UNIVERSITY OF CALIFORNIA, SAN DIEGO

Admixture History of Andean Highlanders

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Shubham Saini

Committee in charge:

Professor Vineet Bafna, Chair
Professor Vikas Bansal
Professor Melissa Gymrek

2017

The thesis of Shubham Saini is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2017

DEDICATION

To Family.

EPIGRAPH

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.
—T. S. Eliot

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I take the opportunity to acknowledge the people who helped shape this thesis. I would like to express my profound gratitude to my advisor and thesis chair Dr. Vineet Bafna for his advise and support throughout the course of the masters program. He is a great inspiration for my continued interest in the field of bioinformatics.

Besides my advisor, I would like to thank my thesis committee members, Dr. Melissa Gymrek and Dr. Vikas Bansal, for taking time to review this work and giving valuable suggestions.

I would like to express a special thanks of gratitude to my professors, at UCSD and VIT, who over the years shaped me into the scientist I am today.

In the end, I reserve a special mention for my friends and family. I owe all my accomplishments to them.

VITA

2014	B. Tech in Computer Science and Engineering, Vellore Institute of Techonology, India
2014-2015	Research Associate, Indraprastha Institute of Information Technology, India
2015-2017	Master of Science in Computer Science, University of California, San Diego

PUBLICATIONS

Afreen Ferdoash, Shubham Saini, Jitesh Khurana, Amarjeet Singh, ”Analytics Driven Operational Efficiency in HVAC Systems” at The 2nd ACM International Conference on Embedded Systems For Energy-Efficient Built Environments, BuildSys 2015, Seoul, South Korea.

Shubham Saini, Pandarasamy Arjunan, Amarjeet Singh, Ullas Nambiar, ”E-Adivino: A Novel Framework for Electricity Consumption Prediction based on Historical Trends” at The 6th ACM International Conference on Future Energy Systems, eEnergy 2015, Bangalore, India.

Amarjeet Singh, Shubham Saini, Sanchit Sharma, Priyank Trivedi, ”Energy Optimization in Commercial Buildings: From Monitoring to Savings Realization” at The 6th ACM International Conference on Future Energy Systems, e-Energy 2015, Bangalore, India.

Shubham Saini, Shraey Bhatia, I. Sumaiya Thaseen, ”sv(M)kmeans - A Hybrid Feature Selection Technique for Reducing False Positives in Network Anomaly Detection” at The 20th International Conference on Management of Data, COMAD 2014, Hyderabad, India.

Bhavesh Kasliwal*, Shraey Bhatia, Shubham Saini*, I.Sumaiya Thaseen, Ch.Aswani Kumar, ”A Hybrid Anomaly Detection Model using G-LDA” at The 4th IEEE International Advance Computing Conference, IACC 2014, Gurgaon, India.

ABSTRACT OF THE THESIS

Admixture History of Andean Highlanders

by

Shubham Saini

Master of Science in Computer Science

University of California, San Diego, 2017

Professor Vineet Bafna, Chair

South American populations have a complex admixture history. The earliest Native Americans are known to have migrated into the Americas as early as 14000 years ago. The region saw large scale migrations from European colonial powers starting early 1500s. In just a couple of centuries the entire Western hemisphere came under the control of these European powers. These colonial powers brought millions of Africans across the Atlantic between the 1500s and 1800s through the slave trade which led to further admixture of African populations into the Americans. Due to multiple migrations events, presence of different colonial powers that followed different slave trade practices, South Americans populations have a highly heterogeneous genetic composition. We investigate the admixture history of one of the South American populations living in Cerro de Pasco, a high altitude mining town in the Andes region of Peru. Studying the admixture history of this population can give us important insights into the origins of several selection forces at play in this population. We found the Andean population to have lesser proportions of European and African ancestry as compared to the other South American populations. The timing of European and African admixture into this population was also found to be significantly different from other populations, which could be a direct result of lower European and African ancestry proportion. These results indicate a higher level of segregation between the Native Americans and the European/African populations in Peru as compared to other regions.

Chapter 1

Introduction

South American populations share an interesting and complex admixture history. The earliest Native Americans entered the Americas from Siberia via the Bering Land Bridge [RSH$^+$15] nearly 14000 years ago. While the timing of this event is well accepted, the number and timings of migration waves is still under research. However, unlike North America, South American indigenous population is derived from a single migration wave that spread South into the Andes and East into the Amazon basin.

Large scale European migration and colonization then began in early 1500s, starting from Caribbean islands and expanding into the rest of the American mainland. With Spanish being the first Europeans to settle the largest areas of North America, Caribbean and South America, English arriving on the North American coast, and Portuguese and French colonizing parts of Americas, eventually the entire Western hemisphere came under the control of European powers.

Things get more complicated because of the African admixture into the American populations. Between the 1500s and 1800s, millions of African were brought into the Americas for working in mines and plantations through the slave trade, which led to further mixing of Africans into the South American populations. It is estimated that nearly four Africans for every one European crossed the Atlantic. Interesting, most of the slaves were brought into the northern South American areas and North America, as evident from the admixture proportion of present day populations.

Due to multiple migration events, presence of different colonial powers that followed different slave trade practices, the South American populations have a highly heterogeneous genetic background. Several factors like diverse geographical features, social divisions between different ethnicities, and higher slave trade than other regions make the Peruvian populations different from other South American populations. We investigate the admixture of history of one of the South American populations living in Cerro de Pasco, a high altitude mining town in Andean region of Peru. This population is shown to have adapted to high altitude regions with low atmospheric oxygen levels in several works [ZUR$^+$13] and multiple samples are diagnosed with Monge disease (Chronic Mountain Sickness). Studying the admixture history of this population can give us important insights into the origins of the selection and the resulting phenotypes. Using publicly available methods we estimate the ancestry proportions of the Andean population. We further estimate the timing of these admixture events and compare it with other American populations.

1.1 Related Work

A number studies on earliest Native American migrations and admixture history of present day South American populations have been done previously. Reich et. al [RPC$^+$12] did an extensive work of collecting ancient and present day samples of Native Americans and studying the genetic diversity of the present American populations. Raghavan et. al [RSH$^+$15] worked on uncovering the migration patterns of earliest Native Americans and proposed some interesting facts about the timing and route of these migrations. Ruiz-Linares et. al [RLAAA$^+$14] did an extensive analysis of genetic ancestry of Latin Americans and found geographic variation in ancestry and its impact on various physical traits. Waldron et. al [Wal16] studied the genetic events and traced the ancestral origins of South Americans from five countries that led to the population structure of present day South America.

1.2 Thesis Organization

The remainder of this thesis is organized as follows: Chapter 2 describes the data sets used for this project and preprocessing steps required for analysis. Chapter 3 describes the algorithms and tools used along with the motivation for selecting those tools. Results and corresponding discussion are presented in Chapter 4. Finally, we give concluding remarks about the results and the future scope of this work in Chapter 5.

Chapter 2

Data Sets

2.1 1000 Genomes Project

The 1000 Genomes Project is an international research effort to catalog the human genetic variation. It consists of reconstructed genomes of 2504 samples from 26 populations. The data has high quality haplotypes from 88 million variants including single nucleotide polymorphisms (SNPs), insertion/deletion (indels) and structural variants (SVs). The regions where each population is sampled from are presented on a world map in Figure 2.1.

In this work we use data of 150 European samples (CEU and GBR), 150 Africans (YRI and MSL), 94 Colombians (CLM), 64 Mexicans (MXL), 85 Peruvians from Lima (PEL) and 104 Puerto Ricans (PUR). The sample IDs and additional information of the selected samples is given in Appendix A.

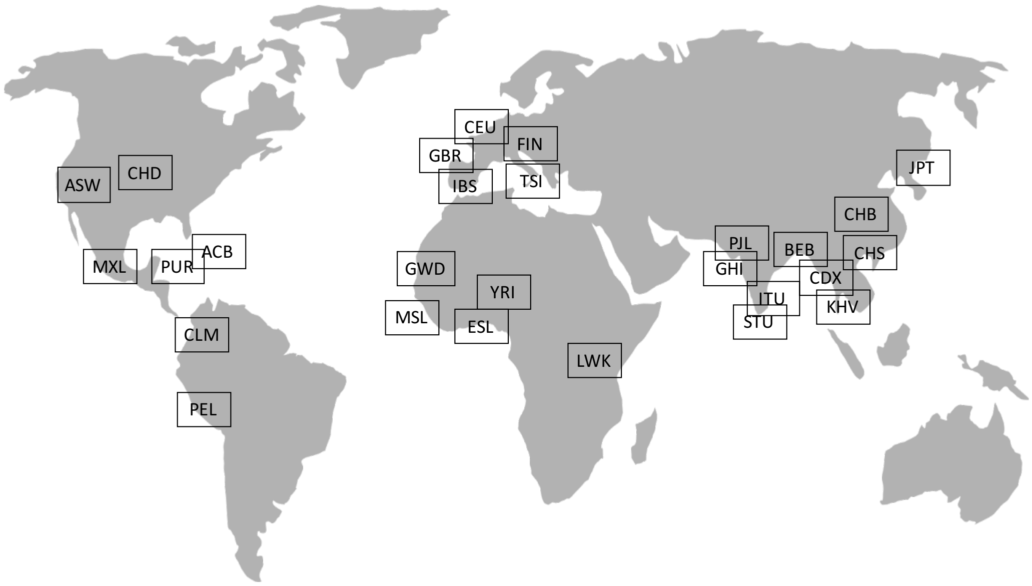

Figure 2.1: 1000 Genomes Project: World wide distribution of the populations sampled as part of the project. Actual names for the population IDs can be found at http://www.internationalgenome.org/category/population/

2.2 Andean Highlanders Data

The Andean Highlanders data, as described in Zhou et. al [ZUR$^+$13], has been previously used to study the genetic mechanisms underlying high altitude adaptation and decode the genetic basis of Chronic Mountain Sickness or Monge disease. The data consists of 120 samples from Andean mountains, residing in Cerro de Pasco, a high altitude mining town of Peru. Whole Genome Sequencing was performed with the Illumina HiSeq2000 platform to a mean per sample depth of 20x-40x. The reads were aligned to the human reference genome hg19 using BWA[LD09], and adjusted using GATK indel realignment. SNVs were finally called and filtered using GATK UnifiedGenotyper with default parameters. We used this data, as prepared by Zhou et. al [ZUR$^+$13], for studying the European and African admixture into the population. We further imputed and phased the missing SNPs in this data set using the 1000 Genomes samples mentioned in the previous section using ShapeIt. This led to retaining only those markers that were present in the 1000 Genomes data.

2.3 Native American Data

In an effort to study the peopling of Americas, Reich et. al [RPC$^+$12] assembled data of 493 individuals from 52 Native American groups, genotyped at nearly 350,000 SNPs. All the samples were genotyped using Illumina arrays at the Broad Institute of Harvard and Massachusetts Institute of Technology. Of the 493 samples, 419 samples were genotyped from genomic DNA, and 74 from whole-genome-amplified material. HAPMIX [PTP$^+$09] was used to model the haplotypes into one of the two ancestral panels: ”Old World” population of Europeans and Africans, and the ”Native” population of Native Americans and Siberians. Genome segments with an expected number of more than 0.01 non-Native American ancestry found using HAPMIX was masked.

We use 100 of the 493 Native American samples for this work (details present in Appendix B). We selected the samples with least proportions of European and African admixture, and in many cases had little to no masking done on them. We also ran IBD analysis to make sure the samples were not related. We further imputed and phased the missing SNPs using the 1000 Genomes samples mentioned previously using ShapeIt. This led to retaining only those markers that were present in the 1000 Genomes data.

Chapter 3

Methods

3.1 Haplotype Phasing using ShapeIt

A haplotype is a sequence of nucleotides along a single strand of a chromosome. With current sequencing technologies it is difficult to obtain sequences for each chromosomal strand separately. We instead obtain genotype information for each position without the strand information. Haplotype phasing is the process of assigning each nucleotide to the correct strand. Haplotype phase information is required for solving problems like detecting chromosomal segments of distinct ancestry, understanding genetic variation and diseases. understanding gene function, and detecting selection.

A number of methods have been developed in response to tackle these problems. We use one of these methods called ShapeIt [DMZ12], a Hidden Markov Model (HMM) based linear time algorithm. ShapeIt collapses the haplotypes into a graph structure and forms an HMM for this graph. The haplotypes are divided into equal sized disjoint segments, with each segment having J distinct haplotypes. The HMM thus formed has J states for each markers. The edges traversing the HMM are weighted by the number of haplotypes that traverse the nodes and edges. This process is presented in Figure 3.1.

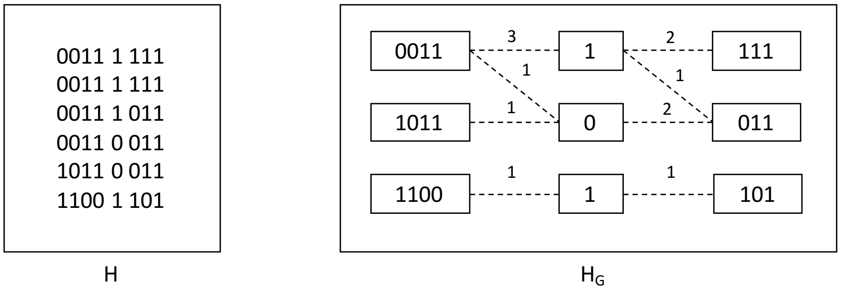

Figure 3.1: ShapeIt haplotype graph. Originally there are six haplotypes with eight markers each, given by H. This is collapsed into a graph H_G, with three unique haplotypes in each segment. The edges are weighted by the number of haplotypes traversing from one node to the next.

Another major improvement of ShapeIt over other methods is linear time sampling of compatible haplotypes. While other methods like Impute2 [HDM09] and MaCH [LWD$^+$10] do this in O(MN^2) where M is the number of markers and N the number of haplotypes, ShapeIt splits the compatible haplotypes into disjoint segments to achieve linear time $O(MJ)$ time sampling of compatible haplotypes. The sampling graph is presented in Figure 3.2.

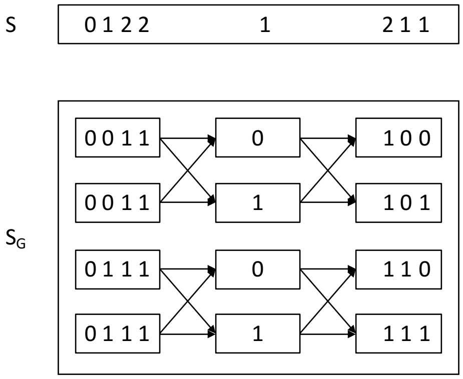

Figure 3.2: Linear time sampling of compatible haplotypes by ShapeIt. Each marker in S can have three possible values: 0/1/2. Compatible haplotypes are sampled from the graph S_G.

Most of the recent methods make use of Identity by Descent (IBD) information to get better phasing accuracy. The key idea behind using IBD information is that even unrelated individuals have distant common ancestors that give rise to IBD segments. This approach was first successfully used [KMF$^+$08] by leveraging long rang ($\geq 10Mb$) IBD information. IBD based phasing requires a identical haplotypes in sufficiently large number of samples to phase heterozygous markers. Alternatively, it can be used to phase related individuals given their pedigree information. Mendelian constraints provide haplotype phase information for many heterozygous markers, since any parent-offspring pair must share at least one allele at every position, and the IBD alleles at different sites on the same chromosome will be on a single strand in both the parent and the child, not accounting for recombination events. Thus, haplotype phase information is unknown only at markers where both the parent and offspring are heterozygous, or where genotype information is missing. But the number of such markers are very few whenever a trio (father-mother-child) information is available.

The ShapeIt method is implemented using C++ programming language. The software takes as input a Variant Call Format (VCF) file of the population genotype data, and a recombination maps file that contains the genomic coordinates of the recombination events for a specie. An additional phased reference file of population close to the target population may be given to help increase the phasing accuracy. If a reference file is supplied, then only the markers present in both the reference and target files are phased, and the markers missing from the target file are imputed (approximated). The output from the ShapeIt software is a phased VCF file.

3.2 Population Stratification using Principal Component Analysis

Principal Component Analysis (PCA) [Pea01] is a statistical method for converting multi-dimensional data of correlated variables into uncorrelated orthogonal variables called Principal Components. PCA is an efficient way to transform a high-dimensional data into low dimensional representation.

PCA has traditionally been used for population stratification [PPP$^+$06]. PCA is applied to genotype data to calculate the principal components that explain the genetic variation in the individuals. The top principal components are continuous axes of variation that represent genetic variation due to ancestry of the samples. We can further apply a clustering algorithm like K-Means for grouping individuals from the same population based on some distance metric.

While the method is easy with little parameters, application of PCA for population stratification has major challenges:

- Deciding the optimal number of principal components.
- Choice of distance metric used for grouping individuals.

3.3 Ancestry Proportion Estimation using *ADMIXTURE*

Admixture is an event when two previously isolated population begin to interbreed. Estimating the proportion of ancestry of an individual from each contributing ancestral population, averaged over the entire genome is known as ”global ancestry”, ”admixture proportion” estimation. *ADMIXTURE* [ANL09] is a Bayesian modeling based admixture estimation algorithm. Unlike other modeling based algorithms like *STRUCTURE* [HFSP09] that uses Markov Chain Monte Carlo (MCMC) for sampling posterior probabilities, *ADMIXTURE* makes use of likelihood function optimization which makes it magnitudes of levels faster. *ADMIXTURE* simultaneously estimates the ancestry proportion and the population allele frequencies, and the likelihood optimization function can accommodate larger number of parameters. The updates to the allele frequency parameters and the ancestry fraction parameter is done alternatively by maximizing the second order Taylor's expansion of the likelihood function. The maximum likelihood estimates are done using a block relaxation approach, which s accelerated using a Quasi-Newton method [ZAL11]. For a population k and individual i, the likelihood function of the ancestry proportion q_{ik} and allele frequencies f_{kj} is given as a function of g_{ij} - number of copies of allele 1 in individual i at SNP j:

$$L(Q,F) = \sum_i \sum_j g_{ij} ln \left[\sum_k q_{ik} f_{kj} \right] + (2 - g_{ij}) ln \left[\sum_k q_{ik(1-f_k j)} \right] \tag{3.1}$$

ADMIXTURE accepts as input genotype data each population in PLINK BED (Binary) [PNTB+07] format. The number of populations K is needed as a parameter.The output from the software is percentage global ancestry proportions of the K populations for each individual.

3.4 Admixture Events Dating using Alder

Alder (Admixture-induced Linkage Disequilibrium for Evolutionary Relationships) [LLP+13] is method for dating admixture events - the number of generations since an admixture event happened. Unlike other admixture events dating and reconstruction methods like *ROLLOFF* that are based on allele frequency divergence or modeling Chromosomal Segments of Distinct Ancestry (CSDA), Alder models the exponential decay of admixture induced linkage disequilibrium (LD) as a function of genetic distance. Assuming a population C is derived from the admixture of populations A and B, Alder uses a weighed LD statistic given as:

$$LD = w(x)w(y)D_2(x,y) \tag{3.2}$$

where $w(x)$ is the allele frequency divergence at site x between admixing populations A and B, and $D_2(x,y)$ is the sample covariance between genotypes at sites x and y in the admixed population C.

The exponential decay of weighted LD is modeled as a function of genetic distance d in cM. The decay constant k of the distribution arises from the heterogeneous mixture computed using pair of SNPs on different chromosome, the amplitude α is the mixture proportion and the branch length, and the decay rate z is the number of generations since the admixture event. The two populations method can be extended to three populations by training multiple models between pairs of populations.

Alder accepts as input genotype data for the ancestral and target population in EIGENSTRAT [PPP+06] format. The output from the software is the weight LD decay for each pair of population, along with the exponential decay parameters.

3.5 Local Ancestry Estimation using LAMP-LD

Local ancestry refers to the ancestry at ever genomic locus in an admixed population. Local ancestry has various applications in medical genetics, recombination rate variation studies and detection of selection events. A number of methods have been proposed to address the local ancestry estimation problem like HapMix [PTP+09], WinPop [PSKH09], LAMP-LD [BPS+12]. While HapMix and WinPop are shown to have good accuracy in two way admixed populations, LAML-LD extends previous methods to accomodate multi-way admixtures like the South American population.

LAMP-LD is a HMM based method, working on non-overlapping windows of genomic data. The hidden states of the HMM represent the local ancestries of each window. For a pair of ancestral states, sub-HMMs within the top HMM emit the genotypes and models the ancestries. The transition probabilities of the HMMs are estimated using reference panel of the ancestral populations. Each window is assigned a pair of local

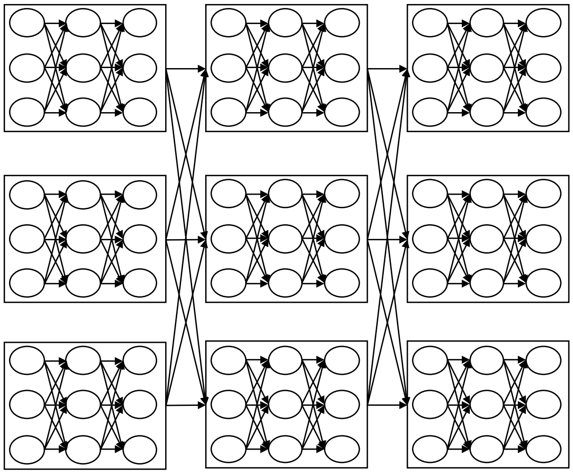

Figure 3.3: HMM structure of LAMP-LD: Top level HMM with smaller HMMs as used by LAMP-LD. Each HMM emits the most likely pair of ancestries.

ancestries, followed by relaxing the restriction on locations of ancestry switches. The top HMM, with smaller sub-HMMs as implemented by LAMP-LD is presented in Figure 3.3.

LAMP-LD accepts as input phased haplotype data for all the ancestral populations, and genotype data for the target population. Additionally, genetic maps are also required to capture the recombination process. The output from the software is local ancestry file for each individual, where the ancestry of each locus is given by 0/1/2 (0: both alleles from population 1, 1: one allele each from two populations, 2: both alleles from population 2).

3.6 Admixture Dynamics

Studying genetic admixture is important for evolutionary and medical studies. Exploring population admixture dynamics is important for admixture mapping, discovering population history, detecting natural selection signals. Typically, admixed populations follow one of the following three admixture models:

- **Hybrid Isolation:** two mutually isolated populations interbreed and the resulting admixed population evolve in isolation.
- **Gradual Admixture:** two mutually isolated populations interbreed and the resulting admixed population evolve with the ancestral populations.
- **Continuous Gene Flow:** two mutually isolated populations interbreed and the resulting admixed population periodically breeds with one of the ancestral populations.

The three models are represented in Figure 3.4.

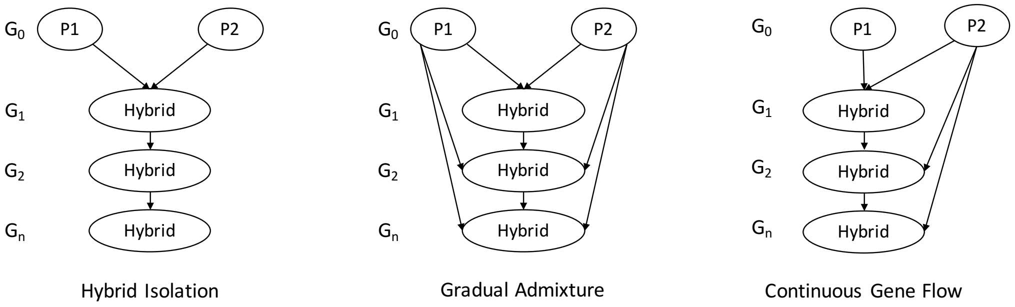

Figure 3.4: Population Admixture Dynamics Models: mutually isolated populations P1 and P2 breed to produce hybrid population. This hybrid population evolves over generations G_i, with or without further admixture from P1 or P2 or both depending on the kind of model.

Individuals of an admixed population have their chromosomes resemble a collection of chromosomal segments of the ancestral populations. The CSDA gets rearranged through recombination which provide important information about the population history. Most of the previous studies explored admixture dynamics through simulated data, and do not generally take into account complexity of admixture processes. Jin et. al [JWW$^+$12] proposed an approach to explore the population admixture dynamics by genome-wide analysis of CSDA. The general idea behind the proposed approach is that CSDA will be spliced into smaller pieces with each generation after an admixture event, sand the chromosomes from recently admixed populations contain longer CSDAs. Thus, typically the average length of CSDA in population following Hybrid isolation model will be much less than the average length of CSDA in populations following Gradual Admixture or Continuous Gene Flow model.

Chapter 4

Results and Discussion

4.1 Global Population Stratification

Principal Component Analysis (PCA) was carried out on Chromosome 19 genotype data using PLINK [PNTB$^+$07] PCA implementation. For performing PCA different world populations available in the 1000 genomes data, 100 Native American samples, and 120 Andean samples were used. 1000 genomes data consists of 5 super populations: Africans, Americans, Europeans, South Asians and East Asians. While merging different panels, only those markers were retained that had at least 70% genotyping rate.

Top two principal components were plotted using Python and is presented in Figure 4.1. The African samples were clustered away from all other populations and is separated by PC1. The remainder of populations formed distinct clusters along the PC2. The Andean cluster superimposes over the American (Mexican, Colombian, Puerto Rican and Peruvian) cluster as expected. Further, the Andean cluster lies between all the other clusters, implying close relationship to the other populations.

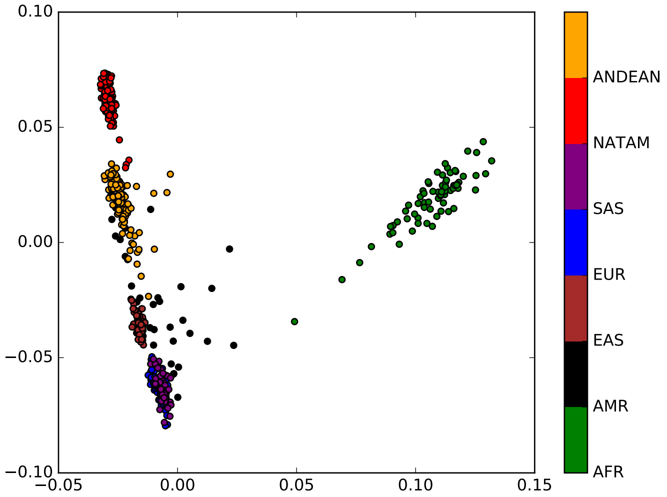

Figure 4.1: Principal Component Analysis of world wide populations: NATAM: Native Americans, SAS: South Asians, EUR: Europeans, EAS: East Asians, AMR: 1000 Genomes South Americans, AFR: Africans. The Andean population is related closest to Europeans, Native Americans and East Asians.

4.2 Ancestry Proportion of the Andean Population

We used the program *ADMIXTURE* [ANL09] to evaluate the proportions of ancestry of three major contributing populations: Africans, Europeans and Native Americans. Four other South American populations: Colombians, Mexicans, Puerto Ricans and Peruvians, were also used to compare with the Andean population.

ADMIXTURE method does not explicitly take LD into consideration, and requires that data be filtered to remove LD. To achieve this, the combined data was processed to remove each SNP that had an R^2 value of greater than 0.1 with any other SNP within a 50 SNP sliding window, advanced by 10 at a time. *ADMIXTURE* was then run in unsupervised mode using number of ancestral populations parameter $K = 3$ since we have three major populations.

As mentioned earlier, the Native American data had European and African markers masked, hence showing no admixture proportion with other populations. Interesting, unlike other South American populations, Peruvians and Andeans do not have high proportions of African admixture despite large scale slave trade history in the region. One possible reason for this could be the social division between the Africans and other groups of Peru. We also notice that unlike other South American populations (Mexicans, Puerto Ricans and Colombians), Peruvians (and Andeans) have significantly lower European admixture.

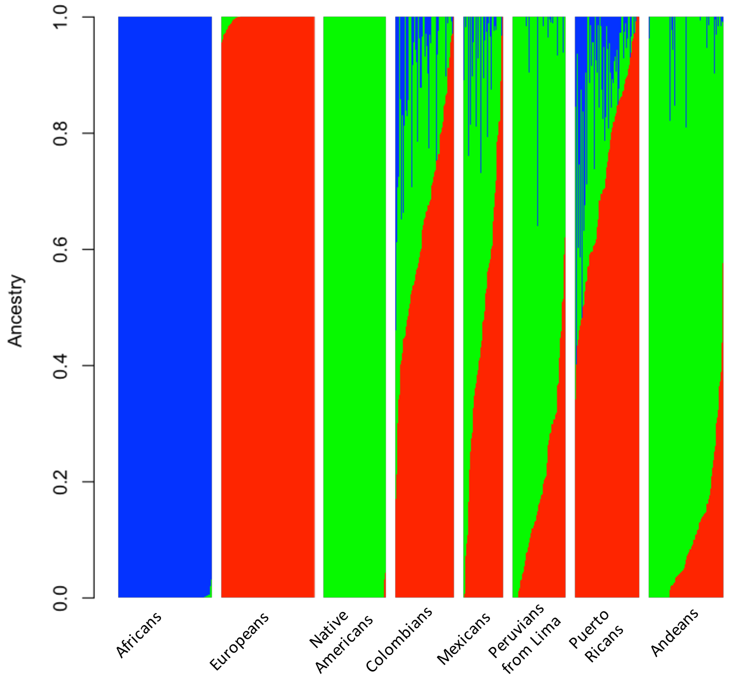

Figure 4.2: Global Admixture of American Populations: ancestry proportion in different South American populations. Blue: Africans, Red: Europeans, Green: Native Americans. Colombians, Mexicans and Puerto Ricans have significantly higher European and African ancestry as compared to Peruvians and Andeans.

4.3 Admixture Events Dating

It is a well accepted fact that South America was first inhabited by Native Americans at least 14000 years ago. Then in the 16th Century European explorers discovered the Americas in soon started to colonize the region. Soon after colonization began, the

Europeans brought Africans into the Americas through the slave trade. Considering the Native Americans as the indigenous population of the region, we estimate the dates of European and African admixture into the Andean population using the Alder [LLP$^+$13] method.

Alder method assumes the following about the population under study:

1. No non-random mating

2. No bottle-neck events

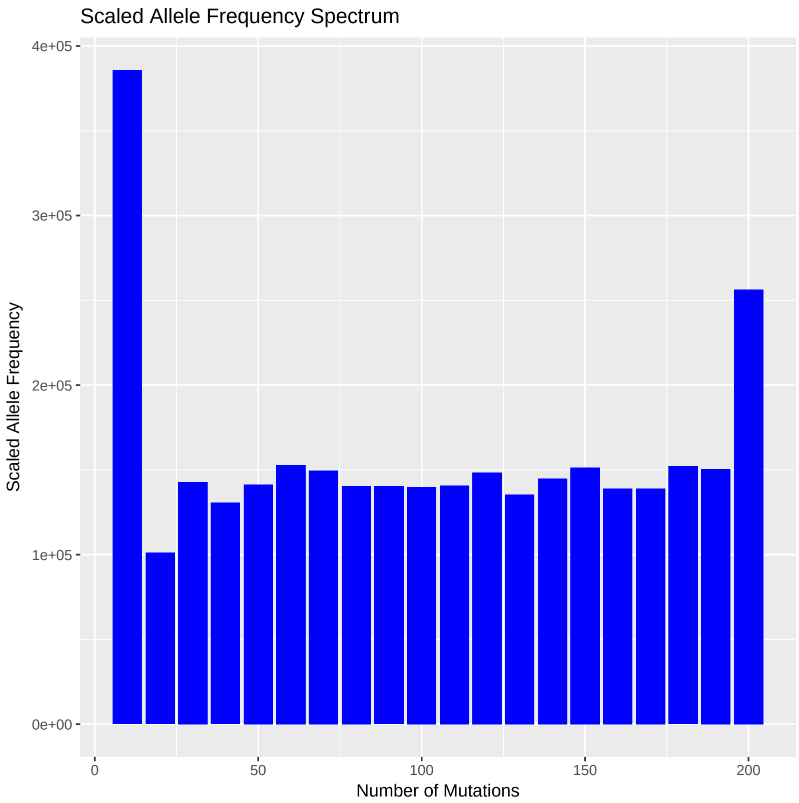

Figure 4.3: Scaled Allele Frequency Spectrum: the number of mutations at each locus and the scaled frequency of the number of mutations. Same scaled frequency values is one of the indicators of random mating.

We applied a simple test to test for random mating: the allele frequency spectrum gives the distribution of allele frequencies of a given set of SNPs in a population. Each entry in the frequency spectrum gives the number of loci with the non-reference allele frequency. We scale the frequencies by multiplying the frequency to the counts (or number of mutations) to obtain the scaled allele frequency spectrum. In case of random mating, the scaled frequencies for each count value are equal. The scaled allele frequency spectrum for the Andean population is given in Figure 4.3. Except for the count value of 1, the remaining count values have similar frequencies, there by giving a strong evidence for random mating in the population. Testing for bottle-neck events in the history of a population is beyond the scope of this work

Assuming Andean population as a three-way admixture between Native Americans, Europeans and Africans, we estimate the admixture events using Alder and compare the number with other South American populations. The results of this experiment are given in Table 4.1.

Table 4.1: Admixture Dating: Number of generations since European and African admixture events in various South American populations as obtained through Alder. The number of generations is given by the decay parameter of the exponential curve of weighted linkage distribution.

Population	Generations Since European Admixture	Generations Since African Admixture
Andean	9.21 ± 3.92	8.5 ± 3.9
Peru from Lima	9.93 ± 4.80	8.66 ± 4.43
Colombian	13.34 ± 1.61	11.72 ± 2.37
Mexican	12.92 ± 1.72	11.69 ± 2.70

Through the Alder runs we found some interesting results about the populations from Peru. Unlike other South American populations like Colombians and Mexicans with European admixture dating back 13.34 generations and 12.92 generations (400 years and 387 years assuming 1 generation to be 30 years) respectively, European admixture in Andeans and other Peruvians happened much later around 9.21 generations (276 years) back. Also the standard deviation for the Peruvian populations is much higher than other populations, which could be due to lesser number of SNPs spanning over the different datasets. While the numbers make sense for Colombians and Mexicans as the European colonization of Americas started in early 1600s, further experiments on Peruvian and Andean populations need to be done to corroborate these results. Detecting for potential bottle-neck event in the history of Peruvian populations could give us a better explanation for these numbers. Also, working with more samples collected from different regions of Peru may capture higher diversity within the Peruvian data.

4.4 Modeling the Admixture Dynamics

We investigated the admixture dynamics of the Andean population on a subset of the entire genome. Using LAMP-LD [BPS$^+$12], we found the chromosomal segments of distinct ancestry (CSDA) on Chromosomes 10, 11, 12, 19, 21 and 22, and compared our results with the ones presented by Jin et. al [JWW$^+$12]. The CSDA lengths and their frequency in the Andean population is presented in Figure 4.4. The average CSDA length in the Andean population was found to be 4.3 cM with a standard deviation of 3.8 cM. These values are much smaller than the ones reported by Jin et. al [JWW$^+$12] for the populations under Hybrid Isolation model. Although prior studies used both simulated and real data to find out the difference in average CSDA lengths of different admixture models, it is unlikely that our population will have average value smaller than 4.3 cM when run with simulated data. We therefore believe that the Andean population is following a Hybrid Isolation model where two ancestral populations (Native Americans and Europeans in this case) admixed several generations back and the resulting hybrid population is evolving in isolation without any external gene flow.

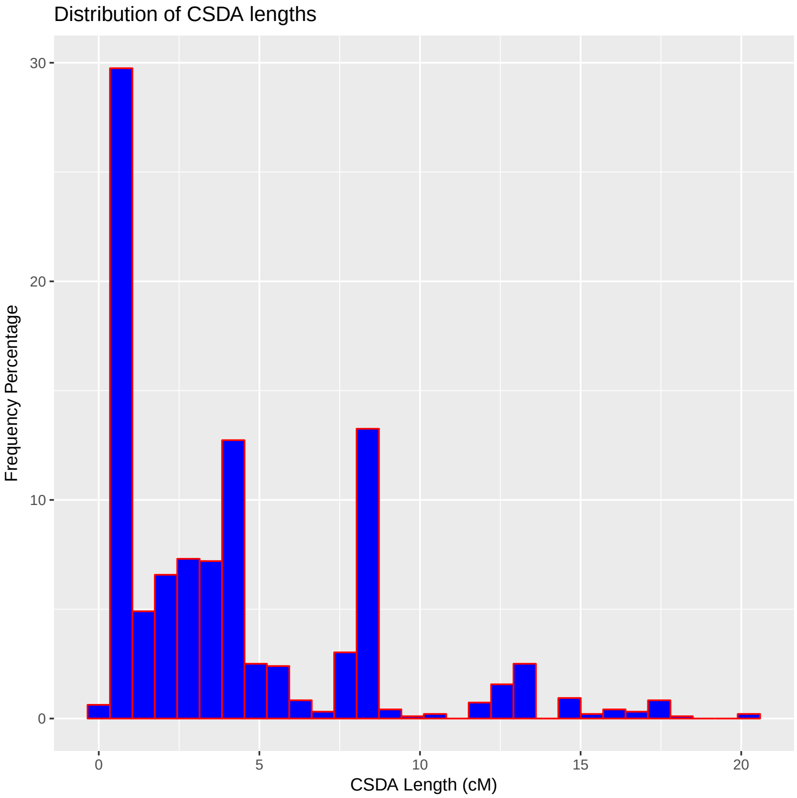

Figure 4.4: CSDA Distribution: Frequency distribution of the CSDA lengths. Average CSDA length is 4.3 cM with 3.8 cM standard deviation.

Chapter 5

Conclusion and Future Directions

South American populations have a complicated history due to multiple migration events from different populations through their history. In this thesis we analyzed the Andean highlander population from Cerro de Pasco, a high altitude town in Peru, as a three way admixture of Native Americans, Europeans and Africans. We performed extensive experiments using previously published methods to validate the admixture history of the Andean population. While most of our results align well with previously established fact, we found some interesting pieces of information that need further analysis. We summarize our findings as follows:

- The Andean population resulted from the admixture between Native Americans, Europeans and Africans. This is supported by historical records on European colonization and African slave trade.

- Unlike other South American populations like Colombians and Mexicans, Peruvians and Andeans have lesser proportions of European and African ancestry; 20% European ancestry in Peruvians compared to more than 60% in Colombians and Mexicans. Several other studies [MGW$^+$16, RSH$^+$15] presented similar results.

- Unlike other South American populations, Peruvians and Andeans witnessed European admixture much later, after almost 130 years. Interesting, no historical records corroborate this fact.
- The Andean population is evolving under a Hybrid Isolation admixture model.

We also identified some major limitations of this work which may have affected our results:

- The admixture events dates for Andeans and Peruvians obtained using Alder varied significantly as compared to other South American populations. One of the limitations of this work is that we assumed the Peruvian population to not have undergone a bottle-neck event in its history. We need to conduct further experiments to identify any such events that might be influencing the admixture dating results.
- The Peruvian data is collected from two regions: Lima and Cerro de Pasco. This does not necessarily capture the diversity of Peruvian population and more samples from different regions can give us better insights into this population.

5.1 Future Work

We found some interesting results about the admixture history of Andeans through this work, and were able to identify several opportunities to advance this project:

- Andean population is shown to have adapted to high altitude regions and undergone natural selection. Several association studies are currently under for different phenotypes. It will be beneficial to establish any ancestral links for these phenotypes to better understand their origins.

- Finally, this entire study can be formulated into a standard framework for reproducibility and application on any new population that has undergone recent admixture.

Appendix A

Native American Samples

Table A.1: 1000 Genomes Samples

Sample ID	Population	Super Population	Sex
NA06984	CEU	EUR	M
NA06985	CEU	EUR	F
NA06986	CEU	EUR	M
NA06989	CEU	EUR	F
NA06994	CEU	EUR	M
NA07000	CEU	EUR	F
NA07037	CEU	EUR	F
NA07048	CEU	EUR	M
NA07051	CEU	EUR	M
NA07056	CEU	EUR	F
NA07347	CEU	EUR	M
NA07357	CEU	EUR	M
NA10847	CEU	EUR	F
NA10851	CEU	EUR	M

NA11829	CEU	EUR	M
NA11830	CEU	EUR	F
NA11831	CEU	EUR	M
NA11832	CEU	EUR	F
NA11840	CEU	EUR	F
NA11843	CEU	EUR	M
NA11881	CEU	EUR	M
NA11892	CEU	EUR	F
NA11893	CEU	EUR	M
NA11894	CEU	EUR	F
NA11918	CEU	EUR	F
NA11919	CEU	EUR	M
NA11920	CEU	EUR	F
NA11930	CEU	EUR	M
NA11931	CEU	EUR	F
NA11932	CEU	EUR	M
NA11933	CEU	EUR	F
NA11992	CEU	EUR	M
NA11994	CEU	EUR	M
NA11995	CEU	EUR	F
NA12003	CEU	EUR	M
NA12004	CEU	EUR	F
NA12005	CEU	EUR	M
NA12006	CEU	EUR	F
NA12043	CEU	EUR	M
NA12044	CEU	EUR	F

NA12045	CEU	EUR	M
NA12046	CEU	EUR	F
NA12058	CEU	EUR	F
NA12144	CEU	EUR	M
NA12154	CEU	EUR	M
NA12155	CEU	EUR	M
NA12156	CEU	EUR	F
NA12234	CEU	EUR	F
NA12249	CEU	EUR	F
NA12272	CEU	EUR	M
NA12273	CEU	EUR	F
NA12275	CEU	EUR	F
NA12282	CEU	EUR	M
NA12283	CEU	EUR	F
NA12286	CEU	EUR	M
NA12287	CEU	EUR	F
NA12340	CEU	EUR	M
NA12341	CEU	EUR	F
NA12342	CEU	EUR	M
NA12347	CEU	EUR	M
NA12348	CEU	EUR	F
NA12383	CEU	EUR	F
NA12399	CEU	EUR	M
NA12400	CEU	EUR	F
NA12413	CEU	EUR	M
NA12414	CEU	EUR	F

NA12489	CEU	EUR	F
NA12546	CEU	EUR	M
NA12716	CEU	EUR	M
NA12717	CEU	EUR	F
NA12718	CEU	EUR	F
NA12748	CEU	EUR	M
NA12749	CEU	EUR	F
NA12750	CEU	EUR	M
NA12751	CEU	EUR	F
NA12760	CEU	EUR	M
NA12761	CEU	EUR	F
NA12762	CEU	EUR	M
NA12763	CEU	EUR	F
NA12775	CEU	EUR	M
NA12776	CEU	EUR	F
NA12777	CEU	EUR	M
NA12778	CEU	EUR	F
NA12812	CEU	EUR	M
NA12813	CEU	EUR	F
NA12814	CEU	EUR	M
NA12815	CEU	EUR	F
NA12827	CEU	EUR	M
NA12828	CEU	EUR	F
NA12829	CEU	EUR	M
NA12830	CEU	EUR	F
NA12842	CEU	EUR	M

NA12843	CEU	EUR	F
NA12872	CEU	EUR	M
NA12873	CEU	EUR	F
NA12874	CEU	EUR	M
NA12878	CEU	EUR	F
NA12889	CEU	EUR	M
NA12890	CEU	EUR	F
HG00096	GBR	EUR	M
HG00097	GBR	EUR	F
HG00099	GBR	EUR	F
HG00100	GBR	EUR	F
HG00101	GBR	EUR	M
HG00102	GBR	EUR	F
HG00103	GBR	EUR	M
HG00105	GBR	EUR	M
HG00106	GBR	EUR	F
HG00107	GBR	EUR	M
HG00108	GBR	EUR	M
HG00109	GBR	EUR	M
HG00110	GBR	EUR	F
HG00111	GBR	EUR	F
HG00112	GBR	EUR	M
HG00113	GBR	EUR	M
HG00114	GBR	EUR	M
HG00115	GBR	EUR	M
HG00116	GBR	EUR	M

HG00117	GBR	EUR	M
HG00118	GBR	EUR	F
HG00119	GBR	EUR	M
HG00120	GBR	EUR	F
HG00121	GBR	EUR	F
HG00122	GBR	EUR	F
HG00123	GBR	EUR	F
HG00125	GBR	EUR	F
HG00126	GBR	EUR	M
HG00127	GBR	EUR	F
HG00128	GBR	EUR	F
HG00129	GBR	EUR	M
HG00130	GBR	EUR	F
HG00131	GBR	EUR	M
HG00132	GBR	EUR	F
HG00133	GBR	EUR	F
HG00136	GBR	EUR	M
HG00137	GBR	EUR	F
HG00138	GBR	EUR	M
HG00139	GBR	EUR	M
HG00140	GBR	EUR	M
HG00141	GBR	EUR	M
HG00142	GBR	EUR	M
HG00143	GBR	EUR	M
HG00145	GBR	EUR	M
HG00146	GBR	EUR	F

HG00148	GBR	EUR	M
HG00149	GBR	EUR	M
HG00150	GBR	EUR	F
HG00151	GBR	EUR	M
HG00154	GBR	EUR	F
HG00155	GBR	EUR	M
NA18486	YRI	AFR	M
NA18488	YRI	AFR	F
NA18489	YRI	AFR	F
NA18498	YRI	AFR	M
NA18499	YRI	AFR	F
NA18501	YRI	AFR	M
NA18502	YRI	AFR	F
NA18504	YRI	AFR	M
NA18505	YRI	AFR	F
NA18507	YRI	AFR	M
NA18508	YRI	AFR	F
NA18510	YRI	AFR	M
NA18511	YRI	AFR	F
NA18516	YRI	AFR	M
NA18517	YRI	AFR	F
NA18519	YRI	AFR	M
NA18520	YRI	AFR	F
NA18522	YRI	AFR	M
NA18523	YRI	AFR	F
NA18853	YRI	AFR	M

NA18856	YRI	AFR	M
NA18858	YRI	AFR	F
NA18861	YRI	AFR	F
NA18864	YRI	AFR	F
NA18865	YRI	AFR	M
NA18867	YRI	AFR	F
NA18868	YRI	AFR	M
NA18870	YRI	AFR	F
NA18871	YRI	AFR	M
NA18873	YRI	AFR	F
NA18874	YRI	AFR	M
NA18876	YRI	AFR	F
NA18877	YRI	AFR	M
NA18878	YRI	AFR	F
NA18879	YRI	AFR	M
NA18881	YRI	AFR	F
NA18907	YRI	AFR	F
NA18908	YRI	AFR	M
NA18909	YRI	AFR	F
NA18910	YRI	AFR	M
NA18912	YRI	AFR	F
NA18915	YRI	AFR	M
NA18916	YRI	AFR	F
NA18917	YRI	AFR	M
NA18923	YRI	AFR	M
NA18924	YRI	AFR	F

NA18933	YRI	AFR	F
NA18934	YRI	AFR	M
NA19092	YRI	AFR	M
NA19093	YRI	AFR	F
NA19095	YRI	AFR	F
NA19096	YRI	AFR	M
NA19098	YRI	AFR	M
NA19099	YRI	AFR	F
NA19102	YRI	AFR	F
NA19107	YRI	AFR	M
NA19108	YRI	AFR	F
NA19113	YRI	AFR	M
NA19114	YRI	AFR	F
NA19116	YRI	AFR	F
NA19117	YRI	AFR	M
NA19118	YRI	AFR	F
NA19119	YRI	AFR	M
NA19121	YRI	AFR	M
NA19129	YRI	AFR	F
NA19130	YRI	AFR	M
NA19131	YRI	AFR	F
NA19137	YRI	AFR	F
NA19138	YRI	AFR	M
NA19141	YRI	AFR	M
NA19143	YRI	AFR	F
NA19144	YRI	AFR	M

NA19146	YRI	AFR	M
NA19147	YRI	AFR	F
NA19149	YRI	AFR	F
NA19152	YRI	AFR	F
NA19153	YRI	AFR	M
NA19159	YRI	AFR	F
NA19160	YRI	AFR	M
NA19171	YRI	AFR	M
NA19172	YRI	AFR	F
NA19175	YRI	AFR	M
NA19184	YRI	AFR	M
NA19185	YRI	AFR	F
NA19189	YRI	AFR	M
NA19190	YRI	AFR	F
NA19197	YRI	AFR	F
NA19198	YRI	AFR	M
NA19200	YRI	AFR	M
NA19201	YRI	AFR	F
NA19204	YRI	AFR	F
NA19206	YRI	AFR	F
NA19207	YRI	AFR	M
NA19209	YRI	AFR	F
NA19210	YRI	AFR	M
NA19213	YRI	AFR	M
NA19214	YRI	AFR	F
NA19222	YRI	AFR	F

NA19223	YRI	AFR	M
NA19225	YRI	AFR	F
NA19235	YRI	AFR	F
NA19236	YRI	AFR	M
NA19238	YRI	AFR	F
NA19239	YRI	AFR	M
NA19247	YRI	AFR	F
NA19248	YRI	AFR	M
NA19256	YRI	AFR	M
NA19257	YRI	AFR	F
HG03078	MSL	AFR	M
HG03079	MSL	AFR	F
HG03081	MSL	AFR	M
HG03082	MSL	AFR	F
HG03084	MSL	AFR	M
HG03085	MSL	AFR	F
HG03086	MSL	AFR	F
HG03088	MSL	AFR	F
HG03091	MSL	AFR	F
HG03095	MSL	AFR	F
HG03096	MSL	AFR	M
HG03097	MSL	AFR	F
HG03209	MSL	AFR	M
HG03212	MSL	AFR	F
HG03224	MSL	AFR	M
HG03225	MSL	AFR	M

HG03376	MSL	AFR	M
HG03378	MSL	AFR	F
HG03380	MSL	AFR	F
HG03382	MSL	AFR	M
HG03385	MSL	AFR	M
HG03388	MSL	AFR	M
HG03391	MSL	AFR	M
HG03394	MSL	AFR	M
HG03397	MSL	AFR	M
HG03401	MSL	AFR	F
HG03410	MSL	AFR	F
HG03419	MSL	AFR	F
HG03428	MSL	AFR	F
HG03432	MSL	AFR	M
HG03433	MSL	AFR	M
HG03436	MSL	AFR	M
HG03437	MSL	AFR	F
HG03439	MSL	AFR	M
HG03442	MSL	AFR	M
HG03445	MSL	AFR	M
HG03446	MSL	AFR	F
HG03449	MSL	AFR	F
HG03451	MSL	AFR	M
HG03452	MSL	AFR	F
HG03455	MSL	AFR	F
HG03457	MSL	AFR	M

HG01112	CLM	AMR	M
HG01113	CLM	AMR	F
HG01119	CLM	AMR	F
HG01121	CLM	AMR	M
HG01122	CLM	AMR	F
HG01124	CLM	AMR	M
HG01125	CLM	AMR	F
HG01130	CLM	AMR	M
HG01131	CLM	AMR	F
HG01133	CLM	AMR	M
HG01134	CLM	AMR	F
HG01136	CLM	AMR	M
HG01137	CLM	AMR	F
HG01139	CLM	AMR	M
HG01140	CLM	AMR	F
HG01142	CLM	AMR	M
HG01148	CLM	AMR	M
HG01149	CLM	AMR	F
HG01250	CLM	AMR	M
HG01251	CLM	AMR	F
HG01253	CLM	AMR	M
HG01254	CLM	AMR	F
HG01256	CLM	AMR	M
HG01257	CLM	AMR	F
HG01259	CLM	AMR	M
HG01260	CLM	AMR	F

HG01269	CLM	AMR	F
HG01271	CLM	AMR	M
HG01272	CLM	AMR	F
HG01275	CLM	AMR	F
HG01277	CLM	AMR	M
HG01280	CLM	AMR	M
HG01281	CLM	AMR	F
HG01284	CLM	AMR	F
HG01341	CLM	AMR	M
HG01342	CLM	AMR	F
HG01344	CLM	AMR	M
HG01345	CLM	AMR	F
HG01348	CLM	AMR	F
HG01350	CLM	AMR	M
HG01351	CLM	AMR	F
HG01353	CLM	AMR	M
HG01354	CLM	AMR	F
HG01356	CLM	AMR	M
HG01357	CLM	AMR	F
HG01359	CLM	AMR	M
HG01360	CLM	AMR	F
HG01362	CLM	AMR	M
HG01363	CLM	AMR	F
HG01365	CLM	AMR	M
HG01366	CLM	AMR	F
HG01369	CLM	AMR	F

HG01372	CLM	AMR	F
HG01374	CLM	AMR	M
HG01375	CLM	AMR	F
HG01377	CLM	AMR	M
HG01378	CLM	AMR	F
HG01383	CLM	AMR	M
HG01384	CLM	AMR	F
HG01389	CLM	AMR	M
HG01390	CLM	AMR	F
HG01431	CLM	AMR	M
HG01432	CLM	AMR	F
HG01435	CLM	AMR	F
HG01437	CLM	AMR	M
HG01438	CLM	AMR	F
HG01440	CLM	AMR	M
HG01441	CLM	AMR	F
HG01443	CLM	AMR	M
HG01444	CLM	AMR	F
HG01447	CLM	AMR	F
HG01455	CLM	AMR	M
HG01456	CLM	AMR	F
HG01459	CLM	AMR	F
HG01461	CLM	AMR	M
HG01462	CLM	AMR	F
HG01464	CLM	AMR	M
HG01465	CLM	AMR	F

HG01468	CLM	AMR	F
HG01474	CLM	AMR	F
HG01479	CLM	AMR	M
HG01485	CLM	AMR	M
HG01486	CLM	AMR	F
HG01488	CLM	AMR	M
HG01489	CLM	AMR	F
HG01491	CLM	AMR	M
HG01492	CLM	AMR	F
HG01494	CLM	AMR	M
HG01495	CLM	AMR	F
HG01497	CLM	AMR	M
HG01498	CLM	AMR	F
HG01550	CLM	AMR	M
HG01551	CLM	AMR	F
HG01556	CLM	AMR	M
NA19648	MXL	AMR	F
NA19649	MXL	AMR	M
NA19651	MXL	AMR	F
NA19652	MXL	AMR	M
NA19654	MXL	AMR	F
NA19655	MXL	AMR	M
NA19657	MXL	AMR	F
NA19658	MXL	AMR	M
NA19661	MXL	AMR	M
NA19663	MXL	AMR	F

NA19664	MXL	AMR	M
NA19669	MXL	AMR	F
NA19670	MXL	AMR	M
NA19676	MXL	AMR	M
NA19678	MXL	AMR	F
NA19679	MXL	AMR	M
NA19681	MXL	AMR	F
NA19682	MXL	AMR	M
NA19684	MXL	AMR	F
NA19716	MXL	AMR	F
NA19717	MXL	AMR	M
NA19719	MXL	AMR	F
NA19720	MXL	AMR	M
NA19722	MXL	AMR	F
NA19723	MXL	AMR	M
NA19725	MXL	AMR	F
NA19726	MXL	AMR	M
NA19728	MXL	AMR	F
NA19729	MXL	AMR	M
NA19731	MXL	AMR	F
NA19732	MXL	AMR	M
NA19734	MXL	AMR	F
NA19735	MXL	AMR	M
NA19740	MXL	AMR	F
NA19741	MXL	AMR	M
NA19746	MXL	AMR	F

NA19747	MXL	AMR	M
NA19749	MXL	AMR	F
NA19750	MXL	AMR	M
NA19752	MXL	AMR	F
NA19755	MXL	AMR	F
NA19756	MXL	AMR	M
NA19758	MXL	AMR	F
NA19759	MXL	AMR	M
NA19761	MXL	AMR	F
NA19762	MXL	AMR	M
NA19764	MXL	AMR	F
NA19770	MXL	AMR	F
NA19771	MXL	AMR	M
NA19773	MXL	AMR	F
NA19774	MXL	AMR	M
NA19776	MXL	AMR	F
NA19777	MXL	AMR	M
NA19779	MXL	AMR	F
NA19780	MXL	AMR	M
NA19782	MXL	AMR	F
NA19783	MXL	AMR	M
NA19785	MXL	AMR	F
NA19786	MXL	AMR	M
NA19788	MXL	AMR	F
NA19789	MXL	AMR	M
NA19792	MXL	AMR	M

NA19794	MXL	AMR	F
NA19795	MXL	AMR	M
HG01565	PEL	AMR	M
HG01566	PEL	AMR	F
HG01571	PEL	AMR	M
HG01572	PEL	AMR	F
HG01577	PEL	AMR	M
HG01578	PEL	AMR	F
HG01892	PEL	AMR	M
HG01893	PEL	AMR	F
HG01917	PEL	AMR	M
HG01918	PEL	AMR	F
HG01920	PEL	AMR	M
HG01921	PEL	AMR	F
HG01923	PEL	AMR	M
HG01924	PEL	AMR	F
HG01926	PEL	AMR	M
HG01927	PEL	AMR	F
HG01932	PEL	AMR	M
HG01933	PEL	AMR	F
HG01935	PEL	AMR	M
HG01936	PEL	AMR	F
HG01938	PEL	AMR	M
HG01939	PEL	AMR	F
HG01941	PEL	AMR	M
HG01942	PEL	AMR	F

HG01944	PEL	AMR	M
HG01945	PEL	AMR	F
HG01947	PEL	AMR	M
HG01948	PEL	AMR	F
HG01950	PEL	AMR	M
HG01951	PEL	AMR	F
HG01953	PEL	AMR	M
HG01954	PEL	AMR	F
HG01961	PEL	AMR	M
HG01965	PEL	AMR	F
HG01967	PEL	AMR	M
HG01968	PEL	AMR	F
HG01970	PEL	AMR	M
HG01971	PEL	AMR	F
HG01973	PEL	AMR	F
HG01974	PEL	AMR	M
HG01976	PEL	AMR	F
HG01977	PEL	AMR	M
HG01979	PEL	AMR	M
HG01980	PEL	AMR	F
HG01982	PEL	AMR	M
HG01991	PEL	AMR	M
HG01992	PEL	AMR	F
HG01997	PEL	AMR	F
HG02002	PEL	AMR	M
HG02003	PEL	AMR	F

HG02006	PEL	AMR	F
HG02008	PEL	AMR	M
HG02089	PEL	AMR	F
HG02090	PEL	AMR	M
HG02102	PEL	AMR	F
HG02104	PEL	AMR	M
HG02105	PEL	AMR	F
HG02146	PEL	AMR	M
HG02147	PEL	AMR	F
HG02150	PEL	AMR	M
HG02252	PEL	AMR	F
HG02253	PEL	AMR	M
HG02259	PEL	AMR	M
HG02260	PEL	AMR	F
HG02262	PEL	AMR	M
HG02265	PEL	AMR	M
HG02266	PEL	AMR	F
HG02271	PEL	AMR	M
HG02272	PEL	AMR	F
HG02274	PEL	AMR	M
HG02275	PEL	AMR	F
HG02277	PEL	AMR	M
HG02278	PEL	AMR	F
HG02285	PEL	AMR	M
HG02286	PEL	AMR	F
HG02291	PEL	AMR	M

HG02292	PEL	AMR	F
HG02298	PEL	AMR	F
HG02299	PEL	AMR	M
HG02301	PEL	AMR	F
HG02304	PEL	AMR	M
HG02312	PEL	AMR	F
HG02345	PEL	AMR	F
HG02348	PEL	AMR	F
HG02425	PEL	AMR	F
HG00551	PUR	AMR	F
HG00553	PUR	AMR	M
HG00554	PUR	AMR	F
HG00637	PUR	AMR	M
HG00638	PUR	AMR	F
HG00640	PUR	AMR	M
HG00641	PUR	AMR	F
HG00731	PUR	AMR	M
HG00732	PUR	AMR	F
HG00734	PUR	AMR	F
HG00736	PUR	AMR	M
HG00737	PUR	AMR	F
HG00739	PUR	AMR	M
HG00740	PUR	AMR	F
HG00742	PUR	AMR	M
HG00743	PUR	AMR	F
HG01047	PUR	AMR	M

HG01048	PUR	AMR	M
HG01049	PUR	AMR	F
HG01051	PUR	AMR	M
HG01052	PUR	AMR	F
HG01054	PUR	AMR	M
HG01055	PUR	AMR	F
HG01058	PUR	AMR	F
HG01060	PUR	AMR	M
HG01061	PUR	AMR	F
HG01063	PUR	AMR	M
HG01064	PUR	AMR	F
HG01066	PUR	AMR	M
HG01067	PUR	AMR	F
HG01069	PUR	AMR	M
HG01070	PUR	AMR	F
HG01072	PUR	AMR	M
HG01073	PUR	AMR	F
HG01075	PUR	AMR	M
HG01077	PUR	AMR	F
HG01079	PUR	AMR	M
HG01080	PUR	AMR	F
HG01082	PUR	AMR	M
HG01083	PUR	AMR	F
HG01085	PUR	AMR	M
HG01086	PUR	AMR	F
HG01088	PUR	AMR	M

HG01089	PUR	AMR	F
HG01092	PUR	AMR	F
HG01094	PUR	AMR	M
HG01095	PUR	AMR	F
HG01097	PUR	AMR	M
HG01098	PUR	AMR	F
HG01101	PUR	AMR	M
HG01102	PUR	AMR	F
HG01104	PUR	AMR	M
HG01105	PUR	AMR	F
HG01107	PUR	AMR	M
HG01108	PUR	AMR	F
HG01110	PUR	AMR	M
HG01111	PUR	AMR	F
HG01161	PUR	AMR	M
HG01162	PUR	AMR	F
HG01164	PUR	AMR	M
HG01167	PUR	AMR	M
HG01168	PUR	AMR	F
HG01170	PUR	AMR	M
HG01171	PUR	AMR	F
HG01173	PUR	AMR	M
HG01174	PUR	AMR	F
HG01176	PUR	AMR	M
HG01177	PUR	AMR	F
HG01182	PUR	AMR	M

HG01183	PUR	AMR	F
HG01187	PUR	AMR	M
HG01188	PUR	AMR	F
HG01190	PUR	AMR	M
HG01191	PUR	AMR	F
HG01197	PUR	AMR	M
HG01198	PUR	AMR	F
HG01200	PUR	AMR	M
HG01204	PUR	AMR	M
HG01205	PUR	AMR	F
HG01241	PUR	AMR	M
HG01242	PUR	AMR	F
HG01247	PUR	AMR	M
HG01248	PUR	AMR	F
HG01286	PUR	AMR	M
HG01302	PUR	AMR	M
HG01303	PUR	AMR	F
HG01305	PUR	AMR	M
HG01308	PUR	AMR	M
HG01311	PUR	AMR	M
HG01312	PUR	AMR	F
HG01323	PUR	AMR	F
HG01325	PUR	AMR	M
HG01326	PUR	AMR	F
HG01392	PUR	AMR	M
HG01393	PUR	AMR	F

HG01395	PUR	AMR	M
HG01396	PUR	AMR	F
HG01398	PUR	AMR	M
HG01402	PUR	AMR	M
HG01403	PUR	AMR	F
HG01405	PUR	AMR	M
HG01412	PUR	AMR	M
HG01413	PUR	AMR	M
HG01414	PUR	AMR	F

Appendix B

Native American Samples

Table B.1: Native American Genomes Samples

Sample ID	Sex	Group	% Genome Masked
4249815035_A	M	Pima	0.00
4249815052_A	F	Pima	0.70
4249815114_A	M	Pima	3.00
4249815287_A	M	Quechua	4.30
4249815288_A	M	Quechua	2.10
4254930178_A	M	Pima	0.60
4254930244_A	F	Pima	2.60
4254930270_A	M	Pima	0.90
4254930496_A	M	Quechua	5.00
4254930595_A	F	Pima	0.20
4256126001_A	M	Surui	0.00
4256126002_A	F	Surui	0.00
4256126004_A	M	Surui	0.00

4256126007_A	M	Surui	0.00
4256126036_A	F	Surui	0.00
4256126086_A	M	Surui	0.00
4256126171_A	F	Surui	0.00
4256126172_A	M	Surui	0.00
4256126173_A	F	Surui	0.00
4256126183_A	M	Surui	0.00
4256126202_A	M	Surui	0.00
4256126311_A	F	Surui	0.00
eastGreenland17	F	EastGreenland	0.50
eastGreenland3	F	EastGreenland	1.40
eastGreenland7	F	EastGreenland	0.40
HGDP00702	F	Piapoco	0.20
HGDP00704	F	Piapoco	0.20
HGDP00706	F	Piapoco	0.00
HGDP00708	F	Piapoco	0.00
HGDP00710	M	Piapoco	0.30
HGDP00832	F	Surui	0.00
HGDP00837	M	Surui	0.00
HGDP00838	F	Surui	0.00
HGDP00843	M	Surui	0.00
HGDP00845	M	Surui	0.00
HGDP00846	F	Surui	0.00
HGDP00849	M	Surui	0.00
HGDP00852	F	Surui	0.00
HGDP00855	F	Maya1	4.20

HGDP00857	F	Maya1	5.40
HGDP00970	F	Piapoco	0.60
HGDP00995	F	Karitiana	0.00
HGDP00998	M	Karitiana	0.10
HGDP00999	F	Karitiana	0.10
Maya_4032_041732	M	Maya1	1.90
PT-8ZVD	F	Hulliche	4.50
PT-8ZVJ	M	Palikur	0.50
PT-8ZVK	M	Palikur	0.10
PT-8ZVS	F	Zapotec1	0.20
PT-8ZVZ	M	Zapotec1	2.40
PT-911I	M	Chipewyan	0.20
PT-912N	F	Mixtec	7.60
PT-912T	F	Mixe	1.70
PT-912U	F	Mixe	3.00
PT-912W	M	Mixe	0.30
PT-912Z	M	Mixe	0.70
PT-9131	F	Mixe	9.60
PT-9133	F	Mixe	1.60
PT-9134	F	Mixe	2.00
PT-9135	F	Mixe	1.50
PT-9136	M	Mixe	0.50
PT-9137	M	Mixe	2.60
PT-913B	F	Mixe	2.00
PT-9172	F	Kaqchikel	2.30
PT-9176	F	Kaqchikel	4.40

PT-917E	M	Guaymi	1.10
PT-917F	M	Guaymi	0.00
PT-917G	M	Guaymi	0.20
PT-918H	M	Cabecar	0.00
PT-918I	M	Cabecar	1.00
PT-918L	F	Cabecar	0.00
PT-918M	F	Cabecar	17.00
PT-918N	F	Cabecar	0.00
PT-918O	M	Teribe	3.30
PT-918P	M	Teribe	0.10
PT-918Q	M	Teribe	0.10
PT-918U	F	Bribri	0.00
PT-918W	F	Bribri	9.60
PT-918X	F	Bribri	0.00
PT-9198	F	Maleku	0.20
PT-9199	M	Maleku	0.20
PT-91CY	F	Ticuna	4.70
PT-91CZ	M	Ticuna	0.20
PT-91D6	M	Kogi	0.00
PT-91D9	M	Embera	1.10
PT-91DA	M	Embera	0.50
PT-91DC	M	Embera	0.10
PT-91DH	M	Waunana	0.10
PT-91DI	M	Waunana	0.40
PT-91YT	M	Aymara	4.00
PT-91YV	M	Aymara	5.50

PT-91Z2	M	Aymara	5.10
PT-91Z5	M	Aymara	4.20
PT-91ZG	M	Quechua	5.40
PT-9GS8	M	Wayuu	8.10
PT-GLH7	M	Guarani	3.90
Tepehuano_10000_102700	F	Tepehuano	5.30
Tepehuano_10028_102728	F	Tepehuano	5.20
Zapotec_20016_201516	M	Zapotec2	5.50
Zapotec_20040_202540	F	Zapotec2	4.90

Bibliography

[ANL09] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[BPS$^+$12] Yael Baran, Bogdan Pasaniuc, Sriram Sankararaman, Dara G Torgerson, Christopher Gignoux, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G Ford, Pedro C Avila, et al. Fast and accurate inference of local ancestry in latino populations. *Bioinformatics*, 28(10):1359–1367, 2012.

[DMZ12] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181, 2012.

[HDM09] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 2009.

[HFSP09] Melissa J Hubisz, Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inferring weak population structure with the assistance of sample group information. *Molecular ecology resources*, 9(5):1322–1332, 2009.

[JWW$^+$12] Wenfei Jin, Sijia Wang, Haifeng Wang, Li Jin, and Shuhua Xu. Exploring population admixture dynamics via empirical and simulated genome-wide distribution of ancestral chromosomal segments. *The American Journal of Human Genetics*, 91(5):849–862, 2012.

[KMF$^+$08] Augustine Kong, Gisli Masson, Michael L Frigge, Arnaldur Gylfason, Pasha Zusmanovich, Gudmar Thorleifsson, Pall I Olason, Andres Ingason, Stacy Steinberg, Thorunn Rafnar, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068–1075, 2008.

[LD09] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[LLP+13] Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K Pickrell, David Reich, and Bonnie Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4):1233–1254, 2013.

[LWD+10] Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

[MGW+16] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *bioRxiv*, page 070797, 2016.

[Pea01] K Peason. On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[PNTB+07] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[PPP+06] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[PSKH09] Bogdan Paşaniuc, Sriram Sankararaman, Gad Kimmel, and Eran Halperin. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):i213–i221, 2009.

[PTP+09] Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, 2009.

[RLAAA+14] Andrés Ruiz-Linares, Kaustubh Adhikari, Victor Acuña-Alonzo, Mirsha Quinto-Sanchez, Claudia Jaramillo, William Arias, Macarena Fuentes, María Pizarro, Paola Everardo, Francisco de Avila, et al. Admixture in latin america: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet*, 10(9):e1004572, 2014.

[RPC+12] David Reich, Nick Patterson, Desmond Campbell, Arti Tandon, Stéphane Mazieres, Nicolas Ray, Maria V Parra, Winston Rojas, Constanza Duque, Natalia Mesa, et al. Reconstructing native american population history. *Nature*, 488(7411):370–374, 2012.

[RSH+15] Maanasa Raghavan, Matthias Steinrücken, Kelley Harris, Stephan Schiffels, Simon Rasmussen, Michael DeGiorgio, Anders Albrechtsen, Cristina Valdiosera, María C Ávila-Arcos, Anna-Sapfo Malaspinas, et al. Genomic evidence for the pleistocene and recent population history of native americans. *Science*, 349(6250):aab3884, 2015.

[Wal16] Denise Waldron. Population genomics: Genomic analysis of south american ancestry. *Nature Reviews Genetics*, 17(2):66–66, 2016.

[ZAL11] Hua Zhou, David Alexander, and Kenneth Lange. A quasi-newton acceleration for high-dimensional optimization algorithms. *Statistics and computing*, 21(2):261–273, 2011.

[ZUR+13] Dan Zhou, Nitin Udpa, Roy Ronen, Tsering Stobdan, Junbin Liang, Otto Appenzeller, Huiwen W Zhao, Yi Yin, Yuanping Du, Lixia Guo, et al. Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in andean highlanders. *The American Journal of Human Genetics*, 93(3):452–462, 2013.