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

4.1 Introduction

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

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

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

As has been noted, anatomically modern humans emerged around 200 KYA. This ancestral human population lived in Africa (with a temporary expansion into the Levant) until around 50-60 KYA, when a subgroup of them with fully modern characteristics migrated out of Africa and populated other parts of the world. Selective sweeps on alleles that contributed to modern human traits should have occurred around the time when modern human emerged, and should have reached fixation before the out-of-Africa migrations. Therefore, if we trace back to the common ancestor of one of these loci, the TMRCA should be around or slightly more than 200,000 years. If we use the unit of $2N_e$ generations, and an N_e of 10,000 for the human population, the TMRCA should be around or a little more than 0.5 and less than 2, as 2 should be the expected value of TMRCA for a diploid Wright-Fisher population. Therefore, the TMRCA of the selected locus should be much less than what we would expect from a neutral region, so we may distinguish these regions that had undergone a complete selective sweep during modern human evolution from neutral regions by calculating TMRCAs of human genomic regions and identifying the most recent ones.

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

4.2 Materials and Methods

4.2.1 Data

To estimate the coalescence time of a particular genomic region, we need the complete set of single nucleotide variation in a set of unrelated samples. According to coalescent theory, in an unstructured population the probability of a sample size n containing the most recent common ancestor of the whole population is $(n\text{-}1)/(n\text{+}1)$, so even with a small sample size of 10, we would still have a more than 80% chance to obtain the TMRCA of the whole population from the sample. However, due to the complex structures of human populations, in order to obtain TMRCAs in all humans, we need samples that can represent at least all the main continental human populations. So in order to conduct a genome-wide survey of TMRCAs in humans, we needed high-coverage whole-genome sequencing data from a diverse collection of human samples. When this project started in 2010, there were 15 personal genomes sequenced at high coverage by different research groups around the world. These include a YRI and a CEU trio from the 1000 Genomes Project pilot 2[40], Venter's[155] and Watson's[156] genomes, one Chinese genome (YH)[157], two Korean genomes[158,159], two European genomes from Complete Genomics Inc.[160], and one Bantu and one Khoisan individual from southern Africa[161]. These individual genomes have diverse population backgrounds, thus formed a good sample of the global human population. We first used 13 out of these 15 individuals (excluding offspring in the two trios) to calculate coalescence times, but found that due to the diversity of platforms used in sequencing those genomes, and different algorithms applied in variation calling, the data quality was not consistent from one genome to another, and when putting these genomes together, there were a lot of genotype gaps and violation of the infinitely-many-sites model. Therefore, it was not useful to calculate coalescence times on these genomes.

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

Low quality sites were removed and missing genotypes were filled before using these data for our analyses. Firstly, trialelic sites, telomere and centromere regions, as well as sites that are not consistent with the Mendelian inheritance in the CGI trios and the pedigree panel were excluded. Because of the highly diverse samples, we avoided using inference algorithms to infer missing genotypes, as inferences from a large number of mixed populations may be inaccurate. Instead, we filled the majority of missing genotypes using the 1000 Genomes Project Phase 1 data in the same samples (34 samples in common) (http://www.1000genomes.org/). We then discarded sites that still had more than two missing genotype calls. For those with one or two missing genotypes, we assigned either the reference or alternative allele as the genotype based on the genotypes of other samples in the same population. After the filtering, around 95% of the SNPs were retained.

Table 4.1 Sample information.

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

4.2.2 Divergence and diversity

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

In order to calculate human divergence from the ancestor, we obtained the inferred ancestral state of each locus across the whole genome from Ensembl (http://www.ensembl.org/). The ancestral states are inferred from the six primates EPO (Enredo-Pecan-Ortheus) pipeline (see Ensembl website for details). We then identified fixed derived alleles in humans based on the 54 CGI genomes and the ancestral alleles. Divergence per site on the same 5-kb non-overlapping windows was calculated as for diversity.

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

4.2.3 TMRCA calculations

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

4.2.4 Simulations

We simulated 1000 independent 100 kb neutral regions in 54 samples, using the *cosi* package[26] and the best-fit demographic model[26]. Due to the limited demographic models, only three main continental populations, i.e. African, European and Asian, were simulated. We categorized the 11 populations in the CGI samples into these three population groups, which gave us 22 Africans, 18 Europeans and 14 Asians. We first used *cosi* to generate a random recombination map using the distribution of recombination rates in autosomes in the deCODE genetic map[168], and then used this recombination map in the simulations. A genome-wide average mutation rate of 1.5×10^{-8} and gene conversion rate of 4.5×10^{-9} were used. All other parameters are the same as in previous simulations.

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

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

In order to investigate whether or not these regions with recent coalescence times calculated from CGI data are likely to have undergone strong selective sweeps during the emergence of modern humans, we compared the variants in the 143 regions with those in two high-coverage southern African genomes – one Bantu and one Khoisan[161]. The Bantu sample ABT was sequenced to over 30-fold coverage using the SOLiD 3.0 platform from Applied Biosystems. The Khoisan sample KB1 was sequenced by two platforms: 10.2-fold coverage using the Roche/454 GS FLX platform, plus 12.3-fold non-redundant clone coverage with long-insert libraries, and 23.2-fold using the Illumina platform[161]. We used the variation data generated by the authors. We also compared the variants with those of a Denisovan genome, sequenced by Reich et al. (http://www.eva.mpg.de/denisova/), with approximately 30-fold coverage using the Illumina GAIIx sequencing platform[12]. First of all, we called variants differing from the human reference genome GRCh37 from the alignment generated by the authors, using SAMtools[169]. We used a maximum read depth of 100 as a filter (~3 times of the average read depth). We then further filtered out heterozygous calls where the ratio of the second-highest:total read depth was less than 0.3:1, or the second-highest read depth was less than 2. Then we obtained all variants in the 54 CGI samples, two southern African genomes and the Denisovan genome within the 143 regions with recent TMRCAs, as well as 100 sets of random windows matching the number of windows in the recent coalescent regions. Firstly, we used the two southern African genomes to validate our human-fixed derived alleles. Only those derived alleles that were fixed in both the CGI and the southern African samples were considered as fixed derived alleles in humans. We then counted the number of the following four types of loci in each set of regions: (1) the derived allele was only seen in the Denisovan genome: a "Denisovan specific variant"; (2) the derived allele was fixed in humans but not seen in the Denisovan genome: a "human specific variant"; (3) the derived allele was seen in both humans and the Denisovan genome, with a frequency in the 54 humans higher than or equal to 50%: "high DAF shared variant"; and (4) the derived allele was seen in both humans and the Denisovan genome, with a frequency in the 54 humans less than 50%: a "low DAF shared variant". In order to test whether or not there was any enrichment, we randomly picked 100 sets of windows with calculated TMRCAs, matching the number of windows in our recent coalescent region set. Then we ranked the numbers of these four types of derived alleles in the recent coalescent regions against the 100 random sets of matched windows to see if any type of alleles was enriched in the recent coalescent windows compared to the random windows.

4.2.6 Phylogenetic network analysis on regions with recent TMRCAs

In order to further understand the relationship between the haplotypes in humans and the Denisovan, we performed phylogenetic network analysis on some regions with recent TMRCAs using the NETWORK software[170] (http://www.fluxus-engineering.com/sharenet.htm). Human haplotypes were inferred using BEAGLE as described before, and heterozygous sites in the Denisovan were assigned to the two chromosomes manually based on the similarities with the human haplotypes. For those Denisovan variants that are not shared with humans, alleles were randomly assigned to the two haplotypes. Then these haplotypes were grouped into African (ASW, LWK, MKK, YRI and southern African), European (CEU and TSI), Asian (CHB, JPT and GIH), other human populations (MXL and PUR), and Denisovan. Phylogenetic networks were built, and each node was marked with colors representing the relevant population group(s).

4.3 Results

4.3.1 Divergence and diversity

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

4.3.2 TMRCA distribution on low and high diversity/divergence regions

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

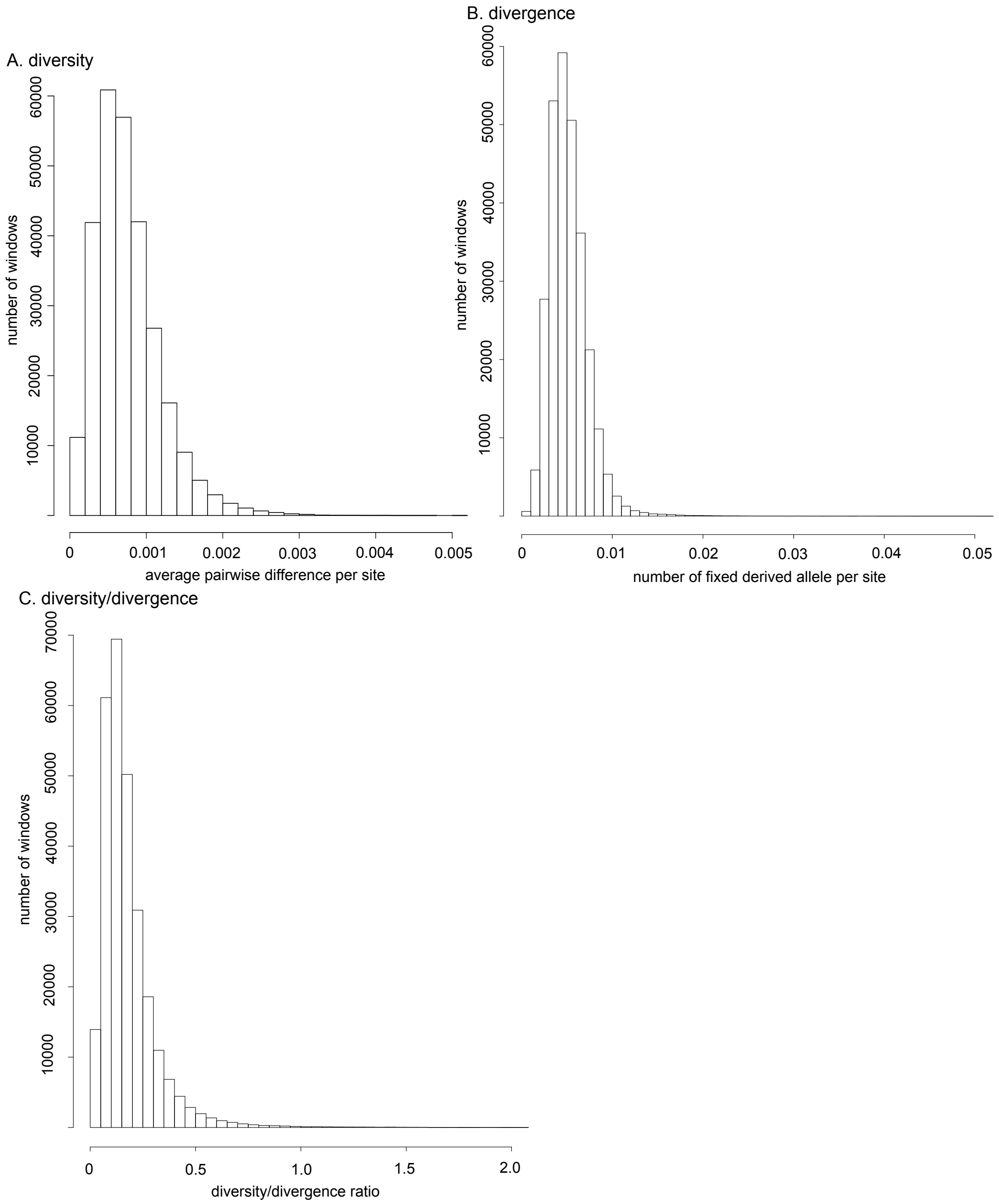

Figure 4.1 Diversity and divergence distributions of the 5-kb windows in the CGI data.

1% highest diversity/divergence ratio, as well as those with the 20% lowest diversity/divergence. As we would expect, the distribution of TMRCAs of the 1% lowest diversity/divergence windows is narrow and sharp, with a median of ~1.5, while that of the 1% highest diversity/divergence windows is much wider and flatter, with a median of ~6.3 (Figure 4.2). The TMRCA distribution of 20% lowest diversity/divergence windows, as we would expect, is slightly fatter and more towards the right, compared to the 1% lowest distribution (Figure 4.2). The median of these TMRCAs is ~1.6, slightly smaller than the expected genome average of 2, which is as we would expect, since in general, lower diversity/divergence regions tend to have a smaller coalescence time.

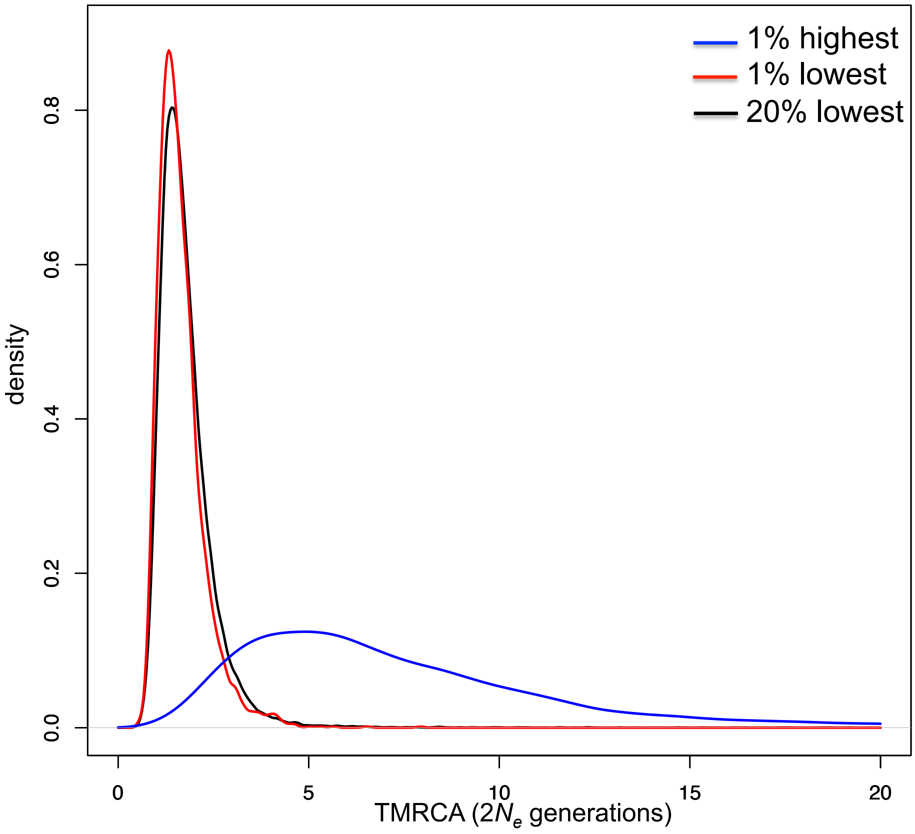

Figure 4.2 TMRCA distributions in the CGI data. This plot shows density distributions of TMRCAs in the 1% highest diversity/divergence windows (blue), 1% lowest diversity/divergence windows (red), and 20% lowest diversity/divergence windows (black).

4.3.3 Validation of TMRCA estimations by simulation

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

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

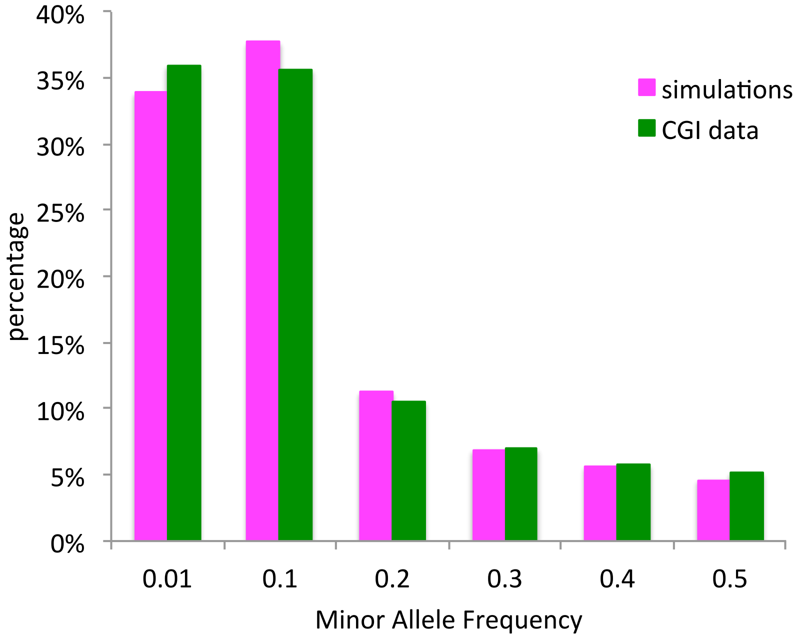

Figure 4.3 Minor allele frequency spectra in the CGI and simulated data.

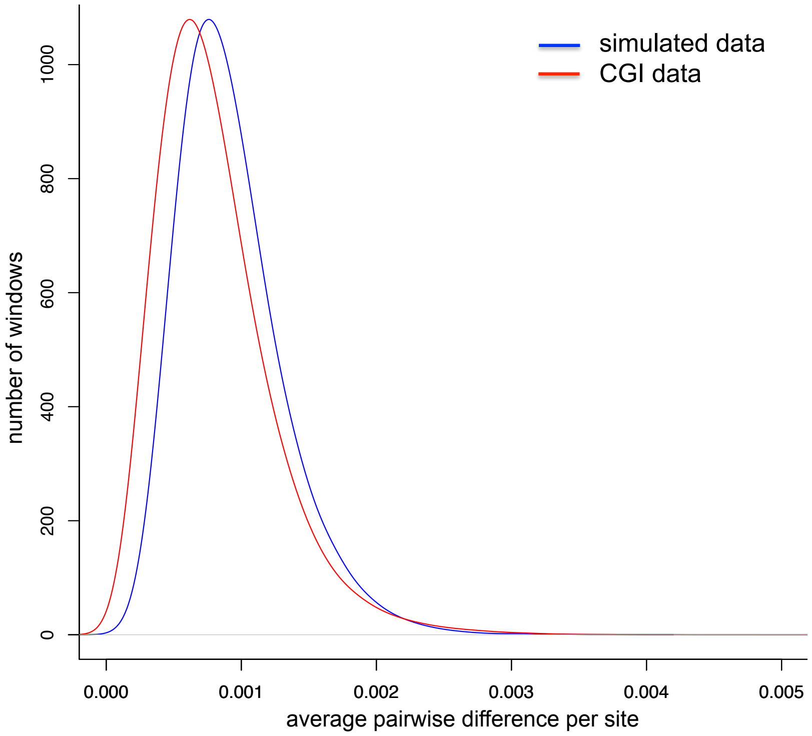

Figure 4.4 Diversity distribution of simulated and CGI data. This plot shows the density distributions of diversity per nucleotide in the CGI data (red) and the simulated data (blue).

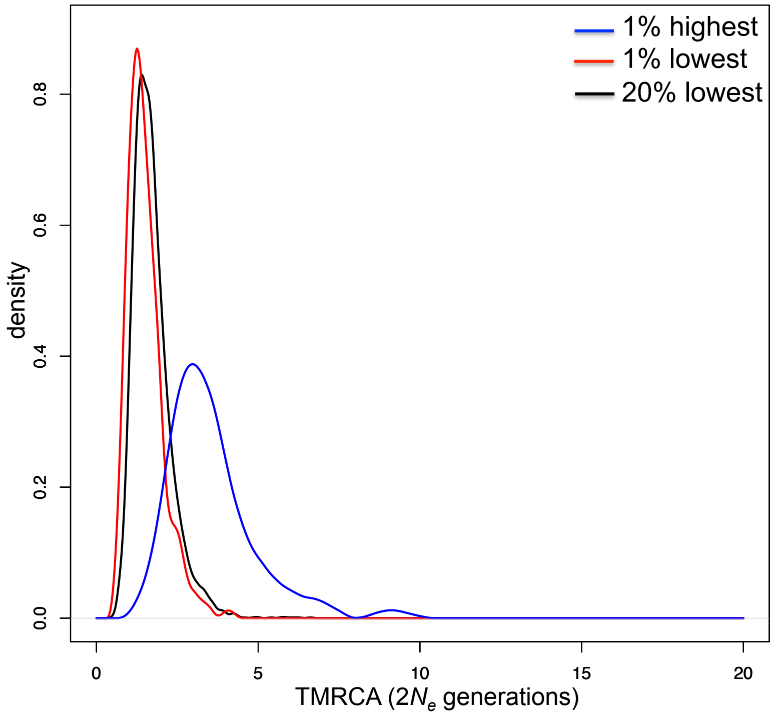

Figure 4.5 TMRCA distributions on simulated windows with different diversity. This plot shows density distributions of TMRCAs in the 1% highest diversity/divergence windows (blue), 1% lowest diversity/divergence windows (red), and 20% lowest diversity windows (black) in the simulated data.

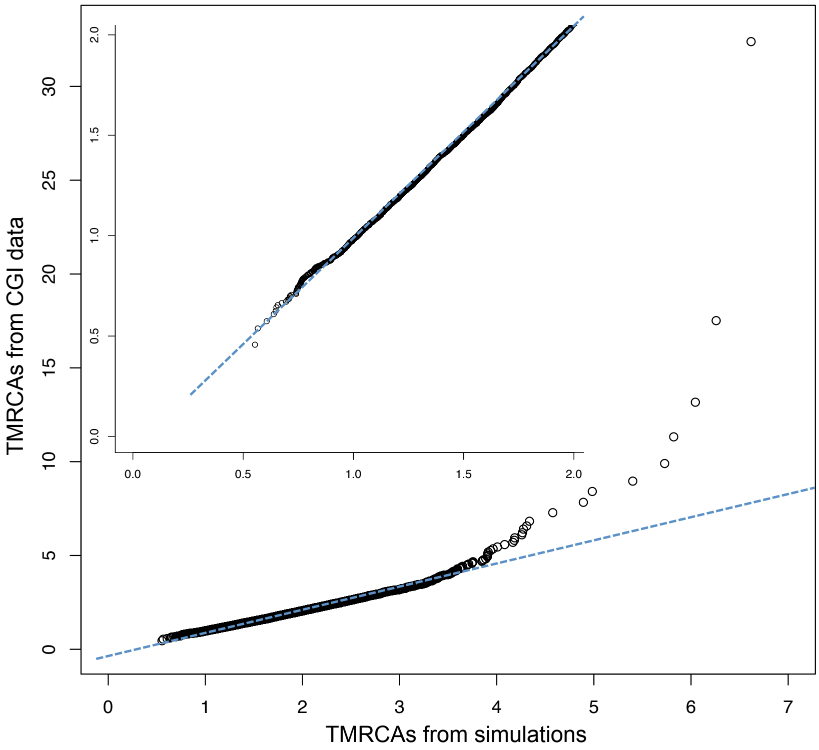

Figure 4.6 Q-Q plot of TMRCAs in simulated data versus the CGI data. This Q-Q plot shows the TMRCAs of simulated data (X axis) versus CGI data (Y axis), blue dashed line is the trend line. The smaller plot on the upper left corner is the magnified Q-Q plot for the part where TMRCAs are less than 2.

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

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

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

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

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

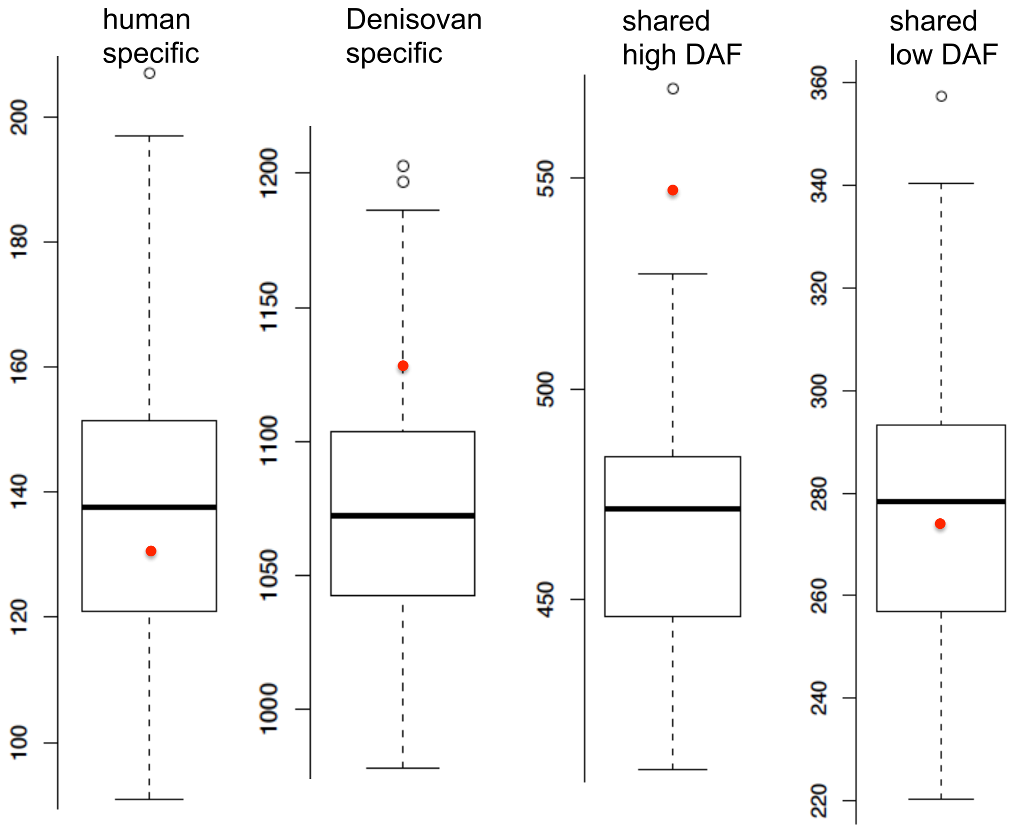

Figure 4.7 Comparison of numbers of four types of derived alleles in humans and the Denisovan genome. These box plots show distributions of numbers of each of the four types of derived alleles in 100 random sets of windows matching the recent-TMRCA windows. Red dots are the corresponding values of the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

4.3.5 Phylogenetic network analysis on regions with recent TMRCAs

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

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

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

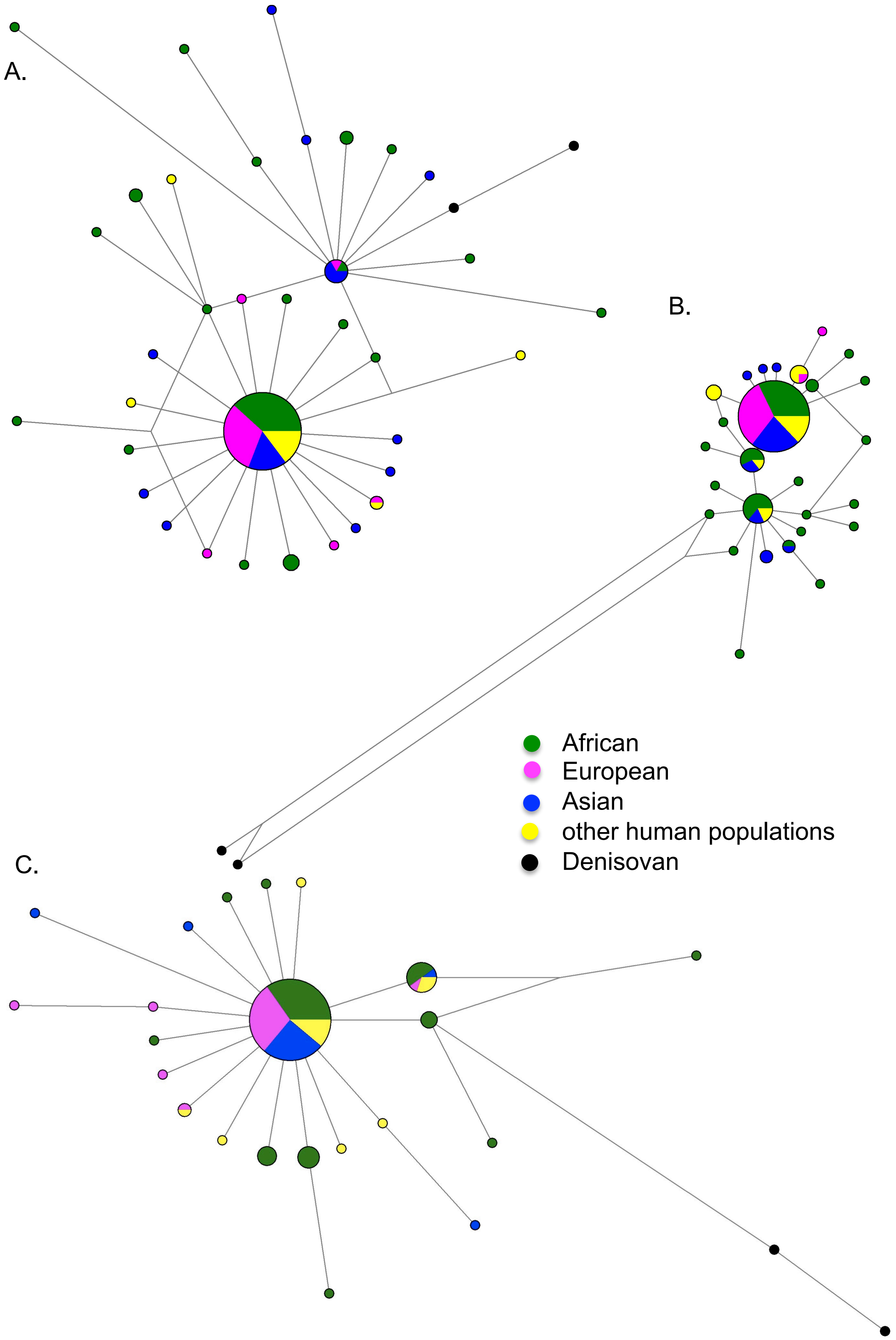

Figure 4.8 Phylogenetic networks of three regions with recent TMRCAs. A: region chr1:32,660,001-32,665,000; TMRCA 0.952 *N_e* generations. B: region chr19:16,465,001-16,470,000; TMRCA 0.992 *N_e* generations. C: region chr11:46,430,001:46,440,000; TMRCA 3.692 *N_e* generations.

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

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

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

4.4 Discussion

This study has for the first time used whole-genome sequencing data from a diverse panel of human samples to systematically estimate coalescence times across the genome in humans, aiming to identify regions that share a very recent common ancestor among all humans, which may indicate positive selection during the early stage of modern human history ~200 KYA. This approach is complementary to the statistical tests used in Chapters 2 and 3, as well as to other LD-based tests, and differs from them in two aspects: one is that it detects selective sweeps that were much older than those statistical tests, and the other is that it only detects complete selective sweeps, where the statistical tests have very limited power.

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

Based on these results, it seemed that we could draw the preliminary conclusion that there is no excess of “human-exclusive” regions spreading to all humans during the early stage of modern human history. However, there are several limitations to this study, which may prevent us from drawing such a conclusion. Firstly, the model used by GENETREE might be too simplistic, so the estimation of TMRCAs might not be accurate. GENETREE assumes a Wright-Fisher population, with no recombination, and an infinitely-many-sites model. Although these may provide a good approximation in most cases, in order to make an accurate estimation of coalescence times, we may need a more realistic model. Secondly, we do not have proper independent sister species to use as outgroups of modern humans in this study. Ideally, we hope to have genomic information from some hominin species that diverged from humans not too long before the modern human emergence, and did not experience much gene flow with modern humans. Although the high-coverage Denisovan genome provided the closest approach to these requirements, it has limitations. For example, there might have been substantial gene flow between Denisovans and humans[12,18], and we only had one Denisovan genome sequence to use. Thirdly, the ancestral alleles were inferred from the primates that split from ancestors of humans several million years ago. This timescale might be too long for our purpose in this study, because multiple mutations will have occurred at some sites. It may be better if the ancestral alleles were inferred from species that are closer to humans.

Nevertheless, despite the limitations mentioned above, the results from this study serve as a first step for the genetic understanding of early modern human evolution. We have seen that strong classic selective sweeps might not have played a major role in the emergence of modern humans. It is more likely that the traits made us modern humans were the results of accumulation of mutations throughout a long period of time, and those genetic changes might have been present in our ancestors for a long time before modern human emerged. However, this does not mean that positive selection did not play a role in shaping early modern humans. Instead, this may indicate in most cases selection might have happened on existing alleles, or in a moderate manner rather than strong selective sweeps. In fact, by drawing gene networks of some regions with TMRCAs of less than $4N_e$ generations in humans and the Denisovan, we found some that showed patterns of rapid expansion of one haplotype specifically in humans. An example is the gene *AMBRA1*. A 10-kb region within this gene had a TMRCA of less than 4 N_e generations, and gene network analysis of this region shows a clear pattern where all human halplotypes in this region were derived from one central haplotype from the African population, which was different from the Denisovan haplotype (Figure 4.8 C). Studies have shown that this gene involved in autophagy and may regulate the development of the nervous system[172]. Some other genes that overlap with the regions with recent TMRCAs listed in Table 4.2 also seem to play important roles in humans. For example, the gene *PTAFR* is a receptor for platelet activating factor, a chemotactic phospholipid mediator that possesses potent inflammatory, smooth-muscle contractile and hypotensive activity (http://www.uniprot.org/uniprot/P25105 - section_comments); another gene *SATB2* may play an important role in palate formation and act as a molecular node in a transcriptional network regulating skeletal development and osteoblast differentiation (http://www.uniprot.org/uniprot/Q9UPW6 - section_comments). These examples indicate that classic positive selection might have shaped some genes or regions that contributed to modern human traits, but the number of such regions is not large. Also, due to our limited knowledge about gene functions in humans, it is often difficult to judge whether a gene is likely to have contributed to the modern human uniqueness. Of course, more studies are needed to understand these processes and to answer the question of what are the critical genetic changes that made us modern humans. Apart from estimating coalescence times of human genomic regions using more realistic models, we could systematically build phylogenetic trees of regions in humans and our sister species, in order to identify regions that are well-separated between the species. Therefore, the availability of additional high-quality genetic information from those hominin groups will be a key factor for the success of this type of study.