IDENTIFYING SOURCES OF GENETIC DIFFERENTIATION IN HUMAN POPULATIONS

by

Brett Jacob Kennedy

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Anthropology

The University of Utah

August 2013

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Brett Jacob Kennedy**

has been approved by the following supervisory committee members:

Henry Harpending, Chair **5-10-2013**
Date Approved

Alan Rogers, Member **5-10-2013**
Date Approved

Dennis O'Rourke, Member **5-10-2013**
Date Approved

Kristen Hawkes, Member **5-10-2013**
Date Approved

Jon Seger, Member **5-10-2013**
Date Approved

and by **Dennis O'Rourke**, Chair of

the Department of **Anthropology**

and by Donna M. White, Interim Dean of The Graduate School.

ABSTRACT

The main concern of human population genetics is to identify and describe genetic differences between groups of people. These differences give insight into the evolutionary processes and unique histories that have shaped these populations. A better understanding of human genetic diversity will lead to a better understanding of the biological systems that underly human phenotypic diversity. Here I explore three processes which have led to population differentiation in modern humans. First, I examine how differential disease risk across continents may have (or may not have) led to differences in allele frequencies immune-related genes. Second, I describe a method for discovering genomic regions in admixed populations that appear more similar to one parent population than the other. This method highlights regions which may have very recently been under selection in these populations. And finally, using the same method I attempt to discern regions of the genome in modern humans that may have been shaped by archaic admixture.

For Beth.

CONTENTS

CHAPTERS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Special thanks to Henry Harpending for his guidance, patience and generosity. Thanks to Alan Rogers, Dennis O'Rourke, Kristin Hawkes, Jon Seger for their advice and help. Also thanks to Ted Coxworth, Chris Parker, Earl Keefe and any other graduate students or faculty I pestered.

CHAPTER 1

NO ETHNIC BIAS IN DISTRIBUTION OF DISEASE ASSOCIATED CYTOKINE POLYMORPHISMS

1.1 Introduction

Among clinical risk factors associated with race, group differences in proneness to inflammation and autoimmune related diseases may be the most important, as well as the most poorly understood [1]. Medical studies have observed differential risk for immune and inflammatory related disorders between Africans and Europeans for decades, but the genetic factors leading to this difference remain something of a mystery [2, 3]. Individuals of Central African descent are subject to a higher risk for a number of autoimmune disorders, including multiple sclerosis, lupus erythematosus, tuberculosis, septicemia, and several types of cancer associated with chronic infection and inflammation [2]. Markers of inflammation such as C-reactive protein and homcysteine, often associated with cardiovascular disease, are also elevated in African Americans [4]. In spite of recent advances in immunosuppressive therapies and better donor matching, African Americans are more likely than either Europeans or Asians to experience renal allograft failure. Individuals of Asian descent, on the other hand, have a higher than expected renal allograft acceptance rate, even with mismatched donors [5].

Taken together these clinical risk factors strongly suggest a genetic basis for observed differences in proneness to inflammation among racial categories. However, the complexity and scale of the human immune system makes it difficult to identify the specific genes and polymorphisms that contribute to the difference. Because of the integral and basic role that immunomodulatory cytokines play in inflammatory response, racially associated differences in cytokine single nucleotide polymorphism (SNP) frequencies have been suggested as a possible cause of the differences we observe in autoimmune disease risk and inflammatory immune response.

Cytokines are humoral proteins or glycoproteins that mediate and facilitate immune response. They bind to specific cytokine receptor ligands in target cells and can induce gene activation, leading to meiotic division, growth and differentiation, migration, or apoptosis. These effects can occur directly as a result of cytokine activity, or indirectly though cytokine mediated responses from other parts of the humoral system [6]. Generally, we can divide cytokines into those that facilitate inflammation and those that mediate it. In other words, there are cytokines that increase the inflammatory immune response, such as Tumor Necrosis Factor (TNF) and Interleukin 1 (IL-1), and those that mediate it directly, either through reduced inflammation or by inhibiting the actions of the proinflammatory type, such as IL-10 and IL-4.

Cytokine genes exon sequences are highly conserved due to their vital role in immune response and the consistent directional selection in immune related gene regions [7]. Nearly all of the polymorphisms found in human cytokine genes are not in the coding region, but rather in or near the 5- and 3- regulatory sequences or introns. While these polymorphisms do not affect the amino acid sequence they can alter the expression of the gene in other ways such as changing transcription rates. This fact, combined with the already flexible, complex, and sometimes-redundant cytokine response system, creates a group of genes with a possible capacity to rapidly adapt to local pathogen pressures though natural selection. A pattern of frequency differences among these cytokine SNPs between populations may be able to inform us not only of population history, but also how recent natural selection has shaped our immune system. As humans dispersed out of Africa, disease pressures and pathogen load changed, and a corresponding change in immune response is likely to have followed. As a species, we exist in a vast array of geographies, each with different immunological challenges. It is no radical assertion to suggest that individual populations' immune response have adapted as a result of local selective pressures. And it has already been demonstrated that recent selection in a number of human genes is not only possible, but also very likely [8]. The more specific questions to be addressed are whether these changes have accumulated into regional—and possibly continental—differences in immunological profile, and what consequences this may have for our current understanding of biological diversity of the human species.

The interaction between cytokines and the rest of the immune system is complex, coordinated, and flexible. One of the primary challenges in discerning the function of a polymorphism in any cytokine gene is that there are significant redundancies and overlap between their individual functions. In other words, even if a single SNP up-regulates or down-regulates transcription of a particular cytokine, the final effect on the immune system could be minimal. Therefore, in order to detect the existence of any type of generalized immunological pattern, we must observe a number of polymorphisms with the same or similar effects as approximated by changes in immune related disease risk changes. An accumulation of pro- or anti-inflammatory variants could hypothetically alter the overall immune response of a population and thereby alter the individual risks for autoimmune or immune related disease.

In the battle between geographically local adaptation to pathogen load and the pathogens themselves, we are likely to see corresponding differences in disease frequency that are reactions to changes in the immunological genetic profile of the local population. In other words, cytokine allele frequencies adjust to local pathogen load, and local pathogens adjust to local immunological adaptations, thus creating a feedback loop where differences may accumulate. There is some expectation for regional or global patterns of variation based on a large number of cytokine SNPs. The difficulty of detecting these patterns is further exacerbated by the rapid and regional selection in immune systems. Because of selection for immune system diversity and consistent, but ever-changing directional selection (similar to the HLA system), few of the cytokine genes show strong signals of recent selection. Most exhibit relatively little linkage disequilibrium (LD) and have low scores on other selection detecting metrics. As Yazici [9] points out, different cytokine genotypes exist in a population, mainly as a result of geographically localized natural selection imposed by invading microbes and hostpathogen interactions. Therefore, association of cytokine gene polymorphisms with a particular disease observed in a single population cannot be extrapolated to other populations with different genetic background. As local human populations and their pathogens co-evolve, the mutations that will give a selective advantage may change rapidly, and what was useful yesterday may be less advantageous today.

That the immune system may be more active in Africans and individuals of African ancestry makes historical sense. Human-adapted and vector-transmitted diseases have been in Africa longer than they have been in any other part of the world. Chronic infectious disease would have been a larger portion of the pathogen load than acute infections, which would have died out quickly given the low population density [10]. As human ancestors left Africa, the reduced pressure from the absence of malaria alone would likely precipitate a tuning down of the inflammatory immune response. Other factors such as dietary shifts and seasonal vector transmission would further the selective advantage for a less active inflammatory immune system. The tradeoff between a strong inflammatory response and the increased risk of autoimmune or inflammatory disease makes the cytokine gene regions likely candidates for evolution as human populations spread across the globe. However, because of the nature of selection in immune gene regions, the traditional signals of selection (such as LD) may have decayed rapidly, as diversification and highly localized selection increased after the initial diversification.

There have been a number of previous studies focusing on population level differences in cytokine polymorphism frequencies. Largely, these studies have utilized relatively small clinical populations and limited their study to only a few cytokine polymorphisms [11, 12, 13]. Only one study has made broad use of the online genome databases such as HAPMAP [3]. The question that these cross-population studies have asked is whether we can describe Africans or Europeans as having a broadly pro- or anti-inflammatory cytokine profile which influences the observed clinical differences in proneness to inflammation. Their conclusions have been mixed. The clinical study by Ness et al. [12] observed that, among the eight cytokine SNPs they tested, African American subjects differed significantly in six of the genes. In all six of these SNPs, the African Americans had higher frequencies of the proinflammatory variant. On the other hand, Van-Dyke [3] looked at a much larger number of SNPs and found that proinflammatory variants were not always found at higher frequencies among African Americans, although their data set was not clearly able to determine function for all of the SNPs.

The goal of this paper is to take a wider view of as many cytokine polymorphisms with discernible function or association as possible and attempt to find a pattern of variation that might explain some of the clinical differences we observe in proneness to inflammation between different populations. The patterns in clinical risk differences between Africans, Europeans, and Asians suggests that, if the key in their divergent risk is found in the cytokine system, then Africans should have a relative abundance of inflammation-linked polymorphisms or dearth of anti-inflammation linked SNPs, while Asian populations should display the opposite and Europeans should fall somewhere in between.

1.2 Methods

To test for population level patterns of variation in cytokine polymorphisms, a large number of autoimmune or inflammatory diseases associated SNPs in regions in or near cytokine genes were identified and typed for function using the Cytokine Gene Polymorphism Database as well as SNPs identified from the genome-wide association study (GWAS) database at GWAScentral.org [6, 14, 15]. Also included were SNPs from recent literature examining differences in cytokine polymorphism frequencies between racial groups [12, 3]. The typing was done based on previous similar works and confirmed by disease associations and function studies using online resources such as SNPedia and GWAScentral.org [12, 3]. SNPs that were identified in only one study population were removed to control for possible differences in function by population. Of course, because relatively few polymorphisms have been subject to genome-wide association studies in all three of the target populations (Asian, African and Europeans), there remains some concern for dissimilar function or association in each group. However, this problem is tempered by the fact that nearly all of the SNPs in question are in regulatory regions or introns rather than in the coding sequence. This suggests that these polymorphisms alter gene expression rather than function, so the difference is more likely to be of degree than kind.

After identification, the polymorphisms were divided into two categories. Cytokine SNPs that were associated with decreased inflammation or decreased risk of an inflammation related or autoimmune disease were typed as anti-inflammatory SNPs, while polymorphisms that were positively associated with inflammation or related diseases were typed as proinflammatory SNPs. The p-value threshold for inclusion in the data set was $-\log(p) \leq 2$ in each GWAS study. Of the 109 SNPs identified and typed in the databases mentioned above, 71 were present in all three target populations in the HAPMAP database. Because of their high SNP density and large sample size the Yoruba (YRI), CEPH European (CEU) and Han Chinese (CHB) samples in the HAPMAP where chosen as representative for their respective regions.

The derived frequencies of this 71 SNP sample were compared in a logistic regression analysis for each population, testing for higher frequencies of either pro- or anti-inflammatory alleles within the population. Comparing only derived frequencies within the population samples controls for any ascertainment bias that may be present. Ascertainment bias in the HAPMAP database is expected to result in general increase of derived allele frequencies outside of Africa due to the SNP discovery methods used in the HAPMAP project [16]. By comparing only derived frequencies differences of pro- and anti-inflammatory polymorphisms within each population, we control for any underlying patterns of allele frequency differences among the populations because there is no reason to expect the differences are affected one way or another. Additionally, because the allele at each locus in the study can be seen as either pro- or anti-inflammatory (i.e., one allele at every locus will be proinflammatory and the other anti-), the division of alleles into these categories is somewhat arbitrary.

In addition to the within-population comparison using logistic regression, the three populations were compared to each other using a number of tests that identify signficant differences in the distribution of pro- and anti-inflammatory alleles between them. The difference between the frequencies of each SNP in two populations were calculated and distributed into pro and anti-inflammatory categories, resulting in three sets of paired lists (pro- or anti-inflammatory) consisting of the difference between the frequencies in each of the paired populations (Europe vs. Asia, Asia vs. Africa, Africa vs. Europe). Each of the paired difference distributions was compared using Students T, Wilcoxon Rank-test, and Kolmogorov-Smirnov. These analyses test whether the frequencies in the two distributions are significantly unlikely to have been drawn from the same distribution. If there were significant bias in pro- or anti-inflammatory SNP frequencies between any of the populations the test would present as significant, meaning one of the two populations had higher or lower frequencies of inflammatory SNPs.

1.3 Results

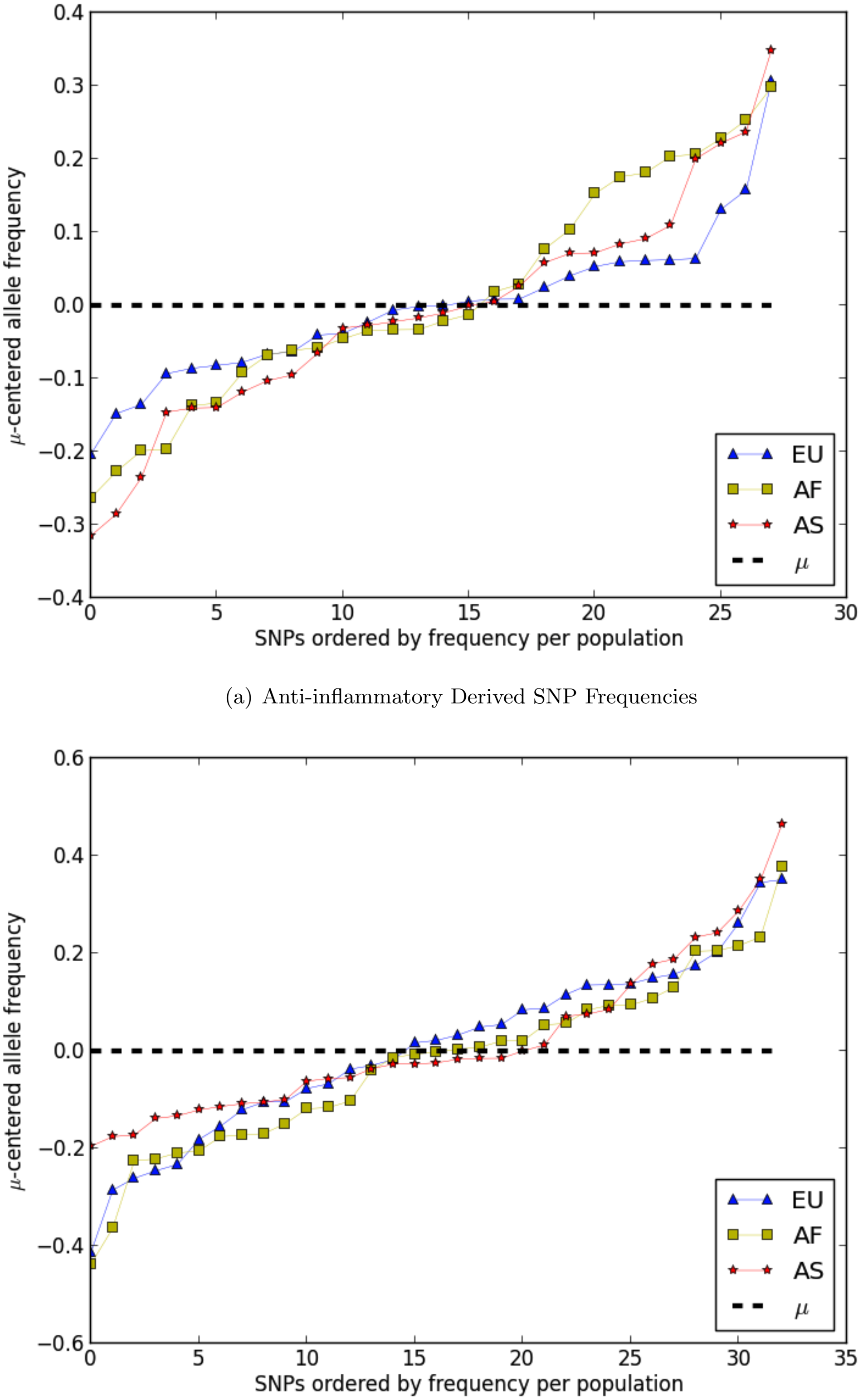
The pattern of derived allele frequencies revealed in the sorted frequency plots (Fig. 1.1) suggests that there are no significant differences in the frequencies of pro- or anti-inflammatory alleles in any of the populations. Similar to a histogram, this plot demonstrates the similarity in the distribution of pro- and anti-inflammatory allele frequencies within populations. The frequencies are mean (μ) centered then sorted, to try and visually find a pattern of increased or decreased allele frequencies of either classification within any population. None of the three populations appears to trend significantly in either direction, and they appear to be more similar to each other than not. However, there appears to be a slight trend for low frequencies of anti-inflammatory alleles in the African sample. These patterns may not hold on a SNP by SNP basis, but overall, the derived frequency histograms demonstrate almost no divergence from the average frequencies.

The same pattern is repeated in the logistic regression analysis. None of the three populations has a statistically significant trend in either direction ($p \gg 0.05$). Likewise, the between-population distribution comparisons showed no significant difference in the distribution of the pro- or anti-inflammatory alleles between the populations. Looking at the plots (Fig. 1.2), there is no appreciable difference between the distributions of cytokine polymorphism differences in any of the three comparisons. In fact, one would struggle to find a data set with less suggestion of a pattern; differences in the distribution of cytokine polymorphism allele frequency difference between any of the three populations is almost nonexistent. With a larger sample size, perhaps a pattern might emerge. But, given the relatively small number of SNPs for which we know function there does not seem to be any bias in the distribution of inflammation related cytokine polymorphisms.

A Students T power analysis for the data set (Fig. 1.3) and effect size shows that, within a reasonable power ($B \leq 0.7$), this sample size should be able to detect allele frequency differences between the populations of 10% or greater. While small differences in allele frequencies cannot be ruled out, these results strongly suggest that there are no large differences in the distribution of known disease-linked cytokine polymorphisms. The lack of differentiation between European, African and Asian populations casts doubt upon some current theories about the cause of clinical difference between the groups, but the result is perhaps unsurprising given the mixed results from previous studies [17, 18, 12, 1, 3]. That we simultaneously observe evidence of selection, yet no evidence of differentiation, suggests that classification of alleles into pro- and anti-inflammatory categories by inflammatory disease association may be insufficient, or these alleles are performing pro- and anti-inflammatory roles simultaneously.

1.4 Discussion

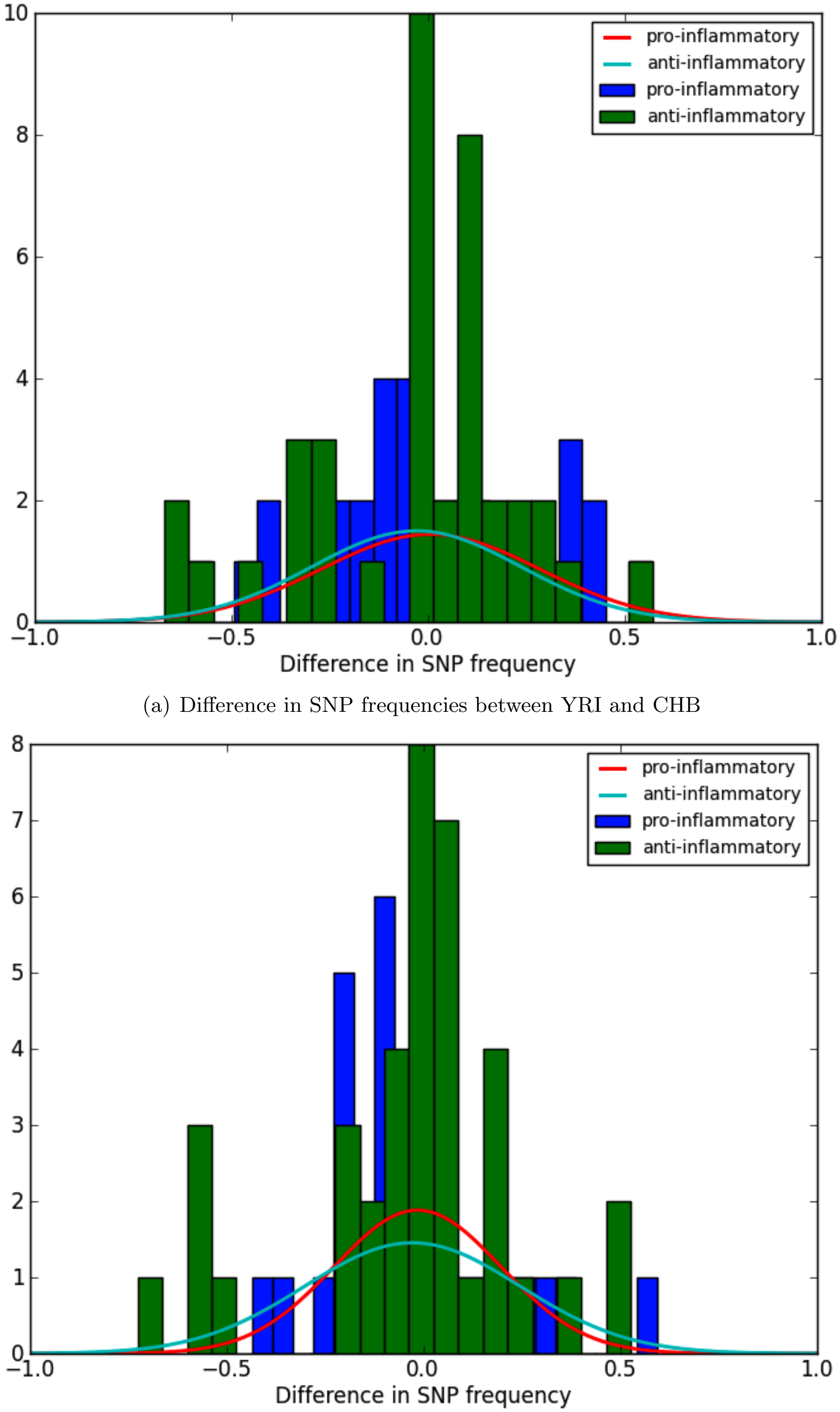
Among the 71 cytokine and cytokine-promoter polymorphisms of known function or association studied here, there appears to be no significant ethnic bias in the distribution of the alleles. Although individuals of African descent are at an observably higher risk for the diseases to which these cytokine polymorphisms are linked, they do not possess the disease-linked alleles in cytokine regions at any higher frequency than do individuals from other ethnicities. Of course, cytokines are only one part of our complex and interconnected immune system, and there are many other genetic factors related to inflammation outside of the direct inflammatory response of the cytokine system. Thus, the key factor in the difference may not be found in these cytokines, but may lie deeper in the immune response system or in a separate but related system. Because of their direct roles in inducing and mediating inflammation cytokines have been suggested as a good candidate for explaining the difference in immune function between individuals and populations. Unfortunately, our dataset of known cytokine polymorphism function is still somewhat limited and the results presented here do not conclusively show that the risk differential does not come from the cytokine system. Furthermore, we cannot easily account for effect size in this dataset. Although a number of GWAS studies from which this dataset was formed did include effect sizes, many of them were not mutually comparable, and others did not include a measure at all. However, the absence of bias in the distribution of those polymorphisms whose functions were most readily determinable as related to inflammation is suggestive that cytokines may not be the key to the ethic differences in clinical risk.



(a) Anti-inflammatory Derived SNP Frequencies

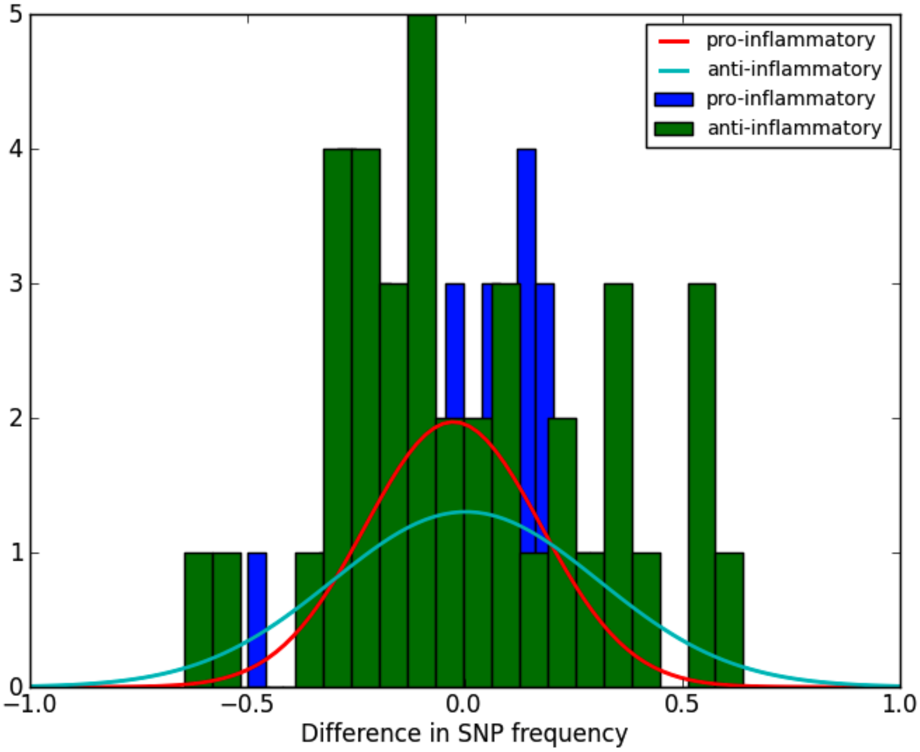
(b) Proinflammatory Derived SNP Frequencies

Figure 1.1: Sorted frequency plots of μ-centered derived allele frequencies.



(a) Difference in SNP frequencies between YRI and CHB

(b) Difference in SNP frequencies between CEU and CHB



(c) Difference in SNP frequencies between CEU and YRI

Figure 1.2: Between population allele frequency comparisons. Histogram of the locus-by-locus difference in each type of allele.

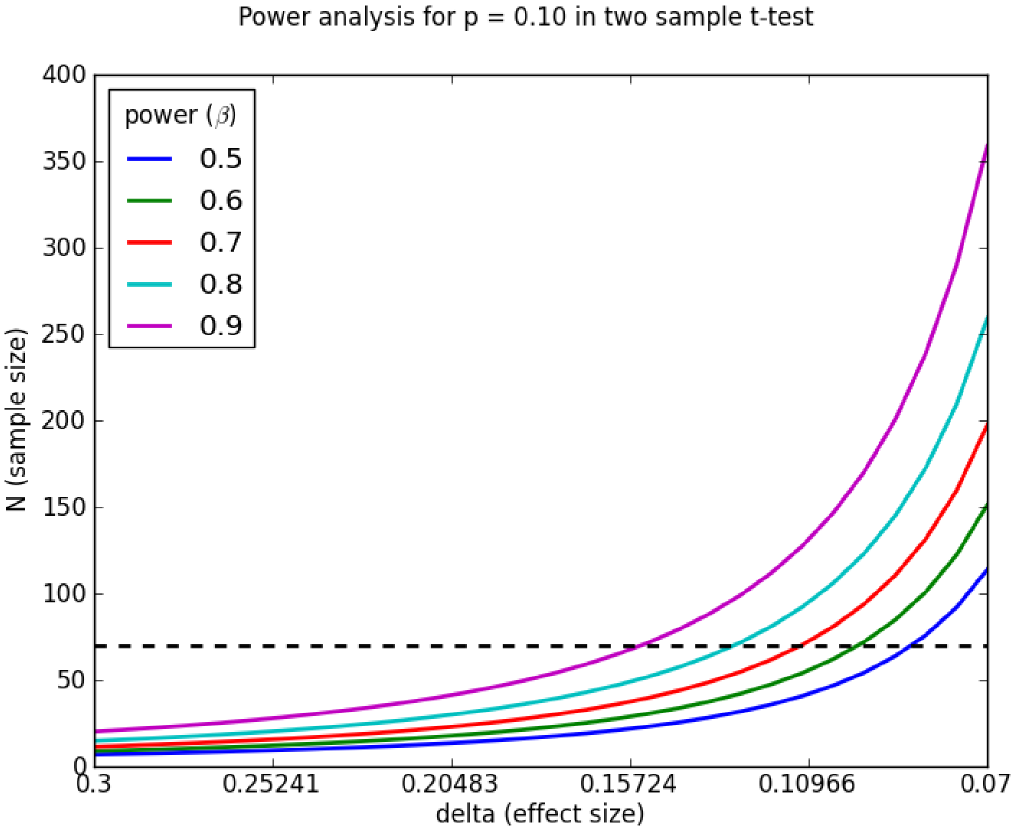


Figure 1.3: Power analysis plot for a dataset with effect range (in this case difference of allele frequency between two populations) δ = 0.30-0.07. The dashed line is the sample size for this study.

CHAPTER 2

MAPPING ADMIXTURE ACROSS THE GENOME USING PRINCIPAL COMPONENT ANALYSIS

2.1 Introduction

Understanding genetic population structure is important both to population genetics and to genetic epidemiology. Recognizing ancestry-associated biases in the distribution of alleles between populations can improve inferences of demographic and evolutionary history and allow for the control of population stratification in genome-wide association studies (GWAS). Admixed populations are of particular interest to these two fields because they can help to elucidate patterns of ancestry association and population history. Admixture occurs when two distinct populations, usually with separate continental origins, exchange genetic material. Individuals in admixed populations have mosaic chromosomes consisting of genomic segments of differing length inherited from either parent population. In other words, some regions of the individual's genome look more similar to one parent than the other. The average length of these segments is largely a result of recombination rates and the time since admixture occurred, but there are other processes that can influence which alleles from which parent population are more likely to be represented at greater frequencies in the population after admixture occurs.

The genetic distance between the parents of an admixed population, though primarily a product of genetic drift, will also be influenced by population-specific adaptations, especially natural selection related to disease, climate, or other factors. A number of studies have shown that substantial natural selection has occurred in human populations within the last few thousand years and has differentiated geographically distinct populations [8, 19]. Evidence has also shown that genetic risk factors for disease vary greatly between distinct human populations. Combined, this suggests that recently admixed populations are likely to have a higher number of functional genetic variants as compared to either parent population [20, 21]. Accordingly they are an important source of information for understanding population structure.

The relative abundance of functional variants in admixed populations makes rapid selection and large changes in allele frequencies likely. As a result, some chromosome regions that harbor functional variants inherited from only one parent will become more similar to the same region in the parent population, while other parts of the genome will do the opposite. Both natural selection and drift play a role in determining which of the variants inherited from the parent populations increase or decrease in frequency. In admixed populations, the distribution of admixture is not the same across all individuals; as a result, admixture may be unevenly distributed throughout the average genomes of the populations. Over time, as recombination breaks up linkage disequalibrium, the allele frequencies of most variants inherited by the admixed population will approach an intermediate frequency determined, on average, by the relative genetic contribution of each parent population (i.e., the proportion of ancestry). Regions of exceptional convergence are of particular interest because they represent regions of possible recent selection as well as regions of considerable ancestry biased population structure. Here we introduce a method for detecting such regions.

There are many approaches used to detect genetic structure in human populations. Some use extended haplotype comparisons (HAPMIX); others use bayesian clustering (STRUCTURE)[22, 23, 24]. But the most common and the oldest approach is principal component analysis (PCA). Principal component analysis (PCA) has a long history in population genetics, from Cavalli-Sforza and Edwards' [25] analysis of blood group frequencies to more recent methods for detailed analysis of population divergence or correcting for structure in GWAS [26, 27]. The PCAs of populations created today from whole genome data are remarkably similar to those created from blood group frequencies 50 years ago. The type and basic shape of the information that we obtain from modern analyses of thousands or millions of variants has not drastically changed but interpretations of the results have. For example, the clinal pattern of allele frequencies between populations in Europe was originally taken as evidence of the expansion of neolithic farming populations. However, more recent work has pointed out that the same clinal patterns result from simple population divergence over distance [28, 26, 29]. This highlights a substantial weakness of PCA: it cannot be used to differentiate between the causes of genetic distance between populations. Instead, it can only identify the patterns thereof [30].

Regardless of arguments over interpretation of patterns in PCs, we still observe that the first and primary axis of variation separates Africans from non-Africans and the second separates Europeans and Asians. With the recent and widespread availability of variant-dense data sets we can perform PCAs on not just populations, but also on individuals within a population and in small regions of the genome. The analytical resolution allowed by modern genome data has not changed the general shape of population distances among human groups, but it has changed the level of detail in the genome on which we are able to detect differentiation and thereby more accurately pinpoint the sources of that variation.

The strength of PCA has been in reducing the complexity of a genetic dataset into a low-dimensional space that can be easily visualized and understood. PCA has seen such continuous use in the field because it is easy to use, computationally inexpensive, and summarizes complex, multidimensional data into an easily comprehendible visual map. For a matrix of allele frequency covariances among populations, the values along principal axis display the amount of genetic variation accounted for by that axis. Because of its reliability and ease of computation and interpretation we use it as a basis for a statistic that summarizes the strength of population structure on a region by region basis in admixed populations.

2.2 Materials and Methods

PCA is a method by which samples can be projected onto a series of orthogonal axes, each of which is made up of a linear combination of values in a number of variable. In our case, the values are allele frequencies and the variables are populations. The range of the orthogonal axes of a PCA are chosen such the projection along the first explains the largest possible variance in the data and each subsequent axis explains a diminishing amount of that variance. The goal of PCA is to find the direction in the data with the most variation, i.e., the eigenvectors that correspond to the largest eigenvalues of the covariance matrix.

The 1000 Genomes project [31] consists of 14 populations and more than 38 million SNPs, making it the deepest sample of human population genomes currently available. Though other sources may draw data from more populations, they have considerably lower SNP density and often introduce problematic ascertainment bias. Looking for broad patterns of population differentiation across all of the SNPs in the 1000 Genomes Project sample, or even a subset, is challenging in part because of the sheer amount of information. Using PCA, we can reduce the data to a smaller number of variables and visualize them in two-dimensional plots. Unlike some other recent works using genetic data, here we use population SNP frequencies rather than individual allele counts [22, 32, 23] because we are interested in patterns of convergence and divergence between populations rather than the structure within the sample of individuals, though the two are closely linked. We lose some information by comparing the population frequencies rather than individuals in the population. However, by using these frequencies, we are able to avoid the problem of uneven sampling [30] and focus on population level processes without having to consider differences within populations .

2.2.1 Measuring Distance in PCA

Consider a matrix X of size $m \times n$, where x_{ij} is the frequency of the j^{th} SNP in the i^{th} population, and $n > m$. In order that X reduces to the principal components of that matrix, X must be mean centered and converted into a normalized data matrix Z [30, 23] where the ij^{th} element of Z is

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sqrt{\mu_j(1-\mu_j)}} \tag{2.1}$$

and where μ_j is the column mean of SNP j calculated by

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}. \tag{2.2}$$

We then calculate the singular value decomposition of Z as

$$Z = USV^T \tag{2.3}$$

where U is an $m \times n$ matrix consisting of the orthonormal eigenvetors of ZZ^T, V is the orthonormal eigenvectors of Z^TZ and S is a diagonal matrix of the square roots of eigenvalues from U and V. The vectors formed by the columns of U represent left singular vectors, which correspond to the space of population frequency values across all SNPs. To find the principal components in coordinate space, we simply multiply the matrix of population eigenvectors, U, by the diagonal matrix of singular values, S, resulting in a matrix of coordinates where each column represents a dimension up to the m^{th} (i.e., the number of populations). Isolating the first two dimensions, we can plot the populations in a two-dimensional scatter where each point represents a population and the distances among them reflects genetic variation. In our case, there are only two interpretable dimensions because are calculating only the distance between two parent populations and a target, ostensibly admixed, population.

PCA flattens SNP and population data into dimensional space where the number of dimensions is equal to the minimum of $m-1$. This transformation allows distance to be measured the same as one would any other two points on a coordinate plane. However, we introduce one small difference, we mediate the distance along each axis by the variance explained by that axis calculated from the first and second eigenvalues. The distance is calculated by

$$\sqrt{(y_1 - y_2)^2 \times V_2 + (x_2 - x_1)^2 \times V_1} \tag{2.4}$$

where x_i and y_i are the component values in the i^{th} dimension and V_i is the variance explained by that axis.

We are interested in three distances: (a) the distance between the first parent population and the target population, (b) the distance between the second parent and the target, and (c) the distance between the the parent populations. Each of these distances can be calculated geometrically from the placement of each population on the first two principal components.

We then use these distances to find regions of divergence in admixed populations. The goal of mapping this distance is to discover regions of the genome where the target population is more similar to one parent than the other. These represent regions that are convergent or divergent after the admixture event due to drift, recombinatory hotspots, or natural selection. However, the key to uncovering regions of interest is in the relationships between the distance measures in each PCA rather than their absolute values. To this end we desire a single statistic that can measure the magnitude of genetic differentiation between the target and each of the parents as well as the distance between the parents.

In order to examine the variation in these distances, and therefore the genetic divergence or convergence of the target population with either parent, we compute a statistic that summarizes the three distances into a single value. We propose

$$R_d = \log \frac{b + \frac{1}{c}}{a + \frac{1}{c}} \tag{2.5}$$

where a, b and c are correspond to the distances described above. For example, if b were the distance from Europeans to African Americans, a would be the distance from an African population to African Americans and c would be the distance between that African population and the Europeans. The statistic, here named ratio distance (R_d) has the following properties:

When $c \to \infty$, $R_d \to \frac{b}{a}$

When $b = a$, $R_d \to 0$

When $c \to 0$, $R_d \to 0$

When $b \to \infty$, $R_d \to \infty$

When $a \to \infty$, $R_d \to -\infty$

We use the log ratio to normalize the distances around 0. In words, when the distance between the parent population is small (i.e., there is little variation in SNP frequency between them), R_d approaches 1. When this distance is large, R_d is closer to the ratio of the distances between the target population and one of the parents. Similarly, when the distance between the target population and both parents is roughly equal, R_d also approaches 0. R_d measures not only the presence or absence of differential ancestry influenced structure in the target population, but it also its magnitude.

Theoretically, R_d could be calculated from raw genetic distances rather than distances calculated from PCA. However, this introduces substantial noise to the analysis because we would be unable to divide the variation among dimensions and use those values to weigh the distance between the populations (eq. 4). McVean [30, p. 6] points out that performing PCA by projecting admixed samples onto axes defined by parent populations gives us the advantage "that other structural features (e.g. admixture from a third population or relatedness) will have little influence on the projection." In other words, PCA subsumes genetic variation that does not contribute to discriminating between the populations in question to lesser axes of variation that are then able to be ignored.

The proximity of two populations in a PCA calculated from a covariance matrix as above can be equated to the correlation of allele frequencies between those populations. This can be extended to the R_d statisitic where R_d is summarizing the relative ratio of distances among three populations and, therefore, the relative correlation between the allele frequences in the comparisons of those three populations. Figure 2.1 illustrates this interpretation of R_d. Each subplot represents an a value of R_d, as described previously, where the target population is close to one parent (subFig. 2.1a), close to the other parent (subfig 2.1b), or where all three population comparisons have highly correlated allele frequencies (subFig. 2.1c). The data used in these plots are from the 1000 Genomes Project. The target population is Mexican Americans living in Los Angeles (MXL) and the parent populations are Northern Europeans in Utah (CEU) and Chinese in Beijing (CHB). In this case, MXL to CEU is distance a and MXL to CHB is distance b from equation 4; this assignment is arbitrary.

2.2.2 Calculating and Interpreting R_d in Genomic Data

The R_d statistic summarizes the genetic distance of a target population from two parent populations. By creating a map summarizing PCAs across each chromosome, we are able to identify regions which are distinct in the admixed target population for being more similar to one of the two comparison populations. In this application, admixture is not necessarily limited to recent interbreeding between two genetically distinct populations, but may include any target population that shares some genetic ancestry with both “parent” populations. This method of summarizing the results of PCA not only picks out conserved or divergent spots in traditionally defined admixed populations, but is also generally useful for identifying convergent or divergent regions' target population that diverged at one time from the two “parent” populations.

This method is powerful regardless of relative expected genetic differentiation between the parents. In other words, R_d can be meaningfully measured even if the target population is substantially closer to one parent population than the other, such as is the case for African Americans, and other admixed populations of interest. The cost of this flexibility is that we are only able to observe variation that is shared with the target and the two parent populations. Any regions that are unique to the target, or share ancestry with populations other than the parents, will remain hidden. This is because R_d necessarily approaches 1 in regions where the target population is equally distant from the parents, as would be the case in regions unique to the target.

Like other principal component based analyses, R_d can be calculated for any reasonable number of SNPs in any dataset of a minimum SNP density. The meaningfulness of the statistic is directly correlated with the interpretability of the region size or SNP number from which theR_d value is calculated. The statistic can even be calculated from a whole chromosome or even a whole genome, though the resulting R_d values would be difficult to interpret and would simply conform to the average distance between the target population and the two parents. The idea of an expectation or null model for the distance between the target population and the parents is a key concept in the R_d statistic .

Based on previous work using PCA to detect population differentiation, the average pattern of population differentiation between any three populations is clear and often easy to predict. For example, in a PCA performed with one African population, one European population and one Asian population, the obvious expectation is that the first dimension will be a split between Africans and non-Africans and the second dimension would divide Europeans from Asians. This is the expected case for each region of the genome, but there will be regional exceptions, places where Asians or Europeans tend closer to Africans than to the other population. This pattern will be exaggerated in recently admixed groups whose expectation of proportional ancestry is closer to the relative ancestry contribution of the parent populations at the time of admixture, especially compared to more anciently divergent populations.

Although R_d can be calculated for almost any number of SNPs there is a minimum threshold of diversity for reliably detecting structure. Defined by Patterson, Price and Reich [23], this minimum is the *BBP threshold* where

$$F_{st} \approx \tau \tag{2.6}$$

is found at

$$\tau = \frac{1}{\sqrt{nm}} \tag{2.7}$$

According to their analysis, divergence between populations should be easy to detect above this threshold, while it would be difficult to detect below. (See [33] and [23] for more complete discussion of the threshold problem.) Following Patterson, Price and Reich, we use this as a minimum chunk size criterion for calculating R_d along chromosomes, and exclude regions that do not cross it.

2.3 Results and Discussion

Here we apply the method of measuring population differences described above to simulated populations as well as three admixed populations in the 1000 Genomes data [31]. R_d is calculated in a rolling window of 100kb regions moving in 25kb steps across the genome. The rolling window allows for fine determination of regions of maximal or minimal divergence. In this case, the size of the window was chosen through trial and error which suggested that 100kb is the smallest region that consistently overcomes the Fst threshold in SNP data of similar density. This was also a major advantage of using the 1000 Genomes Project data as opposed to the HGDP SNP data which have a considerably lower SNP density but a larger and more diverse sample of populations. Across the whole genome, the average number of SNPs per 100kb window was more than 1400, allowing the τ threshold to be crossed even when genetic diversity was low.

In addition to testing how R_d measures population structure and regional divergence in the genome, running R_d across the genome will allow us to see which regions in each admixed population are more like one of the two parent populations than the other. Unlike much previous work using PCA on populations or individuals, the question is not whether there is subdivision between individuals, but where the most drastic divergences between populations appear in the genome. This method of using a rolling window PCA is similar to that of two recently published approaches [34, 35], but here we are focusing on populations instead of individuals, as well as the calculation of the R_d statistic outlined above.

Another advantage of both PCA and the 1000 Genomes data is that sequencing error which may occur in the lower coverage intron regions is not a problem. The SNP density is such that the error they introduce succumbs to the signal from the abundance of data per region. Additionally, because they are not population biased, sequencing errors should not substantially influence the distances between populations in a PCA. This was confirmed using a simulation where sequencing errors up to 5% introduced to a single chromosome resulted in no significant change in the distribution in R_d ($p > 0.95$).

2.3.1 Simulation

To demonstrate how R_d is calculated across the genome, Fig. 2.2 plots chromosome 22 of a simulated partially admixed population consisting of individuals in the 1000 Genomes data [31], samples from Northern Europeans (CEU), and Beijing Chinese (CHB). For the first 15Mb, of the chromosome the population is purely CEU, while the region between 16Mb and the end of the chromosome is a 50/50 mix of individuals from both populations. Simulation was carried out by by combining the individuals from the populations into a single group, randomly selecting half of that group at each variant site then calculating a new allele frequency. As such, this simulation represents the most simple scenario for an admixed population, one that is a single generation of admixture with no reproduction after the event and an equal proportion of both parental populations. Further simulations were conducted using different relative proportions of parental ancestry, and the change in average R_d was exactly proportional to the proportion of ancestry for each parent, as expected.

Important to note in Fig. 2.2 is that despite the first portion of the genome being purely European, R_d varies substantially. This is because the statistic varies as the parent populations vary in relative proximity in the PCA, and in this case the target population is effectively the same as as one of the parent populations for the first part of the chromosome. R_d is simply therefore the measure of the genetic distance between the CHB and CEU populations for that region. This plot demonstrates that the R_d statistic not only accurately emphasizes convergence based on parent population distance, but also accurately controls for equidistance in the parent populations.

R_d is normally distributed across each chromosome relative to the mean value. This is verified through further simulation and described by the quantile-quantile plot in Fig. 2.3 using a simulated population that has randomized the relative proportion of ancestry contributed by each parent at each locus. The random proportion of ancestry is drawn from a normal distribution with mean of 0 and standard deviation of 0.5. Due the biases of differential contributions of ancestry by the parent populations calculations of R_d values across real chromosomes may not be as perfectly normally distributed, but they are close.

Using R_d to summarize PCA not only picks out conserved or divergent chromosomal regions in admixed populations, but also could be generally applied to identify regions that are more or less similar in any target population that shares ancestry with two ancestral populations. The sole caveat is that the target population must be between the two parents in the first dimension of a genome average (or representative) PCA of the three plots. Populations that might fit this profile would include any from regions that lay between known geographically distinct populations. For example, populations in Northern India or the Middle East may share similarity with both European populations and with East Asian populations. While R_d can identify these differences, our knowledge of their relative magnitude is limited to the shared genetic variability between the three populations in question. In the same way that the analysis is blind to regions of uniqueness for the target population, the essential answer we are given is that the target population is more or less like parent population a than parent population b, and vice versa, as we search across the chromosomes.

2.3.2 Admixed Population in 1000 Genomes Project

Here we examine populations with known structure (or at least assumed structure) from the 1000 Genome Project and attempt to discern those regions of the genome that are most similar or dissimilar between populations by comparing admixed populations to their parent populations. The three populations used in our analyses here are Mexican Americans living in Los Angeles (MXL), African Americans living in the American Southwest (ASW), and Colombians in Medellin, Colombia (CLM). Two Latino populations were chosen because of the well known variation among these populations in relative proportion of European ancestry [36], and they can therefore serve as a test of whether R_d can detect this difference. Figure 2.4 shows how R_d varies along chromosome 6 for each of these populations compared to the parent populations. For the Latino populations, we use Northern Europeans living in Utah (CEU) and Beijing Chinese (CHB) as the parent populations. For ASW, we use CEU and Yoruba from Nigeria (YRI).

ecause there are no Native North American populations currently in the 1000 Genomes Project, Chinese, as a subsample of East Asians who share more recent ancestry with Native Americans than Northern Europeans, are used as the second parent population for R_d analysis in the Latino populations. Though Native American populations are available from other sources, such as the HGDP, other databases do not have the SNP density to confidently overcome the diversity threshold described in the methods section (equation 6). Of course, this solution is less than ideal as there are thousands of years separating Native

Americans from Asian populations. However, Native American populations that have been sequenced have been shown to have substantial European admixture and subpopulation specific drift [37], so even if data of sufficient SNP density were available, the results would be similarly difficult to interpret.

Our goal is to determine relative similarities between the target and the two parents, and we are able to achieve this by actively ignoring the unique components of the target and parent population's ancestry. While we may be missing the uniquely Native American aspects of the Latino populations, we can detect similarities shared between Native Americans and East Asians that have persisted in the Latino populations. Furthermore, the average value of R_d across the genome is less than one standard deviation from 0 for MXL ($\mu_{R_d} = -0.36,\ \sigma_{R_d} = 0.56$) and approximately 1.5 standard deviations from 0 in CLM ($\mu_{R_d} = -0.807,\ \sigma_{R_d} = 0.512$). That this difference is very nearly proportional to the difference in average European ancestry between these populations as calculated elsewhere [36, 37] suggests strongly that CHB is an effective stand-in in this analysis. The ability of our method to examine ancestry specific convergence is particularly useful for a population such as Hispanic Americans whose mosaic genetic background shares ancestry with Europeans, African Americans and Native Americans.

In contrast to the results from MXL, the African American population is on average much closer to their African ancestors than to their European ancestors. This is unsurprising and follows previous estimates of the relative contribution of Europeans to the African American genome [38].

Tables 2.1-2.3 highlight 100kb regions in the genomes of the three target populations where the target is exceptionally similar to one parent, and both the target and that parent are dissimilar from the second parent. In other words, these regions have very large or very small values of R_d. Because R_d is normally distributed, these regions were simply identified by being 4 standard deviations or further from the mean R_d for that target population. This threshold of standard deviation is somewhat arbitrary; there were too many regions between 3 and 4 standard deviations to list, and too few below. Only gene regions that contained genes according to the UCSC genome browser (genome.ucsc.edu) [39] are listed. (There were 2 noncoding regions for MXL, 2 for CLM and 8 for ASW.)

Speculating on the phenotypic effects of the genes identified in Tables 2.1-2.3 is beyond the scope of the current analysis. However, one interesting result to note is the presence of the CCDC88A gene in exceptionally CHB-like regions in of the MXL and CLM populations. CCDC88A is a member of the Girdin family of coiled-coil domain containing proteins, and has been associated with cancer metastasis [40]. Both MXL and CLM are exceptionally similar to Europeans in this region. Also of note is that the average values of R_d accurately reflect the relative proportion of ancestry for each population. MXL for example, has a mean R_d closer to 0, suggesting their proportion of European ancestry is smaller than that of CLM, which is nearly a standard deviation closer to the European population. However, despite this difference in mean values, R_d for both of these populations in the CCDC88A region is almost exactly the same.

The regions in Tables 2.1-2.3 represent regions at the most extreme values of R_d across the whole genome. However, R_d can be calculated for smaller regions as well. Figure 2.5 shows the value of R_d for MXL across the HLA region of chromosome 6, an immune system related region known for plasticity and ongoing selection in human populations [41]. Balancing selection and selection for diversity are known the shape the region, and unsurprisingly, the R_d in this area of the genome is a nearly even mix between CHB and CEU like allele frequencies.

Because the focus of R_d analysis is on population distance and not structure between individuals, it cannot be used directly to correct for stratification in GWAS studies. The key to this analysis is that it does not only detect population structure in a sample, or a subset of samples, it gives us an expectation for relationships between populations on a genomic region-by-region basis. The results of calculating R_d on admixed populations from the 1000 Genomes Project inform us as to the regions of the genome which, through evolutionary processes, have become (or been maintained as) more similar to the same region in one of the parent populations than to the other. While other methods, such as EIGENSTRAT and STRUCTURE [23, 24, 22], are useful in correcting for identified structure in a sample, R_d calculates the expectation of that structure across each chromosome. Foreseeably, R_d could be integrated into the current GWAS toolset as a method for verifying or setting the expectation for the structure detected on an individual level in samples of known ethnic origin. In conjunction with programs like EIGENSTRAT, HAPMIX or STRUCTURE, R_d can be used to identify potentially important regions of structure found in EIGENSTRAT that are a result of interpopulation structure in the cases or controls of a GWAS. In other words, R_d could be used to reduce the false negative rate introduced by the correction method in EIGENSTRAT.

2.4 Conclusions

We have described a method for looking at population structure that is computationally simple, mathematically intuitive, and easy to interpret. R_d effectively summarizes the similarity or dissimilarity of a target population in comparison to its parent populations in a single value. Similar to other methods that detect the presence and strength of population structure, R_d, when combined with variant-dense genomic data, can be useful in elucidating the shape of population structure between populations. R_d differs from other PCA methods importantly in that it does not include subpopulation structure, which can influence corrections made in consideration of that structure by methods such as EIGENSTRAT and STRUCTURE. Of course, knowing the substructure among individuals is important to control for as well, but by removing it, R_d determines an expectation of the structure between any individual sampled from the target population as compared to either parent.

The simulations we preformed suggests that R_d is a reliable measure of between population structure and accurately reflects the genetic distance between the target population and its parents. The key to R_d's usefulness is its emphasis on both the distance between the parent populations as well as the target admixed population and to each of the parents. While other analyses, such as HAPMIX, which is formed by haplotype analysis, may be informative as to the likelihood of a region of the genome being inherited from one population or the other, the results of the analysis do not speak as directly to the degree of genetic differentiation that any divergent region represents. R_d mapped across the genome is a representation of the relative genetic differentiation accounted for by each region in the target population within the range of variation in the parent populations.

The 100kb regions we highlight in Tables 2.1-2.3 are not just regions where ASW, MXL or CLM are similar to one parent or the other, but places in the genome where both the target population and one of the parents are substantially genetically different from the other parent. The significant number of noncoding regions with exceptional R_d values in the ASW population suggests that drift is clearly a factor in some regions, but this does not rule out natural selection as a possibility. The increased proportion of functional genetic variants in admixed populations makes them easy targets for selection. Any population with a higher average selection coefficient has the possibility to experience larger changes in allele frequencies per generation. In other words, evolution can happen more rapidly in populations harboring functional variants from two distinct populations [21].

In conclusion, we have shown that R_d can be a useful and novel method for detecting population convergence in admixed groups. Potential uses of the method include highlighting functional genetic differences between populations in divergent regions, and helping to map structure in populations with known ethic origins in order to control for strong population structure in genome wide association studies. Potentially, R_d could be extended to include more than two parent populations. For now, it is clear that PCA, though an old tool, still has great potential in the era of whole genome population genetics.

Analysis and Graphics were completed using the SciPy and MatPlotLib libraries for Python 2.7 [42, 43]

Table 2.1: Exceptionally divergent regions in MXL

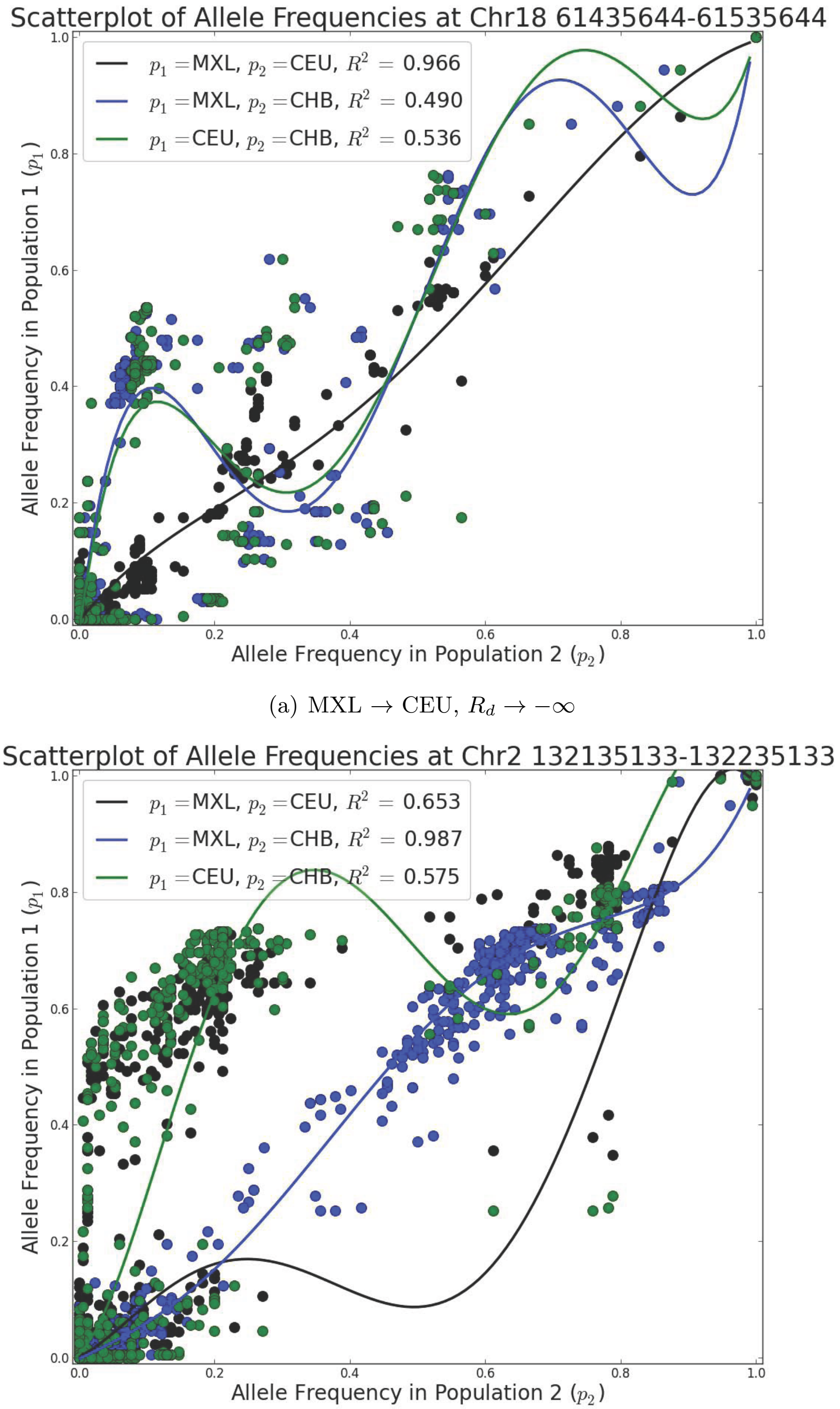
Gene-Containing Regions where $R_d > 4\sigma$ from μ				
Population	Chromsome	HG19 Region	R_d	Genes
MXL $\mu_{R_d} = -0.357$ $\sigma_{R_d} = 0.556$	2	55535133-55635133	-2.855	**CCDC88A**
	2	132135133-132235133	1.998	LOC389043 TUBA3D
	2	242985133-243085133	2.268	LOC728323
	6	111248924-111348924	-2.696	GTF3C6 RPF2
	9	66185023-66285023	-2.707	DQ590378
	10	75010523-75110523	-2.911	MRPS16 C10orf103 BC033983 TTC18
	17	43725056-43825056	-2.842	CRHR1
	17	43900056-44000056	-2.805	MAPT CRHR1 LOC10028977 IMP5
	18	61435644-61535644	-2.749	SRPINB7

Table 2.2: Exceptionally divergent regions in CLM

Gene-Containing Regions where $logR_d > 4\sigma$ from μ				
Population	Chromsome	HG19 Region	R_d	Genes
CLM $\mu_{R_d} = -0.807$ $\sigma_{R_d} = 0.512$	1	25910583-26010583	1.388	MAN1C1
	1	78535583-78635583	-2.928	GIPC2
	1	161760583-161860583	1.533	ATF6
	2	55585133-55685133	-2.912	**CCDC88A**
	4	84935240-85035240	1.353	BC005018 AK095285
	7	57466161-57566161	-3.082	ZNF716
	8	17685422-17785422	1.278	FGL1 PCM1
	12	45736107-45836107	-2.957	AN06
	15	66001200-66101200	-3.16103729562	DENND4A

Table 2.3: Exceptionally divergent regions in ASW

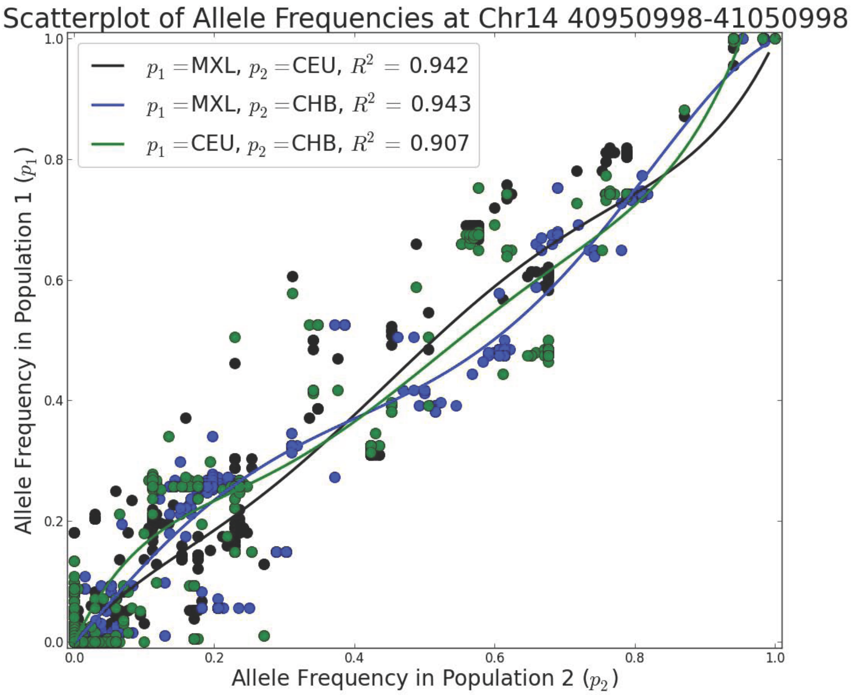
Gene-Containing Regions where $R_d > 4\sigma$ from μ				
Population	Chromsome	HG19 Region	R_d	Genes
ASW $\mu_{R_d} = 0.970$ $\sigma_{R_d} = 0.334$	1	100410583-100510583	-0.444	BC112312 SLC35A3 HIAT1
	1	150010583-150110583	2.328	VPS45
	3	21960157-22060157	2.330	ZNF385D
	4	190810240-190910240	2.401	BC087857 FRG1 TUBB4Q
	5	186940-286940	2.422	PDCD6 SDHA CCDC127 LRRC14B PLEKHG4B
	5	261940-361940	2.643	PDCD6 AHRR
	6	84523924-84623924	2.315	RIPPLY2 CYB5R
	8	106735422-106835422	2.331	ZFPM2
	9	45710023-45810023	-0.390	FAM27A
	10	85160523-85260523	2.640	AK056904
	11	75520855-75620855	2.355	UVRAG
	11	89370855-89470855	2.402	AB231784 FOLH1B TRIM77P
	17	45450056-45550056	-0.420	C17orf57 EFCAB13 MRPL45P2



(a) MXL → CEU, $R_d \to -\infty$

(b) MXL → CHB, $R_d \to \infty$

Figure 2.1: Scatter plots of allele frequencies with best-fit lines in regions of exceptional values of of R_d. Each plot has all three distance comparisons described above where each set of points and their corresponding line represents a population comparison.



(c) $a = b = c, R_d \to 0$

Figure 2.1: Continued

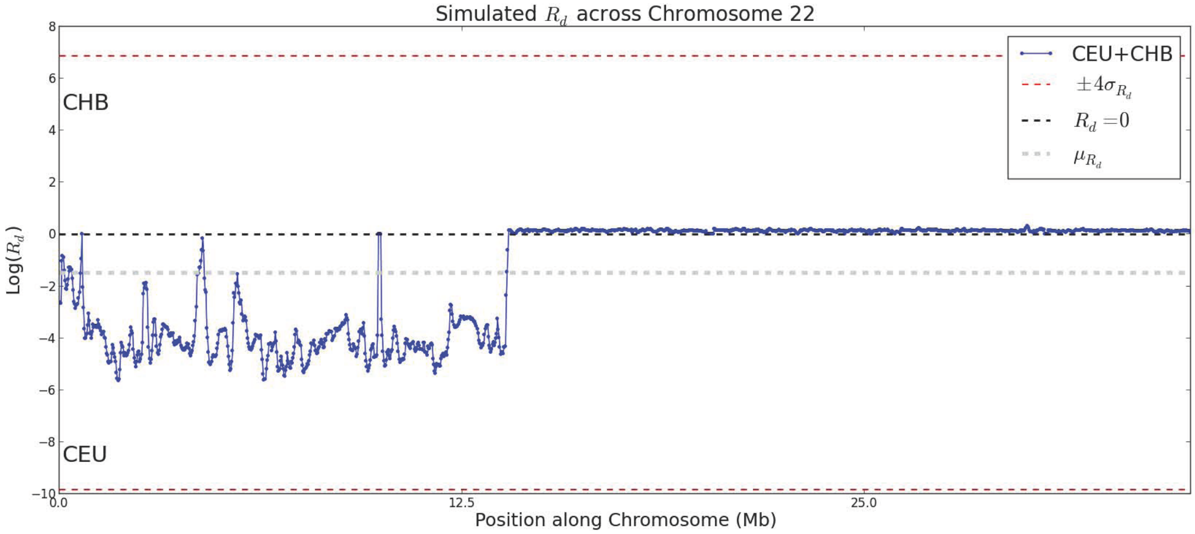


Figure 2.1: R_d calculated from simulated admixture between Europeans (CEU) and Chinese from Bejing (CHB). The first 15Mb consist of purely European allele frequencies while everything after is a 50/50 mix of the two populations.

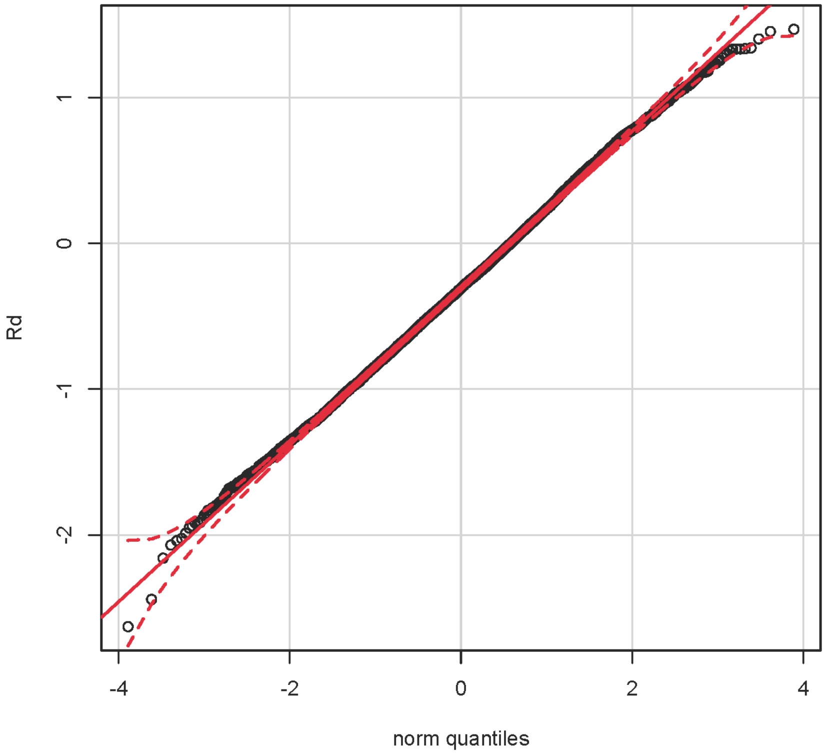


Figure 2.2: Quantile-quantile plot of simulated R_d values drawn from a distribution of randomly assigned ancestry contribution by the parent populations. Red line is the theoretical normal distribution; black circles are the simulated values of R_d across chromosome 22.

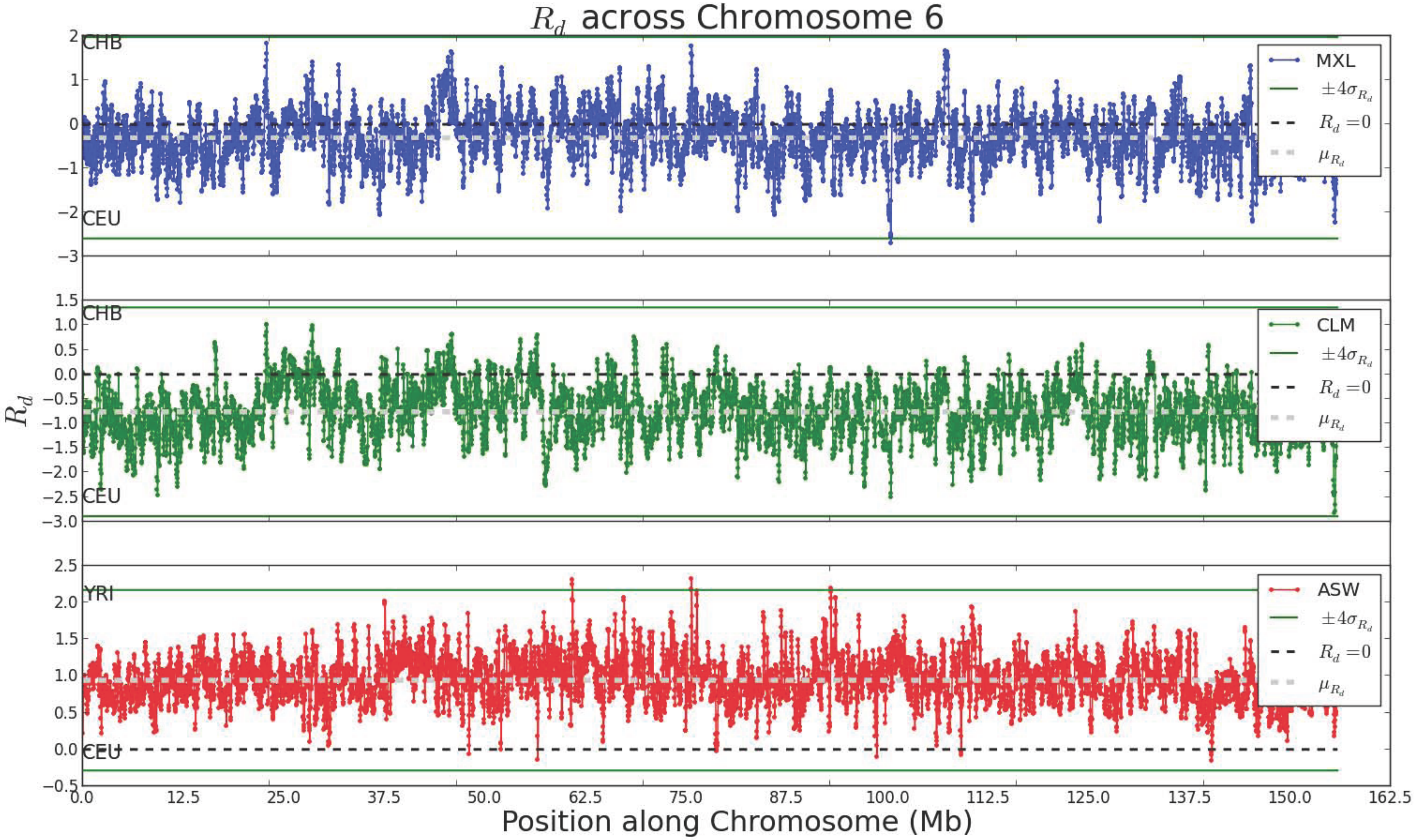


Figure 2.3: R_d across chromosome 6 in three admixed populations from the 1000 Genomes Project.

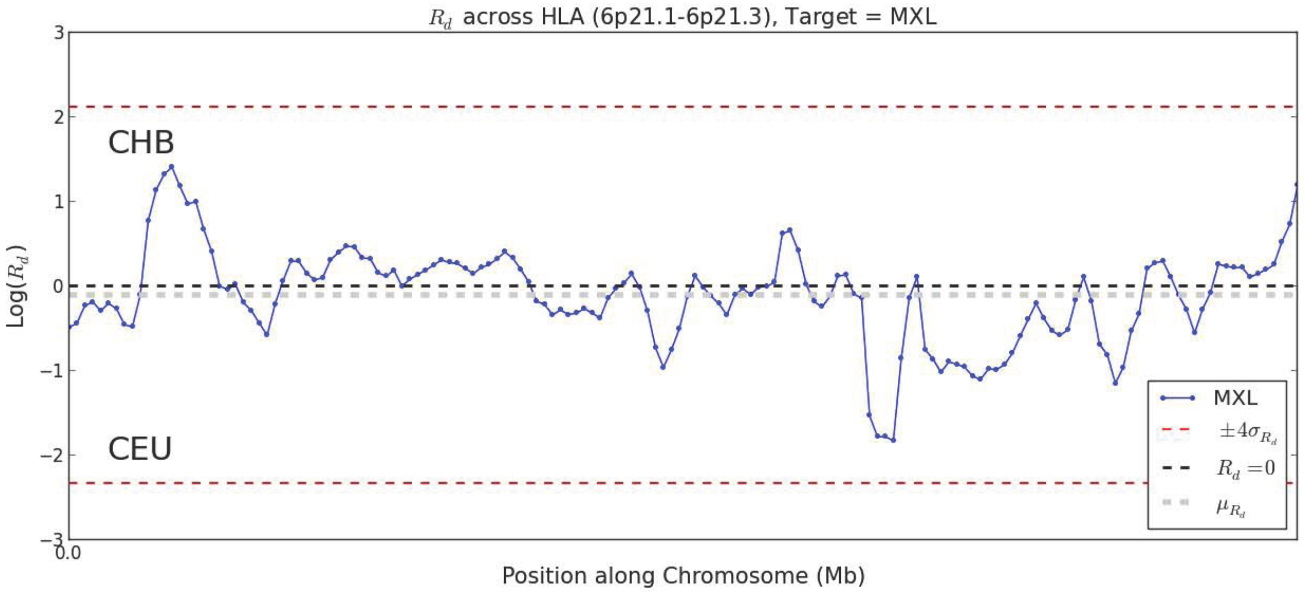


Figure 2.4: R_d in the HLA region for Mexican Americans.

CHAPTER 3

NEANDERTHAL GENOMICS: WHERE DO ARCHAIC HOMININS FIT INTO MODERN HUMAN GENETIC DIVERSITY?

3.1 Introduction

Neanderthals are a branch within genus Homo who have variously been considered a separate species from and a subspecies of *Homo sapiens*. The popularity of each classification has waxed and waned, but the current trend has been to consider Neanderthals to be their own separate species, distinct from modern humans [44]. The species versus subspecies debate over this late hominin may seem trivial, but it reflects the ambivalence surrounding the Neanderthals' classification as either a direct human ancestor who has contributed to modern human diversity through introgression, or a distinct sister clade. Taxonomic delineations of living animals are often contentious. Our closest living relative, chimpanzees, are subject to a similar taxonomic disagreement regarding the robust *Pan troglodytes* compared to the relatively gracile *Pan paniscus* [45]. Of course, when anatomy and behavior can be gleaned only from relatively sparse fossil-skeletal remains and limited genetic data, the problem becomes even more contentious. Central to this debate are several questions: just how similar were Neanderthals to the migrating *Homo sapiens* who left Africa 50,000 to 60,000 years ago [46]? How similar were they to modern human populations? And, perhaps most importantly, what contribution did these extinct hominins make to modern human genetic variation? Because the answers to the first two questions are still unclear, answering the third question is a considerable challenge.

Until very recently, the only evidence with which to address the issue of Neanderthal similarity and possible introgression with humans has been in the form of skeletal remains and stone tools scattered through Europe and some parts of the Middle East. From these, Neanderthals have been described largely on the generally subtle difference between them and later archaic *Homo sapiens.* Neanderthals were stocky and had a robust general bone structure. Their skulls are described as long and low with large cranial capacity, double arched brow ridges, an occipital bun, and protruding midface with a large nose and large dentition [47, 48, 49]. As a group Neanderthals are distinguished by processing this whole suite of traits, though individually many of these morphological characteristics are found both in early *Homo sapiens* and some modern human populations [50]. Because of the anatomical similarities and the shared geographic range, some proposed that Neanderthals and prehistoric Europeans might have interbred long before genetic evidence of the event was available [51, 52, 53]. Some researchers have even described the skeletal remains of a possible Neanderthal-human hybrid [54, 49]. Though morphological comparisons are informative because of the close anatomical similarity between the two hominins, fossil comparisons alone are unlikely to conclusively solve any disputes.

In 1997, the first DNA sequence from a Neanderthal was recovered by [55]. The discovery was remarkable enough for Dan Lieberman to proclaim that this "was proof that there is a God who likes paleoanthropology" [56]. Though the initial sequence was only a short read of a noncoding region of the mitochondria, it changed Neanderthals' place in our phylogenetic tree. The analysis of this small dataset solidified an approach to interpreting the ancient DNA of Neanderthals as well as their place in human evolutionary history. The mitochondrial haplotype recovered by [55] was more similar to humans than to chimpanzees, but it was well outside the range of modern human mitochondrial diversity. This initial mitochondrial sequence and later, more complete, studies of the mitochondrial genome confirmed that humans and Neanderthals separated into two distinct lineages approximately 500,000 years ago [57, 58]. Because the Neanderthal mitochondria were clearly outside of the human range, it seemed that Neanderthals were a very distinct group and that no interbreeding had occurred. In other words, because no Neanderthal-like mitochondria have been found in modern human populations, interbreeding came to be seen as an unlikely scenario.

More recently, successfully recovered Neanderthal nuclear DNA has cast doubt on the estimates of human-Neanderthal divergence inferred from the mitochondrial comparison. The sequencing of Neanderthal nuclear DNA has been a slow process because damage and a low copy number per sample make recovery of large sequences difficult. In addition, because of our close phylogenetic relationship, human contamination in Neanderthal samples is extremely difficult to detect and has caused problems for analysis in the past [59, 60]. However, a draft sequence of a nearly complete Neanderthal genome was completed with meticulous controls for contamination and damage, leading to the creation of a convincingly authentic map for the comparison between humans and Neanderthals in their nuclear genomes [61, 62, 63]. Surprisingly, the results contradict and complicate the story of population history told by the earlier mitochondrial sequences. The relative divergence of Neanderthal nuclear DNA from humans is considerably lower than the divergence previously estimated from mitochondria. Neanderthals shared a last common ancestor with humans approximately 800,000 years ago, but the complete population divergence between them and modern humans did not occur until an estimated 270,000 to 440,000 years ago [61]. This range falls well inside the depth of nuclear DNA sequence diversity within present-day human populations, which is slightly less than 500,000 years [64, 65, 66].

In addition, the Neanderthal genome presented clear evidence for low levels of admixture into Eurasian populations. The interbreeding event with non-African human ancestors is also surprisingly old, estimated to have occurred between 50 and 80 thousand years ago. This initial estimation has been supported by subsequent research, finding evidence of archaic admixture in nearly all human populations [61, 67, 62, 68, 69, 70, 71, 72, 73], possibly even including Africans [74]. Though some have suggested that the detected introgression events may have resulted from ancient population structure [75], such a scenario has been shown to be unlikely given the amount of admixture detected [76]. The publication of the Denisova genome, sequenced from an archaic hominin found in Siberia, has also aided our understanding of prehistoric genetic population differentiation. Denisova's existence is extremely suggestive of complex patterns of both ancient population substructure and late Pleistocene admixture of migrating homo sapiens with contemporary premodern archaic populations [62]. The discovery that Melanesians are more likely to have interbred with Denisova's ancestors than any Neanderthal suggests a complicated mix of Out-of-Africa and Multiregional scenarios that do not conform to any current theories of human expansion based on archaeological evidence.

Perhaps even more unclear is the exact extent of the introgression and its distribution across the genome in modern human populations. A better understanding of the extent of interbreeding between humans and archaic populations will allow insight into what separates modern populations from archaic ones, and it will identify sources of genetic differentiation of living populations.

3.2 Archaic Hominin Divergence

When comparing the genetic resemblance between humans and the Neanderthals, one first needs to discern an expectation of difference based on the level of taxonomic investigation. In considering genetic divergence within the human lineage that includes extinct hominids there are multiple levels of comparison that will inform the search for genetic distinction among the clades. First, looking for differences between modern humans and living nonhuman apes will reveal changes dating back to as far as six million years. These differences are discerned by comparing the chimpanzee genome to modern humans. Polymorphisms that are derived in both humans and Neanderthal with respect to chimpanzees represent the shared ancestry of our lineages and should be common. The chimpanzee genome differs from the human genome by only about six percent, so a Neanderthal is expected to be separated by much less than that [77].

Second, differences between modern humans and the ancient DNA sequences of related hominins, including Neanderthal and Denisova, will inform differences that have arisen since the emergence of the direct ancestral line to modern humans, dating back to about 800,000 years ago at the oldest. Comparing derived polymorphisms between humans and Neanderthals at this level will be the most informative for investigating the similarities between the two, as these polymorphisms are largely comprised of differences that arose since our lineages split. However, a comparison of shared derived alleles that are unique to humans relative to Neanderthal may be confounded by introgression. The admixture between Neanderthals and humans introduces older derived alleles differentially into the human populations that descended from the admixed group.

Third, genetic comparisons between modern human populations allow some insight into the very recent evolution that may have shaped modern humans since emergence from Africa and during continued population differentiation [8, 78]. Each of these inquiries examines a different level of phylogenetic divergence within our lineage.

Figure 3.1 briefly summarizes recent human ancestry in relation to Neanderthal and Denisova, as well as human population differentiation and archaic admixture events. The complexity of identifying genetic differentiation between and among these populations comes from the multiple possible gene genealogies for a given variant. Ancient population structure can result in false positive signals of admixture (labeled Z in the figure). The solid black lines represent only an example of population relatedness at a given allele, though it is the most likely case for any polymorphism among these populations that nearly any combination of shared differences among populations within the grey area are possible.

The exact timing of the introgression events from Neanderthals into Eurasians [73, 76] and from Denisova into Melanesian and Australian Aboriginal populations [79, 63] is not very well known, but estimates suggest that it occurred between 50,000 and 70,000 years ago. The small number of archaic specimens makes it even more difficult to estimate the split between Denisova and Neanderthal. Using the chimpanzee-human split as a reference, Reich et al. [63] estimate that the Neanderthal-Denisova divergence is slightly older than the divergence between the African San and other present day human populations, which occurred 600,000 years ago, making Denisova a sister group to Neanderthal. Part of the difficulty of measuring divergence dates, population similarity, and admixture is the abundance of structural changes in the phylogeny that occured between 450,000 and 550,000 years ago. In this time period, the human and Neanderthal populations diverged, the deepest parts of human population structure diverged, and Denisova diverged from Neanderthal. For some part of this time period Humans, Neanderthals and Denisova may have been interbreeding, causing a number of coalescent events to cluster in this somewhat ambiguous period of our prehistory.

Figure 3.1 also demonstrates the importance of distinguishing between genetic divergence and population divergence in considering the expectation of genetic differences within and between groups. In speciation or population subdivision events, the time from the beginning of the separation event to the completion of population separation may be significant (represented by section C in Fig. 3.1). During this period, there would be some continued interbreeding between two closely related, but separate, populations. Gene genealogies traced back to this period would have a range of coalescent intervals depending on the length of time for the separation event. Additionally, one or both of the offspring populations may inherit any polymorphisms present in the common ancestor. As noted previously, the complete range of nuclear DNA variation present in modern populations has an overall coalescence depth back to about 500,000 years [64]. The initial evidence form mitochondria suggested Neanderthal and human populations diverged around this same time, and so only a relatively small fraction of the variation between the two would also be included in modern human genetic variation. However, if the recently re-estimated divergence time is correct, then a significant portion of the variation between modern humans and Neanderthals may also be contained within modern human variation. In other words, all of us will have some variants that are more closely related to some Neanderthals than they are to variants possessed by other living people [48]. This, of course, complicates the search for Neanderthal genetic distinction, since there may be a significant number of loci that are more different between living populations than they are between any human and Neanderthal. Further complicating this story is the evidence for admixture between Neanderthals and non-Africans, meaning either that some of the differences gained over the length of the divergence were lost in Eurasians, or some of the unique and more recent Neanderthal variants were gained.

Two distinct questions are relevant to the investigation of the genetic differences between Neanderthals, modern humans, and our recent ancestors. First, do Neanderthals fall within the range of general human nuclear DNA variation? And second, how phenotypically distinct were Neanderthals from both modern humans and their archaic ancestors that were contemporary with them? Addressing these questions requires knowledge of functional differences between modern human populations, differences between modern humans and their late Pleistocene ancestors, and significant knowledge of the Neanderthal genome. Until very recently, such an analysis was impossible. However, with an increasing supply of archaic genome sequences and a growing knowledge of recent human evolution [8, 80, 81, 82], these questions are just beginning to be addressed directly.

3.3 Neanderthal Phenotype

Because of our close phylogenetic relationship and the relative abundance of skeletal samples, it is not hard to imagine how a Neanderthal might look. In fact, a Neanderthal should appear, for the most part, very human. This is both an advantage and a disadvantage in comparative genomics. First, it means that accumulated knowledge of human gene function can be used to examine the Neanderthal polymorphisms for functional associations, and possibly even hypothesize on their effects. Second, intragenomic interactions can be assumed to be relatively similar. In other words, variation within polygenic traits is likely to result in similar phenotypic changes. Unlike in deeper phylogenetic comparisons, a polymorphism in a Neanderthal will, more often than not, produce the same change that the polymorphism would cause in a modern human. That being said, the extreme genetic similarity between humans and Neanderthals means that the differences are predictably subtle and may be extremely hard to detect in broad genomic comparisons. Indeed, relatively little direct anatomical knowledge has thus far been gleaned from studying the Neanderthal genome. The first attempts at large scale sequencing, though promising at first, resulted in sequences that were later determined to be up to 80% modern human contamination [59, 83, 60]. However, more targeted studies of specific genes or sequences have been successful, even before the most recent composite nuclear genome was sequenced.

One strategy for analyzing Neanderthal DNA to search for phenotypic differences between our lineages is to focus on genes and regions where the functions are very well known. A number of studies have successfully found some surprising similarities between Neanderthals and modern humans. Perhaps the most informative and controversial of these types of research has been the sequencing of the FOXP2 gene in Neanderthals [84]. FOXP2 is among the most conserved regions of the mammalian genome [85]. The human variant, consisting of two nucleotide substitutions in the 7th exon, is fixed in every known population, and is the only gene currently known to be implicated in speech and language. The inactivation of one FOXP2 copy leads primarily to deficits in orofacial movements and linguistic processing similar to that of individuals with adult-onset Brocas aphasia [84]. In theory, if Neanderthals lacked the human variant of FOXP2, they were much less likely have possessed the capacity for speech.

Krause et al. [84] determined that Neanderthals did share the two substitutions on the 7th exon, as well as much of the surrounding haplotype. This result was surprising because coalescent analysis of the human haplotype surrounding exon 7 of the human FOXP2 gene suggests that the variant arose and swept to fixation within the last 200,000 years, placing it outside of most estimates for Neanderthal divergence. Thus, the expectation was that Neanderthals should not share the human FOXP2. If the divergence dates from Green et al. [61] are more accurate than those estimated from mitochondria this is comparatively less surprising that Neanderthals share our variant, but, the coalescence of the human FOXP2 is near the far low range of the population divergence estimate. Assuming it is not an artifact of contamination, the presence of FOXP2 in the Neanderthal genome is remarkable, even in the context of a later Neanderthal-human divergence. Speech and language are complex traits, which no doubt require a large number of other genes to function. Assuming that FOXP2 is one of many genes involved in language, it is surprising that the coalescence would be so recent and still be shared between Neanderthals and humans [86]. Another, perhaps even more remarkable, possibility is that FOXP2 is the result of introgression, since the human variant and associated haplotype is fixed in all known populations and Green et al. [61] found no evidence for admixture in African populations.

Other similarities between Neanderthals and modern humans have been detected using similar targeted nuclear DNA comparison. Many of these studies were facilitated by the careful extraction of relatively contamination-free samples from Sidrn Cave site in Spain [87]. The research suggests that Neanderthals carry the human specific O01 haplotype for blood type O, as well as an allele of the TAS2R38 gene that is polymorphic in human populations and allows the ability to taste the bitter substance phenylthiocarbamide [88, 89]. That Neanderthals share the polymorphic allele and appear to be heterozygous for the trait suggests that the human TAS2R38 allele is the result of heterozygote advantage, as the polymorphism has existed for hundreds of thousands of years without reaching fixation.

Aside from these two somewhat unsurprising similarities, Neanderthals differ from modern humans in an informative skin pigmentation gene. The gene, MC1R, is implicated in lighter pigmentation of modern Asian populations, but the Neanderthal allele is unlike any known modern variant [90]. Complete and partial function loss in the MC1R gene are known to cause pale skin and red hair, and based on the structure of the change in the Neanderthal variant, there is also "partial loss of function caused by reduced cell-surface expression of receptor protein and altered protein coupling efficency" [90]. Therefore, it is reasonable to suppose that Neanderthals gained lighter skin when they migrated out of Africa, but did so through a different pathway than later humans.

The Neanderthal and Denisova genomes represent a vast amount of information that will take considerable time and effort to interpret and analyze for functional genetic changes. As Green et al. [61, p. 710] point out, "a Neanderthal genome sequence provides a catalog of changes that have become fixed or have risen to high frequency in modern humans during the last few hundred thousand years and should be informative for identifying genes affected by positive selection." There are a few caveats that need to be addressed when analyzing the draft Neanderthal genome presented by Green et al. [61]. First, in comparative genomics, it is always difficult to make conclusive statements about genetic differences because you can never be sure when a genetic variant may exist undetected in low frequencies in one population or another. This problem is compounded in the Neanderthal case because only low-coverage genomes of a few individuals are available. Whatever allelic state the composite Neanderthal genome exhibits cannot be construed as reflecting all Neanderthals, but a sample of one. In other words, comparisons between Neanderthals and modern humans cannot be considered as being conclusive for any single SNP, haplotype, or even any single gene. Indeed, there is little reason reason to suspect that Neanderthals did not share the kind of geographic population differentiation we see in modern humans, as they too were widely dispersed. The Denisova genome further suggests that there is appreciable ancient population diversity that has been, until now, largely hidden [70, 62].

However, rather than looking down the gene tree from humans to Neanderthals, one can instead look across the chromosomes to find general patterns of allelic states that may be informative to population genetic analysis. And in doing just that, the authors of the Neanderthal genome were able to identify regions likely under recent selection in modern humans. Comparing the complete Neanderthal nuclear sequence to the human reference genome and five other complete sequences, Green et al. [61] were able to identify a number of regions as candidates for recent selective sweeps. This investigation speaks to the consideration of modern human change since the split with Neanderthals. Regions that were enriched for ancestral sites in Neanderthals, compared to the modern human samples, contained genes coding for a wide range of function, few of which had obvious phenotypic results. Their analysis found 78 fixed human-derived substitutions. These genes had a statistically significant tendency to be involved in mesoderm development, transcriptional preinitiation, and lipoprotein metabolism. These results match some of the regions of the human genome previously—identified, using modern human genetic variation—as having undergone selective sweeps [80, 82, 91]. However, many of the genes identified as having the most radical change as compared to the Neanderthal sequence have not been identified before as strong candidates for selection. In their companion study, Burbano et al. [68] found no significant function-clusters for the 88 human specific derived SNPs they uncovered using targeted analysis. The functional change implicated by the analysis suggests possible differentiation in terms of diet, muscular development and cognition, though the association is far from clear [67]. These general patterns of functional change are not very informative in a search for a phenotypic comparison between humans and Neanderthals, but they do point the way for future more detailed work.

3.4 Archaic Admixture

The Neanderthal divergence from the modern human lineage is comparable in age to the overall nuclear DNA sequence diversity within present day human populations. This means that there is existing variation between modern populations that is equally as old and divergent as the distance from Neanderthal to any modern human population. Depending on the actual population divergence, the number of existing differences that old may be small, but it does give one pause when thinking about how to put in perspective overall patterns of genetic difference between humans and Neanderthals. Put another way, some of the existing variation between modern human populations stems from genetic differentiation that is nearly as old or older than the divergence of the Neanderthal clade from the prehuman common ancestor. Depending on the patterns of migration and dispersal out of Africa, there may have been population structure existing in the human lineage that is deeper than the division between humans and Neanderthals.

Another major contribution from the publication of the Neanderthal genome was evidence for recent admixture between Neanderthals and non-African human population [61]. The equal distribution of admixture in all Eurasians is somewhat puzzling, and was not predicted by any previous model based on fossil evidence. Thus, to the extent that they ever cohabited in that region—a scenario for which there is some doubt—any admixture hand long been predicted to mostly affect European populations [92]. Many of the Neanderthal-like anatomical traits are found more frequently in Europe, supporting that idea [53]. Based on the Green et al. [61] analysis, the admixture seems to have been older than previously predicted as well, taking place between 47,000 and 65,000 years ago, possibly in the Levant just outside of Africa during an interglacial period [48, 67, 93]. The lower divergence estimates and the evidence for early admixture of Neanderthal and human lineages significantly change our perspective on Neanderthal‘s place in human evolution. Based on the mitochondrial genome, previous evidence had placed Neanderthals well outside of human variation, and showed no evidence for mixing between the two populations. One of the more interesting puzzles presented by the indications of Neanderthal admixture is the absence of any Neanderthal mitochondria in modern humans. Some have suggested this is the result of old, very low levels of interbreeding [71], but the exact cause of the disagreement between the stories told by the mitochondria and nuclear sequences is unclear.

3.4.1 Identifying Possibly Admixed Genome Regions

Previous examinations of genomic differences or admixture between the archaic genomes and modern humans have focused mainly on sites that are fixed and shared between either Denisova or Neanderthal and extant human populations [61, 70, 94]. Few studies have broadly examined genome regional similarities that may have resulted from admixture, with the exception of very recent work finding specific archaic haplotypes in non-African human populations. Here, I apply the method detailed in Kennedy [95] to attempt to discern genome regions in living human populations that are especially similar to the high-coverage Denisova genome published in [70]. In this way I attempt to examine admixture across the whole genome, region by region.

This comparison of modern humans with the Denisova genome allows for the identification of potentially admixed regions which resulted from introgression between Neanderthals and the ancestors of modern Eurasian populations. Of course, performing such a comparison with the Neanderthal genome would be preferable. However, the published Neanderthal genome lacks adequate SNP density and is of considerable lower quality, confidence, and coverage in comparison with the Denisova genome. Also preferable would be to compare

Denisova with Melanesian or Australian Aboriginal populations, but there are no high-coverage whole genome datasets widely available for either population. Because Neanderthal and Denisova share a more recent common ancestor with each other than either does with living human populations, they share considerable genetic diversity. Returning to Fig. 3.1, Neanderthal and Denisova would share any mutations that fall between the divergence of the archaic clade until their lineages split. Therefore, a comparison between Denisova and Eurasian populations should recover at least that component of archaic ancestry shared by Denisova and Neanderthal and introgressed into modern humans.

Because I am using the Denisova genome rather than the Neanderthal genome, and because there is only one individual to represent the archaic population, results from this analysis are far more suggestive than definitive. In addition, similar to other methods for detecting ancient admixture, this analysis can be confounded by ancient population structure persistent human populations that left Africa, though this scenario has been shown to be unlikely to cause the levels of admixture so far observed in Eurasians [76].

The method used here, called Ratio Distance or R_d, is somewhat comparable to a haplotype analysis. R_d utilizes principal component analysis of allele frequencies in a rolling window across the genome to detect regions where the target—in this case, Denisova—is more similar to one of the two comparison populations, while controlling for the relative distance between the those populations. Unlike other admixture analysis that focus on shared fixed derived sites, the PCA-based R_d is able to include sites that are polymorphic in either the archaic genome, the comparison human genomes or both. In the case of the archaic genome, polymorphic would simply equate to heterozygous because the sample size is 1. The statistic compares the region, locus by locus, then summarizes the differences into a single ratio measured along the principal components of the variation (See Fig. 3.1 in [95] for a visualization of the relationship between R_d, allele frequency covariance, and r^2). In this way, the analysis highlights regions where the target population is more similar to one comparison population than the other, and is maximized where the comparison populations are most divergent.

In the original implementation of the method, the algorithm is used to detect differential admixture from the parent populations into the target. Here I use R_d to detect admixture from a target individual into one of the comparison populations. Because current analysis has detected no Neanderthal or Denisova introgression in African populations [61, 62, 63], exceptional similarity between the archaic individual and the Eurasian population may be due to introgression. For the comparison populations I use whole genomes from the

1000 Genomes Project [31], specifically Europeans from Northern Utah (CEU), Chinese from Beijing (CHB), and Yoruba from Nigeria (YRI). The CEU and CHB populations are compared to Denisova with respect to YRI in separate analyses. In other words, one analysis will look for similarity between Denisova and CEU while the other will look for similarity between Denisova and CHB. YRI serves as the second comparison population in both analysis, because I am looking for regions that minimize the similarity between CEU/CHB and YRI, while maximizing the similarly of those populations to Denisova.

A possible source of shared similarity between Denisova and Eurasians—besides introgression—could be recent selection in that genome region in the African population. To control for this, I excluded from the results any regions of exceptional similarity between Denisova and CEU or CHB that also demonstrated higher iHS [78]—a measure of recent selection—in YRI compared to the other populations.

To demonstrate how R_d varies across the genome Figure 3.2 shows R_d values across chromosome 10 in the two comparisons. A similar plot could be generated for any chromosome. In this plot it is easy to observe regions where Denisova is similar to CEU or CHB, but also regions where the archaic is more similar to one of the Eurasian populations and not the other. On average, Denisova is slightly closer to Yorubans than they are to either CEU or CHB. This is somewhat expected because of ancestral variation being more present in African populations than non-African [96]. Between CEU and CHB, the Asian population is slightly more similar to Denisova on average, which is also unsurprising given recent findings of greater archaic ancestry in Asians than Europeans [70, 73].

Table 3.1 describes gene-containing regions of the CHB or CEU genomes that are exceptionally close to the Denisova genome while being divergent from theYRI. For all of these regions iHS is higher in CEU or CHB relative to YRI. R_d was calculated in 200kb rolling windows across the whole genome. The window size was chosen because it maximizes the number of regions that overcome population variation threshold necessary to detect population structure. This window size is larger than the original implementation of theR_d statistic because the Denisova has a lower variant density having only a single individual for variants to be called on.

Because R_d is normally distributed [95] the regions in Table 3.1 are identified as being exceptional by having R_d values that are greater than 5 standard deviations away from mean R_d. All R_d values in the table are negative because I am only interested in similarity to one of the parent populations (CEU or CHB) and not the the other (YRI). Gene regions where R_d is exceptionally positive would represent places where Denisova is similar to Yorubans, which cannot be the result of admixture, but rather may be shared ancestral variation.

Notably, The analysis recovers one region known to harbor a haplotype resultant from introgression. The region containing the OAS immunity genes cluster on chromosome 12 was previously identified by Mendez et al. [97, 72]. There does not seem to be any broad patterns or clustering of function in the regions that are most similar to Denisova in CHB or CEU. Interestingly, the region with the highest value of R_d in the analysis was on chromosome 3, near containing the CADM2 gene, which is implicated in brain development [98]. Some of the gene regions are found in both CEU and CHB, while others are specific to one of those populations. That some regions are shared and others not is unsurprising. The separation of European populations from Asia occurred shortly after the estimated time of the inbreeding event, so as the populations differentiated, regions of archaic admixture would be likely to be more strongly selected for in some populations and more weakly selected in others. Places in the genome where Denisova and YRI are more closely related are a bit more difficult to interpret, but may represent highly ancestral regions in that population.

Because of their genetic distance Denisova is not an ideal stand-in for measuring Neanderthal admixture in modern human populations. Currently there is no evidence of direct introgression from Denisova into any human population outside of Papua New Guinea and Aboriginal Australia [62, 63]. However, until a more complete genome of Neanderthal can be recovered, the Denisovan genome is the only archaic DNA with suitable coverage and depth for the type of analysis presented here. Because of this these results must be viewed as preliminary and need to be confirmed through future comparison with a high coverage, more complete Neanderthal genome.

3.5 Conclusion

The story of Neanderthal population history as told by the Neanderthal nuclear genome has thus far been very different from the previous story told by mitochondria. In the past, evidence suggested two highly divergent populations, splitting more than half a million years ago, where our own lineage eventually replaced the Neanderthals with minimal or absent admixture between the populations. Nuclear DNA, however, suggests that the human-Neanderthal split may have been as recent as 270,000 years ago. It also suggests that the two lineages experienced an interbreeding event less than 200,000 after their populations diverged. Two hundred thousand years is significantly shorter than the depth of the nuclear sequence variation of modern human populations. Even at the upper bounds of the [61] estimate for Neanderthals, the 370,000 years between divergence and introgression would be well within the bounds of modern variation. The disconnect between the mitochondrial story and that told by the nuclear genome is perhaps less than surprising. Nearly all human mitochondrial lineages coalesce into a very small number in within 40,000 years.

Recent work comparing the genomes of Neanderthals and Denisova to that of modern humans has revealed convincing evidence for recent archaic admixture. Here, I have highlighted some regions which may be candidates for closer examination as possible admixed regions. Because of the roughness of the measure and the small sample size, however, the R_d analysis can only be suggestive of haplotypic similarity between archaic Denisova and modern Eurasian populations. Other, more precise measures, such as those used by Mendez et al. [97], may be able to further distinguish these regions as being truly of archaic origin, or not.

The publication of the Denisova genome with remarkably high coverage and careful controls for contamination hints at a bright future for the new field of paleogenomics [70, 99], that is, if more specimens of similar quality can be recovered. These new discoveries shed light not only on the phenotype and population structure recent human ancestors, but also allow better estimation of the timing of changes in recent human evolutionary history and determine genome regions that have been under selection since our split with archaics [61, 69, 70, 94]. Future sequencing of more fossil hominin specimens will only improve these inferences.

Taken together, the sequencing of archaic hominins has substantially muddied the current picture of human evolution, particularity outside of Africa. Neither purely Out-of-Africa nor simple Multiregional Hypothesis based scenarios are able to fully describe our new understanding of human prehistory. New scenarios must be devised that can account for the nearly even spread of Neanderthal admixture across Eurasia, as well as the admixture from Denisova into Melanesian and Australian Aboriginal populations. Denisova in particular presents a new type of challenge, being an almost entirely genetically described human ancestor. Having recovered only a single digit, Denisova cannot be morphologically described, but must be understood using the tools of comparative genomics.

In light of all of this new information from the Neanderthal and Denisovan genomes, I return to the taxonomic question: are Neanderthals really a separate species from our *Homo sapiens* ancestors? As Hofreiter [67, p. 8] points out, "in the end it remains a philosophical question whether the two human forms are assigned to the same or different species or subspecies, which is, moreover, largely irrelevant for understanding the process of human evolution." What is important, however, is a better understanding of Neanderthal functional genetic differences and the contribution that admixture has made to modern human genetic differentiation, an understanding that can be achieved through the careful analysis of archaic genomes.

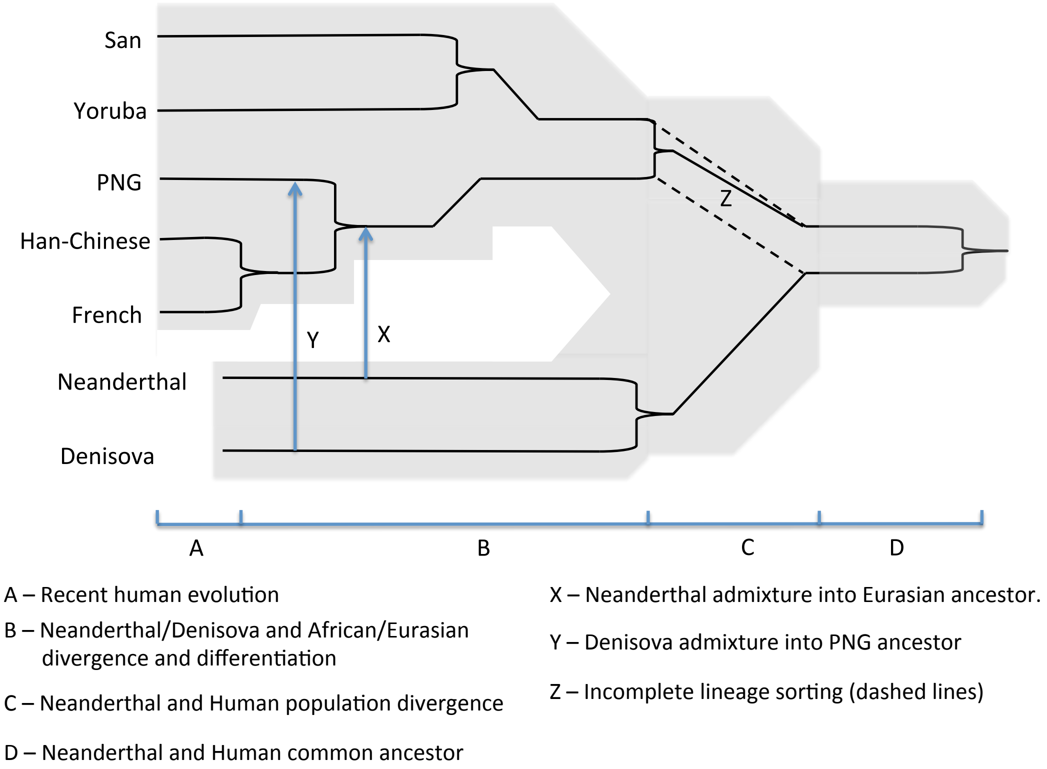

San
Yoruba
PNG
Han-Chinese
French
Neanderthal
Denisova
Y
X
Z
A
B
C
D
A – Recent human evolution
B – Neanderthal/Denisova and African/Eurasian divergence and differentiation
C – Neanderthal and Human population divergence
D – Neanderthal and Human common ancestor
X – Neanderthal admixture into Eurasian ancestor.
Y – Denisova admixture into PNG ancestor
Z – Incomplete lineage sorting (dashed lines)

Figure 3.1: Phylogeny and population subdivision of existing human populations and patterns of archaic admixture. Because of the wide range of dates estimated for the population divergence, the figure is not to scale.

Table 3.1: Regions of exceptional similarity between Denisova and either European (CEU) or Chinese (CHB). Regions that are found to be exceptional in both comparisons are bolded. Consecutive regions of high values are grouped together, and the highest R_d among them is shown.

Gene-Containing Regions where Denisova $R_d > 5\sigma$ from μ				
Population	Chromsome	HG19 Region	R_d	Genes
CEU $\mu_{R_d} = 0.1156$ $\sigma_{R_d} = 0.1821$	**1**	**144410583-144710583**	**-1.450**	**NBPF9**
	4	151510240-151710240	-1.2096	LRBA
	2	242985133-243085133	-0.9821	KIAA1919 REV3L
	6	111573924-111773924	-2.696	PXDNL PCMTD1
	8	52610422-52810422	-1.0165	DQ590378
	10	**47060523-47260523**	**-1.2690**	**PPYR1 ANXA8L1 ANXA8 FAM25B AGAP9 AK309024**
	11	105970855-106170855	-1.1234	BC034795
	12	**113361107-113561107**	**-1.8558**	**OAS1 OAS2 OAS3 DTX1**
	12	20961107-21161107	-1.2926	SLC01B3 SLC01B7
	19	**23780840-23980840**	**-0.9949**	**ZNF675** **RPSA**
CHB $\mu_{R_d} = 0.1017$ $\sigma_{R_d} = 0.1821$	**1**	**144410583-144710583**	**-1.550**	**NBPF9**
	1	169110583-169310583	-1.1128	NME7
	1	161760583-161860583	-1.533	ATF6
	2	24710133-24910133	-1.1379	NCOA1
	3	85460157-85760157	-2.1391	CADM2
	4	68810240-69010240	-1.2960	TMPRSS11A SYT14L
	9	123810023-124010023	-1.0839	FGL1 PCM1
	10	**47060523-47260523**	**-1.2690**	**PPYR1 ANXA8L1 ANXA8 FAM25B AGAP9 AK309024**
	12	**113361107-113561107**	**-1.7044**	**OAS1 OAS2 OAS3 DTX1**
	19	**23780840-23980840**	**-1.3923**	**ZNF675** **RPSA**

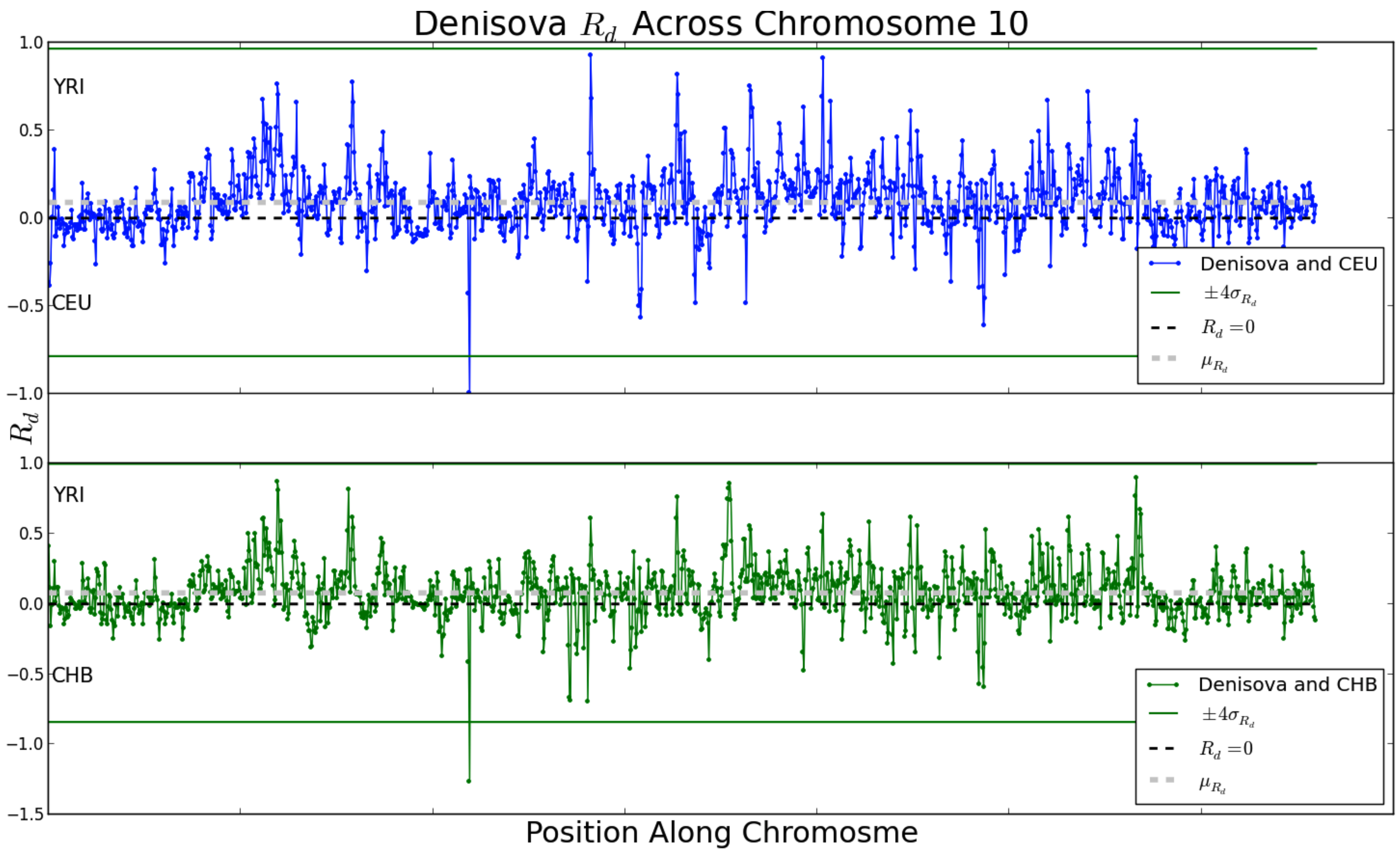


Figure 3.2: R_d for Denisova compared to CEU and CHB against YRI across chromosome 10.

REFERENCES

[1] Pennington R, Gatenbee C, Kennedy B, Harpending H, Cochran G (2009) Group differences in proneness to inflammation. *Infection, Genetics and Evolution* 9:1371–1380.

[2] Richardus JH, Kunst AE (2001) Black-white differences in infectious disease mortality in the United States. *American Journal of Public Health* 91:1251–1253.

[3] Van Dyke AL, Cote ML, Wenzlaff AS, Land S, Schwartz AG (2009) Cytokine SNPs: Comparison of allele frequencies by race and implications for future studies. *Cytokine* 46:236–244 doi: DOI: 10.1016/j.cyto.2009.02.003.

[4] Albert MA (2007) Inflammatory biomarkers, race/ethnicity and cardiovascular disease. *Nutrition Reviews* 65:S234–S238.

[5] Young CJ, Gaston RS (2000) Renal transplantation in Black Americans. *N Engl J Med* 343:1545–1552.

[6] Bidwell J, et al. (1999) Cytokine gene polymorphism in human disease: On-line databases. *Genes and Immunity* 1:3–19.

[7] Hurst LD, Smith NG (1999) Do essential genes evolve slowly? *Current Biology* 9:747–50.

[8] Hawks J, Wang ET, Cochran GM, Harpending HC, Moyzis RK (2007) Recent acceleration of human adaptive evolution. *Proceedings of the National Academy of Sciences* 104:20753–20758.

[9] Yazici AC, Atac FB, Verdi H, Ozbek N (2009) Comparison of IL10 and IL2 genotypes of turkish population with other populations. *Int J Immunogenet* 36:97–101.

[10] Anderson, R.M.; May R (1991) *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford and New York).

[11] Javor J, Bucova M, Ferencik S, Grosse-Wilde H, Buc M (2007) Single nucleotide polymorphisms of cytokine genes in the healthy Slovak population. *International Journal of Immunogenetics* 34:273–280.

[12] Ness RB, Haggerty CL, Harger G, Ferrell R (2004) Differential distribution of allelic variants in cytokine genes among African Americans and White Americans. *American Journal of Epidemiology* 160:1033–1038.

[13] Wirz SA, et al. (2004) High frequency of tnf alleles -238a and -376a in individuals from Northern Sardinia. *Cytokine* 26:149–154.

[14] Hollegaard MV, Bidwell JL (2006) Cytokine gene polymorphism in human disease: on-line databases, supplement 3. *Genes and Immunity* 7:269–276.

[15] Thorisson GA, et al. (2009) Hgvbaseg2p: a central genetic association database. *Nucleic Acids Research* 37:D797–802.

[16] Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* 15:1496–1502.

[17] Deventer SJHv (2000) Cytokine and cytokine receptor polymorphisms in infectious disease. *Intensive Care Medicine* 26:S098–S102.

[18] Harding D, et al. (2003) Is interleukin-6 -174 genotype associated with the development of septicemia in preterm infants? *Pediatrics* 112:800–803.

[19] Grossman SR, et al. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327:883–886.

[20] Price A, Zaitlen N, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* 11:459–463.

[21] Seldin MF, Pasaniuc B, Price AL (2011) New approaches to disease mapping in admixed populations. *Nature Reviews Genetics* 12:523–528.

[22] Price AL, et al. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 5:e1000519.

[23] Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* 2:e190.

[24] Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

[25] Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. models and estimation procedures. *American Journal of Human Genetics* 19:233–257.

[26] Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* 40:646–649.

[27] Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD (2010) Fine-scale population structure and the era of next-generation sequencing. *Human Molecular Genetics* 19:R221–6.

[28] Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456:98–101.

[29] Francois O, et al. (2010) Principal component analysis under population genetic models of range expansion and admixture. *Mol Biol Evol* 27:1257–1268.

[30] McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genetics* 5:e1000686.

[31] Abecasis GR, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.

[32] Ma J, Amos CI (2010) Theoretical formulation of principal components analysis to detect and correct for population stratification. *PLoS ONE* 5:e12510.

[33] Baik J, Ben Arous G, Péché S (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability* 33:1643–1697.

[34] Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M (2011) Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol* 12:R19.

[35] Bryc K, et al. (2010) Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences* 107:8954–8961.

[36] Price AL, et al. (2007) A genomewide admixture map for Latino populations. *American journal of human genetics* 80:1024–1036.

[37] Tian C, et al. (2007) A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *American Journal of Human Genetics* 80:1014–1023.

[38] Tian C, et al. (2006) A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. *American Journal of Human Genetics* 79:640–649.

[39] Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Research* 12:996–1006.

[40] Kitamura T, et al. (2008) Regulation of VEGF-mediated angiogenesis by the Akt/PKB substrate Girdin. *Nature Cell Biology* 10:329–337.

[41] Andrés AM, et al. (2009) Targets of balancing selection in the human genome. *Molecular Biology and Evolution* 26:2755–2764.

[42] Jones E, Oliphant T, Peterson P, et al. (2001) SciPy: Open source scientific tools for Python. http://www.scipy.org/.

[43] Hunter JD (2007) Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* 9:90–95.

[44] Wolpoff MH, Caspari R (1996) *Race and Human Evolution: A Fatal Attraction* (Simon and Schuster, New York), p 462 p.

[45] Fischer A, Pollack J, Thalmann O, Nickel B, Paabo S (2006) Demographic history and genetic differentiation in apes. *Current Biology* 16:1133–1138.

[46] McEvoy BP, Powell JE, Goddard ME, Visscher PM (2011) Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Research* 21:821–829.

[47] Harvati K (2003) Quantitative analysis of Neanderthal temporal bone morphology using three-dimensional geometric morphometrics. *American Journal of Physical Anthropology* 120:323–338.

[48] Hodgson JA, Bergey C, Disotell T (2010) Neandertal genome: The ins and outs of African genetic diversity. *Current Biology* 20:R517–R519.

[49] Tattersall I, Schwartz JH (1999) Hominids and hybrids: The place of Neanderthals in human evolution. *Proceedings of the National Academy of Sciences* 96:7117–7119.

[50] Mellars P (2004) Neanderthals and the modern human colonization of Europe. *Nature* 432:461–465.

[51] Eckhardt RB, Wolpoff MH, Thorne A (1993) Multiregional Evolution. *Science* 262:973–974.

[52] Wolpoff MH, Hawks J, Caspari R (2000) Multiregional, not multiple origins. *American Journal of Physical Anthropology* 112:129–136.

[53] Wolpoff MH, Caspari R (2011) Neandertals and the roots of human recency. *Continuity and discontinuity in the peopling of Europe* pp 367–377.

[54] Duarte C, et al. (1999) The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. *Proceedings of the National Academy of Sciences* 96:7604–7609.

[55] Krings M, et al. (1997) Neandertal DNA sequences and the origin of modern humans. *Cell* 90:19–30.

[56] Kahn P, Gibbons A (1997) DNA from an extinct human. *Science* 277:176–178.

[57] Krings M, et al. (2000) A view of Neandertal genetic diversity. *Nature Genetics* 26:144–146.

[58] Lalueza-Fox C, et al. (2005) Neandertal evolutionary genetics: Mitochondrial DNA data from the Iberian peninsula. *Molecular Biology and Evolution* 22:1077–1081.

[59] Green RE, et al. (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* 444:330–336.

[60] Wall JD, Kim SK (2007) Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genetics* 3:e175.

[61] Green RE, et al. (2010) A draft sequence of the Neandertal genome. *Science* 328:710–722.

[62] Reich D, et al. (2010) Genetic history of an archaic hominin group from Denisova cave in siberia. *Nature* 468:1053–1060.

[63] Reich D, et al. (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics* 89:516–528.

[64] Endicott P, Ho SYW, Stringer C (2010) Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of Neanderthal and modern human origins. *Journal of Human Evolution* 59:87–95.

[65] Paabo S (1999) Human evolution. *Trends in Cell Biology* 9:M13–6.

[66] Schlebusch CM, et al. (2012) Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* 338:374–379.

[67] Hofreiter M (2010) Drafting human ancestry: What does the Neanderthal genome tell us about hominid evolution? commentary on green et al. (2010). *Human Biology* 83:1–11.

[68] Burbano HA, et al. (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* 328:723–725.

[69] Maricic T, et al. (2012) A recent evolutionary change affects a regulatory element in the human FOXP2 gene. *Molecular Biology and Evolution.*

[70] Meyer M, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338:222–226.

[71] Neves AGM, Serva M (2012) Extremely rare interbreeding events can explain Neanderthal DNA in living humans. *PLoS ONE* 7:e47076.

[72] Mendez FL, Watkins JC, Hammer MF (2013) Neandertal origin of genetic variation at the cluster of oas immunity genes. *Molecular Biology and Evolution* 30:798–801.

[73] Wall JD, et al. (2013) Higher levels of Neanderthal ancestry in east asians than in Europeans. *Genetics.*

[74] Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD (2011) Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences* 108:15123–15128.

[75] Eriksson A, Manica A (2012) Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proceedings of the National Academy of Sciences* 109:13956–13960.

[76] Yang MA, Malaspinas AS, Durand EY, Slatkin M (2012) Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Molecular Biology and Evolution* 29:2987–2995.

[77] Demuth JP, Bie TD, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. *PLoS ONE* 1:e85.

[78] Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biology* 4:e72.

[79] Abi-Rached L, et al. (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* 334:89–94.

[80] Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet* 39:1140–1144.

[81] Pollard K, et al. (2006) Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* 2:e168.

[82] Prabhakar S, Noonan JP, Paabo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science* 314:786.

[83] Green RE, et al. (2009) The Neandertal genome and ancient DNA authenticity. *EMBO J* 28:2494–2502.

[84] Krause J, et al. (2007) The derived FOXP2 variant of modern humans was shared with Neandertals. *Current Biology* 17:1908–1912.

[85] Enard W, et al. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418:869–872 10.1038/nature01025.

[86] Coop G, Bullaughey K, Luca F, Przeworski M (2008) The timing of selection at the human FOXP2 gene. *Molecular Biology and Evolution* 25:1257–1259.

[87] Rosas A, et al. (2006) Paleobiology and comparative morphology of a late Neandertal sample from El Sidron, Asturias, Spain. *Proceedings of the National Academy of Sciences* 103:19266–19271.

[88] Lalueza-Fox C, et al. (2008) Genetic characterization of the ABO blood group in Neandertals. *BMC Evolutionary Biology* 8:342.

[89] Lalueza-Fox C, Gigli E, de la Rasilla M, Fortea J, Rosas A (2009) Bitter taste perception in Neanderthals through the analysis of the TAS2R38 gene. *Biology Letters* 5:809–811.

[90] Lalueza-Fox C, et al. (2007) A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. *Science* 318:1453–1455.

[91] Prabhakar S, et al. (2008) Human-specific gain of function in a developmental enhancer. *Science* 321:1346–1350.

[92] Joris O, Street M (2008) At the end of the 14C time scale–the Middle to Upper Paleolithic record of Western Eurasia. *Journal of Human Evolution* 55:782–802.

[93] Sankararaman S, Patterson N, Li H, Pääbo S, Reich D (2012) The date of interbreeding between Neandertals and modern humans. *PLoS Genetics* 8:e1002947.

[94] Paixao-Cortes VR, Viscardi LH, Salzano FM, Hunemeier T, Bortolini MC (2012) Homo sapiens, Homo Neanderthalensis and the Denisova specimen: New insights on their evolutionary histories using whole-genome comparisons. *Genetics and Molecular Biology* 35:904–911.

[95] Kennedy BJ (2013) Ph.D. thesis (University of Utah).

[96] Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

[97] Mendez FL, Watkins JC, Hammer MF (2012) Global genetic variation at oas1 provides evidence of archaic admixture in Melanesian populations. *Molecular Biology and Evolution* 29:1513–1520.

[98] Hunter PR, et al. (2011) Localization of Cadm2a and Cadm3 proteins during development of the zebrafish nervous system. *The Journal of Comparative Neurology* 519:2252–2270.

[99] Wall JD, Slatkin M (2012) Paleopopulation genetics. *Annual Review of Genetics* 46:635–649.