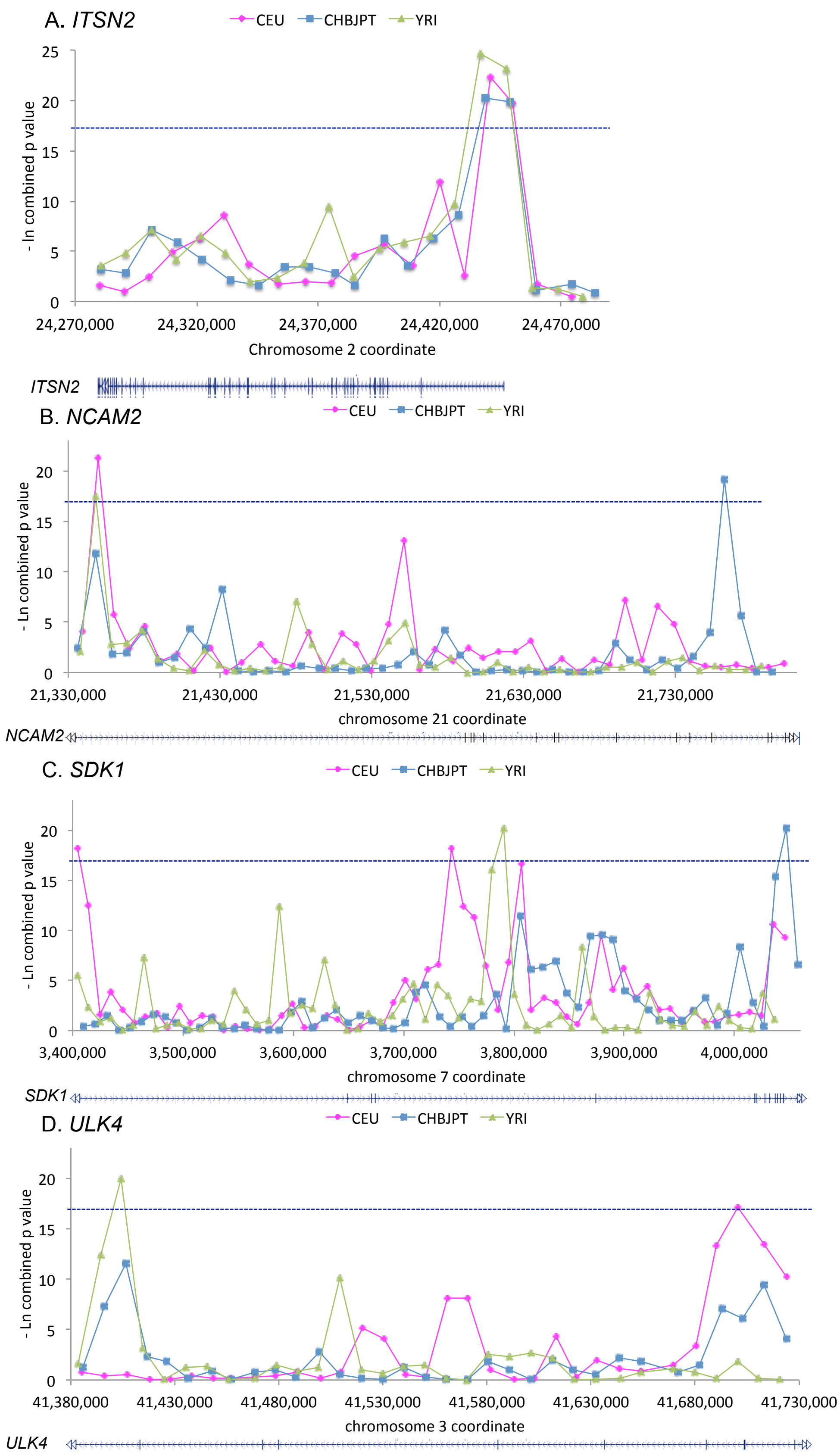


Figure 3.11 Examples of positively selected genes with signals in multiple populations.

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

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

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

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

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

3.6 Discussion

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

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

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

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

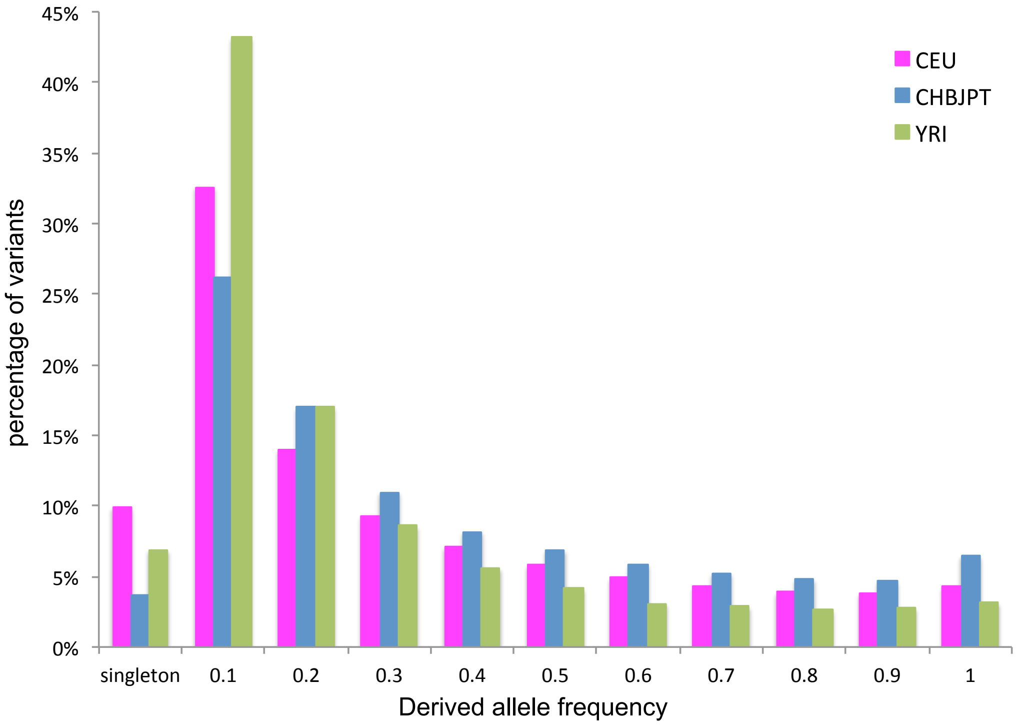


Figure 3.12 Frequency spectra of 1000 Genomes low-coverage pilot data in each population.

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

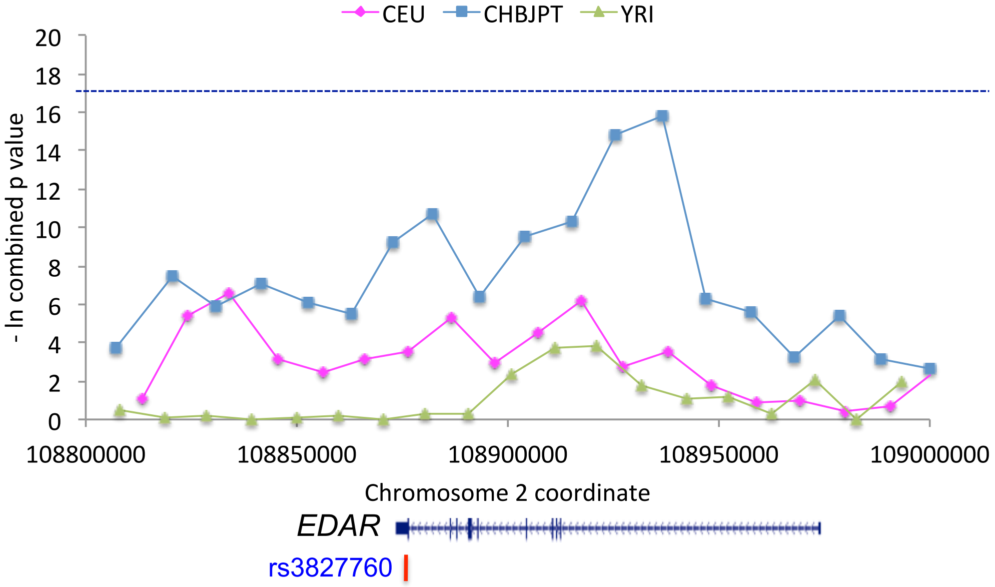


Figure 3.13 Signals of positive selection of EDAR gene in the CHB+JPT population. As observed previously, the peak signal lies in an intron, and not over the non-synonymous SNP rs3827760 often assumed to be the target of selection.

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

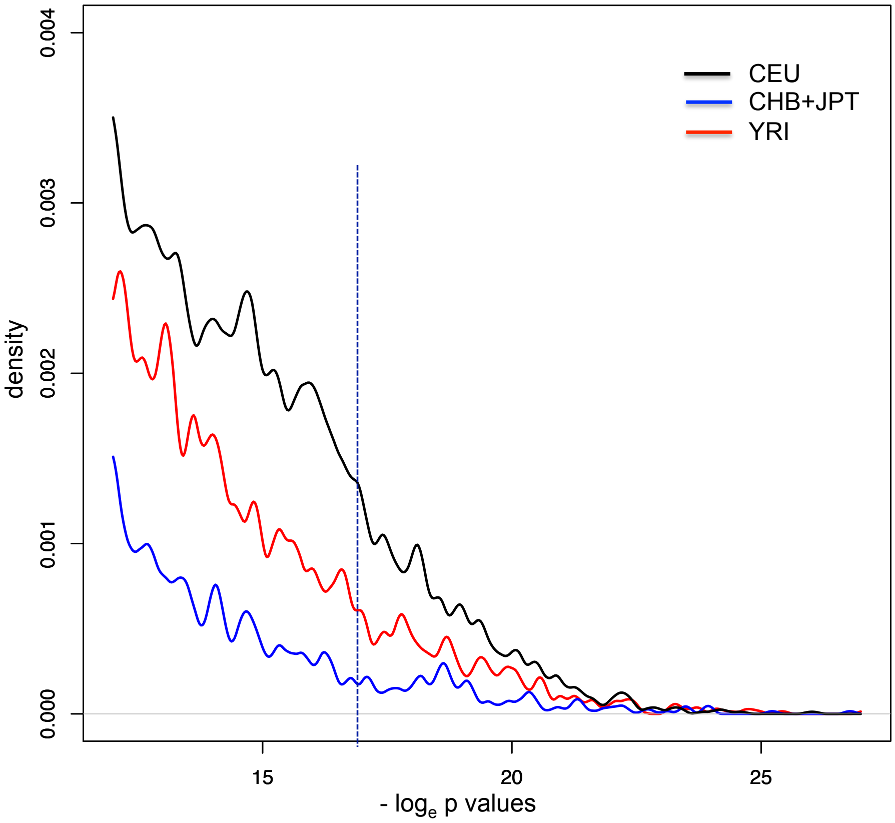


Figure 3.14 Distributions of p values in three populations. In order to show the difference of densities of significant p values in each population, we only showed the distributions of those (-$\log_e$ p) values bigger than 12 (equivalent to p values smaller than 6.1E-6). The blue dashed line is the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

With a much higher density of variants in the sequencing data, we were hoping to achieve a better resolution of signals, which may lead to higher power of localization of selection targets. In our genome-wide scan, we used windows sized about 10 kb, which in general contain enough variants to have the statistical power, and at the same time are small enough for further investigation to identify the selected variant. However, although on average, it is mostly likely that the selected allele will fall into the window with the strongest signal, there is still a high chance that the selected allele is elsewhere. Our simulations suggested that there is about 75% chance that the selected allele will be within the ~100 kb regions centered by the peak signal, though the signal pattern is made more complicated by recombination near the selected allele. Therefore, although it is still not easy to localize the selection target into a very small region or even a variant, by taking into account recombination, we were able to localize the selected region into a reasonable size for further investigations.

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