COMPUTATIONAL METHODS IN POPULATION GENETICS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Aritra Bose

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Petros Drineas, Co-chair

Department of Computer Science

Dr. Peristera Paschou, Co-chair

Department of Biological Sciences

Dr. Alex Pothen, Committee Member

Department of Computer Science

Dr. Hemanta K. Maji, Committee Member

Department of Computer Science

Approved by:

Dr. Christopher W. Clifton

Graduate Committee Chair, Department of Computer Science

To *Amma*

ACKNOWLEDGMENTS

This journey would not have been possible without the help and support of many people. First and foremost, I would like to thank my advisor, Prof. Petros Drineas for his continuous technical, logistical and personal support throughout the journey. I am deeply indebted to Petros for introducing me to the wonderful interdisciplinary world of population genetics, for trusting in my abilities and showing a great deal of patience. Over the years, he has inspired me to think independently about research problems and provided valuable critique while instilling a sense of optimism and confidence in me. He took care of all logistical issues and resources by providing generous funding [1] throughout my years at Purdue. His advice has helped me to become a better researcher and most importantly, a better person. I feel honored to have been advised by him and hope to continue this wonderful relationship.

In almost every respect, Prof. Peristera Paschou has served as a second advisor on this dissertation. She has taught me the art of scientific writing and has helped me with a lot of background in genetics which I lacked. She has patiently discussed ideas, edited write-ups, rehearsed talks and helped me network with other peers and collaborators in genetics conferences. I will always be grateful to her for mentoring me and providing resources in the Department of Biological Sciences. I also want to thank members of Prof. Drineas and Prof. Paschou's group for helping me through different stages of this dissertation.

I want to thank my two other committee members, Prof. Alex Pothen and Prof. Hemanta K. Maji. Prof. Pothen has discussed some key ideas in this dissertation and has nudged me into the right direction whenever I sought help. Prof. Maji enriched

[1] The research in this dissertation was supported by NSF grants IIS-1661756, IIS-1661760 and IIS-1715202.

my understanding of the topics in this dissertation by asking the most insightful questions.

Apart from my committee, I want to thank Dr. Laxmi Parida for hosting me for three consecutive summers at the Computational Biology center in IBM T.J. Watson Research Center, Yorktown Heights, NY. I would also like to thank Dr. Daniel Platt for endless discussions about population genetics and beyond. Dan has been an inspirational figure throughout the past four years that I have known him. His ideas are instrumental to two chapters of this dissertation. Being mentored by Laxmi and Dan had been one of the most fulfilling experiences.

I thank the administrative staff in the Computer Science department for their guidance regarding the process. The folks at Purdue Research Computing also deserves a special mention. They played a key role in maintaining `Brown` and `Snyder` clusters on which most of my code was implemented. They were very helpful with the various tricky issues I have faced when running and installing software in the clusters.

We all need support systems beyond academics in order to push through the ups and downs of the graduate student life. I want to acknowledge that without my parents this wouldn't have been possible. Long before I started this, they showed utmost belief in me and supported me throughout. Their encouragement and optimism in toughest of situations helped me get through. Growing up in a middle-class background, *Maa* and *Babaiya* provided beyond their wits to ensure that I achieve what I seek. For that, and everything else, I will be forever grateful. To say that I would be nowhere without their constant support is a huge understatement. I wish my grandparents were there to see this dissertation come into life. Their influence on my life is immense and there is not a passing day, that I spend without missing them. I dedicate this thesis to *Amma* and my two *Dadubhai*s. It is their blessing that has helped me push through.

TABLE OF CONTENTS

Page

LIST OF TABLES x

LIST OF FIGURES xii

ABBREVIATIONS xxii

ABSTRACT xxiii

1 INTRODUCTION 1

2 BACKGROUND 5

2.1 Genetics Fundamentals 5

2.1.1 Ploidy 5

2.1.2 Single Nucleotide Polymorphisms 6

2.1.3 Human Genetic Variation 6

2.1.4 Population Structure 7

2.1.5 Tests for Admixture 10

2.1.6 Linkage Disequilibrium 14

2.1.7 Ancestral Recombination Graph 15

2.1.8 Association Studies 16

3 RECONSTRUCTING GENETIC POPULATION HISTORY 19

3.1 Genetics of the Peloponnesean Populations 19

3.1.1 Introduction 19

3.1.2 Materials and Methods 21

3.1.3 Results 24

3.1.4 Discussion 32

Page

3.2 Integrating Linguistics, Social Structure and Geography to Model Gene Flow in India . . . 35
3.2.1 Introduction . . . 35
3.2.2 Materials and Methods . . . 37
3.2.3 Implementation . . . 49
3.2.4 Results . . . 49
3.2.5 Discussion . . . 63

4 TERAPCA: A FAST AND SCALABLE SOFTWARE PACKAGE TO STUDY GENETIC VARIATION IN TERA-SCALE GENOTYPES . . . 66
4.1 Introduction . . . 66
4.2 Materials and Methods . . . 69
4.2.1 Simulated Datasets . . . 69
4.2.2 Real Datasets . . . 69
4.2.3 TeraPCA . . . 70
4.2.4 Implementation . . . 76
4.3 Results . . . 76
4.3.1 Synthetic Datasets . . . 77
4.3.2 Real Datasets . . . 78
4.3.3 Multithreading . . . 80
4.4 Discussion . . . 82

5 SSIMRA: MULTIPLE LOCI SELECTION WITH MULTIWAY EPISTASIS IN COALESCENCE WITH RECOMBINATIONS . . . 85
5.1 Introduction . . . 85
5.2 Materials and Methods . . . 87
5.2.1 Modeling Multiple Loci Selection with Multiway Epistasis . . . 87
5.2.2 Backward Simulator Model . . . 88
5.2.3 Forward Simulator Model . . . 93
5.3 Results . . . 99

Page

5.3.1 Implementation 99
5.3.2 Comparison Study 99
5.4 Discussion 102

6 STRUCTURE INFORMED CLUSTERING FOR POPULATION STRATIFICATION AND GENETIC RISK PREDICTION 103
6.1 Introduction 103
6.2 Materials and Methods 106
6.2.1 Simulated Datasets 106
6.2.2 Cochran-Armitage trend χ^2 108
6.2.3 EIGENSTRAT 108
6.2.4 CluStrat 109
6.3 Results 115
6.3.1 BN model 116
6.3.2 PSD model 117
6.3.3 TGP model 119
6.4 Discussion 121

7 CONCLUSION AND FUTURE WORK 126

REFERENCES 129

Appendix A: Supplementary Material for Chapter 3 151
A.1 Genetics of the Peloponnesean Populations 151
A.1.1 Supplementary Information 151
A.2 Integrating linguistics, social structure and geography to model gene flow in India 157
A.2.1 Supplementary Information 157

Appendix B: Supplementary Material for Chapter 4 175
A.3 Supplementary Information 175

Appendix C: Supplementary Material for Chapter 5 180

Page

A.4 Supplementary Information 180

VITA 184

LIST OF TABLES

Table Page

3.1 Correlations between geographic coordinates and principal components . . 25

3.2 Shared ancestry between Peloponnesean populations and Slavic, Italian and other European populations. (The first number for each pair of populations indicates the average shared ancestry for values of K between 4 and 8, while the number in parenthesis indicates the standard deviation) 30

3.3 Shared ancestry between the populations of Mani and Tsakonia and Slavic, Italian and other European populations. (The first number for each pair of populations indicates the average shared ancestry for values of K between 4 and 8, while the number in parenthesis indicates the standard deviation) 31

3.4 Top ten significant ethnic groups in India capturing the genetic structure of the subcontinent as reflected by the RLS statistic (* Vysyas are classified as in between SGA and SGB [81]). 53

3.5 $f3(C; A, B)$ tests highlighting the Steppe and Dravidian mixture in Meghawal and the negative f_3 values and reasonably significant z-scores. This confirms the South India to Gujarat direction of gene flow. Steppe_MLBA: Middle to Late Bronze Age samples from the Steppes [42] 57

4.1 Data sets on which TeraPCA was evaluated (simulated and real) 69

4.2 Wall-clock running times comparisons for the datasets of Table 4.1 using a single thread and 2 GBs of system memory * indicates no convergence after 50 hrs. 80

5.1 Example with three loci under selection and all the possible different epistasis, whether explicitly specified or simply neutral. All the user-specified values are shown in red. The *back-sSimRA* algorithm uses the effective population size as shown here. 90

A.1 Districts of origin of the subjects 151

Table Page

A.2 Top 10% of the significant f_3 statistics ($f_3(C; A, B)$) highlighting the most admixed populations in India. Gounders, Manipuri Brahmins, Tharus and Gonds are the most admixed among all tribes in India. 174

A.3 Accuracy of the ten leading eigenvalues computed by TeraPCA and FlashPCA2. 177

A.4 K-S test statistics with corresponding p-values showing that the probability distributions of H as returned by *fwd-sSimRA* and *back-sSimRA* abstracts each other very closely. 180

LIST OF FIGURES

Figure Page

2.1 A. A population phylogeny with branches corresponding to F_2 (green), F_3 (yellow), and F_4 (blue); B. An admixture graph extends a population phylogeny by allowing gene flow (red, solid line) and admixture events (red, dotted line). 12

2.2 ARG of four populations with three lineages (red, blue and green) showing recombination (nodes splitting into two) and coalescence (nodes merging into one). 15

3.1 Substructure of the Peloponnesean populations. (a) PCA analysis without the Maniot and Tsakones populations showing a partial separation of the population of Laconia. (b) PCA illustrating the separation of Peloponneseans in three groups. On the left is placed the population of Tsakones (north: open circles, south: green dots). On the right are placed the populations of Maniots (Deep Mani, East and West Tayetos.). All the remaining Peloponneseans are clustered in the center. (c) Map of Peloponnese showing the populations studied. Each dot corresponds to the origin of a participant. (d)ADMIXTURE analysis. Notice the distinct structure of the Maniots and the Tsakones and their clear cut separation from all other Peloponneseans in all values of K. 22

3.2 Genetic similarity of Peloponneseans and Europeans showing differentiation from Slavs. (a) Network analysis illustrating the high connectivity between the Peloponnesean populations as well as between the Peloponneseans, the Sicilians and the Italians. Notice the distance between Peloponneseans and the Slavic, and Near Eastern populations. Peloponneseans are connected with the Near Eastern populations through Crete and Dodecanese. (b) Notice the north to south distribution of the populations and that the Peloponneseans are placed to the far left of the graph, overlapping with the Sicilians and distinct from the Slavs (on the right side). 26

Figure Page

3.3 Testing the theory of replacement of medieval Peloponesians by Slavs and Asia Minor settlers. (a) PCA analysis shows the broad separation of Peloponneseans from four populations of the Slavic homeland (Ukrainians, Polish, Russians and Belarusians). (b) PCA comparisons of the Peloponneseans with three Greek-speaking Asia Minor populations shows only partial overlap with the population of the Asia Minor Aegean coast. (c) ADMIXTURE analysis illustrates the wide separation of Peloponneseans from the Slavs in all values of K. 27

3.4 A. Map of India showing the locations of the 835 Indian samples, from 84 well-defined population groups, that were used as the starting point of this study; B. PCA plot of the normalized dataset consisting of 368 individuals, genotyped on 48,373 SNPs shows language groups are clearly significant in the PCA plot and correlate well with the principal components; C. Framework of our approach for Correlation Optimization of Genetics and Geodemographics (COGG). 48

3.5 Population network analysis of all Indian populations reveals four isolated clusters, representing language groups (40% of edges are shown). 54

3.6 A. TreeMix plot with the number of migration edges set to five indicate that the Siberians and Mongols show the most drift from DR_SGA and SGBs (residual plot in Figure A.25). Migration from Uygurs to the Northwestern Frontier populations is also found, making these populations a gateway to the Indian populations; B. Networks formed using the top five PCs (see Methods for the network formation algorithm) and five NNs showing three major paths leading to the two entry points of India; C. Meta-analysis of the ADMIXTURE plot (Figure A.26) quantifies the ADMIXTURE results (darker colors indicate higher pairwise shared ancestry). 61

Figure Page

3.7 Outgroup $f_3(YRI; X, Y)$ gradient map, showing pie charts of the shared affinity between Indian populations (denoted by X) and Eurasian/East Asian populations (denoted by Y). The color coding scheme is represented in the right hand side, signifying the colors attributed to perfect affinity (purple for AA, red for DR, green for IE, and blue for TB). The colors are distributed across gradients with respect to the maximum and minimum significant f_3 values. The population annotations and the detailed f_3 statistics can be found in the supplement (Supplementary Table 7). This gradient map shows the Europeans having more shared genetic drift from the outgroup YRI with the IE speakers of India (specifically, IE_SGA), whereas the East Asians have the maximum shared genetic affinity with TB_SGC. 63

4.1 Projection of the samples of the 1000 Genomes dataset on the top two left singular vectors (PC1 and PC2), as computed by TeraPCA. 77

4.2 Entry-wise relative error of the top ten leading eigenvectors returned by TeraPCA for the HGDP dataset, compared to the eigenvectors returned by LAPACK. The y-axis shows the relative error; recall that each eigenvector has 1,043 entries. We observe that the relative error is roughly the same for each entry of a specific eigenvector. 79

4.3 Speedup of TeraPCA over single-threaded execution. 81

5.1 Outline of the main steps of the forward model. (a)Schematic diagram for simulating the "book of populations" which closely resembles the biological process of evolution. (b) Tracing the ARG from the book of populations (example ARG outlined in red). 95

5.2 Comparing the height of the ARG (H) between the *fwd-sSimRa* and *back-sSimRA* for selection at two-loci with and without epistasis, respectively. We set $g = 25K$, $r = 1.0 \times 10^{-8}$, $N = 100$, $s = \{0.3, 0.3, 0.3\}$, $e_s = \{0, 0.1\}$ and $m = \{10, 20, 30, 40\}$. (i) The box-and-whisker plot summarizes the result for each m. On each box, the central mark is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. (ii) Q-Q plots for each m showing that the distributions of H from *fwd-sSimRa* and *back-sSimRA* agrees (iii) CDFs of *fwd-sSimRa* and *back-sSimRA* also follow each other closely, reconfirming the agreement between them. 100

Figure Page

5.3 Comparison on the height of the ARG (H)for different s1 values in the case of no recombination for $g = 1000$, $N = 400$ and $m = \{20, 50, 80, 120, 150, 200, 250\}$ 101

6.1 Projection of the samples from three populations simulated from BN model on the top two axes of variation. 115

6.2 Box plots for spurious and causal associations on the BN model shows that Armitage trend χ^2 has the maximum number of spurious associations containing about 4-5 causal SNPs whereas EIGENSTRAT has minimum number of spurious associations while detecting almost zero causal SNPs. CluStrat has more spurious associations than EIGENSTRAT and considerably less than Armitage trend χ^2 recovering slightly more number of causal SNPs than the latter. 116

6.3 Projection of the samples from PSD model with varying sets of values of α. We observe that increasing α increases the density between individuals leading to admixture and creates a uniform gradient as all values of α_i are equal. 118

6.4 Box plots for spurious and causal associations on the PSD model ($\alpha = \{0.1, 0.1, 0.1\}$) shows Armitage trend χ^2 has maximum number of spurious associations containing less causal SNPs than the BN model (Figure 6.2) owing to the admixed nature of the individuals in PSD. EIGENSTRAT has minimum number of spurious associations while detecting almost zero causal SNPs. CluStrat has more spurious associations than EIGENSTRAT and less than Armitage trend χ^2 recovering two to three fold more causal SNPs. 119

6.5 Box plots for spurious and causal associations on the TGP model shows Armitage trend χ^2 has the maximum number of spurious associations containing less causal SNPs than both the BN and PSD model (Figure 6.2 and 6.4) owing to the distributions of admixed samples across the world of the individuals. CluStrat outperforms both the methods in this scenario as it has the minimum number of spurious associations as well as the highest number of causal SNPs. 120

6.6 Dendrograms obtained after running AHC with Ward's linkage on PSD model ($\alpha = \{0.1, 0.1, 0.1\}$) shows Mahalanobis distance shows fine grained interactions between the individuals inside a cluster recovering population substructure and cryptic relatedness which Euclidean distance based GRM fails to recover. 123

Figure Page

6.7 Box plots for spurious and causal associations obtained by running AHC with Mahalanobis and Euclidean distances on the PSD model ($\alpha = \{0.1, 0.1, 0.1\}$). We observe similar performance on both the distance metrics in terms of identifying true causal variants. Mahalanobis distance discovers less spurious associations than Euclidean distance. 124

A.1 Locations of the populations listed in Supplementary Table 1 152

A.2 Testing the hypothesis of Armenian ancestry of Peloponneseans. Fallmerayer proposed that Armenians were among the medieval populations moved to Peloponnese by the Byzantines. Comparison of Peloponneseans with the Armenians by, (a) PCA analysis (b) ADMIXTURE analysis, makes this hypothesis unlikely. 153

A.3 Testing the hypothesis of Slavic origin of culturally distinct Peloponnesean populations. PCA comparisons of (a) The Maniots of Deep Mani, Tayetos and Tsakones, with populations of the Slavic homeland (Ukrainians, Polish, Russians and Belarusians). Notice the broad separation between the Slavs and the Peloponnesean populations. (b) ADMIXTURE analysis shows the complete separation of Maniots and Tsakones from the Slavs in all K values. 154

A.4 Testing the hypothesis of Mardaitic origin of Maniots. The Mardaites were a medieval Middle Eastern population considered by some historians to be the ancestors of the Maronites of Lebanon. Comparison of Maniots with Maronites and other Middle Eastern populations by (a) PCA and, (b) ADMIXTURE analysis makes this hypothesis unlikely. 155

A.5 Unique genetic structure of the population of Tsakonia. PCA comparisons of Tsakones with A. the Eastern Europeans. B. North Africans C. Near Eastern populations D. Southern Europeans. 156

A.6 PCA plot of all Indian samples. We note that the formation of the clusters is primarily dominated by language groups, with some populations (Gond, Manipuri Brahmins, Dusadh) showing a certain amount of admixture between the language groups. A few tribal populations across IE and DR languages (Vedda, Madiga, Kol, Bhil, Chamar, Kuruchiyan) cluster together. We also observe that the Irulas, Paniyas, Kurumba and Kadars show divergence from other DR_SGC populations. 157

Figure Page

A.7 ADMIXTURE plot of all Indian populations for values of K between two and eights. Our findings are very similar to the observations in Supplementary Figure 1. The main observation is (again) that the formation of the clusters is primarily dominated by language groups, especially for larger values of K. 158

A.8 An ADMIXTURE plot (for values of K between two and eight) of the normalized data set (368 individuals 48,373 SNPs) clearly shows the four main components related to language groups (Dravidian, Indo-European, Tibeto-Burman, and Austro-Asiatic); see, for example, the plot for K equal to five or six. The plot also shows the divergence of the DR_SGC. We perfomed a meta-analysis of the results of the ADMIXTURE plot (see 3.1.2 for details) to visually and numerically quantify the amount of shared ancestry (as revealed by ADMIXTURE) between any pair of populations. Darker colors indicate larger amounts of shared ancestry; we observe a higher amount of shared ancestry between the IE and DR populations, across all social groups, indicating the existence of significant admixture between the two linguistic groups. The isolation of the DR_SGC samples is primarily due to the isolation of hill SGCs (such as Irula, Kadar, Paniyas, etc.) 159

A.9 Plotting the top two discriminants by (a) region and (b) language groups. Clearly, this follows much what we saw in Figure A.6. However, looking closely we see the following: (a) we see a geographical gradient, starting from IE_SGA and IE_SGB in Northwestern India to the other Indo-European and Dravidian SGA. We also see that the IE_SGC sit closer to the Austro-Asiatic speakers, justifying their geographical location in Central India. This is followed by the Tibeto-Burman speakers forming another cluster, concluding the other spectrum of the gradient. (b) Layers of stratification appears, from right to left. Although the LDA was performed by language groups, we see a two-layer stratification, first by castes and then by languages. The IE_SGA form a separate cline, followed by DR_SGA; then, the IE and DR SGBs follow. Then some DR and AA tribal populations cluster together, followed by a separate cluster of IE tribal populations. 160

A.10 Statistical significance of the COGG output (using random permutations of the features) Clearly, COGG is statistically significant for both the first and the second principal components 161

Figure Page

A.11 (a) COGG-CCA, when run with top 8 PCs, shows statistical significance with $r^2 = 0.94$ when compared against random permutations of the variables with average $r^2 = 0.75$. (b) Varying number of PCs to perform COGG-CCA results in the maximum r^2 when top 6 to 8 PCs are used. 162

A.12 The pairwise shared ancestry matrix of relatedness within DR show high relatedness among a large portion of DR speakers across caste affiliations. The Tribes such as Irula, Kadar, Palliyar, Paniya and Malayan show significant divergence from the others. Among them the Paniyas show absolute divergence, with very less amount of ancestry with all DR speakers, whereas the others tend to form a cluster and show that although they share significant amount of ancestry with each other, than the DR_SGA. The SGB and SGAs tend to cluster together showing high relatedness with some SGCs such as Adi-Dravider, Hakkipikki, Hallaki, Kuruchiyan, etc. 163

A.13 Most significant (Z-score higher than 85) outgroup f_3 statistics of the form $f_3(YRI; A, B)$ where YRI is the outgroup, A are the groups from Table 3.4 and B are all the pan-Indian populations in our data spanning across social groups and language families. 164

A.14 The top two principal components show a long cline of IE and DR speakers with some divergence by few SGCs, such as Tharu, Irula, Palliyar, Paniyas, etc. 164

A.15 The pairwise shared ancestry matrix of relatedness within IE show high relatedness among most of the IE speakers across caste affiliations. The Tharus show divergence from rest of the IE speakers except the Uttaranchal Brahmins, who share close relatedness with the East Asian component in their gene pool. The Brahmin groups (GJR – Gujarati; UP – Uttar Pradesh; UTR – Uttaranchal; WB – West Bengal) show high values of shared ancestry within each other and rest of the IE speakers. Only UTR Brahmins show some divergence. The tribes such as Sahariya, Bhil and Chamar are more closely related to the fellow SGCs than the SGA, but still show around 70% of relatedness with them. 165

A.16 The shared ancestry matrix of relatedness between IE and DR speakers show that high relatedness with some divergent groups, following from the PC plot in Figure A.15. The DR_SGA share very high ancestry with IE SGA and SGC, showing that there was high admixture and contact between these groups prior to endogamy. 166

Figure Page

A.17 The pairwise shared ancestry matrix of relatedness within AA show very high relatedness among almost all AA speakers. Birhors, who are nomadic hunter-gatherer people dwelling in forests share less ancestry than others, probably because of their subsistence nature, where they roam around the forests of eastern and central India. The Khasis also show divergence from the AA speakers because of their location in northeastern India near TB_SGC and presence of admixture from TB speakers. 166

A.18 The top two PCs of AA speakers in India show most of the groups form a cluster with Birhor and Korwa showing divergence from the main cluster. 167

A.19 PCA plot of the first two PCs reveals the Austronesians (Ami and Atayal) and the IE and DR speakers to be distinct from the rest of the southeast Asians along with the Indian AA speakers. 167

A.20 (a) ADMIXTURE plot (for values of K between two and eight) of the Indian dataset merged with Southeast Asian populations shows that the AA and TB speakers do not share a lot of admixture with other Austric speakers from Southeast Asia; (b) The pairwise shared ancestry matrix of AA and TB speakers highlighting that the Khasis share very high amount of ancestry with TB tribals, unlike other AA groups. 168

A.21 (a) Network analysis for top 2 PCs and 5 nearest neighbors show that the Khasis forming a bridge between Indian AA speakers and southeast Asia; (b) TreeMix plot of Indian and Southeast Asian AA speakers with 8 migration edges reveal that there is a migration edge from Cambodian to Bonda, who are Indian AA speakers attributed to southeastern Asian admixture. 169

A.22 (a) Network analysis for top 2 PCs and 5 nearest neighbors show that the Khasis forming a bridge between Indian AA speakers and southeast Asia; (b) TreeMix plot of Indian and Southeast Asian AA speakers with 8 migration edges reveal that there is a migration edge from Cambodian to Bonda, who are Indian AA speakers attributed to southeastern Asian admixture. 170

A.23 (a) PCA plot of the top two principal components of Indian TB speakers and mainland Chinese populations show that the TB_SGC are closer to the southern Chinese; (b) Network analysis show that TB_SGC are closer to Central and Southern China who are geographically closer to northeast India. 171

Figure Page

A.24 Plotting of Indian and Eurasian populations projected on the top two PCs, mirror the geography of Eurasia uncovering a triangular structure with Europeans residing in one corner, the Chinese on another corner and the DR and AA speaking tribal populations of India occupying the third corner. 172

A.25 Residual fit from the maximum likelihood tree in Fig 3. The residuals are normalized over the residual covariance between each pair i and j. Residuals above zero represent populations that are more closely related to each other and are candidates for admixture events. 172

A.26 ADMIXTURE plot (for values of K between two and eight) of the Indian dataset merged with Eurasian populations (1,332 individuals, 42,973 SNPs). Meta-Analysis of this plot in Fig 4a, quantifies the relationship between populations. The IE and DR Forward and Backward Castes share significant amount of ancestry with the Northwestern Frontier populations of Afghanistan and Pakistan, followed by ancestry from Central Asia, Turkey and Caucasia. The TB tribals belong to the same cluster as the Chinese populations along with, Mongolia and Uygurs. 173

A.27 Plots of the three leading eigenvectors returned by TeraPCA and FlashPCA2 for the simulated dataset S_6. 175

A.28 The projection of the HGDP dataset along the two leading eigenvectors computed by TeraPCA. 176

A.29 The wall-clock times achieved by TeraPCA and FlashPCA2 when the number of eigenvectors that we seek to extract (k) ranges from 10 to 500 for the dataset S_6. 177

A.30 The wall-clock times achieved TeraPCA and FlashPCA2 when the number of SNPs ranges from 20K to 100K on for the dataset S_6. 178

A.31 Proportion of variance captured by the ten leading eigenvectors returned by TeraPCA when applied on the 1000 Genomes dataset (FlashPCA2 returns essentially the same values for the proportion of variance captured by the top ten eigenvectors). 178

A.32 Amount of time required to multiply the (normalized) covariance matrix by a set of s vectors using the DGEMM BLAS routine of MKL for different values of s, β and threads, for the datasets S_6 and HGDP. 179

Figure Page

A.33 Comparing the height of the ARG (H) between the *fwd-sSimRa* and *back-sSimRA* for selection at two-loci with and without epistasis, respectively. We set $g = 25K$, $r = 1.0 \times 10^{-8}$ $N = 100$, $s = 0.3$, $e_s = \{0, 0.1\}$ and $m = \{10, 20, 30, 40\}$. (i) The box-and-whisker plot summarizes the result for each m. On each box, the central mark is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. (ii) Q-Q plots for each m showing that the distributions of H from *fwd-sSimRa* and *back-sSimRA* agrees (iii) Plot showing the CDFs of *fwd-sSimRa* and *back-sSimRA* reconfirming the agreement between them. 181

A.34 Comparing the height of the ARG (H) between the *fwd-sSimRa* and *back-sSimRA* for selection at two-loci with and without epistasis, respectively. We set $g = 25K$, $r = 1.0 \times 10^{-8}$ $N = 100$, $s = \{0.3, 0.3\}$, $e_s = \{0, 0.1\}$ and $m = \{10, 20, 30, 40\}$. (i) The box-and-whisker plot summarizes the result for each m. On each box, the central mark is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. (ii) Q-Q plots for each m showing that the distributions of H from *fwd-sSimRa* and *back-sSimRA* agrees (iii) Plot showing the CDFs of *fwd-sSimRa* and *back-sSimRA* reconfirming the agreement between them. 182

A.35 P-P plots of distributions of the height of the ARG (H) between *fwd-sSimRa* and *back-sSimRA* for the neutral model with epistasis and no epistasis at two loci respectively, by setting $g = 25K$, $r = 1.0 \times 10^{-8}$ $N = 100$, $s = 0.3$, $e_s = \{0, 0.1\}$ and $m = \{10, 20, 30, 40\}$. 183

ABBREVIATIONS

AHC	Agglomerative hierarchical clustering
ARG	Ancestral recombination graph
back-sSimRA	backward simulator of sSimRA
BN	Balding-Nichols
CCA	Canonical correlation analysis
COGG	Correlation optimization for genetics and geodemographics
FLOP	Floating point operations
fwd-sSimRA	forward simulator of sSimRA
GRM	Genetic relationship matrix
GWAS	Genome-wide association study
IRA	Implicitly restarted Arnoldi
kya	thousand years ago
LMM	Linear Mixed Model
LD	Linkage disequilibrium
MAF	Minor allele frequency
PCA	Principal component analysis
PSD	Pritchard-Stephens-Donelly
RandNLA	Randomized Numerical Linear Algebra
RLS	Ridge leverage scores
sSimRA	Simulation based random graph algorithms with selection
SMA	Single marker analysis
SNP	Single nucleotide polymorphism
SVD	Singular value decomposition
WF	Wright-Fisher model

ABSTRACT

Aritra Bose Ph.D., Purdue University, December 2019. Computational Methods in Population Genetics. Major Professor: Dr. Petros Drineas.

The field of population genetics has seen an unprecedented growth driven by the advancement of sequencing technologies, resulting in volumes of massive datasets. As a result, efficient computational methods backed by theoretical foundations are required to analyze and understand the intricate details of complex biological processes captured in the genetic code. To this end, we developed novel computational tools to address issues related to population structure, scalability of methods, models of evolution and disease association.

History of a population, in light of genomics, is reconstructed through series of settlements, migrations, adaptations, demographic expansions, mixture, etc. To better understand such a theory of migration for the Peloponnesean Greeks, we analyzed their sub-structure and disproved the theory of their replacement by the Slavs in medieval age. Ecological and environmental factors such as society, language and geographical barriers, among others can influence gene flow in populations resulting in complex structure. We developed a computational framework called COGG (Correlation Optimization of Genetics and Geodemographics) which studies the contribution of these demographic factors shaping the genetic sub-structure of the Indian subcontinent.

Principal Component Analysis (PCA) has profound impact in the study of population structure and a significant challenge is to build scalable software to implement PCA on tera-scale data. To address this issue, we built TeraPCA, an out-of-core, multi-

threaded C++ implementation of the Randomized Subspace Iteration method providing a faster and accurate alternative to the current state-of-the-art packages.

Stochastic models of evolution provides a better abstraction of complex evolutionary processes by simulating generations of random populations and provide foundations to analyze genetic variation among species. We developed the first algorithm that builds multiple loci selection with interacting polymorphic sites in a package called sSimRA. We also provide the first comparison between backward and forward simulators which models the effect of natural selection at multiple loci.

To address the issue of correcting for population structure confounding in Genome Wide Association Studies (GWAS) we developed CluStrat, a structure informed clustering based tool which outperforms the standard PCA based stratification correction approaches. GWAS is used ubiquitously to detect bio-markers predicting disorder traits and estimating heritability underlying phenotypic variation. One of the main challenges in GWAS is to correct for population structure in order to find the true positives. We provide a stratification correction technique called CluStrat, which corrects for complex population structure by performing agglomerative hierarchical clustering on the linkage disequilibrium (LD) induced distances between individuals captured in the Mahalanobis distance based Genetic Relationship Matrix (GRM). We further use CluStrat to outline a comprehensive guide to stratification and subsequent disorder trait prediction or estimation utilizing the underlying LD structure of the genotypes.

1 INTRODUCTION

The field of genetics has seen an unprecedented growth of data in the past years with the development of low-cost, high-throughput methodologies for studying human genome-scale variations. As a result, numerous studies, such as, HGDP [1, 2], HapMap [3], the 1000 Genomes project [4] and most recently, the UK Biobank [5] have made available comprehensive datasets with wide coverage of the human genomes across the world. Availability of such datasets has resulted in better understanding of the evolutionary history of different species by studying population structure [6–8] with effects of migrations, adaptations, population expansions [9–11], etc. The ability to sequence and study DNA by calibrating the rate of accumulation of changes with evolutionary time has enabled robust inferences about how humans have evolved, thus making population genetics an essential tool to reconstruct the human population history. We developed new methods along with standard tools to reconstruct the history of the Peloponnesean peninsula, which has been inhabited since the middle Paleolithic era (100 kya). Ancestry of the present day Peloponnesean Greeks had been a topic of hot debate for over a century when it was proposed that medieval Peloponneseans were totally extinguished by Slavic and Avar invaders and replaced by Slavic settlers during the 6th century CE. We gathered samples from present day Peloponneseans and studied their relatedness with the Slavic populations examining the theory of replacement in light of genomics.

In another problem related to population structure, we developed several methods to study the genetics of the south Asian populations, specifically, the influence of demographic factors on the genetic structure of the Indian subcontinent. Geography has been shown to closely correlate with genetic structure in other parts of the world [7]. However, the strict endogamy imposed by the Indian caste system, and

the large number of spoken languages add further levels of complexity. We set out to explore how these sociolinguistic (social caste and language) and ecological factors have shaped the gene flow in the Indian subcontinent. To this end, we developed COGG (Correlation Optimization of Genetics and Geodemographics), a model that optimally explains the observed population genetic sub-structure and used as a descriptive statistic to explain genetic variation of a population in presence of fixed effects.

As highlighted above, technological and scientific advances have made large-scale, genome-wide projects feasible and cost-effective, thus calling for more sophisticated and efficient computational tools to analyze the data. Principal Component Analysis (PCA) is used ubiquitously across the field of genetics and it's impact is truly massive. The seminal work of Luca Cavalli-Sforza and collaborators in the late 1970s [8, 12] pioneered the application of PCA to the study of human genetic variation. Although PCA is widely used across genetics, it does not scale to datasets with more than a few thousand samples as it has quadratic space and cubic time complexity, respectively. However, in practice, one does not need all the principal components (PCs). We use this relaxation by applying recent advances in the Randomized Numerical Linear Algebra (RandNLA) community to compute a low-dimensional embedding in a package called TeraPCA, a C++ package to approximate the top PCs (and corresponding eigenvalues) of tera-scale genotype datasets. TeraPCA is a multi-threaded, out-of-core implementation of the Randomized Subspace Iteration method, first analyzed in [13, 14]. We demonstrate the advantage of TeraPCA over other standard software suites to compute PCA on genotype data on both simulated and real-world datasets.

Genetic variation in populations arise from complex evolutionary processes such as mutation, recombination, adaptation, natural selection, random genetic drift, etc. The study of genetic variation underwent a paradigm shift with the development of coalescent theory [15, 16] which provided a mathematical foundation of gene ge-

nealogies. With the advent of this theory, a classical forward-time Wright-Fisher model [17] of evolution approach saw a transition to the new, backward-time coalescent approach [18]. Coalescent processes allow fast approximation of the neutral Wright-Fisher model, in which natural selection plays a major role in shaping patterns of variation with the Ancestral Recombination Graph (ARG), a variant of Kingman's coalescent. We provide the first algorithm that modulates the ARG to incorporate multi-locus selection with multi-way interaction between them and give a comprehensive comparison of forward and backward-time approaches. This allows a validation framework for including selection and interacting loci into standard population genetic models, armed with which, we can study the complex evolutionary scenarios.

The study of genetic variation not only helps demystify a population's history and peopling, it has been used extensively to find loci associated with diseases [19–21]. The sharing of genetic data and results from the association studies has been a key factor in mapping genes to diseases. In the past decade, laboratory experiments along with GWAS have led to the discovery of many target genes related to obesity [22], type 2 diabetes [23], inflammatory bowel disease [24], a host of psychiatric disorders [25, 26] among others. Despite the popularity of GWAS to find causal loci for various diseases it has been under scrutiny for the amount of spurious associations or false positives it detects due to a variety of factors such as, number of loci affecting the trait, genetic architecture (distribution of effect size and allele frequency at those loci), sample size, genotyping platforms, heterogeneity of the trait, etc. [21]. Recently, two independent studies [27, 28] failed to replicate the strong evidence for selection for height across Europe as found in previous association studies [29–31], implying that standard population structure correction approaches may not be enough, and that more rigorous, sophisticated methods are required. To address this problem we propose a correction technique for complex population structure while leveraging the linkage disequilibrium (LD) induced distances between individuals. We implemented CluStrat, which performs agglomerative hierarchical clustering using the Mahalanobis distance based Genetic Relationship Matrix (GRM) representing the population-level

covariance (LD) for the genetic markers. This framework harnesses the interaction between the markers to produce structure informed clusters correcting for population stratification. We show that this method produces least amount of spurious association and detects two to three fold more causal loci than other standard Linear Mixed Model (LMM) or PCA based approaches.

In summary, this dissertation extends our understanding of both empirical and theoretical population genetics. It proposes methods to deal with a variety of open problems in genetics such as the reconstruction of genetic history of a population as well as the influence of demographic factors on it's genetic structure. It highlights different computational bottlenecks in these methods and provides a robust, scalable and efficient software package expediting analysis of genotype data. Furthermore, it strengthens our knowledge of theoretical aspects of the field by designing the first validation framework of coalescent models for simulating complex evolutionary scenarios. In what follows, we put together important concepts of population genetics in the Background section and thereafter address each of these above questions in sequential order in each chapter of this dissertation.

2 BACKGROUND

Genetic diversity within and between populations is a result of various evolutionary processes that act on populations. These processes include mutation, recombination, admixture, selection, migration, adaptation, population expansions or contractions, etc. The Main goal of this dissertation is to develop methods to better understand these evolutionary forces from an empirical and theoretical perspective. This chapter provides the required background on foundations of population genetics and methods of capturing genetic variation on topics of interest of this dissertation.

2.1 Genetics Fundamentals

2.1.1 Ploidy

Cells are the foundation of life. Plants, bacteria, human beings and every other living organisms are made up of small, microscopic cells. In biology, ploidy is used to denote the number of sets of chromosomes contained within the nucleus of a cell. The nucleus of a eukaryotic cell is haploid if it has a single set of chromosomes, diploid, if it has two homologous copies of each chromosome, one from each parent and polyploid when cells have multiple sets of chromosomes, usually three or more. Human beings are diploid organisms, containing 46 chromosomes (23 pairs) out of which, 22 pairs are called autosomes, which look the same in both males and females. The 23rd pair is the sex chromosome, which differs between males and females. Females having two copies of the X chromosome and males have one X and one Y chromosome.

2.1.2 Single Nucleotide Polymorphisms

The genetic constitution of an organism is called a genotype. In case of diploid organisms, each chromosomal copy is known as the haplotype, which are jointly called the genotype. The biological processes which copy the genetic material from parents to offspring is not perfect and are influenced by "errors". This imperfect copying of the genetic material is called mutation, which usually happens at a single nucleotide (location) in the genome. If a parental chromosome had Thyamine or 'T' at a specific location, due to imperfect copying, the child might contain a Cytosine or 'C' at that position. The polymorphisms that arise from these single or point mutations are called Single Nucleotide Polymorphisms or SNPs. Each variant at a SNP is called an allele.

If both alleles at a diploid organism are same, the organism is homozygous and if they are different, they are heterozygous at that locus.

2.1.3 Human Genetic Variation

The unit of genetic variation within individuals are SNPs or point mutations. One of the most prominent sources influencing genetic variation apart from mutation is recombination, or exchange of genes involving crossover during reproduction. The genetic material from each parent mixes while getting copied to the child's chromosome and "shuffles" maternal and paternal DNA elements creating new combinations of variants. Natural selection confers an adaptive advantage (or disadvantage) to an allele of an individual in a specific environment, making them more (or less) likely to occur altering the population. Genetic drift, which is the effect of random changes in gene pool of a population is another source which leads to an individual's unique genetic structure. This drift sometimes lead to bottlenecks in small populations, at random. Genetic variation of a population undergoing a bottleneck is very low for a few generations threatening it's existence in some cases.

Changes in genome can affect the phenotype (such as skin color, height, disease traits, etc) of an individual. Phenotypic traits can be passed on to the next generation of a population, if most of the alleles in the population affecting the phenotype are acted by positive selection, increasing it's chances of survival. Thus SNPs, and in turn, traits become heritable.

Migration is another prominent force contributing to genetic variation. When genetically different populations interact with each other to produce offspring, the chromosomes in the resulting population contain genetic contributions from both ancestral populations. This process is called admixture and the resultant population is referred to as an admixed population.

2.1.4 Population Structure

Structure in a population broadly refers to any deviation from random mating, involving inbreeding and/or geographical subdivisions. Effects of natural selection, genetic drift, geographical barriers and other factors which contribute to genetic variation results in structure within a population. Even without barriers of gene flow, organisms do not disperse randomly and tend to practice inbreeding. Genetic population structure can shed light on evolutionary history and migrations of modern populations [6, 32] due to the shared ancestry between related individuals leading to genetic and phenotypic differences within a population. Understanding how and why these partially isolated populations differ in their genetic make-up is one of the fundamental aims of evolutionary biology.

Shared ancestry between populations correspond to relatedness, or kinship and thus population structure can be defined in terms of patterns of kinship among groups of individuals [33]. This relatedness among individuals has a significant impact in case-control association studies which maybe subject to high rate of false positives if there is unrecognized population structure.

Population structure can be visualized by unsupervised clustering algorithms, however they are heavily dependent on the distance metric used for the clusters. Hence, model-based clustering approaches are widely used to detect structure. The earliest method dates back to the *Structure* model [34] which modeled genomes as a mixture of contributions from ancestral populations. This was further developed into a faster algorithm called *Admixture* [35] which is widely used. Another way of detecting population structure uses the eigen-analysis method [36] which computes a singular value decomposition (SVD) on the genotype matrix and project the samples on the top two to three significant PCs to visualize how the data is structured. The widely used tool, *EIGENSOFT* encodes this algorithm and serves as a rough estimate for the intra-population variation and finding outliers in the data. This method is extremely useful when correcting for population stratification to find homogeneous populations before case-control association studies [37]. We discuss these two popular approaches to infer population structure below.

PCA based approaches

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be the genotype matrix where m is the number of observations and n is the number of biallelic markers such as SNPs ($n >> m$ in most cases). For each SNP, we choose a reference and alternate allele, then $\mathbf{A}_{i,j}$ is the number of alternate alleles for individual i and marker j, usually coded as 0, 1 and 2 for homozygous dominant, heterozygous, and homozygous recessive genotypes, respectively. If there are two alleles B and b, we define homozygous major (dominant) and minor (recessive) to be, BB and bb respectively. Similarly, heterozygous allele means a combination of both, Bb or bB. We assume that there is no missing data. The genotype matrix is mean centered, that is we calculate mean of each SNP,

$$\mu_j = \frac{\sum_{i=1}^{m} \mathbf{A}_{i,j}}{m}$$

and subtract it from each entry to get $\mathbf{A}_{i,j} - \mu_j$. We furthermore normalize this with respect to the estimate of the underlying allele frequency, $p_j = \mu_j/2$ to get,

$$\mathbf{X}_{i,j} = \frac{\mathbf{A}_{i,j} - \mu_j}{\sqrt{p_j(1-p_j)}}$$

This normalizing step lets us take into account the frequency change of a SNP due to rate of genetic drift proportional to $\sqrt{p_j(1-p_j)}$ per generation [36], improving the performance of structure detection.

At the "heart" of PCA, we carry out a SVD on $\mathbf{X}$. To improve performance of SVD because of it's cubic computational complexity, we compute a $m \times m$ GRM,

$$\mathbf{M} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top$$

which is a sample covariance matrix of $\mathbf{X}$. We compute an eigenvector decomposition of $\mathbf{M}$ to obtain the eigenvectors corresponding to the large, informative eigenvalues. These top eigenvectors are interchangeably used as PCs in the rest of the text. We note here that with this unsupervised method we can also reconstruct the population labels if they are not available and alternatively, we can also verify the performance of the method if the labels are available. This was first proposed by Cavalli-Sforza [38] revealing population structure.

Model based approaches

The *Structure* model by Pritchard et al. [34] provides a way of stochastically clustering individuals into groups related to their ancestral populations. It uses an allele-frequency profile which are fixed dimensional multinomial distributions from multiple populations and makes up all the SNPs of an individual as independent and identically distributed instantiation of these profiles. These models of admixture identifies each ancestral population by a specific allele frequency profile and displays the frac-

tion of contributions from each profile in a present day individual's genome. Thus, these models generate a matrix $\mathbf{K} \in \mathbb{R}^{m \times k}$ from the original genotype matrix $\mathbf{A}$ where k is the number of user defined ancestral populations. The rows of the matrix $\mathbf{K}$ adds up to 1 as they denote the fractions or probabilities of the individual i belonging to each ancestral population k_j.

2.1.5 Tests for Admixture

A simple, intuitive and popular approach to detect signs of admixture and direction of gene flow are the F-statistics, introduced in [39] and summarized in [40] and [41]. Shared genetic drift between sets of populations is measured to test the hypothesis whether the involved populations share common evolutionary history. F-statistics, namely, f_2, f_3, f_4 and `qpAdm`, `qpWave` are widely used to analyze genetic history of populations using modern as well as ancient DNA [39,42–44]. F-statistics is generally used to answer the following questions among others, which we are interested to study in this thesis.

- Treeness tests: Are populations related in a tree-like phylogeny [39]?
- Admixture tests: Is a particular population descended from multiple ancestral populations [39]?
- Admixture proportions: How much does the ancestral populations contribute into the genetic make-up of a modern population [43]?
- Complex demography: How many mixtures and splits of population explain it's demography [40]?

We use Figure 2.1 from [41] to explain the various statistics discussed above. Under a population phylogeny, three F-statistics labeled as F_2, F_3, F_4 (interchangeably used throughout this thesis as f_2, f_3 and f_4) between two, three and four taxa, respectively. These populations are labeled as P_1, P_2, P_3 and P_4. $f_2(P_1, P_2)$ corresponds to the

path on the tree from P_1 to P_2. The purpose is to measure how much genetic drift occurred between P_1 and P_2, thus, we can define f_2 as

$$f_2(P_1, P_2) = f_2(p_1, p_2) = \mathbf{E}\left[(p_1 - p_2)^2\right] \tag{2.1}$$

p_i is denoted as the allele frequency or the proportion of individuals in P_i that carry a particular allele (minor allele, usually) at a bi-allelic locus as discussed above. The above equation 2.1 assumes haploid individuals, but the deductions hold for diploid organisms as well. The expected values of F-statistics relies on tracing the overlap of genetic drift paths and f_2 can be thought of as the branch length between two populations in a phylogeny with overlaps, $P_1 \longrightarrow P_2$, $P_1 \longrightarrow P_2$.

The three population test which is widely used to detect admixture events and infer direction of gene flow is defined as, $f_3(P_X; P_1, P_2)$ representing the length of the external branch from P_X to the internal vertex containing all three populations. The two parameters P_1 and P_2 can be interchanged keeping the meaning same.

$$f_3(P_X; P_1, P_2) = f_3(p_X; p_1, p_2) = \mathbf{E}\left[(p_X - p_1)(p_X - p_2)\right] \tag{2.2}$$

We seek to test whether P_X is admixed between P_1 and P_2. This can be interpreted as the shared portion of the paths from P_X to P_1 with that of P_X to P_2, with overlaps $P_X \longrightarrow P_1$, $P_X \longrightarrow P_2$. Note that if P_X is admixed, there is a negative term in Equation 2.2 because $P_X \longrightarrow P_1$ and $P_X \longrightarrow P_2$ take opposite directions through the internal vertex connecting them. Thus, the observation of a negative f_3 value provides unambiguous evidence of a population mixture in the history of the target population P_X.

f_3 can also be interpreted on presence of an "Outgroup" (a target population which is very divergent to the populations P_1 and P_2 based on ancestral genetic data). Let's call this population P_O and by calculating $f_3(P_O; P_1, P_2)$, we measure the shared genetic drift between P_1 and P_2. If the f_3 values are high in this case, it means that

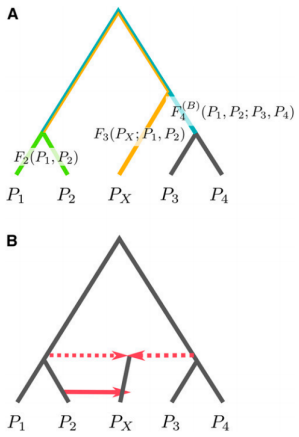


Figure 2.1.: A. A population phylogeny with branches corresponding to F_2 (green), F_3 (yellow), and F_4 (blue); B. An admixture graph extends a population phylogeny by allowing gene flow (red, solid line) and admixture events (red, dotted line).

the populations are very closely related. We extensively use f_3 statistics along with it's outgroup feature in this thesis.

The four population test is similar to it's smaller counterparts but involving the covariance of the allele frequency differences between populations P_1, P_2 and populations P_3, P_4 respectively. $f_4(P_1, P_2; P_3, P_4)$ represents the internal branch from the internal vertex of P_1 and P_2 to the vertex connecting P_3 and P_4.

$$f_4(P_1, P_2; P_3, P_4) = f_4(p_1, p_2; p_3, p_4) = \mathbf{E}\left[(p1 - p_2)(p_3 - p_4)\right] \tag{2.3}$$

The expected value can be computed from the overlap of drifts $P_1 \longrightarrow P_2$ and $P_3 \longrightarrow P_4$. If these paths do not overlap, $f_4 = 0$ and if they overlap, that is, $P_1 \longrightarrow P_3$ and $P_2 \longrightarrow P_4$, then f_4 is equal to the length of the internal branch of the tree and positive because the drift paths overlap in the same direction. Conversely, if it is $P_1 \longrightarrow P_4$ and $P_2 \longrightarrow P_3$, then drift paths are opposite direction and results in negative values. Thus, gene flow directions and admixture scenarios can be interpreted efficiently by studying the allele frequency distributions of the populations under study. The F-statistics derivations shown here are for one site of polymorphism.

But, they hold rigorously after normalizing for all SNPs identifying genetic variation between individuals belonging to arbitrarily structured populations.

`qpWave` or `qpAdm` schemes require us to choose m "right" populations (or outgroups) and n "left" populations (or references). and taking the first population on each side as the point of comparison it builds a $(m-1) \times (n-1)$ matrix of f_4 statistics by testing

$$f_{ij} = f_4(L_i, L_j; R_i, R_j)$$

We need to provide more distinct outgroups ("right", represented as R) than references ("left", represented as L), $m > n$ for accurate f_4 values. The matrix thus formed has maximum rank $n-1$ and minimum zero nontrivial columns.

Assuming this matrix has rank r, therefore r independent columns and rest of them being linear combinations of the r columns, the matrix $\mathbf{F} = \{f_{ij}\}$ can be modeled as product of two matrices $\mathbf{A}$ and $\mathbf{B}$

$$\mathbf{F} = \mathbf{A}.\mathbf{B}$$

where $\mathbf{A} \in \mathbb{R}^{(m-1)\times r}$ representing the r independent f_4 columns and $\mathbf{B} \in \mathbb{R}^{r\times(n-1)}$ representing the weights for combining columns of $\mathbf{A}$ to produce each column of $\mathbf{F}$. The observed f_4 statistic is an estimate of the true parameter and thus contains an error term making the above

$$\mathbf{F} = \mathbf{A}.\mathbf{B} + \mathbf{E} \tag{2.4}$$

$\mathbf{E}$ is the error matrix, $\mathbf{E} = \{\epsilon_{ij}\}$ following a multivariate normal distribution with mean zero. `qpAdm` assumes that the first left populations ("target") is a mixture of the remaining left populations ("references") in presence of the right populations ("outgroups"). It has maximum rank $(n-2)$ with an additional constraint for scaling to make the sum of admixture coefficients (weights) to 1. This is very insightful in representing a "target" population as simply a combination of "reference" ances-

tral populations with weights adding up to 1. We use `qpAdm` to infer how modern populations are a combination of ancient and other modern populations.

2.1.6 Linkage Disequilibrium

Linkage disequilibrium (LD) is a nonrandom association of alleles at two or more loci in a genome. LD is of importance in human genetics because so many factors affect it and are affected by it [45]. It is widely used to provide insight into evolutionary history and is the basis for mapping genes in organisms. The main reason for LD is recombination. These events ensure independent assortment of alleles when they are transmitted across generations. As recombination is a rare event (1 recombination per chromosome per generation) [33], the loci which are linked by LD are also highly correlated. Thus, an extant population inherits a linked allele pair from a remote common ancestor without any intervening recombination site. LD has far reaching implications as stronger LD around a disease causing SNP is easier to detect due to the probability that the causal SNP is in LD with at least one SNP in it's nearby regions is quite high [46]. On the other hand as the correlation is high, there are many markers which is in high LD with the causal variants thus making it harder to identify the causal variant as all of the correlated SNPs show similar strength of association to the phenotype.

Disequilibrium is measured as the difference between the observed and the expected (under independence) frequency of a particular combination of alleles at two loci. This can be represented as

$$D_{AB} = p_{AB} - p_A p_B$$

which is the difference between the frequency of gametes carrying the pair of alleles A and B at two loci and the product of the frequencies of those individual alleles. The quantity D_{AB} is known as the quantity of linkage disequilibrium (D) defined for a specific pair of alleles, A and B. If $D = 0$ there is linkage equilibrium. Due to

this definition, LD is primarily measured as the correlation between alleles [47] and is defined as

$$\rho_{AB} = \frac{D_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

and is often calculated with the Pearson correlation coefficient r^2.

2.1.7 Ancestral Recombination Graph

DNA sequences drawn from one or more individuals are related by a branching structure known as geneaology [48]. Recombination events changes these genealogies, resulting in a complex correlation structure collected together in co-linear orthologous sequences. This can be described by a network called ancestral recombination graph (ARG) [49]. An ARG provides a record of all coalescence and recombination events at all genomic positions for the extant populations under study. ARGs are

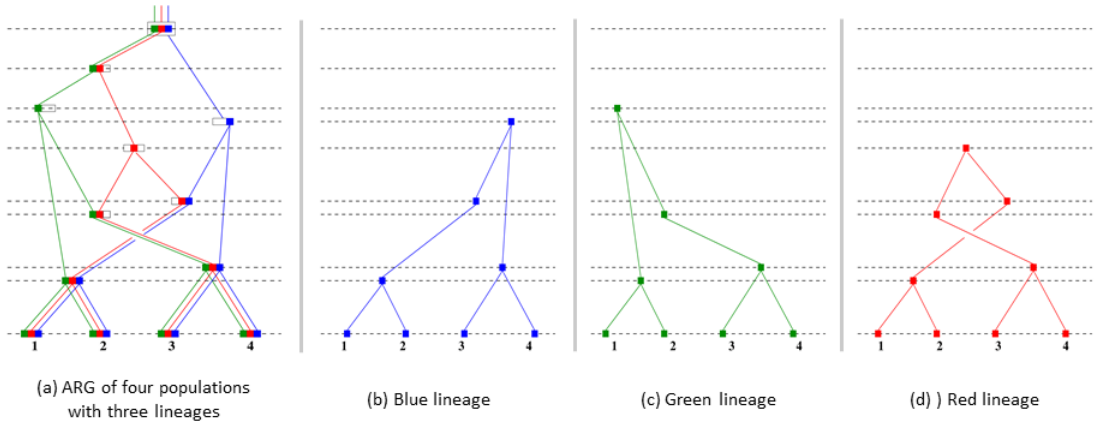


Figure 2.2.: ARG of four populations with three lineages (red, blue and green) showing recombination (nodes splitting into two) and coalescence (nodes merging into one).

mostly simulated by extending the widely used coalescent framework which includes recombination [48] as shown above. However, constructing ARGs from sequence data

and modeling various effects of genetic variation in ARG reconstruction is a challenge.

2.1.8 Association Studies

Genetic association studies are designed to identify loci that contribute to the phenotypic outcome of interest. Traditional methods of association used single marker analysis (SMA) methods, mapping one SNP and one phenotype at a time. Recent methods have shown how to analyze multiple markers simultaneously for association, extending it to eigen-analysis, regression and most recently to mixed models [33, 37, 50–53]. These association studies can involve a quantitative trait locus (QTL with a continuous trait) or case-control binary status, depending on the the disease being studied as well as the goal of the study, which generally involves a complex trait. Cases exhibit this phenotype of interest, whereas controls show no such prevalence. The underlying assumption in an association study is that genotypic differences between cases and controls are likely to be at markers which are causally related to the phenotype. Thus, in an association study the set of markers (LMM or logistic regression) or a single marker (χ^2 tests) is studied with respect to the trait.

Heritability

Phenotypic traits which are quantitative, show a continuous spectrum of variation controlled by a collection of polygenic markers acting in concert. A wide variety of important phenotypic traits are quantitative. Although, these quantitative traits can be converted to binary traits using a liability measure [54]. Quantitative traits can be studied in the terms of variation as well as environmental conditions owing to noise. The variance of a phenotype, therefore, can be partitioned into variances attributable

to the environment and to genetic factors. Heritability, is defined as the proportion of variation in a trait explained by inherited genetic variants, represented as,

$$\mathbf{H}^2 = \frac{\sigma_G^2}{\sigma_P^2}$$

As the phenotypic variance is partitioned, $\sigma_P^2 = \sigma_G^2 + \sigma_E^2$, we can write the above as

$$\mathbf{H}^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}$$

This definition is known as the "Broad-sense" heritability. In practice, $\mathbf{H}^2$ is very hard to estimate without strong assumptions. We are more interested in the amount of heritability explained by the genotypes, which is also known as "SNP-heritability". Given m SNPs we seek to find how much of phenotypic variance it explains. Thus, SNP-heritability is defined as,

$$h_g^2 = \frac{\sigma^2_{\text{SNP}\in m}}{\sigma_P^2}$$

From the above, we can see that $h_g^2 \leq \mathbf{H}^2$ as this is only limited to additive terms of genotyped markers and not the theoretical "broad-sense" heritability due to the entire genome. We usually define the trait, y, for each SNP x_j with an additive effect β_j in a linear model as,

$$y = \sum_{j=1}^{m} x_j \beta_j + \epsilon \tag{2.5}$$

where ϵ is the residual or error term which is not explained by the genotypes. Each x_j is encoded as discussed above with 0/1/2 allele counts and also normalized. We can now define h_g^2 more precisely as,

$$h_g^2 = \frac{\mathbf{Var}\left[\sum_{j=1}^{m} x_j \beta_j\right]}{\mathbf{Var}\left[y\right]}$$

Challenges in GWAS

GWAS has long been plagued by confounding due to the presence of cryptic relatedness owing to the population structure. If the cases disproportionately represent a population in comparison to controls then any SNP with differing allele frequencies between cases and controls will be incorrectly found to be associated with the phenotype resulting in false positives or spurious associations. Another issue in GWAS is the insufficient sample size. Statistical significance tests can fail to identify variants of smaller or moderate effects as causal in studies with small sample sizes. Increasing the sample size has been shown to discover more causal SNPs with respect to the phenotype of interest [55].

As traditional approaches did SMA, multiple hypothesis testing were conducted simultaneously, requiring a correction factor to avoid false positives. A commonly used technique is the Bonferroni correction, by which the test statistic is reduced by a factor of the number of SNPs, assuming all tests performed are independent. However, as discussed above, due to LD this is not always true and hence this technique can be deemed as very conservative and inaccurate. As the widely used genotyping arrays fail to tag rare variants and low frequency markers, we fail to analyze a significant number of markers. Since many traits are complex and multi-factorial, a relatively small number of rare variants with moderate effects could account for a large percentage of variation in the trait.

3 RECONSTRUCTING GENETIC POPULATION HISTORY

3.1 Genetics of the Peloponnesean Populations

This article has appeared in *European Journal of Human Genetics* Published by Springer Nature with DOI: 10.1038/ejhg.2017.18.

3.1.1 Introduction

Peloponnese has been one of the cradles of the European civilization of the classical era and has done distinct contributions to the ancient European history. It has also been in the center of a controversy about the ancestry of its population [56]. This controversy, lasting for about 170 years, has been fueled by historians who try to reconstruct medieval history on the basis of scant written resources. Controversies are not uncommon in historiography and are the source of endless debates among scholars. Controversies concerning population ancestry, however, can potentially be resolved by population genetic analysis. The study of the genetics of the Peloponnesean population provides a test of this premise.

Peloponnese was peopled by a series of migrations that span at least nine millennia. Early migrants arrived from Anatolia ca 9000 BCE [57] and established several Neolithic sites across the peninsula [58]. The Myceneans [59], who established a Bronze era civilization that lasted from the 17th to the 12th centuries BCE (1), were

Greek speaking Indo-Europeans who presumably migrated from the north around 2200 BCE [58–60] or were the descendants of the Anatolian Neolithic migrants [61]. The next known migration took place at the beginning of the first millennium BCE, when the Dorian Greeks arrived in Peloponnese [62] from an area corresponding to Epirus and western Macedonia. The subsequent eight centuries of the Archaic, Classical and Hellenistic periods of Greek history, the four centuries of Roman occupation and the two initial centuries of Byzantine dominance, were marked by quantitative changes of the Peloponnesean population due to wars and epidemics but no qualitative effects from migrations of new population groups. Changes in population structure started in the beginning of the medieval period with the migrations of the Slavs to the Balkans [63, 64]. The effects of these migrations have dominated the historiography of Peloponnese during the last 180 years.

In 1830 CE, the German historian Jacob Philipp Fallmerayer presented his theory of disappearance of the Greek nation and its substitution by Hellenized Slavs [65]. Fallmerayer proposed that during the 6th century CE, large armies of Avars and Slavs overran the Balkans and eliminated the populations of the Hellas, who up to that period had successfully survived the attacks of barbarians and the religious suppression by the Byzantines. The Peloponnesean Greeks, except for few remnants enclosed in coastal castles, were slaughtered or forced to leave from their ancestral lands and Peloponnese was inhabited by Slavic tribes. The Slavs kept their identity for few centuries but eventually they were Hellenized under the influence of the Orthodox Church and interactions with Hellenized Asia Minor populations who were settled in Peloponnese by the Byzantines. Since the time Fallmerayer's theory was published, a debate on the question of the ancestry of Peloponneseans has raged among historians (reviewed in [66]). Of note is that in spite of their diametrically different views, all historians have been using the same medieval written sources.

In this paper we use genome wide data to study the genetic structure of the Peloponnesean populations and compare them with other populations of the world. We

observe characteristic patterns of genetic differentiation within Peloponnese and examine their possible causes. We focus on the question of the impact of Slavic migrations on the genetic structure of the Peloponnesean populations and we test the theory of the extinction of the medieval Peloponnesean Greeks.

3.1.2 Materials and Methods

Study Design

The study has been reviewed by the Institutional Review Board of the University of Washington and the ethical committees of several provisional hospitals. We focused on the rural population. Subjects were included in the study if all four grandparents originated from the same village or from villages that were less than 10 kilometers apart. The ages of the participants ranged between 70 and 90 years (the oldest subject was 107 years old) and hence their grandparents were born between 1860 and 1880. The population of Peloponnese was 578,598 individuals in the 1861 census. At that time the economy of Peloponnese was exclusively agricultural and over 85% of the population was living in small villages and hamlets. We sampled all the districts of Peloponnese (Figure 3.1 and Table A.1) and also focused on two culturally distinct subpopulations, the Tsacones and the Maniots. To compare the Peloponneseans with other European populations, we analyzed samples from published datasets and datasets generated by our studies (Supplementary Table 1 and Figure A.1. Merging genotypes from different sources and quality control were done as described in [67].

PCA

We used TeraPCA [68] as well as our own MATLAB implementation of PCA [67, 69].

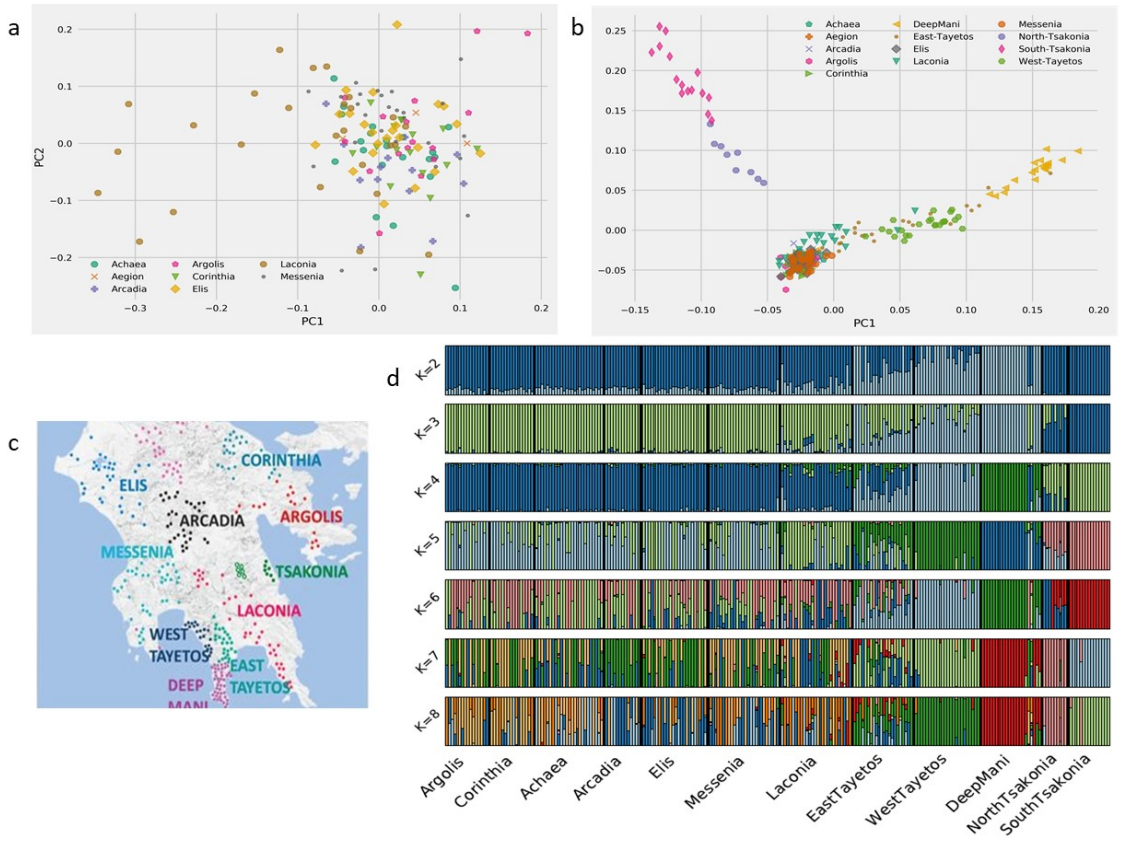


Figure 3.1.: Substructure of the Peloponnesean populations. (a) PCA analysis without the Maniot and Tsakones populations showing a partial separation of the population of Laconia. (b) PCA illustrating the separation of Peloponneseans in three groups. On the left is placed the population of Tsakones (north: open circles, south: green dots). On the right are placed the populations of Maniots (Deep Mani, East and West Tayetos.). All the remaining Peloponneseans are clustered in the center. (c) Map of Peloponnese showing the populations studied. Each dot corresponds to the origin of a participant. (d)ADMIXTURE analysis. Notice the distinct structure of the Maniots and the Tsakones and their clear cut separation from all other Peloponneseans in all values of K.

Estimating Population Admixture

We used the ADMIXTURE v1.22 software for all our admixture analyses [35].

Quantitative Analysis of ADMIXTURE Output

Given a target population $\mathbf{X}$ and reference populations $\mathbf{Y}$, $\mathbf{Z}$, etc., we were interested in quantifying the amount of ancestry of population $\mathbf{X}$ that is captured by populations $\mathbf{Y}$, $\mathbf{Z}$, etc. Towards that end we devised a new approach to quantitatively analyze the output of ADMIXTURE. Recall that ADMIXTURE, for a particular value of K, will represent each sample using K coordinates. Thus, for a particular value of K and for a particular population $\mathbf{Y}$ with n samples, we can represent the output of ADMIXTURE for this population as an n-by-K table. Then, for each reference population $\mathbf{Y}$, we summarize this n-by-K matrix using its top right singular vector only; in all our analyses, the top singular value corresponding to the top right singular vector captured at least 80% of the reference population variance as represented by ADMIXTURE. Let $\mathbf{v}_Y$ be the top right singular vector (a K-dimensional vector) for population $\mathbf{Y}$; similarly, let $\mathbf{v}_Z$ be the top right singular vector (a K-dimensional vector) for population $\mathbf{Z}$, etc. Now that we have represented the ADMIXTURE output for each population as a K-dimensional signature vector, we can apply standard vector space calculus in order to answer our original question: how much of the ancestry of population $\mathbf{X}$ is captured by population $\mathbf{Y}$, or population $\mathbf{Z}$, etc. More specifically, in order to compute the percentage of the ancestry of population $\mathbf{X}$ that is captured by population $\mathbf{Y}$, we compute the percentage of the norm of $\mathbf{v}_X$ that is captured (in projection sense) by $\mathbf{v}_Y$. Formally, we compute

$$\frac{\left\|\mathbf{V}_X - \mathbf{v}_Y \mathbf{v}_Y^+ \mathbf{V}_x\right\|_F}{\|\mathbf{V}_X\|_F}$$

which returns a value between zero and one. In the above, $\mathbf{v}_X$ denotes the m-by-K matrix representing the m samples of population $\mathbf{X}$ with respect to the K coordinates returned by ADMIXTURE. The notation $\mathbf{v}_Y^+$ indicates the pseudoinverse of the vector $\mathbf{v}_Y$, which is equal to the transpose of the vector $\mathbf{v}_Y$, suitably normalized. It is also worth noting that the norm used in the above equation is the standard matrix Frobenius norm. In order to quantify the amount of ancestry of population $\mathbf{X}$ that

is captured by both populations $\mathbf{Y}$ and $\mathbf{Z}$, we form the K-by-2 matrix $\mathbf{v} = [v_Y v_Z]$ whose columns are the vectors $\mathbf{v}_Y$ and $\mathbf{v}_Z$ and we compute

$$\frac{\|\mathbf{V}_X - \mathbf{V}\mathbf{V}^+\mathbf{V}_X\|_F}{\|\mathbf{V}_X\|_F} \tag{3.1}$$

In the above equation, $\mathbf{v}^+$ denotes the pseudoinverse of the matrix $\mathbf{V}$; the matrix $\mathbf{v}V^+$ is a projector on the subspace spanned by the column space of $\mathbf{V}$. Thus, we basically extract from the matrix $\mathbf{v}_X$ the part that is captured by the (subspace spanned by the) vectors $\mathbf{v}_Y$ and $\mathbf{v}_Z$.

Network Analysis

To better visualize and understand the connection between the populations included in our study, we performed a network analysis on the results of ADMIXTURE, using a method presented in [70].

3.1.3 Results

The Substructure of the Peloponnesean Populations

On PCA analysis the populations are arranged in the form of an inverted capital letter V (Figure 3.1b). The left of this formation is occupied by the population of Tsakones who inhabit the east slopes of Mount Parnon and the adjacent costal area (Figure 3.1c). The right of the formation is occupied by the populations of Maniots who inhabit the east and west slopes of mount Tayetos and the southern area of the promontory, the so called Deep Mani. All other Peloponneseans cluster in the tip of the letter V ((Figure 3.1b). Partial separation of some subpopulations of individual districts was also observed by PCA analysis (Figure 3.1a). The ADMIXTURE analysis of Figure 3.1e shows that the Maniots and Tsakones are clearly separated from each other and from all other Peloponnesean populations.Gradients in

gene frequencies from north to south across all Peloponnese, along the Ionian coast, across Arcadia, as well as within Laconia and between the slopes of Tayetos and Deep Mani are suggested by the correlations between geographic coordinates and the two principal components (Table 1)

Table 3.1.: Correlations between geographic coordinates and principal components (respective PC is indicated in parenthesis).

Populations	Latitude Correlation	Longitude Correlation
All Peloponnese	0.50 (PC1)	0.41 (PC2)
Peloponnese minus Tsakonia and Mani	0.49 (PC1)	0.09 (PC2)
Arcadia	0.60 (PC1)	0.12 (PC2)
Laconia	0.45 (PC1)	0.07 (PC2)
Ionian Coast	0.31 (PC2)	0.06 (PC1)
Elis	0.17 (PC1)	0.10 (PC2)
Arcadia and Messenia	0.34 (PC2)	0.16 (PC1)
Arcadia and Laconia	0.36 (PC2)	0.20 (PC1)
Deep Mani	0.15 (PC2)	0.21 (PC1)
East Tayetos and Deep Mani	0.67 (PC1)	0.10 (PC2)
West Tayetos and Deep Mani	0.73 (PC1)	0.42 (PC2)

Comparison with other European Populations

As anticipated from the results of previous studies [38, 71, 72], by PCA analysis the Peloponneseans were placed very close to the Sicilians and Italians (Figure 3.2b) and remotely from all other European populations we compared them with. Network analysis (Figure 3.2a), highlighted the inter connections of Peloponnesean populations as well as the connections between Peloponneseans, Italians and Sicilians; the latter, serve as a bridge between Peloponnese and other Southern European populations (Basque, Andalusians, French). Slavic populations were placed far away from the Peloponneseans as were the Near Eastern populations. The latter were connected to the Peloponnesos via the islands of Crete and the Dodecanese.

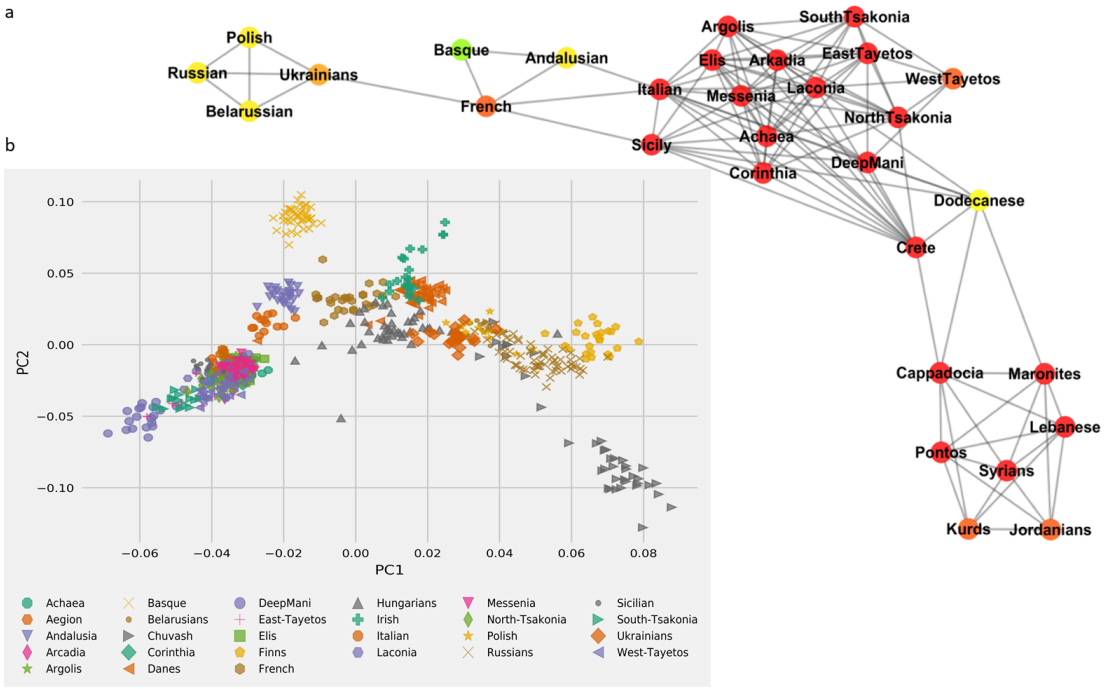


Figure 3.2.: Genetic similarity of Peloponneseans and Europeans showing differentiation from Slavs. (a) Network analysis illustrating the high connectivity between the Peloponnesean populations as well as between the Peloponneseans, the Sicilians and the Italians. Notice the distance between Peloponneseans and the Slavic, and Near Eastern populations. Peloponneseans are connected with the Near Eastern populations through Crete and Dodecanese. (b) Notice the north to south distribution of the populations and that the Peloponneseans are placed to the far left of the graph, overlapping with the Sicilians and distinct from the Slavs (on the right side).

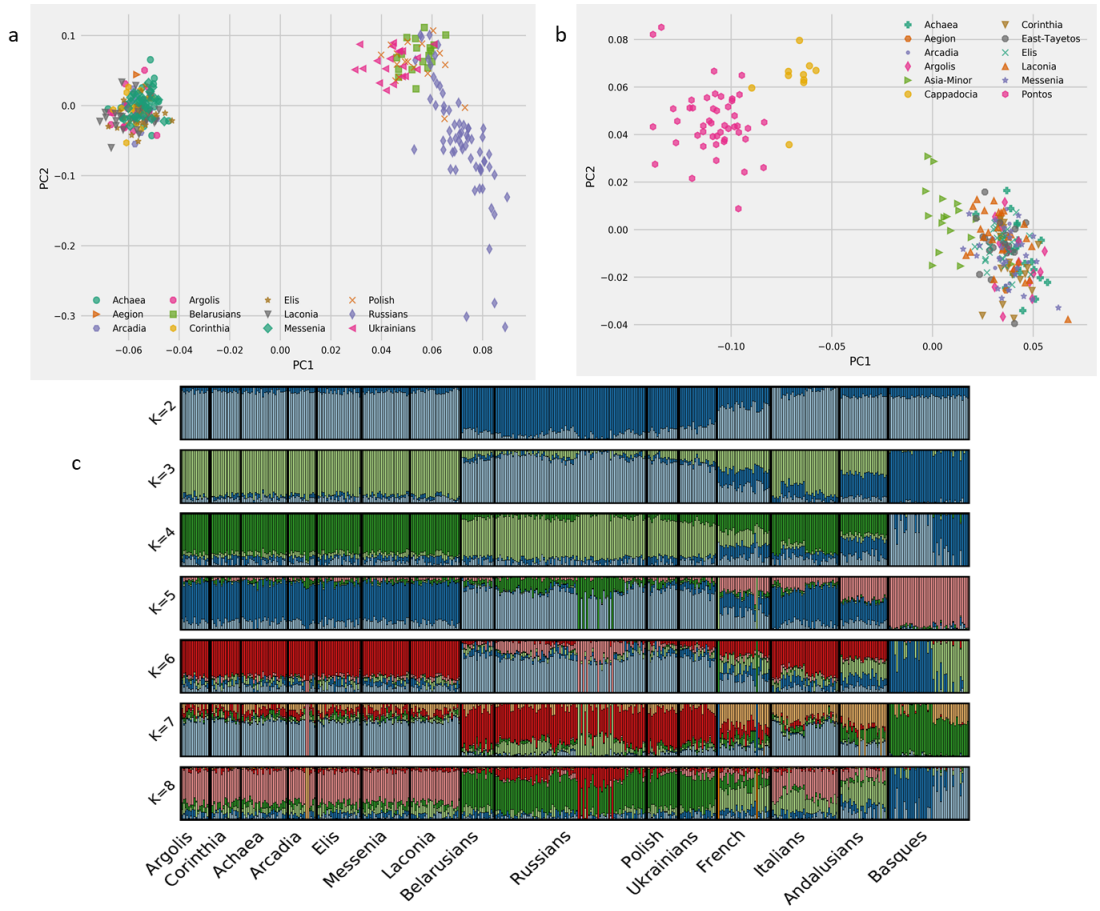


Figure 3.3.: Testing the theory of replacement of medieval Peloponesians by Slavs and Asia Minor settlers. (a) PCA analysis shows the broad separation of Peloponneseans from four populations of the Slavic homeland (Ukrainians, Polish, Russians and Belarusians). (b) PCA comparisons of the Peloponneseans with three Greek-speaking Asia Minor populations shows only partial overlap with the population of the Asia Minor Aegean coast. (c) ADMIXTURE analysis illustrates the wide separation of Peloponneseans from the Slavs in all values of K.

The theory of extinction of the medieval Peloponnesean Greeks allows for specific predictions about the genetic ancestry of the Peloponneseans. The great majority, if not all, Peloponnesean ancestry should be Slavic. We compared, the Peloponneseans (except for Maniots and Tsakones) with populations of the Slavic homeland from which the sixth century Slavs should have originated. The exact location of the

Slavic homeland is debated [63,64] but it is placed north of Danube [63], between the Oder and Dnieper rivers and includes areas inhabited by Polish, Ukrainian, Russian and Belarusian populations. PCA analysis showed a clear separation of Peloponneseans from the Slavs (Figure 3.3a). By ADMIXTURE analysis (Figure 3.3c) the Peloponneseans and the Slavs form separate clusters with a small degree of gene flow from the Slavic to the Peloponnesean cluster.

Fallmerayer hypothesized that the Hellenization of the Peloponnesean Slavs was accelerated by the transfer to the Peloponnese of Hellenized populations from Asia Minor [65]. We tested this hypothesis by comparing the Peloponneseans with three Greek speaking populations of Asia Minor: a western /coastal population sample extending from the Propontis in the north to Alikarnassos (Bodrum) in the south; a northern population from Pontus ie the coast of Black Sea and the Asia Minor interior corresponding to the current northern Turkey; and a central Anatolian population from Cappadocia. All these populations are separated from the Peloponneseans by PCA (Figure 3.3b). The small degree of overlap between Peloponnese and the population of the Asia Minor coast (Figure 3.3b) is expected for Greek populations. The Byzantines frequently moved Armenians to achieve political objectives [73]. Peloponneseans differ from the Armenians by PCA and ADMIXTURE analysis (Figure A.2). Collectively, these results are incompatible with the theory of extinction of the medieval Peloponneseans and their replacement by Slavic and Asia Minor settlers.

The Medieval Ancestry of the Populations of Mani

The Maniots differ from all other Peloponneseans by PCA (Figure 3.1b) and ADMIXTURE (Figure 3.1c) analysis. They also differ from mainland, island and Asia Minor Greek populations (data not shown) and from all the other populations we have compared them. By PCA analysis they overlap partially with Sicilians and Italians (Figure 3.2b).

In his treatise on the administration of the Byzantine Empire [74], the Emperor Constantin Porphyrogenitus describes how two Slavic tribes, the Mellingi and the Ezeritae, were forced by the Byzantines to withdraw to the slopes of Tayetos. Because of the writings of Porphyrogenitus we sampled separately the populations inhabiting the East and the West slopes of the Tayetos and the Deep Mani. By PCA (Figure A.3a) and ADMIXTURE (Figure A.3b) the populations of Tayetos are distinct from the populations of the Slavic motherland. Fallmerayer argued that the inhabitants of Deep Mani are of Slavic origin (9). PCA and ADMIXTURE analysis makes this hypothesis unlikely.

As an alternative origin of the Maniots Fallmerayer proposed that they are descendants of Mardaites [65]. This medieval warrior tribe used to inhabit the mountainous regions between Asia Minor and Syria but in late seventh century CE was resettled by the Byzantines in Asia Minor and other areas of the Empire [73]. The Mardaites have disappeared from the history but oral tradition claims that they are the ancestors of the Maronites of Lebanon, although this claim has been disputed [75]. PCA or ADMIXTURE analyses failed to show any close relationship between Maniots and the Maronites (Figure A.4).

The Question of Slavic Ancestry of Tsakones

The Tsakones of the eastern slopes of Mount Parnon differ from all other Peloponneseans (Figures 3.1b and 3.1d) and from all other populations we have compared them (Figure A.5). They used to speak a dialect of Doric origin (28) which was not comprehended by the other Peloponneseans. Their name was considered by medieval authors to represent a corruption of the word Lacones (Tsakones = Lacones). Fallmerayer argued against a Doric origin of the Tsakones and, instead, proposed that they were the descendants of a Slavic tribe that had migrated to Peloponnese before the flood of the Slavic settlers reached the peninsula. PCA (Figure 3.2b) and ADMIXTURE (Figure 3.3c) analyses argue against Slavic origin of the Tsakones.

Quantitative Assessment of the Ancestry of Peloponneseans

Table 3.2.: Shared ancestry between Peloponnesean populations and Slavic, Italian and other European populations.
(The first number for each pair of populations indicates the average shared ancestry for values of K between 4 and 8, while the number in parenthesis indicates the standard deviation)

Populations	Belarusians	Russians	Polish	Ukrainians	French	Italians	Basque	Andalusians
Argolis	5.4 (1.5)	12.2 (1.2)	5.8 (0.8)	6.8 (1.1)	39.1 (19.2)	94.7 (4.8)	2.8 (1.4)	60.5 (5.9)
Corinthia	5.9 (1.7)	13.0 (1.3)	6.3 (1)	7.5 (1.3)	41.2 (18.5)	94.9 (4.0)	3.1 (1.7)	62.0 (5.9)
Achaea	6.5 (1.7)	13.8 (1.1)	7.0 (0.8)	8.1 (1.1)	41.4 (18.4)	94.8 (4.0)	2.7 (1.4)	61.3 (5.8)
Arcadia	5.3 (1.8)	10.9 (2.4)	5.2 (1.2)	6.2 (1.5)	39.1 (18.2)	85.4 (14.6)	2.4 (1.4)	53.8 (9.1)
Elis	6.1 (1.3)	13.1 (1.2)	6.5 (0.8)	7.6 (1.1)	41.4 (18.3)	95.0 (3.3)	3.3 (1.7)	61.6 (5.6)
Messenia	6.7 (1.7)	14.4 (1.2)	7.3 (0.9)	8.5 (1.2)	42.6 (18.4)	95.2 (4.0)	2.7 (1.3)	61.8 (5.7)
Laconia	4.8 (1.2)	11.4 (1.5)	5.2 (0.9)	6.4 (1.1)	41.1 (14.6)	96.1 (2.3)	2.3 (1.4)	59.8 (5.6)

To quantify the findings of the ADMIXTURE analyses, we employed a method for the meta-analysis of the ADMIXTURE output that treats the output as a set of vectors in a K-dimensional space (for a particular value of K between four and eight). Each population is then summarized by a single vector (using PCA) and vector space calculus is used in order to identify the percentage of ancestry of a target population that is captured by one or more reference populations. It is worth noting that our choice to summarize each population by a single vector is akin to computing the mean ADMIXTURE output for a particular population. In most cases, ADMIXTURE returns a homogenous structure for a particular population and thus the top principal component is a good summary of the sample vectors returned by ADMIXTURE. First we focused on the ADMIXTURE analysis of Figure 3.3c which includes seven Peloponnesean populations (Argolis, Corinthia, Achaea, Elis, Arcadia, Messenia, Laconia), four Slavic populations (Belarusians, Russians, Polish, and Ukrainians), three Southern European populations (Italians, Basque, and Andalusians), and the French.

The results of Table 3.2 show that there is considerably more shared ancestry between the Peloponneseans and the French, Andalusians, and Italians compared to the shared ancestry between the Peloponneseans and the Slavic populations. The average shared ancestry with French ranges from 39 to 42%; with Andalusians 53 to 62%; with the Italians from 85 to 96%. In contrast, the average shared ancestry with the Slavic

populations is always less than 15%. Therefore, the Peloponneseans are genetically much more distinct from the Slavic populations and are much more similar to Southern European populations. We also observe that the Basques, (a population that is well-known to be isolated and genetically different from even its neighboring populations, like the Spaniards and the Andalusians) are very distinct from all populations in our analysis, which is precisely why we included them in these ADMIXTURE meta-analyses: on the average Basques share less than 4% of common ancestry with any Peloponnesean population. Notice that this number is relatively close to the average ancestry shared between the Peloponnesean populations and the Belarusians, Polish, and Ukrainians; all these populations share between 5.2% and 8.5% of common ancestry with the Peloponnesean populations. These Slavic populations are, from a genetic perspective, approximately as far apart from the Peloponneseans as are the Basques.

Table 3.3.: Shared ancestry between the populations of Mani and Tsakonia and Slavic, Italian and other European populations.
(The first number for each pair of populations indicates the average shared ancestry for values of K between 4 and 8, while the number in parenthesis indicates the standard deviation)

	Belarusians	Russians	Polish	Ukrainians	French	Italians	Basque	Andalusians
Deep Mani	0.7 (0.1)	1.6 (0.7)	0.9 (0.4)	1.0 (0.3)	6.4 (3.5)	25.3 (21.7)	0.3 (0.2)	7.6 (5.1)
West Tayetos	4.9 (5.1)	8.6 (6.9)	6.8 (5.4)	6.5 (5.7)	16.4 (12.7)	41.5 (32.5)	0.6 (0.5)	15.2 (11.1)
East Tayetos	5.7 (3.4)	10.9 (4.0)	7.9 (3.7)	8.0 (3.7)	27.7 (4.8)	58.0 (20.7)	2.0 (1.4)	27.0 (4.3)
North Tsakonia	3.9 (1.7)	8.2 (2.1)	5.0 (2.2)	6.0 (2.2)	26.7 (3.5)	51.2 (4.6)	1.5 (1.1)	26.9 (3.5)
South Tsakonia	0.2 (0.0)	0.9 (0.4)	0.4 (0.1)	0.6 (0.2)	4.1 (2.9)	14.2 (11.0)	0.2 (0.1)	5.3 (3.8)

We next determined the shared ancestry between the five distinct Peloponnesean populations (Deep Mani, West and East Tayetos, North and South Tsakonia), and the Slavs, the southern European populations, the French and the Basque. The ADMIXTURE plot of Figure A.3b and the data of Table 3.3 show that the amount of shared ancestry between these five Peloponnesian populations and the Slavic populations is very low. The ancestry Deep Mani shares with Belarusians, Polish and Ukrainians ranges from 0.7 to 1.0%. East and West Tayetos share from 4.9 to 8.6 % ancestry with these three Slavic populations which is five to eight times higher than that of

Deep Mani but slightly lower to the ancestry the other Peloponnesans share with the Slavs. Slightly lower, compared to the other Peloponneseans, is the ancestry shared between West/East Tayetos and the Russians (8.6 to 10.9%). The ancestry North and South Tsakonia share with the Slavs ranges from 4 to 8% and 0.2 to 0.9% respectively. Compared to the very low ancestry shared with the Slavs, South Tsakonia and Deep Mani share 14% and 25% ancestry with the Italians. North Tsakonia, East and West Tayetos share from 41 to 57% ancestry with the Italians. Again, the Basques are isolated from the five Peloponnesean populations.

3.1.4 Discussion

Our analysis of the genetic ancestry of the Peloponnesean populations and their relationships with the Slavs and other Europeans settles a historical controversy that has persisted for over 170 years. This controversy is typical of the problems historians face in their efforts to reconstruct history on the basis of inadequate written sources. Fallmerayer based his theory of extinction of the medieval Peloponnesean Greeks on the writings of early and two middle-medieval Byzantine authors. The early sources were very short comments in texts of sixth and seventh century historians and ecclesiastic authors [65]. The middle medieval documents were a letter by an eleventh century Patriarch of Constantinople and the writings of tenth century Emperor Constantine Porphyrogenitus. Fallmerayer's theory created sensation among historians. An early rebuttal was published by the Greek historian Paparigopoulos who examined the same sources Fallmerayer have used to construct his theory and reached the opposite conclusions i.e. that there was no evidence that the Slavs had reached the Greek proper during the sixth century and, when they arrived, they did not slaughter the local population. The many historians who have contributed to the very extensive literature on this topic during the last century (partially summarized in Curta [63, 66]) usually accept or reject the theory of extinction of the Peloponnesean Greeks. It seems that personal philosophies influence the historians'

judgment. Fallmerayer was an educator and journalist turned historian, a liberal intellectual for his time and a slavophobe who feared the increasing influence of Russia in the Balkans at the expense of the Ottoman Empire. Paparigopoulos was a Greek historian who was promoting the idea of the continuity of the Greek ethnicity during the medieval period. The findings of our study settle these issues and provide a direct test of the theory of the extinction of the medieval Peloponnesean Greeks. It is clear that the Slavs settled in Peloponnese, as the quantitative measurements of Slavic ancestry indicate (Tables 3.2 and 3.3). It also seems that their numbers were relatively small compared to size of the local population as the levels of Slavic ancestry the Peloponneseans indicate.

In his book on the Administration of the Empire [74] Constantin Porphyrogenitua describes the wars between the Byzantines and two Slavic tribes, who initially had settled the lowland Laconia but were forced to withdraw to the security of the slopes of the mount Tayetos, in order to avoid subjugation to Byzantine rule. The writings of Porphyrogenitus leave the impression that the slopes of Tayetos were Slavic lands. However, our analyses show that the levels of Slavic ancestry in the population of Tayetos are very low (Table 3.3). The most reasonable interpretation for the discrepancy between the medieval text and the genetic data is that the size of the Slavic settlements were small and the initial Slavic population was diluted by migration from the Deep Mani during the four centuries of Frankish and the almost three centuries of Ottoman occupation of the Peloponnese. In spite its inhospitable environment, Deep Mani was densely populated as Ottoman and Venetian censuses document. Historical evidence for high mobility and migrations of Maniots is available [76] and a gene flow path from Deep Mani to the slopes of Tayetos is suggested by our PCA analysis and the correlations between geographic coordinates and principal components.

The striking difference between the Tsakones and the remaining Peloponneseans on PCA and ADMIXTURE analysis can be best explained by isolation by distance. Geographic isolation explains the retention of their dialect. It should be mentioned

that in ancient times the area of Tsakonia, then called Cynouria, was inhabited by Doric speaking Ionians [62]. Isolation by distance is also the likely explanation of the findings in the populations of Mani. Porphyrogenetus in his writing about the Slavs of Tayetos also asks what happened to the ancient inhabitants of Laconia, the Hellenes who continued to adhere to the ancient Greek religion [74]. He finds them withdrawn in the inhospitable, agriculturally poor and rocky area of southern Tayetos, the area which we refer to here as the Deep Mani. Ancient DNA studies could perhaps test whether there is any relationship between the Maniots and the ancient Lacons or Tsakones and ancient Ionians.

To precisely determine the shared ancestry between groups of populations we devised a new approach that quantifies the output of methods such as ADMIXTURE. Our approach is linear algebraic in nature and has not appeared in prior work; as a matter of fact, to the best of our knowledge, such meta-analyses of the output of methods such as ADMIXTURE are missing from current literature. Indeed, our method is broadly applicable in determining shared ancestry between populations.

3.2 Integrating Linguistics, Social Structure and Geography to Model Gene Flow in India

3.2.1 Introduction

The genetic structure of human populations reflects gene flow around and through geographic, linguistic, cultural, and social barriers. The intricate tapestry of population substructure and complexity in India undoubtedly showcases the interplay among these evolutionary forces; 3,200 km from North to South, complex topography with elements ranging from the Himalayas to the Thar desert, plateaus and rainforests, almost 800 spoken languages and a strict system of endogamy together with a long history of migrations and invasions are factors that have shaped extant human genetic diversity within India. Numerous studies have attempted to dissect the genetic components and origins of the Indian populations [39, 77–86], including recent thrusts using genome-wide data from ancient individuals from Central and South Asia [87] and present day individuals from Northwest India [88]. However, to date, no study has attempted to model how the evolutionary forces acted in concert and to evaluate the relative contribution of each one towards establishing Indian genomic substructure.

Analysis of genetic structure has shown that Indian ethnic populations when grouped as tribal versus non-tribal, or by geographical region, or by linguistic affiliation, have resulted from admixture of four or five ancestral populations [79, 83, 89]. They represent Indo-European (IE) speakers in Northern India, Dravidian (DR) speakers in Southern India, Austroasiatic (AA) speakers in Central and Eastern India and Tibeto-Burman (TB) speakers in Northeast and the Andaman islanders. These ancestral components are attributed to the four distinct language families prevalent in India, spread over 22 official languages following a distinctive demographic spread.

In addition to the original African source population, West Asia (by demic diffusion of agriculture) and Central Asia have been shown as the major contributors to the

Indian gene pool [87, 90]. At the same time, within India, a rigorous system of social stratification has been in place, governing mate-exchange between social strata [91]. The caste system has been documented since 1500-1000 BC and imposes strict rules of endogamy over the past several thousands of years. Social stratification within India may be summarized into the so-called Forward Castes and the Backward Castes (as connoted in the Indian constitution), while 8.2% of the total population belongs to Scheduled Tribes and represents minorities that lie outside the caste system, still largely based on hunting, gathering and subsistence agriculture, with no written form of language. It has been shown that prior to the establishment of this strict endogamy within social groups, there was wide admixture among them, which came to an abrupt end 1,900 to 4,200 years before present [81]. In this text, we refrain from using broad-brush terms such as forward and backward, and instead define Social Group A (SGA) and Social Group B (SGB) for the so-called Forward Castes and Backward Castes, respectively. For the semi-nomadic tribes in India who are hunter gatherers or depend on subsistence farming for livelihood, we use Social Group C (SGC).

We set out to explore how the complex interplay of geography, spoken language and social structure have shaped the patterns of genetic variation in India. In doing so, we designed a quantitative framework for the evaluation of the relative contribution of different geodemographic, linguistic and social factors to the architecture of the genetic pool of human populations. Earlier attempts to investigate the covariance of allele frequencies and non-genetic factors on genetic structure, either depended heavily on assumptions and a computationally expensive Bayesian framework [92] or did not provide any statistical significance or feature selection to identify the most relevant structure-related factors [93]. Our findings lead to a model that explains human genetic substructure and quantifies the contribution of languages and social factors towards genetic diversity. Our work provides the first model to study the significance of each underlying factor on the genetic substructure of a population. We show that spoken language along with social stratification, rather than geography, appear to be the most significant influences on Indian genetics. We developed COGG (Correla-

tion Optimization of Genetics and Geodemographics), which models the population structure as a function of environmental and ecological factors along with geography. On top of COGG, we used a greedy feature selection technique to identify the most significant factors influencing genetic variation in India.

To further study the interplay between these factors, we propose a simple analytic procedure using the so-called Ridge Leverage Score (RLS) statistic that highlights the most significant population groups in India. This statistic helps us to better understand the intricate details of admixture, sub-structure, and genetic variation across social and language groups in the Indian subcontinent. The ability to correlate genomic background with sociolinguistic and cultural differences opens new avenues to study genomic structure of extant human populations.

Testing an old hypothesis regarding the northward migration of the DR speakers, we ran `qpAdm` [40] tests to find that southern DR speakers with an ancient basal group associated with fellow DR dry-land farmers in the south and admixed with IE around the Gujarat region in the north. This confirms that the Gujarati populations have gene flow *from* southern Indian dry-land farmers with an overlying admixture of IE speakers and a basal group in South Asia which existed prior to the IE arrival. In summary, the relationship between different social groups of India is studied in detail highlighting the autochthonous origin of the caste system in India. Furthermore, we recover ancient routes of migrations into India, for the IE, TB, and AA speakers and a northward movement from the southern DR speakers.

3.2.2 Materials and Methods

Study design and datasets

We used PLINK [94] to assemble genome-wide data for 835 samples from 84 well-defined sociolinguistic groups (see Supplementary Table 2a) genotyped on a 48,225 SNPs. These samples were collected from various sources [39, 80, 81, 84, 95] with the

consent of the corresponding authors. We did not use the Indian samples for the 1000 Genomes [4] project because of unavailability of their geographical coordinates as well as caste and language information. Additionally, three (GIH, STU, ITU) out of the five Indian population groups in the 1000 Genomes project were collected from Indian diaspora living in the USA (Houston) and the UK and might be biased and/or lead to gross underestimation of genetic diversity.

As the consolidated data set was put together from so many varied sources, there was an imbalance of social group and language family representation in the samples. For example, the TB language family has 93 members in the data set, which is considerably smaller than other language families (AA, IE, and DR have 131, 282, 333 members, respectively). To create the normalized data set, we removed the population group Garo from the TB dataset as the social group they belong to were unknown. Thus, the resulting dataset had 89 individuals from TB and we sub-sampled a similar number of individuals from the other three language families. The sub-sampling was done with respect to the social group affiliation and geographical locations. As AA and TB speakers are more homogeneously located in the forests and hills of Central, East, and Northeast India, and, on the other hand, IE and DR speakers are more spread across the northern and southern India, we sampled individuals in order to guarantee a balanced representation of geographical variance. We also made sure that all social groups are equally represented in the normalized data set. This resulted in having 368 individuals sampled across 33 population groups from all over India (Supplementary Table 2b). We created multiple normalized subsets of the original consolidated data set using the same technique to check for the robustness of our results. Indeed, all our analyses returned similar results with very minor changes in the squared correlation values. The normalized subset for which we have reported results for the Indian populations contains 368 samples from 33 populations genotyped for 48,326 SNPs. We converted all data to the same build (hg19) using LiftOver from the UCSC Genome Browser [96] and we merged the data sets and conducted further downstream analyses using PLINK [94,97]. We created the subset of the data

after checking for missing genotypes and filtering out variants with missing call rates exceeding 5%.

We merged reference populations from Eurasia and Southeast Asia, collected from various publicly available sources such as HGDP [1], the Estonian Biocenter [98–104] and the Allele Frequency Database (ALFRED) [105] with our normalized Indian dataset to create a merged data set of 1,516 samples from 73 population groups genotyped on 42,975 SNPs (Supplementary Table 2c).

To test ancient admixture scenarios, we used ancient samples from Near East, genotyped on Illumina Human Origins array [42] and merged them with the Indian samples to form a merged data set of 1,597 individuals across 31,130 markers.

PCA

We used TeraPCA [68] as well as our own MatLab implementation of PCA [67,69], after pruning for LD structure by setting `--indep-pairwise 50 10 0.4` in PLINK [94, 97]. We processed the data as discussed above in 2.1.4. We checked for outliers (using EIGENSTRAT's [37] outlier detection method) in the PCA plot and removed them keeping only autosomal biallelic SNPs with 95% genotyping rate and a minor allele frequency (MAF) of at least 5%.

Linear Discriminant Analysis

We implemented Rao's Discriminant Analysis which is directly based on Fisher's Linear Discriminant Analysis.

Correlation Optimization using Genetics and Geodemographics

Correlation Optimization of Genetics and Geodemographics (COGG) and maximizes the correlation between one of the top two principal components and the Geodemographic matrix, containing geographical coordinates, social groups and language information encoded as indicator variables. We restrict our encoding into three social groups (A,B and C) as described before, instead of the conventional derogatory terms that are widely used. We noticed that the Middle castes are very similar to the Forward castes, such as Kshatriya or Brahmins, hence, we labelled both Forward and Middle castes as SGA. Although the term Backward Class (as well as Scheduled castes and Scheduled Tribes) is used by the Government of India to classify social groups which are socially and educationally disadvantaged, we chose to call them SGB.

Let $\mathbf{u}$ is the m-dimensional vector containing either one of the top two principal components, computed by TeraPCA [68]; the Geodemographic matrix is denoted by $\mathbf{G}$.

$$
\mathbf{G} = \begin{array}{c}
\begin{array}{ccccccccc} G_1 & G_2 & G_3 & G_4 & G_5 & G_6 & G_7 & G_8 & G_9 \end{array} \\
\left[\begin{array}{c|c|c|c|c|c|c|c|c}
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathit{Latitude} & \mathit{Longitude} & \mathit{SGA} & \mathit{SGB} & \mathit{SGC} & \mathit{AA} & \mathit{DR} & \mathit{IE} & \mathit{TB} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}\right]
\end{array}
$$

The social groups (SGA, SGB and SGC) and Language (AA, DR, IE, TB) encoding was done as follows:

$$\textbf{Castes (or Languages)}=\begin{cases}1, & \text{if sample belongs to social group (or Language)}\\ 0, & \text{otherwise}\end{cases}$$

Let $\mathbf{a}$ be the k-dimensional vector whose elements are $a_1 \dots a_k$ (in our case, $k = 9$). COGG solves the following optimization problem:

$$\max_{\mathbf{a}} \textbf{Corr}\left(\mathbf{u}, \sum_{i=1}^{k} a_i \mathbf{G}_i\right). \tag{3.2}$$

Recall that $\mathbf{G}_i$ denotes the i-th column of $\mathbf{G}$ as a column vector. Let

$$d_i = \frac{\mathbf{u}^\top \mathbf{G}_i}{\sqrt{\textbf{Var}\left[\mathbf{u}\right]}}$$

for $i = 1 \dots k$ and let $\mathbf{d}$ be the vector of the d_i's. Also, let $M_{ij} = \mathbf{G}_i^\top \mathbf{G}_j$ for all $i, j = 1 \dots k$ and let $\mathbf{M}$ be the matrix of the M_{ij}'s. Then the optimizer for COGG is given by

$$\mathbf{a}_{\max} = \mathbf{M}^{-1}\mathbf{d}. \tag{3.3}$$

We obtain the above solution for COGG's optimization problem as discussed in Equation (3.2) by recalling the definition of the Pearson correlation coefficient. We can rewrite Equation (3.2) as

$$\begin{aligned}\max_{\mathbf{a}} \textbf{Corr}\left(\mathbf{u}, \sum_{i=1}^{k} a_i \mathbf{G}_i\right) &= \max_{\mathbf{a}} \frac{\mathbf{u}^T\left(\sum_{i=1}^{k} a_i \mathbf{G}_i\right)}{\sqrt{\textbf{Var}\left[\mathbf{u}\right]\textbf{Var}\left[\sum_{i=1}^{k} a_i \mathbf{G}_i\right]}} \\ &= \max_{\mathbf{a}} \frac{\sum_{i=1}^{k} a_i(\mathbf{u}^T \mathbf{G}_i)}{\sqrt{\textbf{Var}\left[\mathbf{u}\right]\sum_{i,j=1}^{k} a_i(\mathbf{G}_i^T \mathbf{G}_j)a_j}}.\end{aligned}$$

By definition, $\mathbf{M}$ is a square, symmetric positive definite matrix and hence its square root $\mathbf{M}^{1/2}$ is well-defined. We can now rewrite the above equation as

$$\max_{\mathbf{a}} \left(\mathbf{u}, \sum_{i=1}^{k} a_i \mathbf{G}_i \right) = \max_{\mathbf{a}} \frac{\mathbf{d}^T \mathbf{a}}{\sqrt{\mathbf{a}^T \mathbf{M} \mathbf{a}}} = \max_{\mathbf{a}} \frac{\mathbf{d}^T \mathbf{a}}{\|\mathbf{M}^{1/2} \mathbf{a}\|_2}.$$

To understand the last equality let $\|\mathbf{x}\|_2$ denote the Euclidean norm of the vector $\mathbf{x}$ and recall that: *(i)* since $\mathbf{M}$ is symmetric positive definite matrix, $\mathbf{M} = (\mathbf{M}^{1/2})^T \mathbf{M}^{1/2}$ and *(ii)* $\sqrt{\mathbf{x}^T \mathbf{x}} = \|\mathbf{x}\|_2$ for any vector $\mathbf{x}$, including $\mathbf{x} = \mathbf{M}^{1/2}\mathbf{a}$. Now assume that $\mathbf{M}$ is invertible and make the change of variable $\mathbf{p} = \mathbf{M}^{1/2}\mathbf{a} / \|\mathbf{M}^{1/2}\mathbf{a}\|_2$. Notice that $\mathbf{p}$ is a unit norm vector (its Euclidean norm is equal to one) and that

$$\mathbf{a} = \|\mathbf{M}^{1/2}\mathbf{a}\|_2 \mathbf{M}^{-1/2} \mathbf{p}. \tag{3.4}$$

Thus, we get:

$$\max_{\mathbf{p}, \|\mathbf{p}\|_2 = 1} \left(\mathbf{u}, \sum_{i=1}^{k} a_i \mathbf{G}_i \right) = \max_{\mathbf{p}, \|\mathbf{p}\|_2 = 1} \mathbf{d}^T \mathbf{M}^{-1/2} \mathbf{p}. \tag{3.5}$$

Using submultiplicativity and the fact that $\mathbf{p}$ is a unit norm vector,

$$\mathbf{d}^T \mathbf{M}^{-1/2} \mathbf{p} \leq \|\mathbf{d}^T \mathbf{M}^{-1/2}\|_2 \|\mathbf{p}\|_2 = \|\mathbf{d}^T \mathbf{M}^{-1/2}\|_2 = \sqrt{\mathbf{d}^T \mathbf{M}^{-1} \mathbf{d}}. \tag{3.6}$$

The last equality follows from the fact that $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T \mathbf{x}}$ for any vector $\mathbf{x}$. The above upper bound is true for any unit norm vector $\mathbf{p}$ and can actually be achieved by the vector $\mathbf{p}_{\max}$:

$$\mathbf{p}_{\max} = \frac{\mathbf{M}^{-1/2}\mathbf{d}}{\|\mathbf{M}^{-1/2}\mathbf{d}\|_2}.$$

Indeed, it is easy to verify that $\mathbf{p}_{\max}$ is a unit norm vector that satisfies

$$\mathbf{d}^T \mathbf{M}^{-1/2} \mathbf{p}_{\max} = \mathbf{d}^T \mathbf{M}^{-\frac{1}{2}} \frac{\mathbf{M}^{-1/2}\mathbf{d}}{\|\mathbf{M}^{-1/2}\mathbf{d}\|_2} = \frac{\mathbf{d}^T \mathbf{M}^{-1} \mathbf{d}}{\sqrt{\mathbf{d}^T \mathbf{M}^{-1} \mathbf{d}}} = \sqrt{\mathbf{d}^T \mathbf{M}^{-1} \mathbf{d}}.$$

Thus, from Equation (3.6), it follows that $\mathbf{p}_{\max}$ is a maximizer for the optimization problem of Equation (3.5). If we let

$$\mathbf{a}_{\max} = \mathbf{M}^{-1}\mathbf{d},$$

it is easy to see that the above values for $\mathbf{a}_{\max}$ and $\mathbf{p}_{\max}$ satisfy

$$\mathbf{a}_{\max} = \|\mathbf{M}^{1/2}\mathbf{a}_{\max}\|_2\mathbf{M}^{-1/2}\mathbf{p}_{\max},$$

as stipulated by the change of variables from Equation (3.4), and thus $\mathbf{a}_{\max}$ is a maximizer for COGG. Hence, we obtain the solution of the optimizer as defined in Equation (3.3).

We also remove the sparsity induced by the zero-one indicator variables by assigning 1, 2, and 3 for SGA, SGB and SGC groups, respectively, in the social category and similarly 1, 2, 3, and 4 for AA, IE, DR, TB affiliations in the language variable and got similar results by fitting the solution of COGG.

We checked for statistical significance of the results obtained by COGG by performing 1,000 iterations with randomly permuted values of the columns related to caste and language encodings in $\mathbf{G}$. We do not permute the columns corresponding to the geographical coordinates in order to maintain a baseline for the comparison. We randomly permuted the rows (individuals) corresponding to the seven columns (variables related to castes and language affiliations) in $\mathbf{G}$ and in each iteration we run COGG to find the optimal $\mathbf{a}_{\max}$ and the respective r^2 value.We find that random permutations return a maximal value which is significantly less than r^2 obtained by COGG (detailed discussion in Results). This clearly indicates the importance of the social group and language encodings in $\mathbf{G}$.

Prior work attempted to disentangle the effects of non-genetic variables such as geography, linguistics, subsistence, social or ecological factors from the genetic variables captured by the top principal components. One such study [93] regressed the top 20 PCs computed from the genotypes of the Khoe-San populations with various combinations of geographic, linguistic and subsistence covariates, and used cross-validation scores to understand which non-genetic variable can predict the observed genetic patterns. They observed that languages improve the predictive capacity of a model that includes only geography in the sub-Saharan and the Southern African dataset. This is similar to the intuition that COGG uses, but COGG provides a conceptually straightforward model to do an in-depth study to account for the factors within the broad generic non-genetic factors, such as which language and social group explain most of the genetic variation captured by the top principal components. Also, in

addition, we do a feature selection procedure to obtain the most significant variables in the geodemographic matrix, unlike previous studies. Another study [92] employed a Bayesian framework to isolate ecological factors from geographic distances. Broadly, COGG tries to achieve the same goal, but it provides the ease of use in this setting, where one can just encode the environmental and ecological factors as covariates and solve the underlying optimization problem to obtain the maximum correlation. Along with this, it is easier to comprehend, as it is closer to a linear regression setting.

Canonical Correlation Analysis

There is no mathematical reason to restrict COGG to the top two principal components and corresponding singular vectors (PC1 and PC2) of the genetic similarity covariance matrix. Prior work has exclusively focused on studying the correlation between longitude and latitude and the top two principal components; COGG goes beyond this by adding geodemographic features to study more general correlations. Our next method applies Canonical Correlation Analysis (CCA, introduced in [106]) to simultaneously study the correlation between the top q Principal Components (where q is a user-defined parameter) and the geodemographic matrix $\mathbf{G}$. CCA extracts linear components that capture correlations between two input datasets, in a manner analogous to PCA. From a statistical point of view, CCA extracts directions of maximal "correlation" between a pair of datasets represented by matrices. From a linear algebraic point of view, CCA measures the similarities between the subspaces spanned by the columns of each of the two datasets, represented by matrices [107]. In our case, we extend the optimization problem of equation 1 to identify the maximal correlation between $\mathbf{U}$, which is now an $n \times q$ matrix containing the top q left singular vectors of the genetic covariance matrix and $\mathbf{G}$, which is the geodemographic matrix described earlier. Formally, we define the following optimization problem, which we call COGG-CCA:

$$\max_{\mathbf{a},\mathbf{b}} \mathbf{Corr}\left(\sum_{j=1}^{q} b_j \mathbf{U}_j, \sum_{i=1}^{k} a_i \mathbf{G}_i\right), \tag{3.7}$$

where $\mathbf{b}$ is a p-dimensional vector whose entries are the b_j's and $\mathbf{a}$ is a k-dimensional vector whose entries are the a_i; $\mathbf{U}_j$ and $\mathbf{G}_i$ represent the j-th and i-th column of $\mathbf{U}$ and $\mathbf{G}$ as

column vectors. Solving COGG-CCA analytically dates back to the work of [106] and allows us to obtain the following closed form solution for the vectors $\mathbf{a}$ and $\mathbf{b}$, the unknown coefficient vectors associated with the matrices $\mathbf{G}$ and $\mathbf{U}$, respectively.

Let $\mathbf{\Sigma}_{UU} = \mathbf{Cov}[U, U]$, $\Sigma_{GU} = \mathbf{Cov}[G, U]$, and $\Sigma_{GG} = \mathbf{Cov}[G, G]$ denote three covariance matrices and construct

$$\mathbf{\Sigma} = \mathbf{\Sigma}_{GG}^{-1/2} \mathbf{\Sigma}_{GU} \mathbf{\Sigma}_{UU}^{-1/2}.$$

Then, $\mathbf{a}$ is the top right singular vector of the matrix $\mathbf{\Sigma}$ and $\mathbf{b}$ is the top left singular vector of $\mathbf{\Sigma}$; it is well-known that the maximum correlation coefficient is equal to the largest singular value of the matrix $\mathbf{\Sigma}$. Applying COGG-CCA on our data we obtain very high, statistically significant r^2 for $q = 8$ showing the prowess of including sociolinguistic factors to construct the population genetic structure of historically diverse populations.

Algorithm 1 OMP Algorithm for Feature Selection

Input: matrix $\mathbf{G} \in \mathbb{R}^{n \times k}$, column vector $\mathbf{U} \in \mathbb{R}^n$, $\epsilon > 0$

Output: matrix $\mathbf{C} \in \mathbb{R}^{n \times p}$ which has columns of $\mathbf{G}$ with indices in τ, $|\tau| = p$, $p < k$

$\tau \leftarrow \phi$; $r \leftarrow 0$; $\mathbf{U}^{(0)} \leftarrow \mathbf{U}$; $\mathbf{G}^{(0) \leftarrow \mathbf{G}}$;$\mathbf{C} \leftarrow \phi$
while $\|\mathbf{U}^{(r)}\|_2 > \epsilon$ **do**
  for $i \in \{1, 2, \cdots, k\} - \tau$ **do**
    choose $\mathbf{i}$ corresponding to maximum $\boldsymbol{corr}\left(\mathbf{U}^{(r)}, \mathbf{G}_i^{(r)}\right)$
  end for
  $\tau \leftarrow \tau \cup \{i\}$; $\mathbf{V} \leftarrow \mathbf{G}_i^{(r)}$
  remove column i from $\mathbf{G}^{(r)}$ to form $\mathbf{G}'^{(r)}$
  project $\mathbf{G}'^{(r)}$ onto the subspace orthogonal to $\mathbf{V}$, i.e., $\mathbf{G}^{(r+1)} \leftarrow \mathbf{G}'^{(r)} - \left(\mathbf{V}\mathbf{V}^{\dagger}\right)\mathbf{G}'^{(r)}$
  project $\mathbf{U}^{(r)}$ onto the subspace orthogonal to $\mathbf{V}$, i.e., $\mathbf{U}^{(r+1)} \leftarrow \mathbf{U}^{(r)} - \left(\mathbf{V}\mathbf{V}^{\dagger}\right)\mathbf{U}^{(r)}$
  $r \leftarrow r + 1$
end while
$\mathbf{C} \leftarrow \mathbf{G}_{\tau}$

Feature selection using Orthogonal Matching Pursuit (OMP)

We used a greedy feature selection algorithm described in [108] to select features in the Geodemographic matrix $\mathbf{G}$. It selects the column which results in the maximum r^2 value from $\mathbf{G}$ and then projects $\mathbf{G}$ (and $\mathbf{u}$) on the subspace perpendicular to the selected column in order to form $\mathbf{G}'$ (and $\mathbf{u}'$). We iterate the process until we remove the required number of features from $\mathbf{G}$. The precise algorithm is described in Algorithm (1) We obtain two sets of the three most significant features from the nine features in $\mathbf{G}$, one for PC1 and the other for PC2. All the values returned by this method are statistically significant, as random permutations of the elements of the features in S_1 and S_2 recover almost nothing. We also checked all $\binom{9}{3}$ possible sets of three features exhaustively and concluded that (for both PC1 and PC2) S_1 and S_2 return the maximum correlation.

Ridge Leverage Scores

We devised a simple method based on the Ridge Leverage Score (RLS) statistic in order to identify Indian populations that maximally contribute to the genetic diversity within the Indian sub-continent. We considered the genotype data, denoted by mean-centered (by SNPs) matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ where m is the number of individuals and n is the number of markers in the normalized subset of 33 Indian populations (approx 48K markers). We also considered the mean-centered Geodemographic matrix $\mathbf{G} \in \mathbb{R}^{m \times k}$, which includes k features across m individuals in 33 Indian populations.

The ridge leverage score of the i-th row of the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined as

$$\tau_i{}^{\lambda}(\mathbf{A}) = \left(\mathbf{A}\mathbf{A}^{\top} \left(\mathbf{A}\mathbf{A}^{\top} + \lambda \mathbf{I}_n \right)^{-1} \right)_{ii} \tag{3.8}$$

where $\lambda > 0$ is the regularization parameter.

Our analysis procedure based on the RLS statistic has four steps:

- We apply the RLS algorithm (see **Supplementary Note** for details) separately to the matrices $\mathbf{M}$ and $\mathbf{G}$ to find their corresponding row ridge leverage scores, denoted by $\tau_i^{\lambda}(\mathbf{M})$ and $\tau_i^{\lambda}(\mathbf{G})$, respectively, for $i = 1 \ldots m$.

- We grouped the RLSs by population groups to obtain a single score per group, defined as the median of the respective RLSs. If there are $T = \{t_1, t_2, \ldots, t_T\}$ populations in the normalized set of the Indian populations ($|T| = 33$ in our case), then we obtain $|T|$ RLSs in this manner, one per population t_i, defined as the $|T| \times 1$ vectors $\bar{\tau}^{\lambda}(\mathbf{M})$ and $\bar{\tau}^{\lambda}(\mathbf{G})$.

- Next, we compute an additive ridge leverage score for each population after normalizing the vectors obtained in the last step. This additive RLS highlights the significant rows (in our case, Indian populations), across both the genotype and the Geodemographic matrices. We define this consolidated additive RLS as,

$$\widetilde{\tau} = \bar{\tau}^{\lambda}(\mathbf{M}) + \bar{\tau}^{\lambda}(\mathbf{G}).$$

- Finally, we sort the entries of $\widetilde{\tau}$ in descending order to obtain a set of representative populations.

Estimating population admixture

We used the ADMIXTURE v1.22 software [35] for all admixture analyses and used our in house script to plot the admixture estimates. Before running ADMIXTURE, we pruned for LD using PLINK [94] by setting `--indep-pairwise 50 10 0.8`. To determine the optimal number of ancestral populations (K), we varied K between two and eight performing iterations until convergence for each value of K. We also performed a quantitative analysis (Section 3.1.2) of ADMIXTURE's output using a method described and implemented in [56]. To visualize the results of this quantitative analysis, we designed a color-coding scheme, where the highest shared ancestry between two populations is black and the lowest shared ancestry is white. All intermediate values of shared ancestry follow a gradient from white to black.

Three population statistics, `qpAdm`, network analysis, and TreeMix

We used ADMIXTOOLS [40] to compute f_3 statistics and `qpAdm` for our data sets to find signs of admixture using the `qp3Pop` and `qpAdm` programs respectively. To better visualize and understand the connection between the populations included in our study, we performed a network analysis on the results of ADMIXTURE, using a method presented by a previous study [70]. The parameters to generate the networks are the number of nearest neighbors (NN) and the number of Principal components (PC) to use. We varied NN from four to eight and PC from two to five. We report the network (**Figure 3.7**) for NN=5 and PC=5. Finally, TreeMix [109] was used to analyze the population divergence, mainly for the IE language dispersal into the Indian subcontinent. We used migration values from zero to eight and the `-k` flag to allow LD and set it to 1,000 SNPs to infer language dispersal routes.

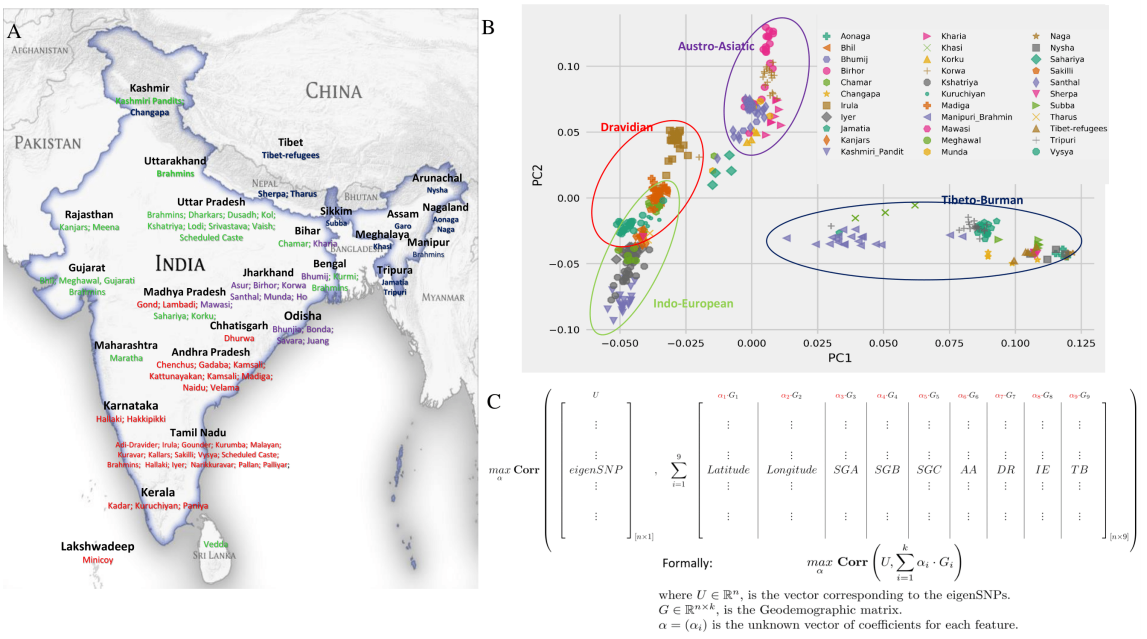


Figure 3.4.: A. Map of India showing the locations of the 835 Indian samples, from 84 well-defined population groups, that were used as the starting point of this study; B. PCA plot of the normalized dataset consisting of 368 individuals, genotyped on 48,373 SNPs shows language groups are clearly significant in the PCA plot and correlate well with the principal components; C. Framework of our approach for Correlation Optimization of Genetics and Geodemographics (COGG).

3.2.3 Implementation

COGG was implemented in `MATLAB` and is available to download with a GNU GPL 3.0 license at `https://github.com/aritra90/COGG`.

3.2.4 Results

Geography versus population structure within India

Starting from all publicly available autosomes from the Indian subcontinent (835 individuals, see Figure 3.4 and Supplementary Table 2) and unlike prior studies [80], we created a normalized data set (see Section 3.2.2 for details) over social groups, geographical locations, and language families that guarantees an approximately equal representation of each group (a total of 368 individuals from 33 populations genotyped across $48,373$ SNPs).

In other regions of the world, it has often been observed that individuals from the same geographical region cluster together with the top two principal components (PCs) being well-correlated with geography, namely longitude and latitude [7]. For instance, within Europe, the squared Pearson-correlation coefficient r^2 between the top singular vector of the genetic covariance matrix vs. latitude (north-south) was equal to 0.77 and 0.78 for the second singular vector of the same matrix vs. longitude (east-west). In order to explore whether Indian genetic information mirrors geography, we computed the top two PCs using smartpca [37] and plotted the top two left singular vectors of the resulting genetic covariance matrix (Figure 3.4 and Figure A.6 for the entire dataset), with the first and second PC explaining 32% and 15% of the total variance, respectively. It is straight-forward to observe that the IE and DR speaking populations form a long cline, while the AA and TB speakers form separate clusters. We computed the Pearson correlation coefficient (r^2) between the top two left singular vectors (we will denote them by PC1 and PC2) of the covariance matrix and the geographic coordinates (longitude and latitude) of the samples under study and we observed $r^2 = 0.604$ for PC1 vs. longitude and $r^2 = 0.065$ for PC2 vs. latitude. Thus, PC1 correlates well with longitude, but PC2 essentially entirely fails to correlate with latitude.These findings are in sharp contrast with findings within the European continent [7,

110] and highlight the need for social and linguistic factors to be accounted for, as noted in prior work [77,80,83,86,89], which argued that genetic stratification in India is particularly influenced by endogamy as well as language groups. ADMIXTURE analysis is consistent with previous studies (Figure A.7), showing high degrees of shared ancestry across all the social groups (Figure A.8), thus supporting the notion that a demographic shift from wide admixture to endogamy occurred recently in Indian history; indicating the autochthonous origin of the caste system in India. f_3 statistics show further evidence that most of the social groups in India are admixed across languages affiliations (Supplementary Table 3). The geographically isolated Tibeto-Burman SGC (TB_SGC) and the Dravidian speaking SGC (DR_SGC) appear to be the most isolated in India. Linear Discriminant Analysis (LDA) on the normalized data set clearly supports genetic stratification by social structure and languages in the Indian sub-continent (Figure A.9). Separate clines resembling IE, DR and TB SGA, respectively appear in order, followed by SGB and SGC. Thus, we see a two layer stratification, when LDA was run with language-caste groups.

Correlation Optimization of Genetics and Geodemographics

In order to understand the genetic substructure of India, considering the strongly endogamous social structure as well as the presence of multiple language families, we developed COGG (Correlation Optimization of Genetics and Geodemographics). COGG is the first deterministic method that correlates genomewide genotypes, as represented by the top two principal components, with geography (longitude and latitude) and sociolinguistic factors (caste and language information in this case). The need for such methods has been pointed out by many studies [77, 79, 80, 86, 89]. Given information on m samples, the objective of COGG is to maximize the correlation between the genetic component as represented by the top singular vectors of the genetic covariance matrix formed by the genotypic data and a matrix containing information on geography, castes, tribes, and languages for each sample (Figure 3.4).

Solving the optimization problem underlying COGG (see 3.2.2) and plugging in the solution, we obtain a Pearson correlation coefficient $r^2 = 0.93$ for PC1 vs. $\mathbf{G}$ and $r^2 = 0.85$ for PC2

vs. **G**. Thus, we observe almost perfect correlation with PC1 and PC2 representing the genetic structure of the Indian subcontinent using the Geodemographic matrix **G** instead of just longitude and latitude: the values of r^2 increase from 0.6 to 0.93 for PC1 and from 0.06 to 0.85 for PC2. This massive improvement came from considering endogamy and language families, two attributes that are pivotal in studying the genetic stratification of Indian populations. The results are statistically significant (Figure A.10) over 1,000 iterations with permutation of the variables related to social factors and languages (see 3.2.2 for details). We randomly permuted the rows (individuals) corresponding to the seven columns (variables related to castes and language affiliations) in **G** and in each iteration we run COGG to find the optimal $\mathbf{a}_{\max}$ and the respective r^2 value. We find that the random permutations return a maximal value of r^2 equal to 0.6422 for PC1 and 0.1679 for PC2 (Supplementary Figure 5). This is a minor increase from 0.6 and 0.06 respectively for PC1 and PC2, clearly indicating the importance of the caste and language encodings in **G**.

We further explored an extension of COGG in order to jointly analyze multiple PCs simultaneously and not just each component individually. To do this we employed Canonical Correlation Analysis (CCA), a well-studied statistical technique, which maximizes the correlation between the genetic and the Geodemographic matrices by jointly finding linear combinations of the variables in each matrix. We used the top eight PCs of the genetic matrix as the results did not improve significantly, beyond that (Figure A.11). We note that these eight PCs capture, collectively, 88.9% of the variance of the genetic matrix. Let **U** denote the matrix containing the top eight principal components and let **G** be the same Geodemographic matrix as before. Running COGG-CCA on these inputs returns a statistically significant (Figure A.11) r^2 equal to 0.94 (which is well above the $r^2 = 0.6$ obtained when COGG-CCA was ran without including the sociolinguistic factors.

Identifying the features that drive population structure within India

In order to formally investigate which of the nine features (columns) in the Geodemographic matrix **G** contribute more in the optimization problem posed by COGG, we used the sparse approximation framework and the Orthogonal Matching Pursuit (OMP) algorithm from

applied mathematics [108] (Algorithm 1). Running OMP on our dataset we obtain two sets of three features each, S_1 and S_2, for PC1 and PC2 respectively:

$$S_1 = \text{AA, TB, SGA, } and$$
$$S_2 = \text{AA, Latitude, SGA.}$$

Plugging in S_1 as the reduced feature space in COGG resulted in $r^2 = 0.92$ for PC1 with S_1 and $r^2 = 0.85$ for PC2 with S_2, respectively; these values are capturing approximately over 99% of the values returned by COGG when all the features in G are included.

The feature selection algorithm identifies the AA and TB language groups to be the significant features. These language groups (AA and TB) consist of mostly tribal nomadic hunter gatherers who dwell in the hills and forests of Central-eastern and North-eastern India, respectively. Thus, the AA and TB language groups automatically capture SGC. Another significant feature was SGA; SGA spans across most of the IE and DR speakers found across northern and southern India. Thus, these three features encompass most of the geographical, social and linguistic diversity found in the Indian subcontinent and highlight the demographic interplay.

Significant ethnic groups capturing genetic diversity in India

We developed a simple approach to identify influential (from a genetic perspective) Indian populations, based on the Ridge Leverage Score (RLS) statistic of [111] (see 3.2.2). We applied our approach to the genotype matrix of the normalized dataset, as well as on the corresponding Geodemographic matrix, to identify population groups capturing the genetic variation of the Indian subcontinent. Pan-Indian nomadic hunter gatherers, represented as SGC and SGB, across language families, are found to encapsulate much of the genetic structure of the subcontinent.

The following ethnic groups are all found to be significant in the light of our analysis: TB speaking Changpas, who are semi-nomadic pastoralits dwelling in the high altitudes

of Tibet and Ladakh in India; AA speaking Mundas spanning the forests of Central and Eastern India; IE speaking Meghawals situated in the northwestern states of India as well as in Pakistan and continue to live in mud-brick huts; DR speaking Madigas in the southern states of India, who are also listed as *Dalits* (belonging to SGB). Table 3.4 shows the most significant populations returned by the RLS statistic when applied on the normalized set of the Indian populations. These populations spread across the entire subcontinent and

Table 3.4.: Top ten significant ethnic groups in India capturing the genetic structure of the subcontinent as reflected by the RLS statistic (* Vysyas are classified as in between SGA and SGB [81]).

Population group	State/Territory	Language family	Social group
Changpas	Jammu and Kashmir	TB	SGC
Vysya	Andhra Pradesh	DR	SGA *
Munda	Jharkhand and Odisha	AA	SGB
Mawasi	Madhya Pradesh	AA	SGB
Meghawal	Rajasthan	IE	SGB
Sahariya	Uttar Pradesh	IE	SGB
Sakilli	Tamil Nadu	DR	SGB
Korku	Madhya Pradesh	AA	SGB
Madiga	Andhra Pradesh	DR	SGB
Sherpa	Nepal	TB	SGC

consist of mostly SGC and SGB, who are semi-nomadic groups dwelling in forests and remote areas in India. Some AA_SGC listed here (such as the Mawasi and the Korku) are northern Mundari speakers and have an Ancestral North Indian (ANI) component, an Ancestral South Indian (ASI) component, and an ancestral South-East Asian component (SEA) [112]. Vysyas have been shown to have a founder event going back 100 generations, due to the strong imposition of endogamy [39]. They are almost equally admixed between DR_SGC and IE_SGA (Supplementary Table 3 and Figure A.7), indicating that they were mixing across social groups prior to becoming endogamous. The RLS statistic highlights these populations and indicates that these groups have significant contributions in shaping the genetic diversity of India and are thus important candidates to be studied further in more detail.

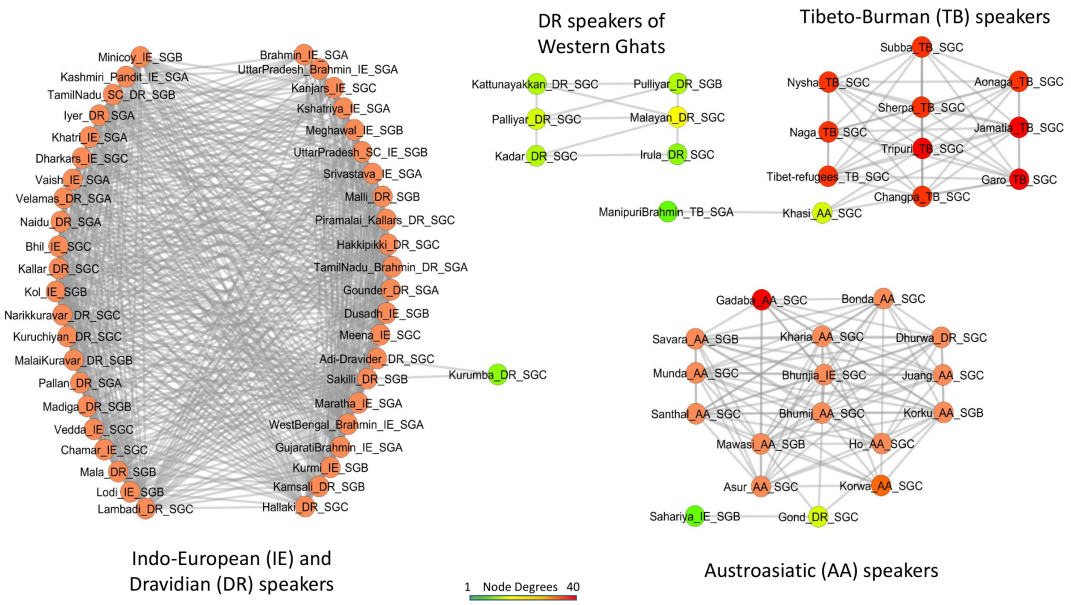


Figure 3.5.: Population network analysis of all Indian populations reveals four isolated clusters, representing language groups (40% of edges are shown).

Running COGG with the significant ethnic groups as shown in Table 3.4 further confirmed the importance of of these populations in shaping Indian genetics. The r^2 value between geographical coordinates and the PCs came out to be 0.21 for PC1 and 0.08 for PC2, when ran with populations from Table 3.4. When the same populations were ran with COGG, the values returned were $r^2 = 0.853$ for PC1 and the geodemographic matrix $\mathbf{G}$ and $r^2 = 0.794$ for PC2 and $\mathbf{G}$. Thus, COGG returns very high correlations using only the populations selected using the RLS statistics, capturing most of the variance reflected by the top PCs of the genetic matrix.

Relationship between social groups

Our analyses using COGG clearly support the fact that language families and endogamy within social groups have played a significant role in shaping the genetic structure of the Indian subcontinent. We further explored the relationship between SGC with the endogamous SGA and SGB in order to reconstruct the population history of these populations.

The SGA populations across languages share approximately 85% average ancestry with SGC belonging to the same language. The IE and DR language groups show more homogeneity in shared ancestry than the TB and AA groups. This supports the notion that there was mixture between IE and DR speakers across SGA and SGB (Figure A.8) around 1,900 to 4,200 years ago [81] and that the caste system originated from a "classless" society which became hierarchical with the knowledge of agriculture [78,113]. Some DR_SGC populations such as Irula, Kadar and Paniyas show divergence from the rest of the Indian population (Supplementary Figure 8b, Supplementary Table 3). Irula and Kadar, who are nomadic people residing in forests in southern India share 90% ancestry between themselves, forming an isolated cluster among the DR speaking groups. Paniyas show isolation from other DR_SGC groups with considerably smaller amounts of shared ancestry with all IE and DR speakers. To better illustrate the intricacies in the relationships between the social groups in India, we constructed a network of all the population groups under study (**Figure 3.5**), with their weights resembling the shared ancestry between them.

The shared ancestry network (Figure 3.5), revealed four clusters: (i) A cluster of IE and DR groups across social groups resembling a nearly complete graph with over 60% of all possible edges present. (ii) Few DR_SGC populations such as Kadar, Irula, Malayan, Palliyar, etc. formed a connected component, isolated from the main IE-DR cluster. (iii) All AA populations formed an almost complete graph where few groups such as Mawasi, Korku, and Korwa were connected to the Gonds and Sahariya (who contain ∼71% of AA ancestry). The Gonds are a candidate mosaic Indian population, containing ∼51% AA, ∼36% DR, and ∼13% IE ancestry (Supplementary Table 4), which can be attributed to their central location in India [114]. They act as the bridge between the AA clique merging into the IE-DR cluster when we allow 60% edges in the network. (iv) A cluster formed by all TB speakers, with Naga, Garo, and Tripuri being connected to the Khasis, who are an AA speaking group residing in northeastern India. It is connected to the Manipuri Brahmins, who are known to have significant admixture from IE_SGA (Supplementary Table 3). The clusters confirm that languages play a very significant role in shaping the Indian gene pool.

We try to dissect each cluster to further analyze the divergent populations as well as the cohesive forces acting towards making the Indian subcontinent a melting pot of different demographics.

Homogeneity of IE and DR speakers

Network analysis within India (Figure 3.5) revealed that the IE and DR cluster resembles an almost complete graph, indicating the significant amount of shared ancestry between all involved populations. This is further corroborated by running outgroup f_3 statistics [40,41] (with Yorubans in Nigeria (YRI) from the 1000 Genomes phase3 dataset as the outgroup [4]) to find the shared genetic drift between all Indian populations (Figure A.13). In this analysis, we focused on all of 84 ethnic groups from India consisting of 835 individuals. We observe that IE and DR populations across social group affiliations share substantial ancestry. Focusing on just the IE and DR speakers across social groups, we created a subset of our data encompassing 510 individuals across 22K markers. They form a long cline when plotting the first two PCs (Figure A.14), with only few populations showing divergence. This follows from the network analysis (Figure 3.5) as well as the f_3 tests (Supplementary Table 6) described above, which also showed a separate cluster formed by DR_SGC populations(also shown in Figure A.12). A visibly divergent groups in DR speakers are Paniyas, owing to their remote location in Wayanad, Kerala, a southern Indian state. Another divergent group in all statistics were Tharus, who are admixed between TB_SGC and IE speakers Figure A.15) and has dual ancestry with with one-half of their gene pool being East Asian, whereas the other half is South Asian [115]. The homogeneity between the IE and DR speakers across social groups is best reflected in the quantitative assessment of the ADMIXTURE analysis (Figure A.16).

Archaeologically, India has been influenced by a period of population movements with a number of demic characteristics such as agriculture, husbandry, ashmounds, etc. The interaction between groups prior to the imposition of endogamy is layered [116]. We sought to identify evidence of genetic immigration which is archaeologically associated with the expansion of dry-land farming in Gujarat (nortwestern India) during the post-last Glacial

Period aridification. This would be far earlier than IE dispersal. Given the tendency of subsistence-associated endogamy, we sought to test whether northern Gujarati (Meghawal) and southern dry-land farmer groups (such as Piramalai Kallar) showed differential admixture by IE lineages via `qpAdm`. In the south, a model of admixture for Meghawal, Gujarati Brahmins (GB) and Paniyas accounted for Piramalai Kallar (PK) genetics to within sampling variation. However, GB appeared with a negative coefficient, subtracting IE contributions to produce the fit. This suggests that a basal population (possibly a sub-population among DR, termed as Ancient Ancestral South Indian in [87]) mixed with Paniyas (southern DR hill dwellers) prior to absorption of IE lineages by GB. West European Hunter Gatherers (WHG) showed no significant admixture coefficients when added to these regression analyses (Supplementary Table 4). Thus, we observe a relatively recent IE admixture into a Gujarati pool of lineages, with more IE admixture towards the north, supporting a chronology of relatively recent genetic arrival of IE lineages. The direction of the gene flow was confirmed by the f_3 tests as shown in Table 3.5 where we see significant negative f_3 values when Meghawals are target and are admixed between European ancient DNA samples and PK (DR_SGC).

Table 3.5.: $f3(C; A, B)$ tests highlighting the Steppe and Dravidian mixture in Meghawal and the negative f_3 values and reasonably significant z-scores. This confirms the South India to Gujarat direction of gene flow.
Steppe_MLBA: Middle to Late Bronze Age samples from the Steppes [42]

A	B	C	F3	Err	Z
PK	GB	Meghawal	0.002634	0.000631	4.18
Paniyas	Meghawal	PK	0.004474	0.000809	5.53
Paniyas	GB	PK	0.003062	0.000678	4.517
Steppe_MLBA	GB	PK	0.020187	0.00074	27.27
WHG	Meghawal	PK	0.015089	0.001208	12.48
PK	Paniyas	GB	0.00188	0.00086	4.789
Steppe_MLBA	PK	Meghawal	-0.002393	0.000815	-3.935
WHG	PK	Meghawal	-0.002016	0.00114	-2.895

AA speakers of India and Southeast Asia

All AA speakers in India form a dense graph in Figure 3.5. The AA_SGCs share high ancestry values among the northern and southern Munda speakers, but not as much among them (Figure A.17). Southern Munda speakers such as Juang, Bonda, Savara, etc share high ancestry between themselves, whereas the Northern Munda speakers, such as Santhal, Mawasi, Ho, etc. are more homogenous than others. This is owing to separate admixture events for these two Munda speaking groups and the northern Munda speakers receiving longer admixture pulses from IE, DR and southeast Asian groups [112]. Khasis, Birhors, and Korwa are the divergent groups in PCA (Figure A.18) and f_3 (Figure A.13) analyses, mainly due to their geographical location and admixture from other TB, DR, and IE speakers, respectively (see Supplementary Table 3 for details).

There are two rival hypotheses of dispersal of AA languages across Asia. The former being that AA languages originated in Southeast Asia and later migrated to India, whereas the latter postulates that AA originated in South Asia and dispersed to the southeast [95]. As described below, our analysis supports the first hypothesis (southeastern origin of AA languages) and is concordant with the findings of [95,112]. We merged the samples collected from southeast Asian countries such as Myanmar, Laos, Vietnam and Cambodia with the normalized Indian data set to investigate the origin of the AA languages, forming a new data set of 624 samples spanning across 48,252 markers and 38 populations. The first two PCs (Figure A.19) showed the southeast Asian speakers such as Cambodians, Vietnamese and Laotians are closer to the Indian TB speakers followed by the AA speakers.

A closer look on the population genetic network, showed that the Khasis (Figure A.21a) indeed form a bridge between the Southeast Asian populations and the Indian AA speakers. The TB speakers share a large amount of ancestry with the Southeast Asian AA speakers, whereas the Indian AA speakers share less amount of ancestry with them (Figure A.20a). This is indicative of the fact that the genetic history of the Indian AA speakers and their southeast Asian counterparts are not homogeneous and probably they split very long ago. TreeMix analysis (Figure A.21b) and a recent study [112] revealed that AA speakers in India contain a significant ancestral component from southeast Asian populations. Although caution should be taken in interpreting TreeMix plots with weak migration edges, we also

found signs of admixture when we ran f_3 tests with Indian AA speakers as target and Southeast Asian speakers as the source along with DR speakers (Supplementary Table 6).

Relationship between Indian TB speakers and East Asia

The Indian TB speakers form a cluster in Figure 3.5 with Khasis, who are highly admixed among IE, AA and TB speakers(Supplementary Table 3 and Table A.2.1), becoming a link between the group of TB speakers and Manipuri Brahmins, who belong to the TB_SGA (Figure A.22). The TB speakers do not show a lot of homogeneity with each other (Figure A.23b) most likely due to their wide geographical dispersal across the Himalayan mountain ranges, ranging from east to west and acting as a barrier of gene flow.

To study the origin of the TB language family as well as the relationship between the Indian TB speakers and their East Asian counterparts we focused on samples from publicly curated data sets such as HGDP and other sources [117] to form a data set comprising of 347 individuals, sampled from 27 population groups, spanning across 38,667 SNPs. The first two PCs (Figure A.22a) show that the Indian TB speakers lie in close proximity with the Chinese mainland speakers. ADMIXTURE plots (Figure A.23a) and meta-analysis for shared ancestry (Figure A.23b), show that TB_SGC share significant ancestry with the mainland Chinese people.

Network analysis (Figure A.22b) reveals that the TB_SGC are close to central and southern Chinese, whereas, the SGAs are closer to Uygurs and the Burmese. This shows that, Himalayas although acted as a major barrier for gene flow from the north to south of Asia, had some level of permeability across the Himalayas in northeast India. TB speakers in India show signs of admixture from East Asia (Figure 3.6) confirming the gene flow from China to northeastern India as also shown in a recent study [118] and in previous studies using Y chromosomal markers [79,119]. The inferences drawn from demographically smaller groups, especially for the AA and TB language groups should be interpreted with caution, as random genetic drift might contribute to the variation of allele frequencies for these

groups. We do note that, in order to mitigate this bias as much as possible, we used SGCs that have maximum number of individuals in our study.

Routes of migrations into India

We proceeded to explore how migrations might have influenced the genomic structure of the Indian sub-continent in relation to the rest of Eurasia and Southeast Asia. Towards that end, we analyzed a dataset of 1,516 individuals over 42,975 SNPs (Supplementary Table 2c), sampled from 79 populations. PCA plots uncover a structure that resembles a triangle, with Europeans residing in one corner, the Chinese on another corner, and the DR and AA SGCs of India occupying the third corner (Figure A.24). IE, TB, and AA_SGCs are major nodes connecting to multiple populations. TB_SGC stand at the Northeastern gateway from China to India, while IE_SGA are at the entry-point from the Northwestern frontier (Figure 3.6). TreeMix [109] and f_3 statistics (Supplementary Table 5) show signs of admixture, with significant allele sharing between populations in Eurasia, revealing directions of gene flow. Meta-analysis of the ADMIXTURE output reveals that, overall, Indian populations share a great proportion of ancestry with the so-called Indian Northwestern Frontier populations, namely the SGC populations spanning Afghanistan and Pakistan (Figure 3.6). In concordance with previous studies we find higher degrees of shared ancestry in Central Asian populations with IE and DR SGA [82, 87–89]. In particular, IE_SGA share large amounts of ancestry with other IE speaking populations (i.e., Europeans). However, IE, TB, and DR speakers also share considerable amounts of ancestry with the Uygurs. On the other hand, AA speakers, who have been suggested as the earliest settlers of India [89], appear more isolated.

The time of admixture between Indian agriculturalist-related ancestry and the ASI component in the *Indus Periphery* samples in [87] is shown to be around 4700 - 3000 BCE. Subsequently, the timing of admixture between the ANI and ASI is known to be around 1900-4200 YBP. To test the hypothesis of whether modern IE and DR speakers show signs of admixture prior to the arrival of the Steppe ancestral component in the Indian subcontinent, we merged the ancient and modern humans from Near East [42,120] to form a merged

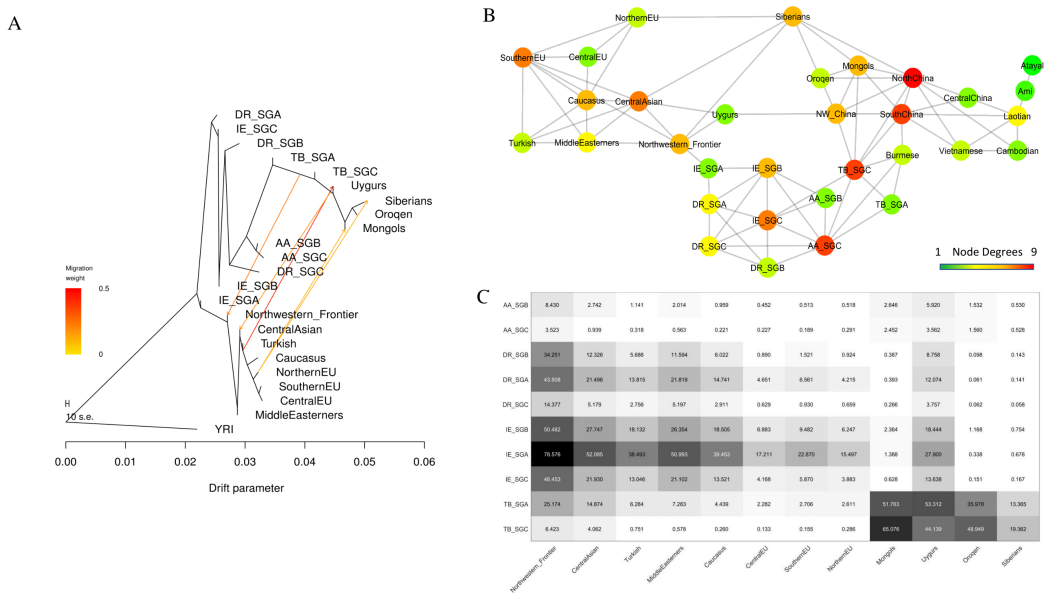


Figure 3.6.: A. TreeMix plot with the number of migration edges set to five indicate that the Siberians and Mongols show the most drift from DR_SGA and SGBs (residual plot in Figure A.25). Migration from Uygurs to the Northwestern Frontier populations is also found, making these populations a gateway to the Indian populations; B. Networks formed using the top five PCs (see Methods for the network formation algorithm) and five NNs showing three major paths leading to the two entry points of India; C. Meta-analysis of the ADMIXTURE plot (Figure A.26) quantifies the ADMIXTURE results (darker colors indicate higher pairwise shared ancestry).

dataset of 1,597 individuals across ∼31K markers. We tested for signs of mixture in the DR_SGB and SGC (denoting ASI) with IE_SGA and SGB (denoting ANI) in the presence of WHG or (equivalently) Scandinavian Hunter Gatherers (SHG) by running `qpAdm` [40]. The DR_SGB showed signs of admixture from both IE_SGA and WHG with very low standard errors and reasonably high p-values (Supplementary Table 4). Given the suggested timings of arrival of IE in the Indian subcontinent [39, 42, 43, 88], we note that the `qpAdm` tests indicate that IE speakers in north India brought WHG ancestry with them when they mixed with DR_SGB relatively recently after the ancient basal group contributed to both the modern Gujarati populations (IE speakers) as well as to PK and other DR_SGCs dwelling in the hills of southern India.

PCA and subsequent network analysis show that the genetic structure of Indian AA speakers and southeast Asian Austric speakers mimics geography. TreeMix analysis (Figure A.21b) of AA speakers and their southeast Asian counterparts (with six migration edges) shows that there is a migration edge from Cambodian to Bonda, who are northern Munda speakers. `qpAdm` tests find Mundas and Birhors to be admixed between DR_SGC and the Vietnamese (with ∼17% ancestry), as also observed in [112] (Supplementary Table 6). This was further validated by outgroup f_3 statistics (**Figure 3.7**). In order to interpret the allele sharing between Indian populations with that of their Eurasian counterparts, we analyzed the outgroup f_3 statistics, using YRI as the outgroup. We test our hypothesis of shared genetic affinity of the refined social groups of AA, DR, IE and TB speakers in India with Eurasian populations by running f_3 tests such as $f_3(YRI; X, Y)$, where X is an Indian group and Y being an Eurasian group. Outgroup f_3 statistic reveals European populations showing greater affinity, i.e., shared genetic drift with the IE social groups along with DR_SGA, whereas, the East Asian populations have larger shared affinity with the TB speakers along with some affinity with AA speakers as well. This clearly shows a gradient of gene flow from Siberia, then Mongolia, splitting towards China and Northeast India on one hand and the Uygurs, Central Asia mixing with the Europeans and Middle Easterners towards India, on the other. This also corroborates our findings from the network analysis and the entry points towards the Indian subcontinent (Figure 3.6). Previous studies have also supported a north-western and north-eastern corridor of migration towards India [39, 43, 81, 87, 88].

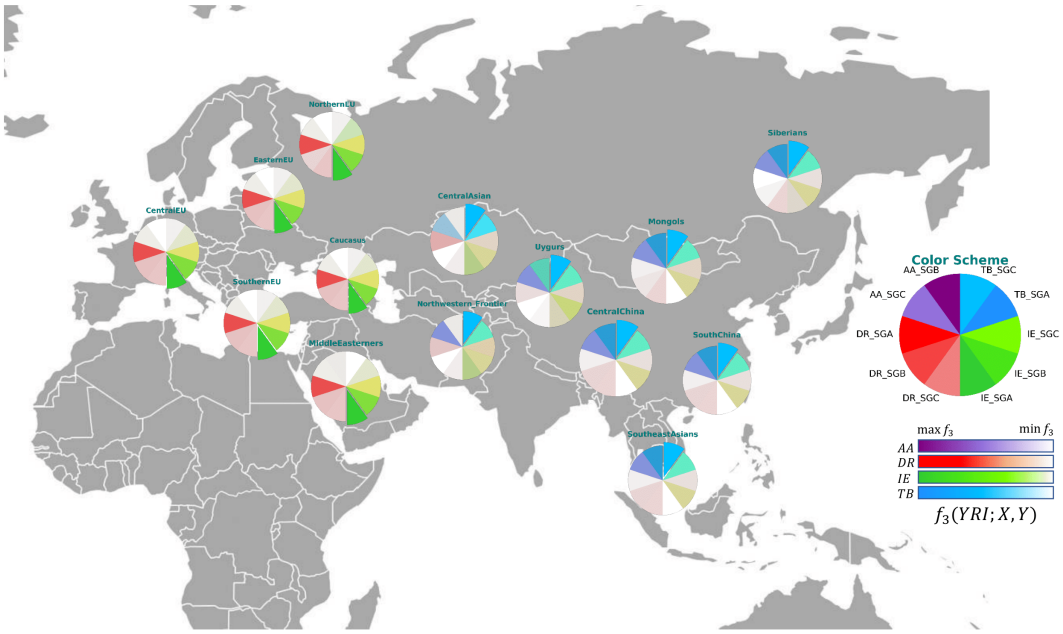


Figure 3.7.: Outgroup $f_3(YRI;X,Y)$ gradient map, showing pie charts of the shared affinity between Indian populations (denoted by X) and Eurasian/East Asian populations (denoted by Y). The color coding scheme is represented in the right hand side, signifying the colors attributed to perfect affinity (purple for AA, red for DR, green for IE, and blue for TB). The colors are distributed across gradients with respect to the maximum and minimum significant f_3 values. The population annotations and the detailed f_3 statistics can be found in the supplement (Supplementary Table 7). This gradient map shows the Europeans having more shared genetic drift from the outgroup YRI with the IE speakers of India (specifically, IE_SGA), whereas the East Asians have the maximum shared genetic affinity with TB_SGC.

However, this is the first study connecting both paths through the populations of Siberia and Mongolia.

3.2.5 Discussion

India represents a country of great social and linguistic complexity. We attempted to dissect this complex structure and reveal how these forces have shaped the Indian gene pool. Furthermore, putting India in a worldwide context, we integrated multiple representative populations from Eurasia, drawing paths of human migrations and gene flow throughout Eurasia. To do this, we investigated a comprehensive dataset that brought together all pub-

licly available data on the Indian sub-continent. Importantly, we established a quantitative deterministic and non-parametric framework aiming to evaluate the relative contribution of language, social structure and geography and identify the degree of impact of each factor. Earlier attempts to investigate geography and non-genetic factors in relation to population genetic structure did not provide statistical significance measures or feature selection or were based on multiple assumptions under a complex Bayesian framework [92, 93].

In concordance with previous studies, we find evidence for wide mixture across all the social groups. As shown previously, this wide admixture came to an abrupt end around 1,900 to 4,200 years ago [81] and the caste system originated from a semi-nomadic society which became hierarchical with the knowledge of agriculture [78, 113]. We time this event with the arrival of the steppe ancestral component in the Indian subcontinent around 3000 BC, and show that IE speakers mixing with southern Indian DR already contained a steppe-like western hunter gatherer ancestral component. Linguistic analyses also support a history of contacts between divergent populations in India. Indo-European languages (primarily spoken in northern India) are part of a larger language family that includes the great majority of European languages. In contrast, Dravidian languages (primarily spoken in southern India) are not closely related to languages outside of South Asia. Nevertheless, the earliest Hindu text (the Rig Veda, written in archaic Sanskrit) contains Dravidian loanwords that are not found in Indo-European languages outside the Indian subcontinent [81,121,122]. Further supporting the long contact between IE and DR speakers in India, our network analysis and f_3 tests identifies a large cluster consisting of IE and DR populations which resembles an almost complete graph with almost all pairs of populations connected to each other. Our findings indicate that the DR_SGC are indigenous to the Indian subcontinent with the knowledge of domesticated crops, conducting dry-land farming in the shades of Western Ghats in the southern peninsular India. `qpAdm` tests support the hypothesis of an outward migration from the south showing a south-to-north gradient of mixture with Gujaratis, which has similar archaeobotanical, as well as agricultural, footprints as southern India. This seems to indicate that the basal group admixed with the DR hill SGC lineages to produce dry-land farmers (PK), while that group admixed with IE speakers in Gujarat to produce Meghawals.

India has served as a major corridor for both Paleolithic and Neolithic migrations of anatomically modern humans [80, 87, 123]. An early dispersal of modern humans from Africa into India through the southern coastal route [124–126] and migration from West and Central Asia through the northwest corridor [77, 79, 127, 128] have been supported both by archaeological findings and genetic studies [129]. The proportions of ancestry derived from the western Eurasian gene pool has been found to be greater in populations inhabiting northern India than those inhabiting southern India. On the other hand, TB speakers in India seem to have arrived through the northeast corridor [79].

Language, social structure and geography create channels of gene flow across populations. However, to date, no study had attempted to establish a quantitative framework in order to dissect the relative contribution of each factor and translate it into a model that correlates with observed population genetic structure. Here, we establish such an analytic framework allowing the quantitative assessment of different evolutionary factors as well as the interplay among them. Applying this novel method on a comprehensive dataset from the Indian subcontinent, we are able to uncover the major forces that have shaped population genetic structure within India. In other parts of the world, geography has been found as the major contributor to shaping population genetic structure [1, 4]. Our results within India are in sharp contrast to what has been seen, e.g., in Europe [7], highlighting the importance of population specific studies around the world. Intriguingly, our study shows that spoken language seems to have been the major force bringing people together in India, across geographical and social barriers. The possibility to correlate genomic background to geographic, social and cultural differences opens new avenues for understanding how human history and mating patterns translate into the genomic structure of extant human populations.

4 TERAPCA: A FAST AND SCALABLE SOFTWARE PACKAGE TO STUDY GENETIC VARIATION IN TERA-SCALE GENOTYPES

This article has been accepted for publication in *Bioinformatics* Published by Oxford University Press with DOI: `10.1093/bioinformatics/btz157`.

4.1 Introduction

Principal Component Analysis (PCA) is perhaps the most fundamental unsupervised linear dimensionality reduction technique. It was invented by Pearson in the early 1900s [130]; and later reinvented and named by Hotelling in the 1930s [106, 131]. In statistical parlance, PCA converts a set of observations of possibly correlated variables into a set of linearly uncorrelated (orthogonal) variables called principal components (PCs). The seminal work of Luca Cavalli-Sforza and collaborators in the late 1970s [8, 12] pioneered the application of PCA for the study of human genetic variation.

PCA analyses and plots appear in virtually *every single paper* that analyzes human genetic variation in order to make inferences about population structures. Given m samples genotyped on n genetic loci, it is well-known that applying PCA on the $m \times m$ covariance matrix that emerges by computing any reasonable notion of genotypic distance between every pair of samples using the n genotyped loci results in the observation that the leading PCs mirror geography, e.g. see [7, 70, 132] for detailed discussions and examples. This observation was leveraged by [36, 37, 133] to derive one of the most established methods to account (and correct) for the confounding effects of population stratification in genome-wide association studies (GWAS). The method in [36, 37, 133] is essentially equivalent to using a small number of leading PCs as covariates in order to check for associations between genetic loci and affection status in statistical tests, and is implemented in the EIGENSTRAT software

package which is routinely used in GWAS analyses to correct for population stratification. Other applications of PCA include the identification of sets of genetic loci that are ancestry-informative or are under selective pressure [37,67,134]; and, when combined with other lines of evidence such as social structure and linguistics, the extraction of complex population histories and demographic structures [135]. We also note that PCA extracts the fundamental features of a dataset without complex computational modeling. Interestingly, even the output of model-based, more complex, methods to detect population structure (such as ADMIXTURE [35]) typically exhibits high correlation with the output of PCA, rendering further support to the significance of PCA in the analysis of human genetics data.

From a computational viewpoint, PCA essentially amounts to computing eigenvectors of the $m \times m$ (normalized) covariance matrix associated with the dataset at hand. When m does not exceed a few thousands, all eigenvectors can be computed by appropriate dense linear algebra routines in LAPACK, a Fortran 90 matrix factorization-based library which is widely used for solving systems of linear equations, least-squares problems, eigenvalue problems, and singular value problems [136]. Matrix factorization-based dense eigenvalue solvers return all m eigenvectors with a time complexity in the order of $O(m^3)$, which becomes impractical as m, the number of samples, increases. Practical applications of PCA in population genetics only require the computation of those principal components (PCs) determined by the eigenvectors associated with only a few (say 10-20) of the largest eigenvalues. Computing a few of the leading eigenvalues and associated eigenvectors of large (sparse or dense) matrices is typically achieved by first projecting the original eigenvalue problem onto a low-dimensional subspace which includes an invariant subspace associated with the relevant eigenvectors. This low-dimensional subspace can be formed in many different ways, e.g., by means of subspace iteration or Krylov projection schemes and much work in the Numerical Analysis community has been devoted in understanding the theoretical properties of such approaches [137, 138]. In particular, a variant of the family of Krylov projection schemes, the so-called Implicitly Restarted Arnoldi method (IRA), is the projection scheme of choice in FlashPCA2 [139], a software package which has been shown to outperform other PCA software packages, both in terms of memory usage and wall-clock time. On the other hand, recent advances in the design and analysis of Randomized Numerical Linear Algebra (RandNLA) [140] algorithms have yielded novel insights as well as

fast and efficient alternatives to approximate the leading principal components of large matrices [14, 141–143]. Indeed, FastPCA [144] applied such randomized algorithms to perform PCA analyses in population genetics data.

This paper presents TeraPCA, a C++ software package to perform PCA of tera-scale genotypic datasets that can not fully reside in the system memory. TeraPCA is essentially an out-of-core implementation of the Randomized Subspace Iteration method [13, 14] and features minimal dependencies to external[1] libraries. As the amount of time spent on I/O typically dominates the wall-clock time in out-of-core scenarios, TeraPCA builds a high-dimensional initial approximation subspace by loading the dataset from secondary storage exactly once. The dimension of this initial approximation subspace can be controlled directly by the user. Each subsequent iteration of Randomized Subspace Iteration "corrects" the initial subspace so that an invariant subspace associated with the leading target eigenvectors is computed. The dataset needs to be accessed twice in each iteration, but, fortunately, a few steps of Randomized Subspace Iteration are typically sufficient in practice in order to get highly accurate approximations to the leading eigenvectors. Note here that the above idea is somewhat orthogonal to the ideas underlying IRA, which builds the approximation subspace in a vector-by-vector manner, thus necessitating a large number of dataset fetches from secondary storage to even form an approximation subspace whose dimension is equal to or slightly larger than the number of PCs that we seek to approximate.

TeraPCA was tested extensively on both real (Human Genome Diversity Panel, 1000 Genomes, etc.) and synthetic datasets. Our synthetic datasets were generated via the Pritchard-Stephens-Donelly (PSD) model [34, 145]. Our results suggest that TeraPCA is both fast and accurate and in most cases outperforms other out-of-core PCA libraries such as FlashPCA2. Specific highlights include the computation of the ten leading principal components of a dataset of one million samples genotyped on one million genetic markers (this dataset exceeds 3.5 TBs in uncompressed format) in about 13 hours (using a single thread) and in less than 4.5 hours (using 12 threads).

[1] In contrast to FlashPCA2 which relies on the IRA implementation on the Spectra C++ library, TeraPCA comes with an in-house implementation of the Randomized Subspace Iteration algorithm.

4.2 Materials and Methods

4.2.1 Simulated Datasets

The first group of the datasets used for our experiments was generated using the Pritchard-Stephens-Donelly's (PSD) model of simulating genotypes. In particular, a recent study [145] simulated genotypic data by obtaining individual ancestry proportions from the PSD model to fit the 1000 Genomes dataset and then modelling the per-population allele frequencies using Wright's F_{ST} and the Weir & Cockerham estimate [146]. We developed a multi-threaded C++ package which is essentially an efficient implementation of the `R` code developed in Tera-Structure [145]. We generated various datasets in order to evaluate TeraPCA's performance, with the number of markers ranging from 100,000 to 1,000,000 and the number of samples ranging from 5,000 to 1,000,000.

Table 4.1.: Data sets on which TeraPCA was evaluated (simulated and real)

Dataset	Size (.PED file)	Size (.BED file)	# Samples	# SNPs
S_1 (simulated)	19 GB	120 MB	5,000	1,000,000
S_2 (simulated)	38 GB	239 MB	10,000	1,000,000
S_3 (simulated)	373 GB	24 GB	100,000	1,000,000
S_4 (simulated)	1.9 TB	117 GB	500,000	1,000,000
S_5 (simulated)	3.7 TB	233 GB	1,000,000	1,000,000
S_6 (simulated)	38 GB	2.4 GB	100,000	100,000
S_7 (simulated)	150 GB	9.4 GB	2,000	20,000,000
HGDP	615 MB	39 MB	1,043	154,417
1000 Genomes	8.4 GB	483 MB	2,504	808,704
PRK	2 GB	126 MB	4,706	111,831
T2D	1.8 GB	111 MB	6,370	72,457

4.2.2 Real Datasets

The Human Genome Diversity Panel (HGDP) dataset consists of 1,043 individuals genotyped at 660,734 SNPs, across 51 populations across Africa, Europe, Middle East, South and Central Asia, East Asia, Oceania, and the Americas [1]. We ran Quality Control (QC) on

the data by filtering SNPs with minor allele frequency below 0.01 and subsequently pruning for LD using a window size of 1000 kb. Moreover, we set the variance inflation factor to 50 and set $r^2 > 0.2$, thus retaining 154,471 variants. We applied the same parameters for LD pruning on the 1000 Genomes dataset which has 2,504 individuals sampled from 26 different populations across all continents genotyped at 39 million SNPs. After QC, we retained approximately 808,704 SNPs and ran our experiments on the pruned dataset.

We also tested the performance of TeraPCA on case-control data, which are ubiquitous in population genetics. We used the Wellcome Trust Case Control Consortium's (WTCCC) Type 2 Diabetes (T2D) and Parkinson's (PRK) datasets. The T2D dataset had 6,371 individuals (1,816 cases and 4,555 controls) genotyped on 313,654 SNPs and the PRK dataset had 5,000 individuals (2,000 cases and 3,000 controls) genotyped on 500,000 SNPs. We removed related samples from these datasets and pruned them using the aforementioned QC parameters resulting in datasets with 6,370 individuals genotyped on 72,457 SNPs for T2D and 4,706 individuals genotyped on 111,831 SNPs for Parkinson's.

4.2.3 TeraPCA

TeraPCA first normalizes the genotypes using the same procedure that was used by both FlashPCA [147] and FastPCA [144] (also discussed in Section 2.1.4) and then applies Randomized Subspace Iteration in an out-of-core fashion.

Randomized Subspace Iteration

This section describes Randomized Subspace Iteration (or Randomized Simultaneous Iteration), a commonly used technique for the computation of invariant subspaces associated with the largest (in magnitude) eigenvalues of matrices. More specifically, given a square $m \times m$ matrix B (in our case, $B = AA^\top$), a positive integer ρ, and an $m \times s$ matrix X_0 representing a basis of the initial approximation subspace, Subspace Iteration extracts an approximation of the invariant subspace associated with the $k \leq s$ largest eigenvalues of B by projecting the problem onto a subspace formed by the range space of the matrix $B^\rho X_0$.

Here we have assumed that the rank of the matrix $X_0^T U$ is at least k, where U is the $m \times k$ matrix whose columns are formed by the eigenvectors[2] associated with the k largest eigenvalues of B. In the absence of any initial approximation of the target invariant subspace, a reasonable choice is to draw the entries of X_0 from the (standard) normal distribution $\mathcal{N}(0,1)$. A practical implementation of Subspace Iteration applied to the computation of

Algorithm 2 Randomized Subspace Iteration

Input: $n \times m$ matrix $A^\top$, $\rho > 0$, $m \times s$ guess matrix $X_0 \in \mathcal{N}(0,1)$, $k \geq 1$, and $s \geq k$
Output: The k leading approximate eigenvectors of matrix A

1: $C = A(A^\top X_0)$
2: **Repeat**
3: **for** $i = 2 : \rho$
4: $Q = \texttt{orth}(C)$
5: $C = AA^\top Q$
6: **end for**
7: $Q = \texttt{orth}(C)$
8: $C = AA^\top Q$
9: $M = Q^\top C$
10: Compute the eigenvalue decomposition $M = XDX^\top$.
11: Set $C := QX$
12: **Until Convergence**

an invariant subspace associated with the k largest eigenvalues of the matrix $AA^\top$ is listed in Algorithm 2. In practice, the least sufficient number of iterations ρ required to compute the target invariant subspace can not be determined a priori (at least in the absence of an estimate of the distribution of the $k+1$ largest eigenvalues of the matrix $AA^\top$) and we perform ρ steps at a time instead. The procedure is then repeated with the most recent approximation of the target invariant subspace as the new approximation (see Line 11). By default we set $\rho = 1$.

Each iteration of Algorithm 2 requires $\rho + 1$ Matrix-MultiVector (MMV) multiplications with matrix $AA^\top$ (the second MMV product is needed for the Rayleigh-Ritz projection), and thus the dataset must be loaded $\rho + 1$ times from the secondary storage. The matrix X multiplying the matrix Q is formed by the eigenvectors of the Rayleigh-Ritz eigenvalue problem shown in Line 10. This $s \times s$ eigenvalue problem is dense and symmetric, and

[2]Without loss of generality we assume that the eigenvectors are normalized to have unit length

is solved by calling the `DSYEV` routine in LAPACK [136]. After each MMV multiplication of the form $C = A(A^\top Q)$, the resulting product C is orthonormalized (calling the routine `orth`(.)) to avoid a loss in the numerical accuracy due to overflowing. The TeraPCA library performs this orthonormalization by calling the `DGEQRF` and `DORGQR` routines in LAPACK. As a final remark, we note that the Rayleigh-Ritz matrix M in Line 9 could be also computed as $M = C^\top C$ where $C = A^\top Q$. However, this approach requires storing the $n \times s$ matrix $C = A^\top Q$ which is not be feasible when the number of SNPs is very large. Instead, we chose to use a slightly more expensive approach in terms of Floating-Point Operations (FLOPs) to form $M = Q^\top C$, where $C = AA^\top Q$ is of size $m \times s$.

The convergence rate of Algorithm 2 depends on the value of s as well as the distribution of the eigenvalues of $AA^\top$. In particular, let us order the eigenvalues of $AA^\top$ as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. Since $AA^\top$ is positive-semidefinite, its eigenvalues are non-negative. The rate of convergence of Algorithm 2 towards an invariant subspace associated with the ith dominant direction, i.e., the direction associated with the eigenvector corresponding to eigenvalue λ_i, is then governed by the ratio λ_{s+1}/λ_i. As a result, fast convergence should be expected when either a sufficiently large value of s is used or the k leading eigenvalues $\lambda_1, \ldots, \lambda_k$ are much larger than the trailing eigenvalues $\lambda_{k+1}, \ldots, \lambda_m$.

An Out-of-Core Implementation of Randomized Subspace Iteration

Typically, the number of individuals and SNPs will be such so that only a certain block of rows of matrix $A^\top$ will fit in the system memory. It is thus necessary to develop an out-of-core implementation of the MMV product with matrix $AA^\top$ where only a few lines of the binary PLINK file, i.e., rows of the matrix $A^\top$, reside the system memory at any given time. An algorithm for such an implementation is provided in Algorithm 3.

Algorithm 3 Out-of-core MMV $C = A(A^\top X)$

Input: $\zeta > 0$, $m \times s$ matrix X
Output: $m \times s$ matrix C
1: $C = 0$
2: **for** $i = 1 : \zeta$
3: Fetch the i-th row-block of $A^\top$
4: $C = C + A_i(A_i^\top X)$
5: **end for**

Let $\beta \in \mathbb{Z}^*$ be the integer denoting the maximum number of rows of matrix $A^\top$ that can reside in the system memory. Matrix $A^\top$ can be then written in a block row form as

$$A^\top = \begin{pmatrix} A_1^\top \\ A_2^\top \\ \vdots \\ A_\zeta^\top \end{pmatrix},$$

where $\zeta = \left\lceil \frac{n}{\beta} \right\rceil$. Note that when $\beta = n$, i.e., $\zeta = 1$, TeraPCA executes in-core and the entire dataset resides in the system memory.

Following the block partition of $A^\top$, the MMV product between $AA^\top$ and a MultiVector X can be written as

$$A(A^\top X) = \sum_{i=1}^{\zeta} A_i(A_i^\top X). \tag{4.1}$$

Each row block $A_i^\top$, $i = 1, \ldots, \zeta$, needs to be loaded from the secondary storage exactly once. As soon as $A_i^\top$ becomes available, we compute the product $A_i(A_i^\top X)$ and update $C = C + A_i(A_i^\top X)$. This computation can be achieved by a single call to the `DGEMM` BLAS routine [148].

By accounting for all ζ different row blocks of $A^\top$ in (4.1), we can easily determine that the computational cost to compute the MMV product $A(A^\top X)$ is equal to $sn(2m - 1)$ floating-point operations, and this cost is independent of the value of ζ (and thus β as well). Moreover, the value of ζ (as long as $\zeta > 1$) does not greatly affect the amount of time spent on loading each independent row block $A_i^\top$ from the secondary storage. On the

other hand, the value of β affects the performance of `DGEMM` applied to $C = C + A_i(A_i^\top X)$. In particular, low values of β can lead to cache conflicts and thus lower performance of `DGEMM`.

Figure 6 plots the amount of time spent on a single call to Algorithm 3 for different values of s, β and threads used in `DGEMM`, for the datasets S_6 and HGDP. The speedups obtained over single-thread executions are also shown. A few remarks are in order. First, while the computational complexity of Algorithm 3 is linear with respect to s, in practice the time complexity is sublinear, i.e., the amount of time required to multiply $AA^\top$ with a $m \times s$ matrix X is less than the amount of time required to multiply $AA^\top$ with a single vector s different times, especially for larger values of s, since the memory bandwidth cost of accessing each row block of $A^\top$ is amortized over s vectors. For the same reason the speedups obtained by using more threads in `DGEMM` are higher for larger values of s. Finally, increasing the value of β until it becomes greater than a certain threshold had a positive effect in the performance of `DGEMM` as it led to better cache utilization.

Convergence Criteria of Randomized Subspace Iteration

Different convergence criteria are possible to monitor the convergence of Randomized Subspace Iteration. TeraPCA considers two different criteria. The first criterion monitors the relative change of the sum of the (target) leading approximate eigenvalues between two consecutive iterations. When this difference becomes smaller than a user-given threshold, Randomized Subspace Iteration terminates. An alternative criterion is based on monitoring the relative error between successive approximations of each target approximate eigenvalue independently, and terminate the algorithm as soon as all relative errors associated with the k largest eigenvalues of $AA^\top$ drop below a user-specified threshold.

The main parameters influencing performance of TeraPCA are as follows:

1. Number of PCs to be computed (denoted by k). Default value is set to $k := 10$.

2. Number of contiguous rows of the SNP-major input matrix fetched from the secondary storage at each time unit (denoted by β). This can be user-defined or automatically determined based on the available system memory.

3. Dimension of the initial approximation subspace (denoted by s). Default value is set to $s := 2k$.

4. Convergence tolerance (denoted by tol). Default value is set to tol $:= 1e-3$.

The wall-clock time of TeraPCA is affected by all of the above parameters. Clearly, reducing tol or increasing k results in an increase of the wall-clock time. Using a higher-dimensional approximation subspace, i.e., increasing s, might reduce the corresponding wall-clock time as it typically enhances convergence towards the k-leading eigenvectors. On the other hand, increasing the value of s also increases the amount of floating-point operations performed. Finally, since only a part of the dataset can fit in the system memory at any time unit, the choice of β is typically determined automatically by TeraPCA based on the size of the system memory. The total amount of time spent on I/O is largely independent of the value of β but we have observed that the value of β has an effect on the wall-clock time of the LAPACK routines.

Setting the Value of β

TeraPCA allows the users to choose their own value of β. Error checking is included to determine the user-given value of β is inbounds (i.e., whether it satisfies $1 \leq \beta \leq n$). If β is out of bounds, TeraPCA determines an alternative value based on the amount of available system memory. Similarly, if no value of β is provided, TeraPCA will determine one on its own. In both cases, this value of β is set as

$$\beta = \frac{(\text{available amount of RAM}) - (\text{memory buffer})}{8 \times (\#\text{ samples})}.$$

Herein, the term "memory buffer" denotes the precomputed size of memory that TeraPCA needs for the rest of its variables except $A_i^\top$. By default, we fix the amount of available system memory to only 2 GiBs in order to make TeraPCA as flexible as possible. We

observed that increasing the amount of the available system memory did not lead to significant changes in the wall-clock time achieved by TeraPCA, unless the size of the available system memory became large enough to load the entire dataset in RAM (in-core). We also observed that very small values of β are likely to penalize the performance of DGEMM due to non-optimal cache utilization.

Sketching Dimension of X_0

TeraPCA employs the Randomized Subspace Iteration method to approximate the top k PCs of the normalized genotype matrix A. As discussed in detail in the next section, the Randomized Subspace Iteration method requires an initial "guess" matrix X_0, of dimensions $m \times s$ (see Algorithm 2). The choice of s is important for the performance of TeraPCA.

We chose to set s to $2k$. This conservative choice of s is rooted on the fact that the magnitude of all eigenvalues (except for the leading three-four ones) of the normalized covariance matrix in our datasets are typically clustered. From a geometrical viewpoint, the latter means that the variance of the dataset along the trailing PCs is roughly the same. Choosing a large value for s directly increases the matrix-matrix multiplication overhead but could improve the convergence rate. An exhaustive analysis of this complicated trade-off is beyond the scope of this paper.

4.2.4 Implementation

TeraPCA, implemented in `C++` using standard Linear Algebra libraries such as `BLAS` and `LAPACKE` is available to download with a GNU GPL v3.0 license at `https://github.com/aritra90/TeraPCA`.

4.3 Results

The performance of TeraPCA was tested on both simulated and real-world genotypic datasets. All our experiments were performed at Purdue's `Brown` cluster on a dedicated

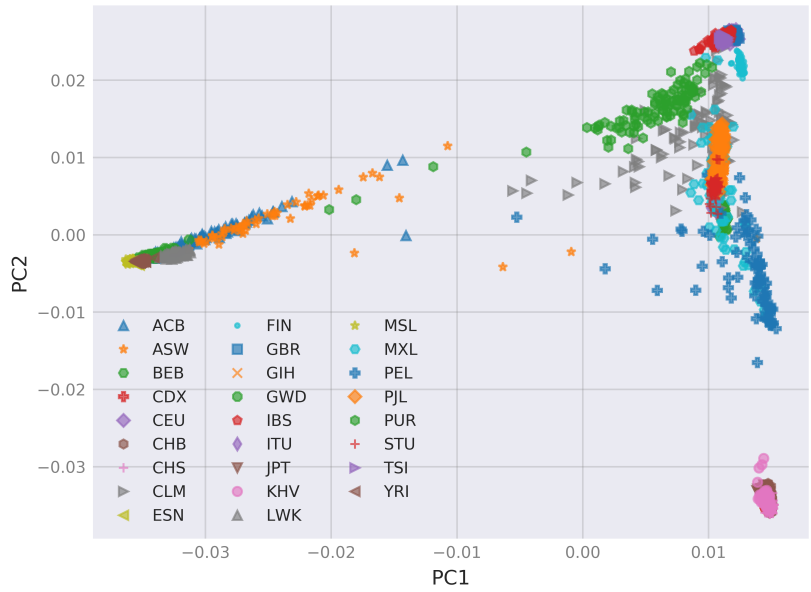


Figure 4.1.: Projection of the samples of the 1000 Genomes dataset on the top two left singular vectors (PC1 and PC2), as computed by TeraPCA.

node which features an Intel Xeon Gold 6126 processor running at 2.6 GHz with 96 GB of RAM and a 64-bit CentOS Linux 7 operating system. Table 4.1 lists the number of samples, number of SNPs, and size of each dataset. Datasets S_1 through S_7 are synthetic datasets and the remaining ones are real-world datasets. This section provides comparisons between TeraPCA and FlashPCA2. The latter has already been shown to be faster than previous methods such as FlashPCA [147], FastPCA [144], etc. The results reported throughout the remainder of this section were obtained by setting the amount of system memory made available to TeraPCA (as well as FlashPCA2) to 2 GBs. This is precisely the amount of memory allowed to FlashPCA2 in prior work.

4.3.1 Synthetic Datasets

Datasets S_1 through S_5 in Table 4.1 have a fixed number of SNPs (equal to one million) and a varying number of samples (from 5,000 to one million). On the other hand, dataset S_6 was used to fine-tune prior state-of-the-art methods and contains 100,000 samples genotyped on

100,000 SNPs. S_7 was used to test the performance of TeraPCA on extremely rectangular matrices, where the number of SNPs heavily outnumbers the number of individuals.

We first consider the plots of the three leading principal components returned by both TeraPCA and FlashPCA2 for dataset S_6 (see Figure 1 in supplementary material). TeraPCA and FlashPCA2 show a complete visual agreement with each other and both libraries agree with the expected outcome of the PSD model. For this particular example, TeraPCA terminated in just under 40 minutes, while FlashPCA2 required 141 minutes[3].

Table 4.2 lists the wall-clock times achieved by TeraPCA when applied on datasets S_1 through S_7. For datasets S_4 and S_5, which were the largest ones in our collection, TeraPCA terminated after 7.3 and 13.2 hours respectively. On the other hand, FlashPCA2 did not terminate within the 50 hours limit that we imposed. TeraPCA outperformed FlashPCA2 on all synthetic datasets, with a speedup that ranged between 1.3 and 4.5, at least for those datasets where FlashPCA2 terminated within our 50 hour limit. We note that for all synthetic datasets the leading PCs returned by TeraPCA and FlashPCA2 showed perfect correlation as measured by the Pearson correlation coefficient (equal to one in all cases). To further test TeraPCA's performance on datasets where the number of SNPs heavily outnumbers the number of individuals, we applied it to S_7 and observed that even in a heavily under-determined system, TeraPCA outperformed FlashPCA2 by a factor of 2.9, with similar accuracy guarantees.

4.3.2 Real Datasets

We first considered the Human Genome Diversity Panel (HGDP) dataset [1]. TeraPCA was marginally faster than FlashPCA2 and both libraries required about seven seconds. A plot of the projection of the HGDP dataset along the two leading PCs computed by TeraPCA

[3]To be fair in our comparisons between TeraPCA and FlashPCA2, we performed multiple runs of FlashPCA2 on dataset S_6 in order to explore and understand its properties. In particular, we varied the convergence criterion in FlashPCA2 and recorded the resulting trade-off between wall-clock time and digits of accuracy for the top ten computed eigenvalues. Fixing the convergence tolerance in FlashPCA2 to three digits of accuracy and the maximum number of iterations of FlashPCA2 to 100 was the best choice in terms of the tradeoff between running time and accuracy

is shown in Figure A.28. Given the relatively small size of this dataset, we were able to compute the exact ten leading eigenvectors using LAPACK.

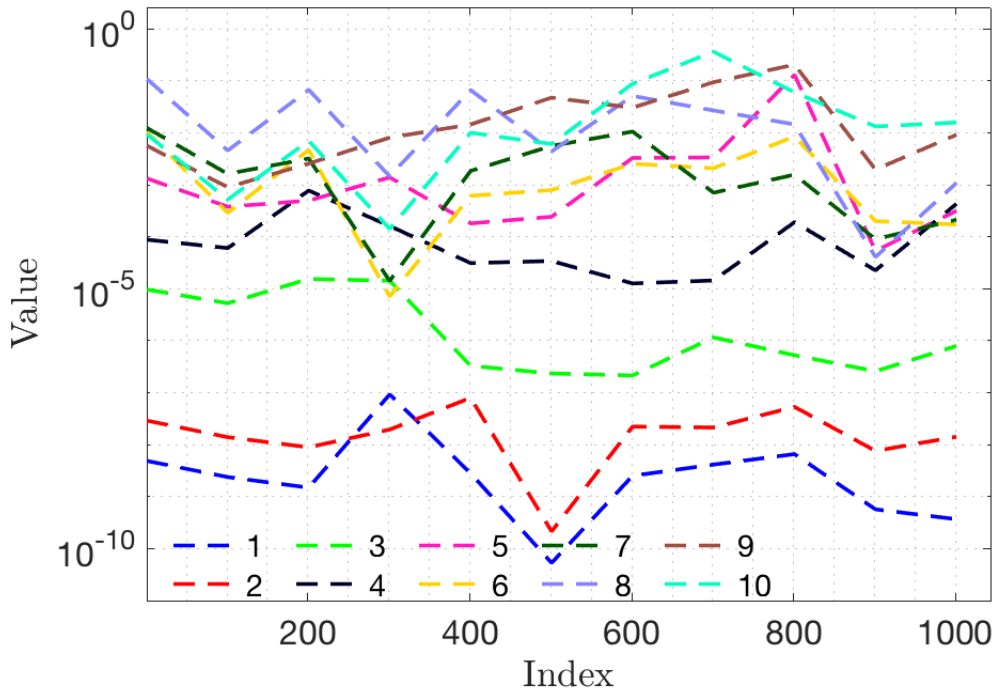


Figure 4.2.: Entry-wise relative error of the top ten leading eigenvectors returned by TeraPCA for the HGDP dataset, compared to the eigenvectors returned by LAPACK. The y-axis shows the relative error; recall that each eigenvector has 1,043 entries. We observe that the relative error is roughly the same for each entry of a specific eigenvector.

Figure 4.2 reports the entry-wise error of the ten leading eigenvectors returned by TeraPCA. As expected, eigenvectors associated with the largest eigenvalues are captured more accurately since they converge faster. In addition, Table A.3 reports the relative and absolute errors of the ten leading eigenvalues returned by TeraPCA and FlashPCA2. For TeraPCA, the (much) higher accuracy in the approximation of the three-four leading eigenvalues is due to the fact that these approximate eigenvalues kept improving as Randomized Subspace Iteration kept iterating to approximate the trailing eigenvalues and eigenvectors. On the other hand, the accuracy in the approximation of the eigenvalues returned by FlashPCA2 was somewhat uniform for all eigenvalues.

TeraPCA and FlashPCA2 showed similar qualitative and computational performance on the pruned 1000 Genomes dataset (see Figure 4.1), with FlashPCA2 terminating slightly faster than TeraPCA. Notice that this dataset is also the one in which the number of SNPs outnumbered the number of individuals by the largest factor. PCA is an essential tool to detect population stratication in GWAS. In order to evaluate TeraPCA's performance on real-world case-control studies, we applied it on WTCCC's T2D and PRK datasets. Like other real-world datasets, both FlashPCA2 and TeraPCA performed similarly, needing roughly the same wall-clock time. Execution of TeraPCA on these datasets can also be done in-core, as they fit in the system memory, leading to comparatively faster computation.

Table 4.2.: Wall-clock running times comparisons for the datasets of Table 4.1 using a single thread and 2 GBs of system memory
* indicates no convergence after 50 hrs.

Dataset	TeraPCA	FlashPCA2	Speed-up
S_1	26.2 mins	33.3 mins	1.27
S_2	39.3 mins	87.5 mins	2.22
S_3	7.9 hrs	35.6 hrs	4.50
S_4	7.3 hrs	n/a*	∞
S_5	13.2 hrs	n/a*	∞
S_6	39.5 mins	141.1 mins	3.57
S_7	37.3 mins	106.5 mins	2.86
HGDP	6.5 secs	7.7 secs	1.22
1000 Genomes	4.3 mins	3.5 mins	0.81
T2D	96 secs	119 secs	1.24
PRK	76 secs	73 secs	0.96

4.3.3 Multithreading

The wall-clock times of TeraPCA and FlashPCA2 can significantly improve by executing the associated linear algebra computations using more than one threads. This is indeed the most obvious way to speed up software such as ours. To test the performance of TeraPCA as a function of the number of threads, we focused on datasets S_1, S_2, S_4, S_6, S_7, and the 1000 Genomes dataset. The number of threads was set to 4, 8, and 12 and the speedups reported in Figure 4.3 are against the single-thread execution of TeraPCA. Generally speaking, we

observed a 1.6x-2.8x speedup, which is somewhat sub-optimal. The reason underlying this non-optimality is that we used multithreading only for the linear algebraic operations. However, much of the wall-clock time is spent on I/O operations in order to load the dataset from secondary memory, a procedure that cannot be multithreaded. We emphasize that FlashPCA2 did not demonstrate comparable improvements when multi-threading was enabled. In particular, when applied to the dataset S_6, the wall-clock time of FlashPCA2 reduced only by two minutes, i.e., from 141 minutes to 139 minutes.

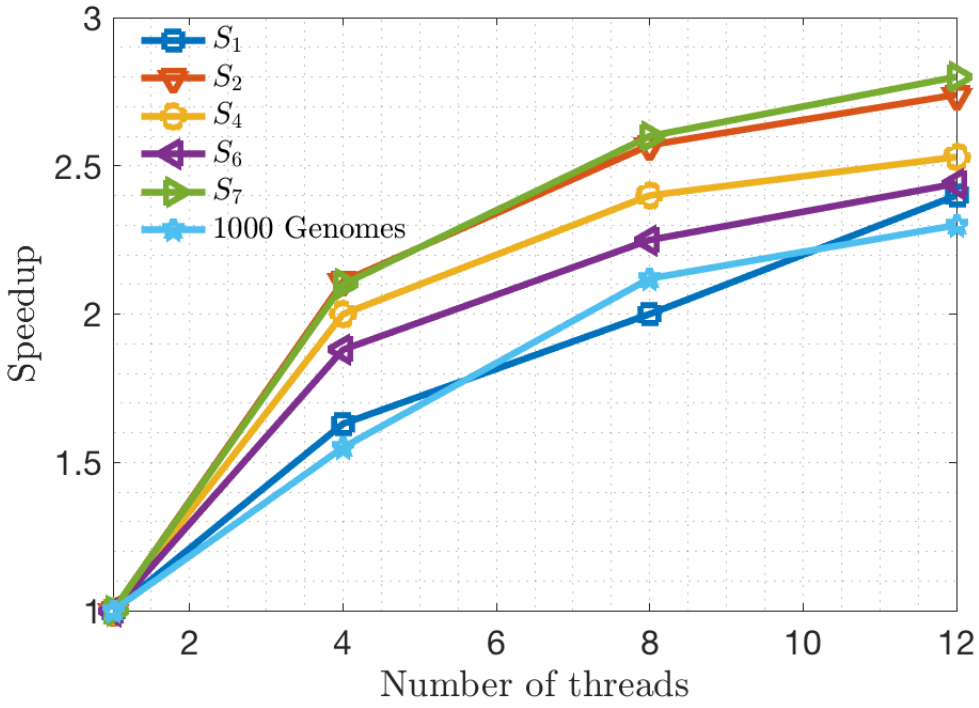


Figure 4.3.: Speedup of TeraPCA over single-threaded execution.

In all of the above experiments we set $s := 2k$ and $k := 10$. Finally, Figure A.3 reports the amount of time required to multiply the (normalized) covariance matrix by a set of s vectors using the DGEMM BLAS routine of MKL and a varying number of threads for different values of s and β for datasets S_6 and HGDP. It is worth noting that while an exhaustive analysis lies outside the goals of this paper, it is easy to verify that doubling the value of s does not double the amount of time required to perform the multiplication, while larger values of s also lead to higher speedups when multiple threads are used. Similarly, very

small values of β are likely to penalize the performance of DGEMM due to non-optimal cache utilization.

4.4 Discussion

In this paper we presented TeraPCA, a C++ library to perform out-of-core PCA analysis of massive genomic datasets. It is based on Randomized Subspace Iteration, building upon principled and theoretically sound methods to approximate the top principal components of massive covariance matrices. TeraPCA returns highly accurate approximations to the top principal components, while taking advantage of modern computer architectures that support multi-threading and it has minimal dependencies to external libraries. TeraPCA can be applied both in-core and out-of-core and is able to successfully operate even on personal workstations with a system memory of just a few gigabytes. Numerical experiments performed on synthetic and real datasets demonstrate that TeraPCA performs similarly or better when compared to state-of-the-art software packages such as FlashPCA2, on a single thread and significantly better with multi-threading.

Similar to FlashPCA2, the main goal of TeraPCA is to make the application of PCA feasible for genotype datasets whose size is (much) larger than the available amount of system memory, and to do so, both techniques apply a projection-based eigenvalue solver to compute the leading eigenvectors of the normalized covariance matrix. FlashPCA2 is based on Implicitly Restarted Arnoldi (IRA), a widely-used Krylov subspace eigenvalue solver. IRA builds a subspace whose dimension is increased by one at each iteration until the algorithm restarts while retaining an approximate invariant subspace of the target eigendirections (i.e., the PCs of interest). IRA essentially creates a "gap" between eigenvalues associated with target/unwanted PCs which allows the eigenpairs of interest to converge faster altogether. We used dataset S_6 of Table 4.1 to evaluate the performance of FlasPCA2 on a single thread by varying the convergence criteria. We varied the number of iterations from 20 to 100 and the digits of accuracy from two to six. The best performance was noted (with a trade-off between accuracy and running time) when the convergence criteria was set to three digits of accuracy and 100 iterations. On the other hand, Randomized Subspace Iteration keeps

the dimension of the approximation subspace fixed, and the convergence rate of the invariant subspace associated with the k leading eigenvectors is practically determined by the distance of the kth largest eigenvalue from the next largest eigenvalue.

The main rationale for exploiting Randomized Subspace Iteration in TeraPCA is based on the fact that the latter allows for the construction of an high-dimensional approximation subspace by loading the genotype dataset from the secondary storage exactly once. Indeed, one iteration of Randomized Subspace Iteration will generate an approximation subspace whose dimension is equal to s. Each subsequent iteration will then try to correct the approximation subspace so as the target eigenvectors are approximated more accurately. On the other hand, IRA builds the approximation subspace vector by vector, thus requiring the dataset to be loaded multiple times to just form an approximation subspace whose dimension is equal to the number of target eigenvectors. While IRA is typically more efficient that Randomized Subspace Iteration in terms of computational cost per target eigenvector, in the vast majority of experiments we tested the latter was not sufficient to offset the fact that Randomized Subspace Iteration required fewer iterations and thus fetched the dataset from the secondary storage fewer times.

The amount of time spent on I/O typically dominates the overall wall-clock times for both TeraPCA and FlashPCA2, due to loading the data set multiple times per iteration. Thus, loading the data set as few times as possible is the main priority. TeraPCA has an advantage over FlashPCA2 as the former multiplies the covariance matrix by more than one vectors at each iteration. In terms of complexity, each iteration of FlashPCA2 requires $\mathcal{O}(nm)$ floating-point operations, where m and n denote the number of individuals and SNPs, respectively. On the other hand, each iteration of TeraPCA requires $\mathcal{O}(nms)$ FLOPs where the variable s denotes the dimension of the initial approximation subspace (shown as X_0 above). For practical purposes, we have $s << \min(m, n)$, and s is set to $2k$ as discussed above. Since the number of sought PCs is usually a small constant, e.g., $k << \min(m, n)$, the asymptotic complexities of TeraPCA and FlashPCA2 with respect to the input dataset are practically the same. Recall also that the time required to multiply the coefficient matrix by s vectors simultaneously is typically much less than the time needed to multiply the same matrix with one vector s separate times due to better bandwith and cache utilizations.

In summary, if there is enough system memory to fully load the genotype dataset, then the amount of time spent on FLOPs is typically the dominant part of the wall-clock times achieved by TeraPCA and FlashPCA2, and one should probably use a Krylov projection scheme (such as IRA in FlashPCA2) to compute the sought PCs. On the other hand, when the dataset can not fully reside in the system memory, fetching the dataset from the secondary storage is typically the main bottleneck, and one should opt for block methods (e.g. Randomized Subspace Iteration). We also note that Randomized Subspace Iteration is based on BLAS3 routines, thus allowing better cache utilization and reduced intra-processor communication when performing the MMV products. The latter also becomes advantageous when multiple computational threads are used during execution. Figure A.29 plots the wall-clock times (in hours) required by TeraPCA and FlashPCA2 to compute a varying number of leading PCs of the dataset S_6. Similarly, Figure A.30 plots the wall-clock times (in minutes) required by TeraPCA and FlashPCA2 to compute the ten leading PCs of a dataset with $m = 10^5$ individuals and a varying number of SNPs. The top 10 PCs returned by both the software suites captured similar proportions of variance (see Figure A.31), as a testament to their nearly identical qualitative performance discussed in the main text.

Future work will focus on implementing a distributed memory version of TeraPCA using the Message Passing Interface (MPI) standard. Another interesting research direction would be to combine TeraPCA with block Krylov subspace techniques.

5 SSIMRA: MULTIPLE LOCI SELECTION WITH MULTIWAY EPISTASIS IN COALESCENCE WITH RECOMBINATIONS

5.1 Introduction

Nothing in Biology Makes Sense Except in the Light of Evolution[1] and simulating the evolution process of, whether multi-cellular humans, unicellular micro-organisms or even cancer-tumor continues to be an important device in understanding the observed molecular profiles of populations. Molecular profiles are captured by the genetic variability generated by mutations and the change in frequency of alleles within populations over time. The selectively neutral infinite-sites model [149] is often the basis for the analysis of this variation [150]. Many systems [151–154] simulate the generation of realistic random populations and the reader is directed to [155] for an efficient algorithm and a comprehensive survey of literature.

The Ancestral Recombination Graph (ARG) [156] is a variant of [157]'s coalescent, which is used to reconstruct the grand most recent common ancestor (GMRCA), backwards starting from the leaves, using recombination and coalescent operations (described in 2.1.7). Here we provide the first coalescent simulation framework called *back-sSimRA*, which allows for multilocus selection with multiway epistasis. To validate the findings of *back-sSimRA*, we built a forward-time simulator, *fwd-sSimRA* which has similar setup as it's backward counterpart.

Coalescent processes allow fast approximation of the neutral Wright-Fisher (WF) model which accounts for the effects of various evolutionary forces such as random genetic drift, mutation, selection on allele frequencies. Efficient simulation algorithms such as *ms* [49], *fastsimcoal* [158], *msprime* [159], *MaCS* [160], *SMC* [161] exist and are fast as they track ancestral lineages to extant populations going backward in time. As natural selection influ-

[1] This quote is attributed to evolutionary biologist Theodosius Dobzhansky.

ences the evolutionary process design of coalescent simulators with selection on a haploid or diploid locus is of more general use and has piqued interest in development of a host of software suites such as, *msms* [162], *discoal* [163], *cosi2* [164], *mbs* [165] etc. Most of these takes into account the demographic histories and population structure information to study selective sweeps and footprints of local adaptation. These methods approximates the Markovian coalescence by tracking the number of coalesceable pairs and none of them allows interaction between alleles in multiple loci under selection. Epistasis has long been recognized as a significant component in understanding genealogies and evolution of complex genetic systems [166]. Only a few forward simulators exist which provides a framework to model epistasis, such as *SELAM* [167] allowing for pairwise epistatic selection to model the process and consequences of admixture or *SLiM* [168], which constructs ecologically realistic scenarios while accounting for a host of complex biological processes and *SLiM 3* [169] provides scenarios beyond the WF framework. It also efficiently simulates epistatic scenarios similar to *fwdSimRA* whose main purpose is to provide a validation framework to the coalescent simulator. Apart from *SLiM*, forward-time simulations are captured in other packages such as *msms* [162] and *ForwSim* [170], all of which takes into account geographical population structure and demographic information along with selection at a single locus. Some account for selection on polygenic quantitative traits along with recombination in a small number of generations [171, 172]. *SelSim* [173] differs from the other packages by following a Moran model based approach, while the others [162, 168, 170, 171] including our algorithm follows WF model for forward simulations of complex evolutionary processes. *simuPOP* [171] is an individual based genetics simulation program and has a lot of parallels to *fwd-sSimRA* in an individual level, but our model focuses on the holistic view of the populations as well as the individual statistics in a WF framework. *ForwSim* also draw parallels with the forward simulator proposed here, but it accounts for the "book of populations" differently such as removing the SNPs and lineages that die out, whereas *fwd-sSimRA* keeps them for tracing the ARG and comparing hallmarks with *back-sSimRA*. Although, time complexity is a trade-off by accounting for all those lineages, but, we get all the benefits of recording every mutation and recombination event while simulating evolutionary history of each individual across generations along with the selected alleles at a mutated site.

Our goal was to design separate models for backward (coalescent) and forward simulation and compare their results in studying complex evolutionary scenarios. This is the first algorithm that modulates the ARG to incorporate multilocus selection with multiway epistasis in a coalescent as well as in a forward setting. *back-sSimRA* determines whether the next event in time is a recombination or coalescent by taking the event which takes minimum time to occur, unlike all available coalescent simulators, most of which approximates the Markovian coalescence by tracking the number of coalesceble pairs. Another significant advantage of *back-sSimRA* is that allows for interaction between loci providing an efficient framework to model realistic ecological scenarios which accounts for linkage disequilibrium (LD). *fwd-sSimRA* on the other hand validates the results obtained from the coalescent simulator and is capable of handling multilocus selection with multiway epistasis in a simple and efficient model which approximates the evolution process closely. The comparison between the two models is done systematically by using the hallmarks of an ARG [155]. We make available to the user both the forward and backward simulators.

5.2 Materials and Methods

5.2.1 Modeling Multiple Loci Selection with Multiway Epistasis

Overview of the Forward and Backward schemes. A forward simulator represents the basic biological processes, such as diploid inheritance from two parents, recombination that occurs in that context, where children pick their parents according to WF with probabilities modified by selection. Such a model would not explicitly include a coalescent, but the impact of the WF model spontaneously produces a Kingman coalescent among ARGs for subsets selected from among the a current population [48, 49].

Fitting selection and epistasis into the "backward" scheme is complex [174, 175]. The backward simulator is very targeted and constructs only the ARG. On the other hand, a forward model is a generation-by-generation simulation of the population of diploids. To compare this with the backward simulator, the subgraph (ARG) of interest that is embedded in the complete "book of populations", is traced. The accuracy of the backward simulator is

demonstrated by the concordance of the distributions of the hallmarks between the forward and the backward models.

Our aim is to overcome the challenges of modeling and designing the algorithms for a backward simulator that incorporates all the features of a flexible forward simulator. But how "accurate" is the ARG? To answer this rather difficult question, we adopt the mechanism that used by [155] to compare different algorithms. These provide a means for a comparison of different algorithms even under different regiments forward, or, backward.

5.2.2 Backward Simulator Model

The algorithm works back-in-time starting from the present (time 0), moving back into the past. See [155] for a detailed exposition on neutral scenario. Let the number of loci under selection be l, possibly with multiway epistasis. As an illustration let l be 3 with selection values s_1, s_2 and s_3. The algorithm will assign three random locations on the genetic segment for the three, unless the locations are explicitly specified. It is assumed that the minor allele is under selection while the major is neutral. The possible multiway epistasis are e_{12}, e_{13}, e_{23} and e_{123}. If no value is specified then the epistasis is assumed to be neutral. Given this, we get 2^l possible types of lineages as shown in Table 5.1, denoted as l_z.Let l_0 be the lineage type with no selection. For the running example, the other lineage types are l_1, l_2, l_3, l_{12}, l_{13}, l_{23}, l_{123}. For two lineage types z_a and z_b, let

$$l_{z_a} \prec l_{z_b} \text{ when } z_a \supset z_b.$$

For example, $l_{12} \prec l_1$ and $l_{12} \prec l_2$. Also, $l_{123} \prec l_{12}$. For the lineage type z, let N_z be the effective population size.

Isolated (single) locus selection. The fitness $1 + s$ is the ratio of the probabilities that the selected allele produces an offspring to the unselected allele, which relates to the proportions in generation $t + 1$ given proportion in generation t. See Supplement for a detailed exposition. Let N_s be the *effective* population size with the allele under selection

and $N_{\tilde{s}}$ $(= N - N_s)$ is the *effective* population size with the reference or ancestral allele which is not under selection, giving:

$$\frac{N_s}{N_s + N_{\tilde{s}}} : \frac{N_{\tilde{s}}}{N_s + N_{\tilde{s}}} = 1 + s : 1 \implies N_s = \frac{1+s}{2+s} N = f_s N. \tag{5.1}$$

Thus $-1 < s$. This is extensible to multiple loci with or without epistasis. Continuing the running example of three loci under selection with possible epistasis is shown in Table 5.1. TABLE IN

back-sSimRA: Algorithm to Generate the Topology with Multiple Locus Selection & Multiway Epistasis

If s_i and s_j are two locations with the minimum (or derived) allele under selection at locus i and j respectively, then e_{ij} denotes the epistasis between the two. If it is not explicitly specified then a neutral epistasis is assumed. The algorithm randomly chooses the location of the SNPs on the genetic segment being simulated. `INPUT`:

	Parameters	example values	user-specified units	units in bp for the algorithm	scaling factor
g	seqment length	$25; 75$	Kb	$\times 10^3$ bp	$\times 10^3$
m	extant units	$10; 20; 30; 40$	–	–	$\times 1$
N	population size	$100; 200$	–	–	$\times 1$
	rates/generation				
r	recombination rate	1	bp/gen $\times 10^{-7}$	bp/gen	$\times 10^{-7}$
μ	SNP mutation rate	1.5	mut/bp/gen $\times 10^{-8}$	$\times 1$ mut/bp/gen	$\times 10^{-8}$
	selection, epistasis parameters				
s_i	fitness	0.3	–	$\times 1$	
e_{ij}	epistasis	0.1	–	–	$\times 1$

`ASSUMPTION`: Not more than one event, coalescent or recombination, occurs at a generation. Also, no back mutations, i.e., a position (base) undergoes no more than one mutation in the entire ARG. The mutation rate and recombination rate are uniform over the segment being simulated.

`ALGORITHM`:

I. Initialization:

3 SNPs *Tableau*

lineage types l_z	(3 SNPs under selection) s_1	s_2	s_3	epistasis: user defined e_s(explicit) $\left[f_s = \frac{1+s}{2+s}\right]$	epistasis: no epistasis or neutral (implicit)	effective pop size $2Nf'$ for l_z-coalescence in backward algo f'
No alleles under selection lineage (immortal)						
l_0	×	×	×	$f_{\bar{s}}$ / −	$f_{\bar{s}}$	$f_{\bar{s}} = 1 - \sum_i f_{s_i} + \sum_{i,j} f_{s_{ij}} - f_{s_{123}}$
Main effect lineages						
l_1	s_1	×	×	f_{s_1}	−	$f_{s_1} - \sum_i f_{s_{1i}} + \sum_{i,j} f_{s_{1ij}}$
l_2	×	s_2	×	f_{s_2}	−	$f_{s_2} - \sum_i f_{s_{2i}} + \sum_{i,j} f_{s_{2ij}}$
l_3	×	×	s_3	f_{s_3}	−	$f_{s_3} - \sum_i f_{s_{3i}} + \sum_{i,j} f_{s_{3ij}}$
2-way epistasis lineages						
l_{12}	s_1	s_2	×	$f_{s_{12}}$ / − / $f_{(e_{s_{12}})}$	$f_{s^*} = f_{s_1} f_{s_2}$ / −	$f_{s_{12}} - \sum_i f_{s_{12i}}$
l_{13}	s_1	×	s_3	$f_{s_{13}}$ / − / $f_{(e_{s_{13}})}$	$f_{s^*} = f_{s_1} f_{s_3}$ / −	$f_{s_{13}} - \sum_i f_{s_{13i}}$
l_{23}	×	s_2	s_3	$f_{s_{23}}$ / − / $f_{(e_{s_{23}})}$	$f_{s^*} = f_{s_2} f_{s_3}$ / −	$f_{s_{23}} - \sum_i f_{s_{23i}}$
3-way epistasis lineage						
l_{123}	s_1	s_2	s_3	$f_{s_{123}}$ / − / $f_{(e_{s_{123}})}$	$f_{s^*} = f_{s_1} f_{s_2} f_{s_3}$ / −	$f_{s_{123}}$
						[Ex-/In-clusion principle]

Table 5.1.: Example with three loci under selection and all the possible different epistasis, whether explicitly specified or simply neutral. All the user-specified values are shown in red. The *back-sSimRA* algorithm uses the effective population size as shown here.

1. The genetic material, I_v, of each of the m leaf nodes, v, is set to $I_v = \{[0,1]\}$. For $r > 0$, randomly assign the lineage types to the lineages.

   For $r = 0$, the lineage-types are so assigned that no pair of types of lineages straddle (either they are disjoint or one is contained in the other). Note that lineage l_0 corresponds to lineage with no alleles under selection.

   For each lineage type l_z:

   (a) Count the number of lineages L_z. If $l_z > 0$, then the lineage is ACTIVE.

   (b) Set time T_z to 0.

   (c) Set a list C_z to empty. This is list of nodes, each with a time $t > T_z$.

   Append C_z: This occurs only when a lineage type changes during the iterative process (at a recombination event or node). At iteration i with lineage l_z, if $t_x > T_z$ then a new node, with time t_x, is appended to list C_z.

2. For each lineage l of type l_z, incident on leaf node v, the recombination rate

$$r'_l = N_z g r \text{len}(I_v). \tag{5.2}$$

   For each lineage type l_z, $r'_l = \alpha_z$

$$\alpha_z = N_z g r. \tag{5.3}$$

II. Loop: The stochasticity of the method allows for two possible regimens: *Pooled* and *Round Robin* .

Pooled Loop. In this regimen, all the lineages are pooled together. At each iteration, i, the minimum of t_{z_i}, over all the lineages l_z, is computed using Eqn 5.4. In the pseudocode below, the "FOR" loop is not required in *pooled* regimen.

Round Robin Loop. In this regimen, the lineage types are processed separately and in any order (the "FOR" loop in the pseudocode below).

In both the regimens, the lineages intermingle at recombination events and when the label of lineage type is changed. The latter occurs when all the lineages of a singleton type coalesce into one lineage.

REPEAT

Round Robin Loop. FOR each lineage type l_z with $L_z > 0$, do the following:

1. Let $l = l_z$.

2. Compute the recombination rate r'_l of each lineage l using Eqn 5.3 as:

$$r'_l = \alpha_* \times \text{len}(I_v),$$

using the effective population size N_z as described in Table 5.1. Then compute the time $t_{z_i} = N_z \times t$ to the next event using

$$\begin{aligned} t &= \min\left(\overbrace{\min_{1\le a<b\le L_z}(t_{ab}^{\text{coal}})}, \underbrace{\min_{1\le i\le L_z}(t_i^{\text{rcmb}})}\right) \\ &= \text{Exp}\left(\overbrace{1+1+...+1}+\underbrace{r'_1+r'_2+..+r'_{L_z}}\right) \\ &= \text{Exp}\left(\overbrace{L_z}+\underbrace{r'_1+r'_2+..+r'_{L_z}}\right) \end{aligned} \tag{5.4}$$

For $L_z > 1$, this gives either coalescence or recombination for the next step when $L_z > 1$. For $L_z = 1$, this gives only recombination for the next step.

3. T_z is updated as $T_z + t_{z_i}$.

4. **Coalescence event**: If the operation is coalescence then L_z is decremented by 1; two random lineages of type l_z are coalesced into one at time T_z and the outgoing edge of the coalesced node is labeled by lineage l_z.

If $L_z = 1$, z is a singleton label (such as s_1 but not s_1s_2 or $s_1s_2s_3$), and, there exist no ACTIVE lineage l'_z such that $z' \prec z$, THEN the mutation(s) corresponding to lineage l_z is assigned to this edge and the label of the outgoing edge of the new node

is changed to l_0 and L_0 is incremented by 1. Next, L_z is set to 0 and thus the lineage l_z is made INACTIVE.

Recombination event: If the operation is recombination then, randomly pick a lineage of type l_z and create node v at T_z Then label of z is randomly split into two lineage labels that is compatible with the location of the SNPs on the segment carried by the node v.

L_z is decremented by 1 and the recombination split is best explained by an example.

Illustrative Example: If a lineage $l_z = s_1 s_2 s_3$ is split up due to recombination to produce two lineages $l_{z'} = s_1$ and $l_{z''} = s_2 s_3$ at time t_{z_i}, then L_z is decremented by 1 and:
If $T_{z'} < T_z$, THEN *append* a new node, with time T_z, to list $C_{z'}$, ELSE increment $L_{z'}$ by 1.
If $T_{z''} < T_z$, THEN *append* a new node, with time T_z, to list $C_{z''}$, ELSE increment $L_{z''}$ by 1.

Let there be k_z nodes in list C_z, such that each has a time $< T_z$. Then each of the k_z nodes are removed from the list C_z, L_z is incremented by k_z and these k_z nodes will participate in the next iteration with lineage l_z.

UNTIL l_0 is the only ACTIVE lineage and $L_0 = 1$ (or, $\max_z(T_z) >$ a predefined threshold).

5.2.3 Forward Simulator Model

The model simulates evolution for a full population, forward in time with each generation containing N equal number of males and females, each carrying two chromosomes (see Supplement I for a detailed discussion and extension to selection on multiple loci). The

complex evolutionary relationships between generations yields a number of mutations, recombinations, selected allele inheritance, LD, etc. along the length of chromosome for each individual. This data is recorded in a data structure, which we call the "book of populations". We trace the lineage of each site along the chromosome while tracing the 'book' and constructing the ARG. Inheritance follows the convention of a standard WF model applied to diploid organisms [49], with children randomly picking their parents corresponding to the fitness coefficients when selection is in effect. The stages of the model is described as follows:

fwd-sSimRA: Simulating the "book of populations"

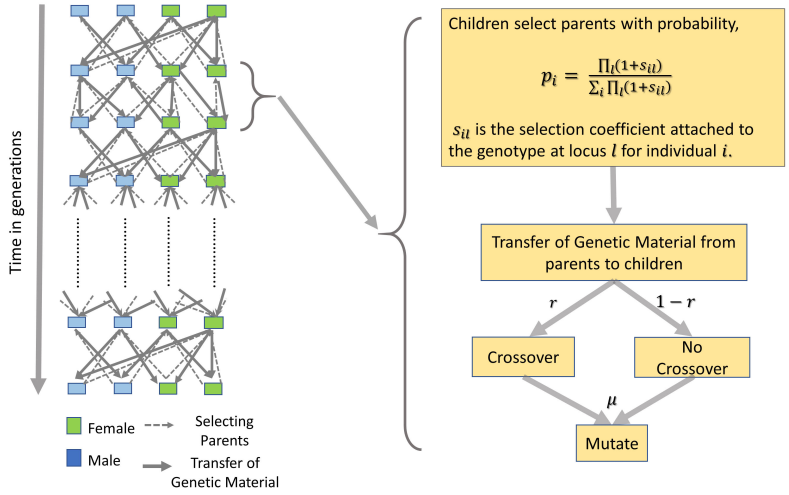
Each chromosome is represented by the alleles at each locus $l \in [1, g]$, which is randomly assigned initially. We use same notations as defined in section 5.2.2 to describe *fwd-sSimRA*. The model assumes that each locus l has a fitness function $s_l(a) \in \mathbb{R}$, where a is an allele comprising the genotype. An individual i with allele a_{il} at locus l is assigned a selection coefficient $s_{il} = s(a_{il})$ which is user-defined, similar to *back-sSimRA*.

The function $s(.)$ denotes the selective pressure and can be varied by intentional specification of recessive, dominant, additive, and other configurations, including homozygous advantage. This function encompasses selection at both single and multiple loci allowing flexible user-defined variations. When selection is not present, we set $s_{il} = 0$.

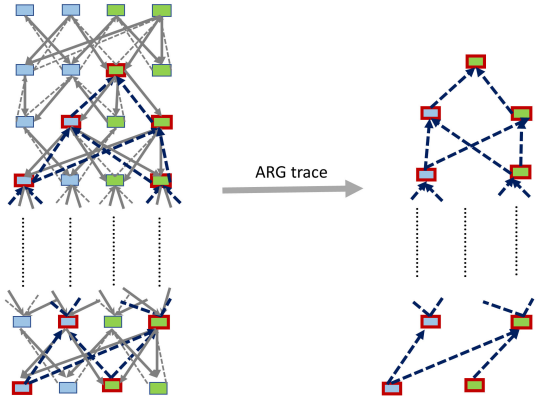
For an individual i, the probability that it has children is given by

$$p_i = \frac{\prod_l (1 + s_{il})}{\sum_i \prod_l (1 + s_{il})} \tag{5.5}$$

In each new generation, as in the WF model, the N children pick their parents with replacement according to the parent probabilities p_i. The simulation is run for $t = \{0, 1, \ldots, G\}$ discrete generations with the $t = 0$ being the base generation, outlined in Figure 5.1(a).



(a) Simulating the "book of populations"



(b) Outline of the building blocks of the forward model.

Figure 5.1.: Outline of the main steps of the forward model. (a)Schematic diagram for simulating the "book of populations" which closely resembles the biological process of evolution. (b) Tracing the ARG from the book of populations (example ARG outlined in red).

Modeling Multiway Epistasis

Multiway epistasis requires multiple interacting loci with similar selection effects. We assign selection coefficients to interacting sites for k-way epistasis, where k is the maximum number

of interacting sites. Let there be q groups of loci, each containing at most k elements and we re-compute equation 5.5 accounting for fitness related to interacting sites as,

$$p_i = \frac{\prod_q (1 + S_{iq})}{\sum_i \prod_q (1 + S_{iq})} \tag{5.6}$$

If a group only has one element, that is if the selected locus is non-interacting, then we allow $S = s$, the user defined selection input. For all other cases, we select S from a matrix or tensor of all possible allele combinations with respect to the number of interacting sites. S is calculated by taking the fitness product of each interacting site as,

$$S = \prod_j \left(1 + \sum_i s_i^{(j)}\right) + e_q \tag{5.7}$$

e_q is the epistatic interaction coefficient for each combination of interacting sites as mentioned Table 5.1 and $s_i^{(j)}$ is the selection coefficient at allele j in individual i's chromosome.

`ALGORITHM`:

1. **Initialization**:
   (a) N individuals ($\frac{N}{2}$ males and $\frac{N}{2}$ females) in the base generation, which remains constant throughout the simulation.
   (b) Number of Generations, $G = c * N$, where c is a constant.
   (c) Randomly allocate genetic material along the length of chromosome, g.
   (d) Assign selection coefficients for interacting sites for two-way epistasis (0 for neutral).
   (e) Set flag, f, for allele(s) under selection on a mutated site (0 for neutral).
2. If f is set, randomly select an individual among N and a site, g_s along g which underwent mutation. Select an allele randomly in g_s and set f to 1.
3. **Loop** For each generation, $t \in \{1, \cdots, G\}$
4. **Loop** For each individual i in $\{1, \cdots, N\}$, in $(t-1)^{th}$ generation.

5. Compute $p_i = \frac{\prod_k (1+S_{ik})}{\sum_i \prod_k (1+S_{ik})}$, where any group k of loci could contain a single locus under selection, for which $S = s$ is defined as the user input. It can also contain a locus interacting with another locus, in a two-way epistasis. In this case s is populated from a matrix formed by the all possible alleles at each loci, from the following form, $S = \prod_j \left(1 + \sum_i s_i^{(j)}\right)$. $s_i^{(j)}$ is the selection cofficient at allele j in individual i's chromosome.

6. Select parents for each child in t^{th} generation based on p_i from $(t-1)^{th}$ generation.

7. **End**

8. For each child i in t^{th} generation, compute scaled recombination rate $r' = r * g$ and select a value, $r_{val} \in [0, 1]$.

9. If $r_{val} = \begin{cases} [0, (1\text{-r'})), & \text{No recombination event} \\ [(1\text{-r'}), 1], & \text{recombination event} \end{cases}$

10. If No recombination event: Randomly pick a chromosome from the parent and assign it's genetic material to the child.

11. Else Randomly pick a crossover index $z \in [1, g]$. Get the genetic material from $[1, z]$ in the first chromosome of the parent and $[(z + 1), g]$ in the second, combine them and assign it to the child.

12. In the child's genetic material, randomly select locations along the chromosome length, g for mutation according to the Poisson distribution and the scaled mutation rate $\mu' = \mu * g$. Change the alleles randomly to other bases. For example, if the allele was A, change it randomly to one of the other bases $\{G, T, C\}$.

13. Update the Chromosomes of the current generation with the new genetic information obtained from the previous generation and continue until the last generation, G.

14. **End**

fwd-sSimRA: Tracing the ARG

Detecting the past recombination events from extant sequences and specifying the place of each recombination is well studied [176–178]. The ARGs define a genealogical graph for all of the chromosomes in a population. Recent advances in population genetics simulators have resulted in tree-sequence recordings which obtains the geneaological history of all genomes in a simulated population [179]. However, no natural ARG is recorded for the interacting loci resulting in LD in forward simulators, in contrast to the backward simulator. Hence, it has to be traced from the "book of populations" from a number of extant haplotypes, outlined in Figure 5.1(b).

We start from m randomly selected extant populations and trace the recombination and coalescent events back each generation. We keep a track of each lineage corresponding to every site along the chromosome and stop when we have found a convergence for all lineages. This final coalescent event along the entire 'book' is known as GMRCA and we output the corresponding ARG.

`ALGORITHM`:

1. **Initialization**:
   (a) Randomly select m number of extant individuals from N in the last generation.
   (b) Select one chromosome out of the two in these m extant samples, randomly. Compute the active lineages, j by comparing the genetic material g in each of the m chromosomes selected.
2. **Loop** for each generation, t going backwards from $\{G, \cdots, 1\}$
3. Identify each chromosome from the previous generation $(t-1)$ which contributed to each chromosome in the current generation, following the book of populations.
4. Check to see if multiple children in the g^{th} generation share the same parent in the previous generation.
5. Iterate and Count the number of active samples, m' in each generation.
6. **Until** $m' = 1$

7. Compute the Height of the GMRCA from the height of convergence.

5.3 Results

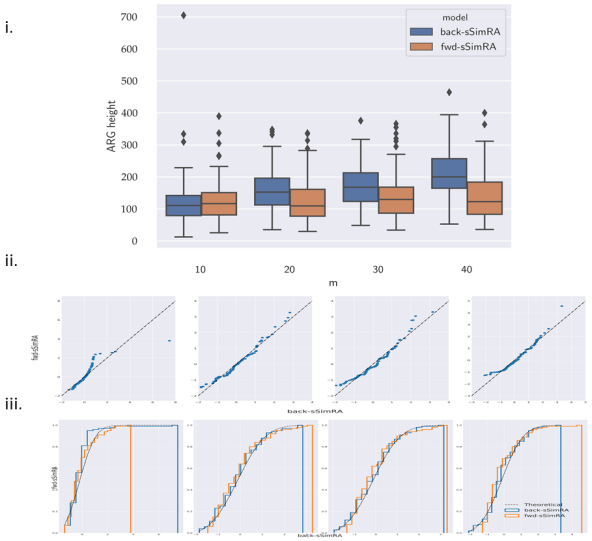
5.3.1 Implementation

Both *back-sSimRA* (implemented in `JAVA` and *fwd-sSimRA* (implemented in `C++`) with `OpenMP` for multithreading are available with an Apache License v2.0 at `https://github.com/ComputationalGenomics/SimRA`.

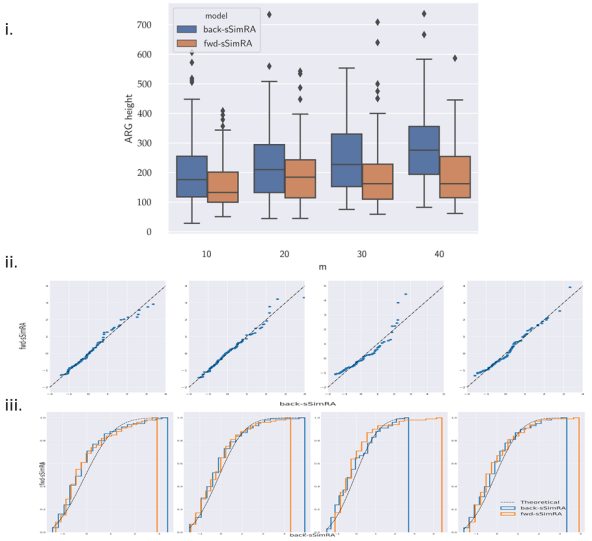
Selection in a diploid heterozygous sample can boost, for one generation, the non-selected chromosome. This can entangle the impact of selection on lineages in the diploid forward model, but not the haploid. We expected the impact of boosted preference to be minimal along any given lineage, since such a boost only occurs for dominant or additive alleles, and then for only one generation. In combinations in a population over time, this effect could be more significant and thus, we sought to test this.

5.3.2 Comparison Study

Comparing the two models under selection calls for an assessment of the values. In both the models, common phenomena such as faster coalescence, decreasing diversity, decreasing number of recombination events occur when we study the individuals under selection. Hence, we compare the height of the ARG or the time to GMRCA, as it is the most significant hallmark of the common history of a sample. We run simulations for different parameter set-ups for the forward and backward model. Each experiment was run 100 times. We demonstrate the accuracy of the two algorithms by comparing the depth of the GMRCA (also known as time to GMRCA, TGMRCA) under different simulation scenarios allowing at most three interacting loci. The simplest scenario in this case is when there is no selection in effect i,e. the neutral coalescent model and when there is selection at a single locus. We show that the two proposed models *back-sSimRA* and *fwd-sSimRA* show agreement in this basic case (Figure A.33). The results for the complex scenario in this setting, accounting for epistasis with three loci are shown in Fig. 5.2, where we show the



(a) Without epistasis



(b) With epistasis

Figure 5.2.: Comparing the height of the ARG (H) between the *fwd-sSimRa* and *back-sSimRA* for selection at two-loci with and without epistasis, respectively. We set $g = 25K$, $r = 1.0 \times 10^{-8}$, $N = 100$, $s = \{0.3, 0.3, 0.3\}$, $e_s = \{0, 0.1\}$ and $m = \{10, 20, 30, 40\}$. (i) The box-and-whisker plot summarizes the result for each m. On each box, the central mark is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. (ii) Q-Q plots for each m showing that the distributions of H from *fwd-sSimRa* and *back-sSimRA* agrees (iii) CDFs of *fwd-sSimRa* and *back-sSimRA* also follow each other closely, reconfirming the agreement between them.

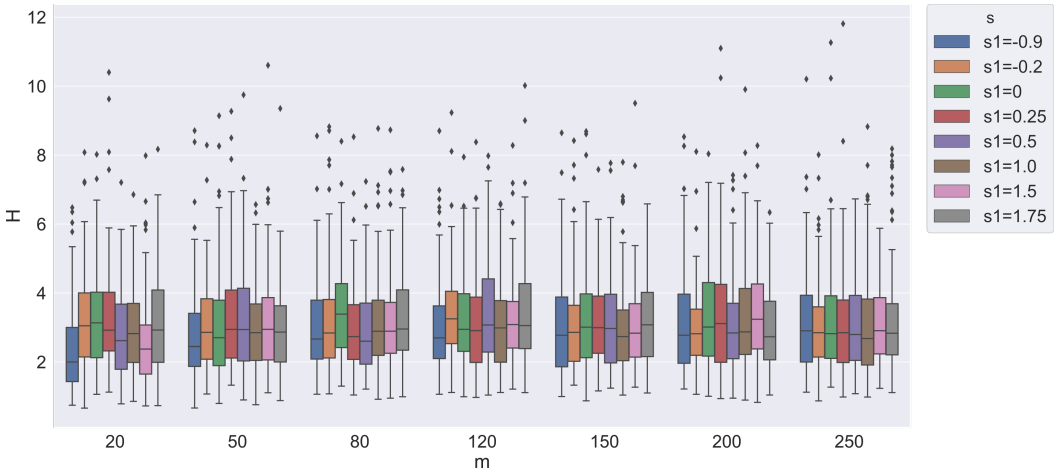


Figure 5.3.: Comparison on the height of the ARG (H)for different s1 values in the case of no recombination for $g = 1000$, $N = 400$ and $m = \{20, 50, 80, 120, 150, 200, 250\}$

concordance for the forward and backward simulation in the case of with/without epistasis (see Figures A.34 and A.35 for detailed comparisons).

To obtain further validation we observed similar agreement in the P-P plots (Figure A.35) between the two simulators in all scenarios. We ran Kolmogorov-Smirnov test on the distributions of H as returned by *fwd-sSimRA* and *back-sSimRA* for all scenarios. We found that for each, the null hypothesis that the two samples are drawn from the same distribution is never rejected and the test statistic is very small (Table A.4).

Furthermore, we compare the height of the lineages under selection in *back-sSimRA* and show how different scenarios impact the height of the ARG (see Fig. 5.3)

Time and Space Comparisons

The coalescent simulator is extremely fast in finding approximations to TGMRCA, in comparison to the forward simulator, as the latter has to build the entire "book of populations" and trace it, as discussed above. That requires it to store every coalescent or recombination event for each site along the chromosome, for each individual in each generation. That increases both the computational and storage overhead of the algorithm. *fwd-sSimRA* makes use of this information and accurately finds the TGMRCA of each complex evolutionary

scenario and provides an exact framework to validate the TGMRCA returned by the coalescent simulator is indeed accurate when complex evolutionary scenarios such as epistasis is in effect.

On the other hand *back-sSimRA* is extremely efficient as it does not require a detailed bookkeeping approach as the forward simulator. It starts from the extant populations and builds the ARG by taking into effect the event (coalescent or recombination) taking minimum time to occure and approximates the TGMRCA of the extant populations.

5.4 Discussion

We present the first algorithm that builds multilocus selection and multiway epistasis into the backward coalescent model with recombinations, as well as, in a forward scheme. Moreover, to the best of our knowledge, this is the first model which took a backward simulator and compared it nose-to-nose with its forward counterpart. Through extensive comparison studies, we show that for complex scenarios with selection and epistasis (or even under neutral scenarios) the hallmark values by the backward and the forward schemes approximately abstract each other. This allows a validation framework for including selection and epistasis into standard population genetic models where we can now study the different scenarios when all the diploids associated with mutated sites along the chromosome with differing fitness values corresponding to the alleles. As the distributions of both the schemes are concordant, we conclude that any one of the simulators (*back-sSimRA* or *fwd-sSimRA*) can be used to understand the effects of negative and positive selection, with multiway epistasis, along with selective sweeps across generations.

6 STRUCTURE INFORMED CLUSTERING FOR POPULATION STRATIFICATION AND GENETIC RISK PREDICTION

6.1 Introduction

The basic principle underlying GWAS is testing for association between genotyped variants for each individual and the trait of interest. GWAS have been extensively used to estimate the signed effects of trait-associated alleles, mapping genes to disorders. GWAS uses LD between genotyped and potentially not typed causal markers, to identify loci implicated in traits and diseases. LD between genetic variants (calculated as squared correlation r^2) can be large only if the allele frequencies at the two loci match [180]. GWASs from common SNP arrays are not as powerful due to this phenomenon as causal variants are usually rare, leading to large number of spurious associations. The power to detect a variant-trait association from LD between an unobserved causal variant and an observed genotype is also largely dependent on the number of observations in the cohort under study. GWAS results have been reported for hundreds of complex traits including both common and rare diseases across various domains such as quantitative traits (as well as binary), brain imaging phenotypes, gene expression, and social and behavioral traits, etc. Over the past decade about 10,000 strong associations between genetic variants and one or more complex traits have been reported [55, 181]. One unambiguous conclusion from GWASs is that for almost any complex trait that has been studied so far, genetic variation is linked with many loci contributing to the polygenic nature of the traits. Hence, on average, the proportion of variance explained at the single marker is very small [21]. The polygenic nature of traits is best explained by heritable height in humans, which is estimated to be modulated by as much as 4% of human allelic variation [182, 183].

Polygenicity of complex traits is known to be one of the potential sources for "missing heritability" [184]. Heritability is defined as the fraction of phenotypic variance explained by additive genetic effects and is related to the coefficient of determination (R^2) of linear

models. Heritability was traditionally estimated by regressing the parental trait against the trait in offspring [185]. The phenomenon of "missing heritability" refers to the positive difference between these heritability estimates and the proportion of phenotypic variance explained by the additive effects of GWAS loci. So far, GWAS on large cohorts has made progress in recovering some missing heritability in height [91] and Schizophrenia [25]. Jointly modeling the effects from multiple genetic variants and fixed effects has shown to recover large proportions of missing heritability for many complex traits [55,186,187]. Linear Mixed Models (LMMs) are used widely to aggregate genetic effects across multiple variants as random effects which work in concert with fixed effects such as environmental and ecological factors. This approach was proposed by Fisher to model inheritance of complex traits [188] and applied heavily on plant and animal breeding [189, 190] before humans.

Another challenge in GWAS is confounding factors such as population structure, which can lead to spurious genotype-trait associations [19, 34, 36]. If a dataset consists of individuals from different ethnic groups then the genotype data will be characterized by genome-wide LD between variants as alleles at different loci tend occur together in individuals from the same ethnic group (as discussed in Section 2.1.6). Population structure cause genuine genetic signals in causal variants to be mirrored in numerous non-causal loci in LD [191], resulting in spurious associations. These are caused by two types of relatedness in population structure: ancestry differences and cryptic relatedness. Ancestry differences is observed when individuals with different ancestral and ethnic background are studied together. Cryptic relatedness is caused by individuals who are closely related and often grouped together by population structure correction strategies posing a more serious confounding problem than ancestry differences [192]. Two popular approaches for population stratification correction involves including the PCs of genotypes as adjustment variables [36, 37] and fitting a LMM with an estimated kinship or GRM from the individual's genotypes [55]. These two widely used approaches are shown to be related to a common model with differing arguments and approaches of building the GRM [33, 52]. The LMM approach requires normally distributed additive small effects of the genetic markers and thus is an approximation. In most cases, the PCs are also used as covariates in LMM based approaches and considered as fixed effects (environmental factors).

A series of recent studies have reported evidence of polygenic adaptation at alleles associated with height in Europeans from the GIANT consortium (253,288 individuals [193]). They observed that alleles related to increasing height are systematically more in frequency in northern compared to southern European populations [28, 29, 31, 194, 195]. More recently, three independent studies [27,28,196] tried to replicate the results found by all of the studies in the more recent and comprehensive UK Biobank cohort (500,000 individuals [197]) which has become a key resource for GWAS with relatively unstructured populations. They found that the previously reported signals of directional selection on height in European populations do not replicate using GWAS effect estimates from the UK Biobank [28]. They further show that the GIANT GWAS is confounded due to stratification along north to south where signals of selection were previously reported. These recent studies highlight the need for more sophisticated tools for population structure confounding correction. Here, we propose an algorithm which corrects for complex arbitrarily structured populations while leveraging the LD induced distances between individuals. We implement CluStrat, which performs agglomerative hierarchical clustering (AHC) using Mahalanobis distance based Genetic Relationship Matrix (GRM), which represents the population-level covariance (LD) matrix for the SNPs.

With growing size of data, computing and storing the genome wide covariance matrix is non-trivial and we get around this overhead by computing the GRM directly using a connection between statistical leverage scores and the Mahalanobis distance. For biobank-scale datasets, we also implement a fast algorithm to approximate all leverage scores, therefore approximating the GRM. We test CluStrat on a large simulation study of discrete and admixed, arbitrarily-structured subpopulations with allele frequencies simulated from widely used Balding-Nichols (BN) and PSD models and to replicate real-world scenarios we simulated genotypes with allele frequencies from HGDP and 1000 Genomes datasets respectively for 500,000 SNPs and 1,000 individuals across 9 different scenarios. We simulated a quantitative (and its binary equivalent) trait with genetic effects at causal loci drawn from the normal distribution and varied the genetic, environmental and noise variances. CluStrat not only observed the lowest number of spurious associations for all the scenarios, but also identified two to three-fold more rare variants at causal loci as obtained by the ubiquitously used Principal Component (PC) based stratification method. CluStrat returned similar

results when applied on the Parkinson's Disease data set from WTCCC cohort, identifying less spurious associations than PCA-based approaches. Harnessing the LD structure by fast approximation of the Mahalanobis distance is also useful in calculating the kinship matrix in LMM for heritability estimation in tera-scale datasets as well as large GWAS summary statistics. Here, we provide a comprehensive guide to stratification and subsequent disorder trait prediction and estimation leveraging the underlying LD structure of the genotypes.

6.2 Materials and Methods

6.2.1 Simulated Datasets

We generated an extensive set of simulations to demonstrate the robustness to different real-world scenarios and power to detect less spurious associations when compared to the standard population stratification correction approaches. We included two widely used methods in the study: (i) without adjusting for population structure in Armitage trend χ^2 association statistic serving as the control; (ii) method of adjusting the trait and genotypes by PCs computed from full set of genotypes.

For each of the 9 simulation configurations, we simulated and analyzed 100 GWAS datasets from a quantitative trait model

$$y_j = \alpha + \sum_{i=1}^{m} \beta_i x_{ij} + \lambda_j + \epsilon_j \tag{6.1}$$

where β_i is the genetic effect of SNP i on the trait, λ_j is the random non-genetic effect and ϵ_j is the random noise variation for individual j. Let Z be a latent variable which captures environmental factors contributed by population structure. Equation 6.1 allows interdependence of structure, lifestyle and environment. We assume $\mathbf{E}\left[\epsilon_j | z_j\right] \sim \mathcal{N}(0, \sigma^2(z_j))$ allowing for heteroskedasticity of the random noise variation [52]. Therefore, $x^j = (x_{1j}, x_{2j}, \cdots, x_{mj})^\top$, λ_j and σ^2 can be thought of as functions of z_j where $Z = (z_1, z_2, \cdots, z_m)$. λ_j is unspecified but along with z_j, they are assumed to be dependent, random variables. Thus, the population

genetic model is dependent on the structure variable z_j for each individual. Similar to the continuous trait model described in Equation 6.1, we define the binary trait model as

$$\log\left(\frac{\Pr\left(y_j=1\right)}{\Pr\left(y_j=0\right)}\right)=\alpha+\sum_{i=1}^{m}\beta_i x_{ij}+\lambda_j \tag{6.2}$$

using the Odds Ratio (OR) as the classifier for disease status from the continuous variable y.

The complete simulation study on quantitative traits with population structure latent variable is constructed in 3 different ways for 3 different proportions of variance among genetic effects, non-genetic effects and random noise, all of which contributing to the trait. Therefore $\mathbf{Var}\left[\sum_{i=1}^{n}\beta_i x_{ij}\right]$, $\mathbf{Var}\left[\sum_{j=1}^{n}\lambda_j\right]$ and $\mathbf{Var}\left[\epsilon_j\right]$ are assigned in proportions of (5%,5%,90%), (10%,0%,90%) and (10%,20%,70%), respectively. Thus, we varied the amount of genetic contribution to the trait for each simulation scenarios and capture variable amount of population structure confounding. We simulated ten truly associated SNPs whose effect sizes were distributed according to a Normal distribution and we set $\beta_i = 0$ for all other non-causal SNPs. We simulated data for $m = 5,000$ and $n = 1,000$ for 100 iterations, spanning $500,000$ SNPs.

The genotype matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ consisting of the simulated allele frequencies was simulated using the algorithm from a previous study [52, 198]. Specifically, we set $\mathbf{F} = \mathbf{TS}$ where $\mathbf{T} \in \mathbb{R}^{m\times d}$ and $\mathbf{S} \in \mathbb{R}^{d\times n}$ where $d \leq n$ is the number of population groups. $\mathbf{S}$ is the matrix containing the population groups encompassing the structure for the individuals shared across all SNPs. On the other hand, $\mathbf{T}$ maps how the structure is manifested in the allele frequencies of each SNP [198]. Finally, projecting $\mathbf{S}$ onto the column space of $\mathbf{T}$ we obtain the allele frequency matrix $\mathbf{F}$. We sample $\mathbf{X}$ as a special case of $\mathbf{F}$ for BN, PSD and TGP (1000 Genomes Project), respectively. We formed $\mathbf{T}$ and $\mathbf{S}$ for the above 3 simulations with 3 scenarios each and continuous traits, resulting in, 9 different evaluation scenarios. The algorithm for constructing $\mathbf{T}$ and $\mathbf{S}$ is detailed in reference [52, 198].

For BN, the allele frequency matrix is simulated from the HapMap phase 3 dataset [199] using three unrelated populations. The final genotype matrix $\mathbf{X}$ is drawn independently at random from the Binomial distribution with parameters n set to 2, denoting the allele status

(0,1 or 2) corresponding to homozygous major/minor or heterozygous with probability p set to the simulated allele frequency for each individual-SNP pair. For PSD, the allele frequency matrix was drawn from the BN frequency distribution. However, it differs from BN in simulating $\mathbf{S}$ by i.i.d draws from Dirichlet distribution with varying α which denotes the parameter influencing the relatedness between the individuals. We show results for $\alpha = 0.1$ here and conducted simulations on a wide range of α values from 0.01 to 1.

6.2.2 Cochran-Armitage trend χ^2

The Armitage trend χ^2 statistic [200] is shown to be more appropriate than a simple χ^2 test for association [192]. We compute the Armitage trend χ^2 in a similar way as done in [37]. The Armitage trend χ^2 is equal to m times the squared Pearson correlation coefficient r^2 between genotype (0,1, or 2) and phenotype (binary or continuous traits), where m is the number of samples.

6.2.3 EIGENSTRAT

EIGENSTRAT [36, 37] involves adjusting the genotypes and phenotypes by the ancestry captured by each axes of variation as described below:

Algorithm 4 Eigen analysis for population stratification correction

Input: $\mathbf{X} \in \mathbb{R}^{m \times n}$, $y \in \mathbb{R}^m$, k number of significant PCs, p-value threshold p
Output: Set of significantly associated SNPs M

1: Compute reduced SVD for as $\mathbf{X} = \mathbf{U}_k \mathbf{\Sigma_k} \mathbf{V}_k^\top$
2: $\mathbf{X}_{adj} = \mathbf{X} - \mathbf{U}_k \mathbf{U}_k^\top \mathbf{X}$
3: $y_{adj} = y - \mathbf{U}_k \mathbf{U}_k^\top y$
4: Compute Armitage trend χ^2 for association between y_{adj} and $\mathbf{X}_{adj}$ and obtain p-values $P = \{p_1, p_2, \cdots, p_m\}$
5: Obtain significant set of SNPs $M = \{m_i |\ p_i < p\}$

After adjusting the genotype and phenotype for population structure, we compute the χ^2 statistic which is equal to $(m - k - 1)$ times r^2 between the adjusted genotype and phenotype, where m is the number of samples, n is the number of SNPs and k is the

number of the axes of variation used to adjust for ancestry. This is a generalization of Armitage trend χ^2 statistic as described above. Correlation between two vectors projected into a lower dimensional embedded subspace, namely the space orthogonal to the $k << n$ axes of variation is tested.

6.2.4 CluStrat

CluStrat provides a LD based clustering framework to capture the population structure and tests for association within each cluster, as described in Algorithm 5:

Algorithm 5 Structure informed clustering to correct for population stratification

1: **Input:** Genotype matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, trait vector $y \in \mathbb{R}^m$, p-value threshold p, number of clusters k
2: **Output:** Set of significantly associated SNPs M
3: $\mathbf{D} = MahDist(\mathbf{X})$
4: $\mathbf{C}$: Cluster membership vector (output of agglomerative hierarchical clustering on $\mathbf{D}$, k clusters)
5: **for** $i = 1 \dots k$
6: $Y_i = y_{C_i}$ and $\mathbf{X}^{(C_i)} = \mathbf{X}_{C_i *}$
7: Find $\hat{\beta}_i^{ridge} = \left(\mathbf{X}^{(C_i)^\top}\mathbf{X}^{(C_i)} + \lambda I\right)^{-1}\mathbf{X}^{(C_i)^\top}Y_i$.
8: Obtain set of significant p-value indices P_i from $\hat{\beta}_i^{ridge}$.
9: **end for**
10: $P = \bigcup_{i \in C} P_i$ and get $\mathbf{X}^{(P_1)} = \mathbf{X}_{*P}$
11: Find $\hat{\beta}^{ridge} = \left(\mathbf{X}^{(P_1)^\top}\mathbf{X}^{(P_1)} + \lambda I\right)^{-1}\mathbf{X}^{(P_1)^\top}y$.
12: Obtain set of p-values P_2 for $\hat{\beta}^{ridge}$.
13: Return M, set of markers corresponding to significant p-values from P_2.

CluStrat computes the distance matrix $\mathbf{D}$ from the normalized genotype matrix $\mathbf{X}$ and performs AHC for a number of clusters k, selected by a cross validation. For each cluster, it runs an association test using ridge regression and obtains p-values for each marker. Thereafter, it computes P_1 the union of intersections of significant associations across all clusters and select the corresponding markers from $\mathbf{X}$ to form $\mathbf{X}^{(P_1)}$. We can interpret this step as a scheme for variable selection. We run another association test with ridge regression

on $\mathbf{X}^{(P_1)}$ to obtain M, the final set of significant associations for all meta-analysis p-values below p.

We now briefly discuss the use of the Mahalanobis distance at the first step of the proposed algorithm. In an arbitrarily structured breeding population, correlation between loci due to LD often results in block-diagonal structures in the genetic relationship matrix. Thus, it is important to account for this LD structure in the computation of the distance matrix [201]. One way to account for the LD structure is to use the squared Mahalanobis distance [202, 203] (denoted as $\mathbf{D}$ in eqn. 6.3). Given a matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ which contains the covariance structure of LD (covariance due to LD between genetic markers), the LD-corrected GRM implementing the Mahalanobis distance is defined as

$$\mathbf{D} = \mathbf{X}\mathbf{G}^{-1}\mathbf{X}^{\top} \tag{6.3}$$

We perform the association test in CluStrat by running ridge regression on each cluster. The regularizer, λ, is chosen by 5-fold cross validation. It is worth noting that we use ridge regression for each cluster as the number of samples is significantly smaller than the number of SNPs, thus making the overall system under-determined. We find the ridge-estimates as follows:

$$\hat{\beta}^{ridge} = (\mathbf{X}^{\top}\mathbf{X} + \lambda \mathbf{I}_n)^{-1}\mathbf{X}^{\top}y = \mathbf{X}^{\top}(\mathbf{X}\mathbf{X}^{\top} + \lambda \mathbf{I}_m)^{-1}y \tag{6.4}$$

We emphasize that the above operation is run for each cluster. We simply dropped the superscripts from $\mathbf{X}$ in the above equation for simplicity. Then, we find the standard error of the estimates in order to calculate the p-values associated with each marker to compute the significance of its association with the trait. The standard error for each marker i in ridge regression is given by

$$SE(\hat{\beta}_i^{ridge}) = \frac{\sigma}{\nu}\left\|(\mathbf{X}\mathbf{X}^{\top} + \lambda \mathbf{I}_m)^{-1}\mathbf{X}_{*i}\right\|_2. \tag{6.5}$$

Recall that $\mathbf{X}_{*i}$ is the i-th column of $\mathbf{X}$ and ν is known as the residual degrees of freedom. We set ν as shown in previous work [204] to the following,

$$\nu = m - c\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top \tag{6.6}$$

for a small scalar constant $c > 0$.

For biobank-scale datasets requiring terabytes of memory, computing the standard error can be a challenge. However, we can use random projection based approaches to sketch the input matrix $\mathbf{X}$ in order to approximate the standard error for each marker. This is indeed a novel contribution of our approach. We delegate details to S1 Appendix. We do note that our work is heavily based on previous work on Randomized Linear Algebra (RLA) [205–208]). To the best of our knowledge, this is the first approximation of the standard error in penalized regression using a sketching based framework and is of independent interest; see also [209] for related work.

Mahalanobis Distance and Leverage Scores

Mahalanobis distance is known to be connected to statistical leverage [210], which is extended in the RandNLA framework as leverage scores. We show this relationship by first noting that Mahalanobis distance is invariant to linear transforbIons, which means the Mahalanobis distance between two vectors,

$$\mathbf{D}(\mathbf{X}_{i*}, \mathbf{X}_{j*}) = (\mathbf{X}_{i*} - \mathbf{X}_{j*})\mathbf{G}^{-1}(\mathbf{X}_{i*} - \mathbf{X}_{j*})^\top \tag{6.7}$$

can have zero means for each vector. The genotype matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, ($n$ markers; m observations) with which we intend to fit the model, must contain an intercept and thus we refer to $\mathbf{X}$ here as the design matrix containing the intercept column followed by one column for each SNP for all the individuals in rows. Furthermore, as we compute the Mahalanobis distance with respect to the *low-rank* genotype matrix $\mathbf{X}_k$, we only consider the *low-rank leverage scores* (rather than the leverage scores of the original matrix $\mathbf{X}$) which are essentially the diagonal elements of the following projection-matrix:

$$\mathbf{H} = \mathbf{X}_k \left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1} \mathbf{X}_k^\top \tag{6.8}$$

and similarly, the off-diagonal elements of $\mathbf{H}$ are called *cross-leverage scores* of $\mathbf{X}_k$.

Now, we will give a clean connection between Mahalanobis distance and these leverage and cross-leverage scores. First, consider the diagonal elements of $\mathbf{H}$ *i.e.* when $i = j$, we have

$$\mathbf{H}_{ii} = (1; \mathbf{X}_{k_{i*}}) \left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1} (1; \mathbf{X}_{k_{i*}})^\top . \tag{6.9}$$

Exploiting the structure of $\left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1}$, we can reformulate it in terms of a block matrix as follows

$$\mathbf{X}_k^\top \mathbf{X}_k = m \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$$

where $\mathbf{C}_{ij} = \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{X}_{k_{\ell i}} \mathbf{X}_{k_{\ell j}} = \frac{m-1}{m} Cov(\mathbf{X}_{k_{*i}}, \mathbf{X}_{k_{*j}}) = \frac{m-1}{m} \mathbf{\Sigma}_{ij}$. $\mathbf{\Sigma}$ here is the corresponding sample covariance matrix. Thus,

$$\left(\mathbf{X}_k^\top \mathbf{X}_k\right)^{-1} = \frac{1}{m} \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{C}^{-1} \end{pmatrix} = \begin{pmatrix} \frac{1}{m} & \mathbf{0}^\top \\ \mathbf{0} & \frac{1}{m-1} \mathbf{\Sigma^{-1}} \end{pmatrix}$$

From Equation 6.9 we obtain

$$\mathbf{H}_i = (1; \mathbf{X}_{k_{i*}}) \begin{pmatrix} \frac{1}{m} & \mathbf{0}^\top \\ \mathbf{0} & \frac{1}{m-1} \mathbf{\Sigma^{-1}} \end{pmatrix} (1; \mathbf{X}_{k_{i*}})^\top \tag{6.10}$$

$$= \frac{1}{m} + \frac{1}{m-1} \mathbf{X}_{k_{i*}} \mathbf{\Sigma^{-1}} \mathbf{X}_{k_{i*}}^\top \tag{6.11}$$

$$= \frac{1}{m} + \frac{1}{m-1} \mathbf{D} \left(\mathbf{X}_{k_{i*}}, 0\right) \tag{6.12}$$

Solving for

$$\mathbf{D}_i = \mathbf{D}(\mathbf{X}_{k_{i*}}, 0)$$

yields,

$$\mathbf{D}_i = (m-1) \left(\mathbf{H}_i - \frac{1}{m}\right)$$

Similarly, we can prove the cross-leverage scores

$$\mathbf{H}_{ij} = \frac{1}{m} + \frac{1}{m-1}\mathbf{X}_{k_{i*}}\mathbf{\Sigma^{-1}}\mathbf{X}_{k_{j*}} \tag{6.13}$$

To prove the relationship of $\mathbf{H}_{ij}$ with $\mathbf{D}_{ij}$ we see,

$$\begin{aligned}
\mathbf{D}(\mathbf{X}_{k_{i*}}, \mathbf{X}_{k_{j*}}) &= (\mathbf{X}_{k_{i*}} - \mathbf{X}_{k_{j*}}^{\top})\mathbf{\Sigma}^{-1}(\mathbf{X}_{k_{i*}} - \mathbf{X}_{k_{j*}}) \\
&= \mathbf{D}(\mathbf{X}_{k_{i*}}, 0) + \mathbf{D}(\mathbf{X}_{k_{j*}}, 0) - 2\mathbf{X}_{k_{i*}}\mathbf{\Sigma}^{-1}\mathbf{X}_{k_{j*}} \\
&= (m-1)(\mathbf{H}_i - \frac{1}{m}) + (m-1)(\mathbf{H}_j - \frac{1}{m}) - 2(m-1)(\mathbf{H}_{ij} - \frac{1}{m}) \\
&= (m-1)(\mathbf{H}_i + \mathbf{H}_j - 2\mathbf{H}_{ij})
\end{aligned}$$

If we take $\mathbf{X}_{k_{i*}} = \mathbf{X}_{k_{j*}}$ then we find $\mathbf{D}(\mathbf{X}_{k_{i*}}, \mathbf{X}_{k_{j*}}) = 0$. Thus, we show that Mahalanobis distance between two vectors can be computed by the corresponding vector's leverage scores.

Now, recall that the rank-k leverage scores of the genotype matrix ($n \gg m$) are defined by the row norms of the matrix of its top k left singular vectors $\mathbf{U}_k \in \mathbb{R}^{m\times k}$. Let $(\mathbf{U}_k)_{i*}$ denote the i-th row of the matrix $\mathbf{U}_k$. Then the rank-k statistical leverage scores of the rows of $\mathbf{X}$, for $i \in 1, \cdots, n$ are given by

$$\mathbf{H}_i = \|(\mathbf{U}_k)_{i*}\|_2^2.$$

Similarly, the rank-k (i,j)-th cross-leverage score, $\mathbf{H}_{ij}$, is equal to the dot product of the i-th and j-th rows of $\mathbf{U}_k$, namely

$$\mathbf{H}_{ij} = \langle(\mathbf{U}_k)_{i*}, (\mathbf{U}_k)_{j*}\rangle. \tag{6.14}$$

Here, $\mathbf{H} \in \mathbb{R}^{m\times m}$ is the matrix of all leverage and cross-leverage scores. We note that $\mathbf{H}_i = \mathbf{H}_{ii} = \|(\mathbf{U}_k)_{i*}\|_2^2 = \left(\mathbf{U}_k\mathbf{U}_k^{\top}\right)_{ii}$ is a special case of the dot product in eqn. 6.14 for the diagonal leverage scores.

Algorithm 6 MahDist : Compute Mahalanobis distance based GRM

1: **Input**: $\mathbf{X} \in \mathbb{R}^{m \times n}$ where $n > m$, k number of PCs to retain
2: **Output**: Mahalanobis GRM $\mathbf{D}$
3: Compute $\mathbf{U}_k$, the matrix of the top k left singular vectors of the genotype matrix $\mathbf{X}$
4: $\mathbf{H} = \mathbf{U}_k \mathbf{U}_k^\top$
5: $\mathbf{D}(\mathbf{X}_{i*}, \mathbf{X}_{j*}) = (m-1)\left(\mathbf{H}_{ii} + \mathbf{H}_{jj} + 2\mathbf{H}_{ij}\right)$
6: Return $\mathbf{D}$

One of the key computational bottlenecks of Mahalanobis distance is computing the inverse of the SNP covariance matrix $\mathbf{G}$ as required in Equation 6.3. In real datasets, with the improvements in genotyping and sequencing technologies, the number of SNPs can be in the millions, thereby making $\mathbf{G}$ in the order of million times million and infeasible to store in secondary memory. Here, we propose the first approxibIon of Mahalanobis distance by computing leverage and cross-leverage scores in a faster and efficient way. As we have shown in Equation 6.13 and 6.10 following up from previous work [210], Mahalanobis distance can be written in terms of leverage scores. Advances in RandNLA community have brought about faster computations for leverage scores as well as cross-leverage scores; hence, we can compute approximations to these scores using random sampling algorithms with theoretical guarantees [205]. For our purposes of demonstrating the proof-of-concept, we work with simulated data as described above for 1,000 individuals and 500,000 SNPs which could be feasibly processed in a personal workstation to compute the deterministic leverage and cross-leverage scores. We note that running SVD on $\mathbf{X}\mathbf{X}^\top$ can be computationally infeasible as the matrix $\mathbf{X}$ in Algorithm 6 will be in the order of m^3 where m is in millions. However, methods such as TeraPCA [68] as detailed in Chapter 4 and other randomized SVD methods [14] can find the approximate invariant low-rank subspace of the higher dimensional space accurately and efficiently.

Agglomerative Hierarchical Clustering

We perform AHC using the LD induced Mahalanobis distance with varying number of clusters. We set the expected number of clusters as $d + k$ where d is the number of populations

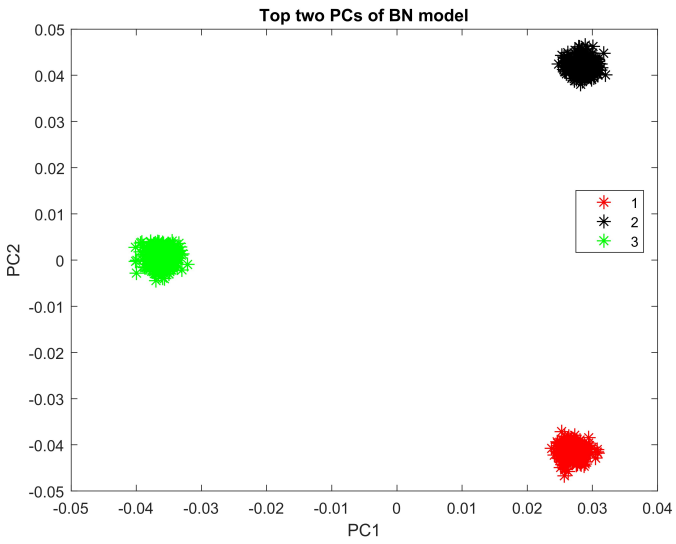


Figure 6.1.: Projection of the samples from three populations simulated from BN model on the top two axes of variation.

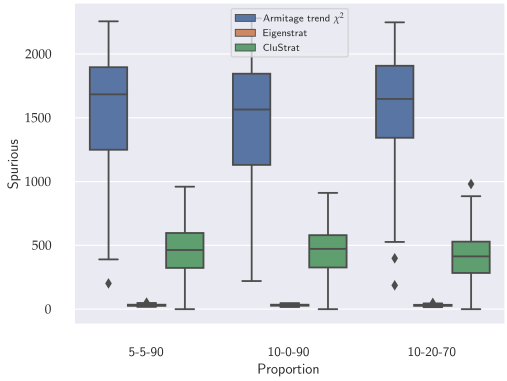
in the data and k ranges from 0 to 5. Therefore, we run the clustering with five different number of clusters and retain the cluster which has the maximum intersection of spurious associations across all the clusters. The observed number of clusters is obtained by the inconsistency method of pruning according to the depth of the dendrogram.

6.3 Results

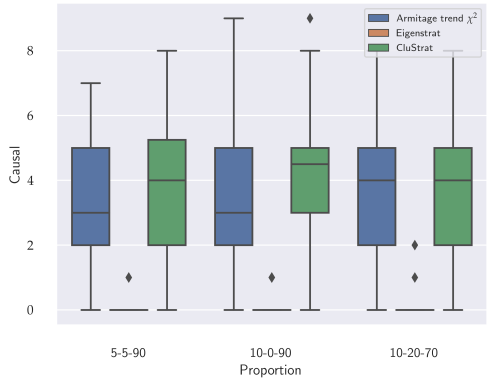
We tested CluStrat on three simulation scenarios spanning from isolated unrelated population structure, arbitrarily structured populations and admixed populations emulating structure of worldwide populations. The genotype data was simulated following the same procedure as described in a prior work [52, 198] with three variance settings regulating the contribution of genetics, environmental variables and noise to the target trait resulting in 9 simulation scenarios.

6.3.1 BN model

The BN model simulates scenarios with unrelated isolated populations (Figure 6.1) and serves as the basic case for arbitrarily structured population with no admixture.



(a) Spurious associations



(b) Causal associations

Figure 6.2.: Box plots for spurious and causal associations on the BN model shows that Armitage trend χ^2 has the maximum number of spurious associations containing about 4-5 causal SNPs whereas EIGENSTRAT has minimum number of spurious associations while detecting almost zero causal SNPs. CluStrat has more spurious associations than EIGENSTRAT and considerably less than Armitage trend χ^2 recovering slightly more number of causal SNPs than the latter.

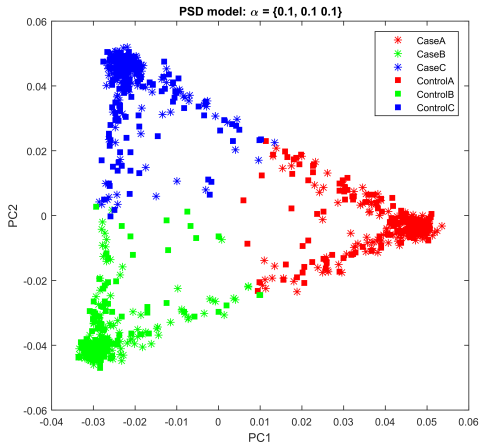
The samples when projected on the top two PCs clearly resembles three isolated clusters with no connections between them. This is an ideal case when the populations are not mixing due to environmental factors acting as barriers of gene flow between populations.

GWAS has shown to be robust in these settings [21], however, the cryptic relatedness for each cluster remains a plaguing issue [28]. We ran CluStrat on this scenario with p-value threshold set to $p = \frac{25}{m_i} = 0.005$ (m_i is the number of SNPs in each iteration, set to 5,000 for 100 iterations). The expected number of spurious association as mentioned in [52] is $m_0 \times p$ where $m_0 = m-$ number of causal SNPs. In our case, as we set the number of causal SNPs to 10 as per [52], $m_0 = 4990$ and therefore, the number of spurious associations to be approximately 25 with degree of freedom set to 1 for genotypes.

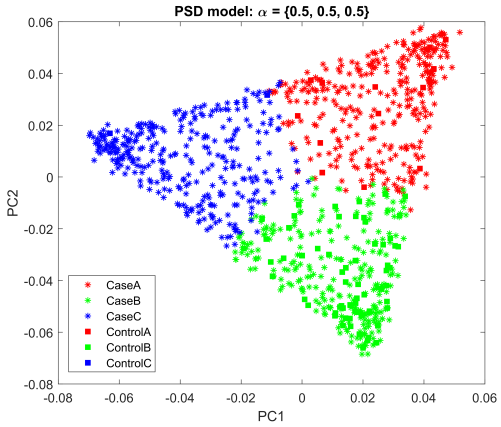
Armitage trend χ^2 with no population structure correction renders almost half of the SNPs in the simulation study as true associations resulting in considerable amount of spurious associations highlighting the need for population structure correction. EIGENSTRAT on the other hand results in the expected number of spurious associations as also shown in previous work [37]. But, it behaves stringently and detects zero causal SNPs almost all of the time (Figure 6.2). CluStrat, however, strikes a balance between the two and generates far more spurious associations than the expected value but about 5 folds less than Armitage trend χ^2 recovering slightly higher number of causal SNPs. This shows that in the ideal case of population structure correction, CluStrat can identify more causal SNPs due to the structure informed clustering setup which widely used stratification correction methods lack.

6.3.2 PSD model

The PSD model emulates real world datasets more closely than BN model. It allows for admixing individuals and gradients across the populations. It is sampled from the Dirichlet distribution parameterized by a concentration parameter $\alpha \in \mathbb{R}^d$ where $d = 3$ (the number of populations for all simulations conducted). Higher value of α_i corresponds to greater weight of i^{th} population. We ran CluStrat on the PSD model with varying number of α from 0.01 to 1, keeping equal α_i for a symmetric distribution. We report the boxplots of spurious and causal associations (Figure 6.4) for $\alpha = \{0.1, 0.1, 0.1\}$ and observe that for the first case of variance $(5\%, 5\%, 90\%)$ Armitage trend χ^2 and CluStrat performs almost similarly in terms of spurious associations. This is due to the fact that only 5% of the trait is explained by



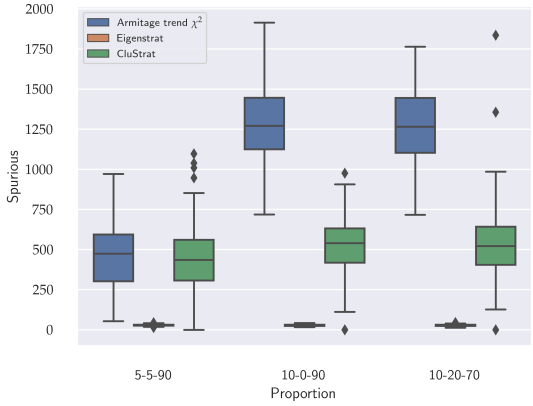
(a) $\alpha = \{0.1, 0.1, 0.1\}$



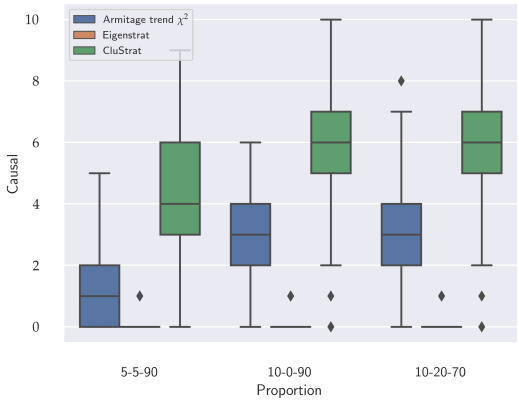
(b) $\alpha = \{0.5, 0.5, 0.5\}$

Figure 6.3.: Projection of the samples from PSD model with varying sets of values of α. We observe that increasing α increases the density between individuals leading to admixture and creates a uniform gradient as all values of α_i are equal.

true genetic associations in presence of LD and the rest is noise and environmental factors. However, CluStrat outnumbers EIGENSTRAT in terms of causal associations and detects four to six fold more true causal SNPs. For the other two variance proportions CluStrat performed better than the other methods in detecting the causal associations and strikes a balance in terms of spurious associations.



(a) Spurious associations

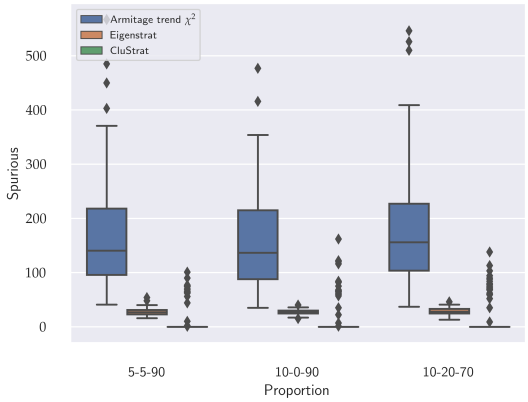


(b) Causal associations

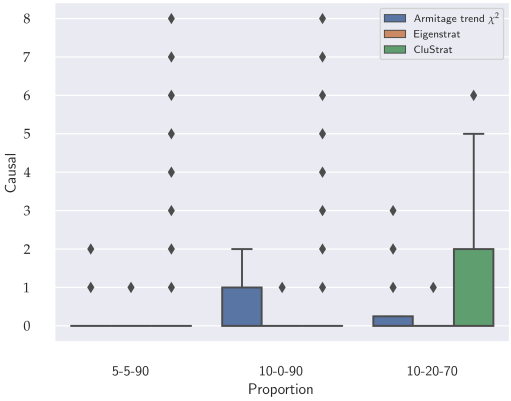
Figure 6.4.: Box plots for spurious and causal associations on the PSD model ($\alpha = \{0.1, 0.1, 0.1\}$) shows Armitage trend χ^2 has maximum number of spurious associations containing less causal SNPs than the BN model (Figure 6.2) owing to the admixed nature of the individuals in PSD. EIGENSTRAT has minimum number of spurious associations while detecting almost zero causal SNPs. CluStrat has more spurious associations than EIGENSTRAT and less than Armitage trend χ^2 recovering two to three fold more causal SNPs.

6.3.3 TGP model

The TGP model is more realistic, drawing from allele frequency distributions from the 1000 Genomes Phase 3 dataset [4]. Projection of individuals from the 1000 Genomes (TGP) dataset on the top two axes of variations shows the distribution of samples across the world (Figure 4.1).



(a) Spurious associations



(b) Causal associations

Figure 6.5.: Box plots for spurious and causal associations on the TGP model shows Armitage trend χ^2 has the maximum number of spurious associations containing less causal SNPs than both the BN and PSD model (Figure 6.2 and 6.4) owing to the distributions of admixed samples across the world of the individuals. CluStrat outperforms both the methods in this scenario as it has the minimum number of spurious associations as well as the highest number of causal SNPs.

CluStrat performs better than EIGENSTRAT (Figure 6.5) in correcting for population structure in real world scenarios such as the TGP data. It captures the minimum number of spurious associations while observing the highest number of true causal SNPs. This shows that structure informed clustering of the genotype data and subsequently performing association tests with regularization outperforms adjusting the genotype and phenotype with the top k PCs explaining the variance of the genetic data.

6.4 Discussion

CluStrat provides a structure informed clustering approach to correct for population structure in a wide variety of simulation scenarios as shown above. We observed that CluStrat outperforms the widely used EIGENSTRAT approach in all of the above scenarios by detecting five to six folds more causal SNPs. Although, it detects more spurious associations than EIGENSTRAT, it is considerably less than the uncorrected Armitage trend χ^2 tests. EIGENSTRAT has been under scrutiny recently as independent studies [27, 28] on UK Biobank [197] failed to replicate the genetic associations of heritable height in Europeans where a positive selection signal was observed in a north to south gradient [10, 29, 195] in the GIANT [193] cohort. These studies attributed the failure to replicate the results in UK Biobank to cryptic relatedness among individuals which the PCA based approaches for population structure correction does not always capture, among other reasons. CluStrat provides a fine structure based clustering approach to tackle cryptic relatedness and ancestral differences among the individuals between and within populations.

As discussed above we chose the Mahalanobis distance metric for CluStrat because it captures the LD induced structure information in the GRM. Thereafter, we established a link between leverage and cross-leverage scores and the Mahalanobis distance. We get around the computational and storage bottlenecks of Mahalanobis distance by computing the leverage and cross-leverage scores. However, we do note that CluStrat do not scale well for realistic datasets of terabyte scale. In our prior work we developed TeraPCA [68] to address this issue of computing the top k left singular vectors of the genotype matrix with number of individuals and markers in the order of millions. We can use TeraPCA to find approximation of the top PCs by performing an out-of-core PCA analysis of massive genomic datasets. Advances in RandNLA community has resulted in faster calculations of leverage scores using random projection methods [205] which can be used to approximate the scores and therefore approximate the Mahalanobis distance. These promising avenues of further work on making CluStrat scalable can be very useful in detecting rare causal variants in various traits as well as common and rare diseases and disorders using GWAS summary statistics or in biobank-scale datasets.

CluStrat with Euclidean distance metric based GRM (sample covariance matrix) also contains structure information as part of the relationships between the individuals within and between population groups. The GRM with Euclidean distance is straightforward to compute as shown below

$$\mathbf{D} = \mathbf{X}\mathbf{X}^\top$$

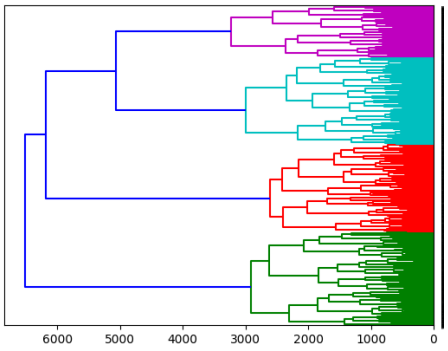
where $\mathbf{X} \in \mathbb{R}^{m \times n}$ $(n >> m)$, but it fails to distinguish fine-grained distances between individuals in the same cluster owing to cryptic relatedness. This is highlighted after performing AHC using Ward's linkage method which minimizes the increase in sum of squares between two cluster centroids in order to decide when to merge them (Figure 6.6).

The Mahalanobis distance in contrast is useful in high-dimensions where Euclidean distance falls short. The Cholesky factorization of the covariance matrix $\mathbf{G} = \mathbf{L}\mathbf{L}^\top$ where $\mathbf{L}$ is the lower diagonal matrix known as the Cholesky factor of $\mathbf{G}$ [201]. We can represent equation 6.3 as

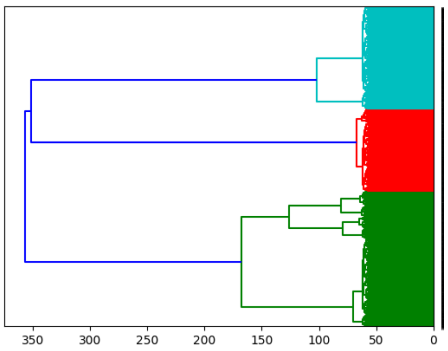
$$\begin{aligned}
\mathbf{X}\mathbf{G}^{-1}\mathbf{X}^\top &= \mathbf{X}\left(\mathbf{L}\mathbf{L}^\top\right)^{-1}\mathbf{X}^\top && (6.15)\\
&= \mathbf{X}(\mathbf{L}^\top)^{-1}(\mathbf{L})^{-1}\mathbf{X}^\top && (6.16)\\
&= \left(\mathbf{X}(\mathbf{L}^{-1})^\top\right)\left(\mathbf{L}^{-1}\mathbf{X}\right)^\top && (6.17)\\
&= \left(\mathbf{L}^{-1}\mathbf{X}^\top\right)^\top\left(\mathbf{L}^{-1}\mathbf{X}^\top\right) && (6.18)\\
&= \mathbf{Q}^\top\mathbf{Q} && (6.19)
\end{aligned}$$

$\mathbf{Q} = \mathbf{L}^{-1}\mathbf{X}^\top$ represents the transformed variables and $\mathbf{Q}^\top\mathbf{Q}$ is the squared Euclidean distance between the transformed variables. Thus Mahalanobis distance accounts for covariance between variables by transforming the data into an uncorrelated form and computing the euclidean distances between them.

When Mahalanobis distance based GRM is used instead of Euclidean distance in AHC on PSD model with 1,000 individuals and 10,000 SNPs across 3 admixed arbitrarily structured ethnic groups, it reveals four broad clusters with various fine-grained sub-clusters revealing how Mahalanobis distance help recover cryptic relatedness and substructure within a population. Due to admixture in the PSD model ($\alpha = \{0.1, 0.1, 0.1\}$) as shown in Figure 6.3



(a) Mahalanobis distance

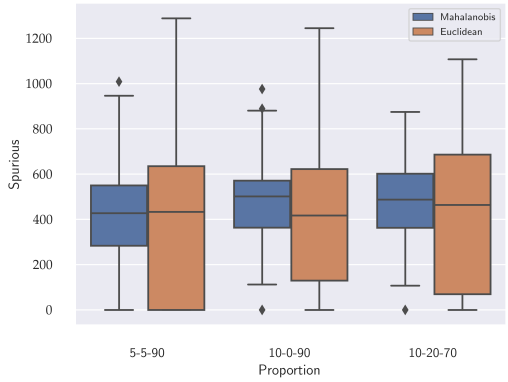


(b) Euclidean distance

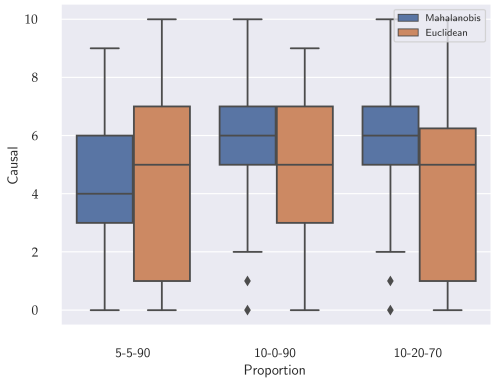
Figure 6.6.: Dendrograms obtained after running AHC with Ward's linkage on PSD model ($\alpha = \{0.1, 0.1, 0.1\}$) shows Mahalanobis distance shows fine grained interactions between the individuals inside a cluster recovering population substructure and cryptic relatedness which Euclidean distance based GRM fails to recover.

the dendrogram finds three broad clusters owing to the three populations in the simulation. It subsequently finds different sub-clusters at different depth on the horizontal axis. Thus, identifying interaction between individuals inside a cluster. This is a significant advantage of using Mahalanobis distance over it's Euclidean counterpart as the latter only reveals three broad clusters with indistinguishable interactions in each cluster (Figure 6.6).

When we ran AHC with both the distances, we observe similar performance on the PSD model with Mahalanobis distance based GRM performing slightly better with respect to



(a) Spurious associations



(b) Causal associations

Figure 6.7.: Box plots for spurious and causal associations obtained by running AHC with Mahalanobis and Euclidean distances on the PSD model ($\alpha = \{0.1, 0.1, 0.1\}$). We observe similar performance on both the distance metrics in terms of identifying true causal variants. Mahalanobis distance discovers less spurious associations than Euclidean distance.

it's Euclidean counterpart (Figure 6.7). We note that, as we increase the scale of admixed genotype data with more complex structure, Mahalanobis distance is better suited as it is known to project correlated high dimensional data to an uncorrelated lower dimensional space where it recovers the hidden Euclidean distances [202].

In this thesis we provide a proof-of-concept of CluStrat and argue that structure informed clustering methods are better suited to capture the cryptic relatedness among individuals within ethnic groups. Availability of higher dimensional datasets such as UK Biobank will

lead to a better evaluation of CluStrat, scrutinizing whether it fails to replicate the north to south gradient of positive selection of height in Europeans as found in prior work [27,28]. Another future direction of CluStrat is to extend it to compute Polygenic Risk Scores (PRS) on a discovery or validation data which is held out and compare with widely used packages such as PRSice2 [211] and LDPred [212] which computes PRS from GWAS summary statistics as well as raw genotypes. LMM methods such as EMMAX [53], GEMMA [213], etc. are ubiquitously used to correct for population structure as well performing almost similar to PCA based methods. Another method, GCAT [52], also performs similar to or better than PCA and LMM approaches however it is not as popular as the former. We have not evaluated performance of CluStrat with respect to these methods in this thesis and in future, we plan to conduct a comprehensive comparison with these methods for population structure correction.

In summary, here, we have highlighted the advantages of biologically inspired distance metrics such as Mahalanobis distance based GRM which captures the cryptic interactions within populations induced by the presence of LD. We evaluated CluStrat on three distinct simulated scenarios of structured populations. We outline how CluStrat outperforms the current widely used PCA based population stratification correction technique in all the scenarios by detecting more true positives. CluStrat detects more false positives than EIGENSTRAT, which can be due to the latter being more stringent in finding genotype-phenotype associations. We also propose various computational challenges in scaling CluStrat and methods to overcome those in order to efficiently compute the GRM deterministically. We highlight the advantages of randomized algorithms to approximate the GRM for tera-scale genotype dataset. Therefore, a comprehensive study on CluStrat and it's performance on other complex simulation scenarios as well as scalable real world datasets would be of particular interest.

7 CONCLUSION AND FUTURE WORK

In this dissertation, we presented various computational methods catering different facets of population genetics. In Chapter 3.1 we analyse the genetic ancestry of the Peloponnesean populations and their relationships with the Slavs and other Europeans, settles a historical controversy that has persisted for over 170 years. Language, social structure and geography create channels of gene flow across populations. However, to date, no study had attempted to establish a quantitative framework in order to dissect the relative contribution of each factor and translate it into a model that correlates with observed population genetic structure. In Chapter 3.2, we establish such an analytic framework called COGG allowing the quantitative assessment of different evolutionary factors as well as the interplay among them. Applying this novel method on a comprehensive dataset from the Indian subcontinent, we are able to uncover major forces that have shaped population genetic structure within India. We seek to extend our computational armamentaria to analyse factors contributing to genetic stratification of a population and reconstruct it's history. As technological advances are allowing us to sequence ancient DNA from thousands of years ago and even from sediments of bones and fossils, we want to develop efficient statistical frameworks to infer patterns of migrations before the last glacial maximum. The software is available to use by GNU GPL-3.0 license with open-source collaborations at https://github.com/aritra90/COGG.

PCA is a statistical workhorse in population genetics, but it does not scale well to modern, massive datasets that are emerging and the ones expected to be generated by large-scale projects in the next few years. In Chapter 4, we present TeraPCA, a multi-threaded, out-of-core implementation of the Randomized Subspace Iteration method compares favourably to current state-of-the-art software tools. TeraPCA builds upon principled and theoretically sound methods to approximate the top principal components of massive covariance matrices, returning highly accurate approximations to the top principal components,

while taking advantage of modern computer architectures that support multi-threading. The software is available to use by GNU GPL license with open-source collaborations at https://github.com/aritra90/TeraPCA and we seek to extend it's usage to cater to gene expression data as well as non-genetic datasets.

We address the task of modeling and simulating complex scenarios of related multiple populations under the effect of natural selection at multiple loci with interacting alleles by building a coalescent simulator, sSimRA and it's forward counterpart, fwdSimRA in Chapter 5. This allows a validation framework for including selection and epistasis into standard population genetic models where we can now study the divergent scenarios when all the diploids associated with mutated sites along the chromosome with diverging fitness values corresponding to the alleles. As the distributions of both the schemes are concordant, we conclude that any one of the simulators (sSimRA or fwdSimRA) can be used to understand the effects of negative and positive selection, with multi-way epistasis, along with selective sweeps across generations. We have successfully built and tested the selection for both models at multiple loci with and without interactions between them. This is the first model to account for epistatic interactions between forward and coalescent simulators and conduct a comprehensive comparison between the two. The code for both the forward and backward simulators is available to use by Apache license with open-source collaborations at https://github.com/ComputationalGenomics/SimRA.

We address the issue of cryptic relatedness in arbitrarily structured isolated and admixed populations respectively in Chapter 6. We implemented CluStrat which uses structure information captured by LD induced GRM computed using Mahalanobis distances between individuals as the distance metric for AHC. Thereafter, it runs ridge regression by cross validation to find association between genotype and phenotype. We compute the Mahalanobis distance by showing a connection of statistical leverage and cross-leverage scores with them, accounting for the computational and storage bottlenecks. We show that CluStrat outperformed the widely used, PCA based, population stratification correction technique in all the three simulation scenarios spanning from isolated, admixed and real world population structure. We also propose various future directions of scaling CluStrat to process on GWAS summary statistics and biobank-scale datasets. Code for

CluStrat is available to use by GNU GPL-3.0 license with open-source collaborations at https://github.com/aritra90/CluStrat.

REFERENCES

REFERENCES

[1] Howard M. Cann, Claudia de Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F. Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, Zhu Chen, Jiayou Chu, Carlo Carcassi, Licinio Contu, Ruofu Du, Laurent Excoffier, G. B. Ferrara, Jonathan S. Friedlaender, Helena Groot, David Gurwitz, Trefor Jenkins, Rene J. Herrera, Xiaoyi Huang, Judith Kidd, Kenneth K. Kidd, Andre Langaney, Alice A. Lin, S. Qasim Mehdi, Peter Parham, Alberto Piazza, Maria Pia Pistillo, Yaping Qian, Qunfang Shu, Jiujin Xu, S. Zhu, James L. Weber, Henry T. Greely, Marcus W. Feldman, Gilles Thomas, Jean Dausset, and L. Luca Cavalli-Sforza. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

[2] L. Luca Cavalli-Sforza. The human genome diversity project: past, present and future. *Nature Reviews Genetics*, 6(4):333–340, 2005.

[3] Richard A. Gibbs, John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, Paul Kwong-Hang Tam, Lap-Chee Tsui, Mary Miu Yee Waye, Jeffrey Tze-Fei Wong, Changqing Zeng, Qingrun Zhang, Mark S. Chee, Luana M. Galver, Semyon Kruglyak, Sarah S. Murray, Arnold R. Oliphant, Alexandre Montpetit, Thomas J. Hudson, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Michael S. Phillips, Andrei Verner, Pui-Yan Kwok, Shenghui Duan, Denise L. Lind, Raymond D. Miller, John P. Rice, Nancy L. Saccone, Patricia Taillon-Miller, Ming Xiao, Yusuke Nakamura, Akihiro Sekine, Koki Sorimachi, Toshihiro Tanaka, Yoichi Tanaka, Tatsuhiko Tsunoda, Eiji Yoshino, David R. Bentley, Panos Deloukas, Sarah Hunt, Don Powell, David Altshuler, Stacey B. Gabriel, Houcan Zhang, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R. Macer, Eiko Suda, Charles N. Rotimi, Clement A. Adebamowo, Toyin Aniagwu, Patricia A. Marshall, Olayemi Matthew, Chibuzor Nkwodimmah, Charmaine D. M. Royal, Mark F. Leppert, Missy Dixon, Lincoln D. Stein, Fiona Cunningham, Ardavan Kanani, Gudmundur A. Thorisson, Aravinda Chakravarti, Peter E. Chen, David J. Cutler, Carl S. Kashuk, Peter Donnelly, Jonathan Marchini, Gilean A. T. McVean, Simon R. Myers, Lon R. Cardon, Gonçalo R. Abecasis, Andrew Morris, Bruce S. Weir, James C. Mullikin, Stephen T. Sherry, Michael Feolo, Mark J. Daly, Stephen F. Schaffner, Renzong Qiu, Alastair Kent, Georgia M. Dunston, Kazuto Kato, Norio Niikawa, Bartha M. Knoppers, Morris W. Foster, Ellen Wright Clayton, Vivian Ota Wang, Jessica Watkin, Erica Sodergren, George M. Weinstock, Richard K. Wilson, Lucinda L. Fulton, Jane Rogers, Bruce W. Birren, Hua Han, Hongguang Wang, Martin Godbout, John C. Wallenburg, Paul L'Archevêque, Guy Bellemare, Kazuo Todani, Takashi Fujita, Satoshi Tanaka, Arthur L. Holden, Eric H. Lai, Francis S. Collins, Lisa D. Brooks, Jean E. McEwen, Mark S. Guyer, Elke Jordan, Jane L. Peterson, Jack Spiegel, Lawrence M. Sung, Lynn F. Zacharia, Karen Kennedy, Michael G. Dunn, Richard Seabrook, Mark Shillito, Barbara Skene, John G. Stewart, David L. Valle (chair), Ellen Wright Clayton (co chair), Lynn B. Jorde (co chair), Mildred K. Cho, Troy Duster, Marla Jasperse, Julio Licinio, Jeffrey C. Long, Pilar N. Ossorio,

Patricia Spallone, Sharon F. Terry, Eric S. Lander (chair), Eric H. Lai (co chair), Deborah A. Nickerson (co chair), Michael Boehnke, Julie A. Douglas, Richard R. Hudson, Leonid Kruglyak, Robert L. Nussbaum, +The International HapMap Consortium, Genotyping centres: Baylor College of Medicine BioScience, ParAllele, Chinese HapMap Consortium, Illumina, McGill University Centre, Génome Québec Innovation, University of California at San Francisco University, Washington, University of Tokyo RIKEN, , Wellcome Trust Sanger Institute, Whitehead Institute/MIT Center for Genome Research, Community engagement/public consultation Institute, sample-collection groups: Beijing Normal University, Beijing Genomics, Eubios Ethics Institute Health Sciences University of Hokkaido, Shinshu University, Howard University Ibadan, University of, University of Utah, Analysis Groups: Cold Spring Harbor Laboratory, Johns Hopkins University School of Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics University of Oxford, US National Institutes of Health, Legal Ethical, Social Issues: Chinese Academy of Social Sciences, Genetic Interest Group, Howard University, Kyoto University, Nagasaki University, University of Montréal, University of Oklahoma, Vanderbilt University, Wellcome Trust, SNP Discovery: Baylor College of Medicine, Washington University, Scientific Management: Chinese Academy of Sciences, Chinese Ministry of Science Technology, , Genome Canada, Génome Québec, Culture Sports Science Japanese Ministry of Education, Technology, The SNP Consortium, Initial Planning Groups: Populations, Legal Ethical, Social Issues Group, and Methods Group. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[4] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Jun Wang, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Paul Flicek, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R. Mardis, Richard K. Wilson, Lucinda Fulton, Robert Fulton, Stephen T. Sherry, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yip-

ing Zhan, Adam Auton, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Richard A. Gibbs, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Donna Muzny, Aniko Sabo, Zhuoyi Huang, Jun Wang, Lachlan J. M. Coin, Lin Fang, Xiaosen Guo, Xin Jin, Guoqing Li, Qibin Li, Yingrui Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, David M. Altshuler, Eric Banks, Gaurav Bhatia, Guillermo del Angel, Stacey B. Gabriel, Giulio Genovese, Namrata Gupta, Heng Li, Seva Kashin, Eric S. Lander, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Andrew G. Clark, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Jan O. Korbel, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Paul Flicek, Kathryn Beal, Laura Clarke, Avik Datta, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Richard E. Smith, Daniel Zerbino, Xiangqun Zheng-Bradley, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, David R. Bentley, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Sudbrak, Vyacheslav S. Amstislavskiy, Ralf Herwig, Elaine R. Mardis, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, Anand Swaroop, Emily Chew, Tuuli Lappalainen, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Jared T. Simpson, Mark D. Shriver, Jeffrey A. Rosenfeld, Carlos D. Bustamante, Stephen B. Montgomery, Francisco M. De La Vega, Jake K. Byrnes, Andrew W. Carroll, Marianne K. DeGorter, Phil Lacroute, Brian K. Maples, Alicia R. Martin, Andres Moreno-Estrada, Suyash S. Shringarpure, Fouad Zakharia, Eran Halperin, Yael Baran, Charles Lee, Eliza Cerveira, Jaeho Hwang, Ankit Malhotra, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, Fiona C. L. Hyland, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Ahmet A. Kurdoglu, Shripad A. Sinari, Kevin Squire, Stephen T. Sherry, Chunlin Xiao, Jonathan Sebat, Danny Antaki, Madhusudan Gujral, Amina Noor, Kenny Ye, Esteban G. Burchard, Ryan D. Hernandez, Christopher R. Gignoux, David Haussler, Sol J. Katzman, W. James Kent, Bryan Howie, Andres Ruiz-Linares, Emmanouil T. Dermitzakis, Scott E. Devine, Gonçalo R. Abecasis, Hyun Min Kang, Jeffrey M. Kidd, Tom Blackwell, Sean Caron, Wei Chen, Sarah Emery, Lars Fritsche, Christian Fuchsberger, Goo Jun, Bingshan Li, Robert Lyons, Chris Scheller, Carlo Sidore, Shiya Song, Elzbieta Sliwerska, Daniel Taliun, Adrian Tan, Ryan Welch, Mary Kate Wing, Xiaowei Zhan, Philip Awadalla, Alan Hodgkinson, Yun Li, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Gil A. McVean, Jonathan L. Marchini, Simon Myers, Claire Churchhouse, Olivier Delaneau, Anjali Gupta-Hinch, Warren Kretzschmar, Zamin Iqbal, Iain Mathieson, Androniki Menelaou, Andy Rimmer, Dionysia K. Xifara, Taras K. Oleksyk, Yunxin Fu, Xiaoming Liu, Momiao Xiong, Lynn Jorde, David Witherspoon, Jinchuan Xing, Evan E. Eichler, Brian L. Browning, Sharon R. Browning, Fereydoun Hormozdiari, Peter H. Sudmant, Ekta Khurana, Richard M. Durbin, Matthew E. Hurles, Chris Tyler-Smith, Cornelis A. Albers, Qasim Ayub, Senduran Balasubramaniam, Yuan Chen, Vincenza Colonna, Petr Danecek, Luke Jostins, Thomas M. Keane, Shane McCarthy, Klaudia Walter, Yali Xue, Mark B. Gerstein, Alexej Abyzov, Suganthi Balasubramanian, Jieming Chen, Declan Clarke, Yao Fu, Arif O. Harmanci, Mike Jin, Donghoon Lee, Jeremy Liu, Xinmeng Jasmine Mu, Jing Zhang, Yan Zhang, Yingrui Li, Ruibang Luo, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee,

Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Steven A. McCarroll, Robert E. Handsaker, David M. Altshuler, Eric Banks, Guillermo del Angel, Giulio Genovese, Chris Hartl, Heng Li, Seva Kashin, James C. Nemesh, Khalid Shakir, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Jeremiah Degenhardt, Jan O. Korbel, Markus H. Fritz, Sascha Meiers, Benjamin Raeder, Tobias Rausch, Adrian M. Stütz, Paul Flicek, Francesco Paolo Casale, Laura Clarke, Richard E. Smith, Oliver Stegle, Xiangqun Zheng-Bradley, David R. Bentley, Bret Barnes, R. Keira Cheetham, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, Richard Shaw, Eric-Wubbo Lameijer, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Li Ding, Ira Hall, Kai Ye, Phil Lacroute, Charles Lee, Eliza Cerveira, Ankit Malhotra, Jaeho Hwang, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, David W. Craig, Nils Homer, Deanna Church, Chunlin Xiao, Jonathan Sebat, Danny Antaki, Vineet Bafna, Jacob Michaelson, Kenny Ye, Scott E. Devine, Eugene J. Gardner, Gonçalo R. Abecasis, Jeffrey M. Kidd, Ryan E. Mills, Gargi Dayama, Sarah Emery, Goo Jun, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Gil A. McVean, Ken Chen, Xian Fan, Zechen Chong, Tenghui Chen, David Witherspoon, Jinchuan Xing, Evan E. Eichler, Mark J. Chaisson, Fereydoun Hormozdiari, John Huddleston, Maika Malig, Bradley J. Nelson, Peter H. Sudmant, Nicholas F. Parrish, Ekta Khurana, Matthew E. Hurles, Ben Blackburne, Sarah J. Lindsay, Zemin Ning, Klaudia Walter, Yujun Zhang, Mark B. Gerstein, Alexej Abyzov, Jieming Chen, Declan Clarke, Hugo Lam, Xinmeng Jasmine Mu, Cristina Sisu, Jing Zhang, Yan Zhang, Richard A. Gibbs, Fuli Yu, Matthew Bainbridge, Danny Challis, Uday S. Evani, Christie Kovar, James Lu, Donna Muzny, Uma Nagaswamy, Jeffrey G. Reid, Aniko Sabo, Jin Yu, Xiaosen Guo, Wangshen Li, Yingrui Li, Renhua Wu, Gabor T. Marth, Erik P. Garrison, Wen Fung Leong, Alistair N. Ward, Guillermo del Angel, Mark A. DePristo, Stacey B. Gabriel, Namrata Gupta, Chris Hartl, Ryan E. Poplin, Andrew G. Clark, Juan L. Rodriguez-Flores, Paul Flicek, Laura Clarke, Richard E. Smith, Xiangqun Zheng-Bradley, Daniel G. MacArthur, Elaine R. Mardis, Robert Fulton, Daniel C. Koboldt, Simon Gravel, Carlos D. Bustamante, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Stephen T. Sherry, Chunlin Xiao, Emmanouil T. Dermitzakis, Gonçalo R. Abecasis, Hyun Min Kang, Gil A. McVean, Mark B. Gerstein, Suganthi Balasubramanian, Lukas Habegger, Haiyuan Yu, Paul Flicek, Laura Clarke, Fiona Cunningham, Ian Dunham, Daniel Zerbino, Xiangqun Zheng-Bradley, Kasper Lage, Jakob Berg Jespersen, Heiko Horn, Stephen B. Montgomery, Marianne K. DeGorter, Ekta Khurana, Chris Tyler-Smith, Yuan Chen, Vincenza Colonna, Yali Xue, Mark B. Gerstein, Suganthi Balasubramanian, Yao Fu, Donghoon Kim, Adam Auton, Anthony Marcketta, Rob Desalle, Apurva Narechania, Melissa A. Wilson Sayres, Erik P. Garrison, Robert E. Handsaker, Seva Kashin, Steven A. McCarroll, Juan L. Rodriguez-Flores, Paul Flicek, Laura Clarke, Xiangqun Zheng-Bradley, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Carlos D. Bustamante, Fernando L. Mendez, G. David Poznik, Peter A. Underhill, Charles Lee, Eliza Cerveira, Ankit Malhotra, Mallory Romanovitch, Chengsheng Zhang, Gonçalo R. Abecasis, Lachlan Coin, Haojing Shao, David Mittelman, Chris Tyler-Smith, Qasim Ayub, Ruby Banerjee, Maria Cerezo, Yuan Chen, Thomas W. Fitzgerald, Sandra Louzada, Andrea Massaia, Shane McCarthy, Graham R. Ritchie, Yali Xue, Fengtang Yang, Richard A. Gibbs, Christie Kovar, Divya Kalra, Walker Hale, Donna Muzny, Jeffrey G. Reid, Jun Wang, Xu Dan, Xiaosen Guo, Guoqing Li, Yingrui Li, Chen Ye, Xiaole Zheng, David M. Altshuler, Paul Flicek, Laura Clarke, Xiangqun Zheng-Bradley, David R. Bentley, Anthony Cox, Sean Humphray, Scott Kahn, Ralf Sudbrak, Marcus W. Albrecht, Matthias Lienhard, David Larson, David W. Craig, Tyler Izatt, Ahmet A. Kurdoglu, Stephen T. Sherry, Chunlin Xiao, David Haussler, Gonçalo R. Abecasis, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, Thomas M. Keane, Shane McCarthy, James

Stalker, Aravinda Chakravarti, Bartha M. Knoppers, Gonçalo R. Abecasis, Kathleen C. Barnes, Christine Beiswanger, Esteban G. Burchard, Carlos D. Bustamante, Hongyu Cai, Hongzhi Cao, Richard M. Durbin, Norman P. Gerry, Neda Gharani, Richard A. Gibbs, Christopher R. Gignoux, Simon Gravel, Brenna Henn, Danielle Jones, Lynn Jorde, Jane S. Kaye, Alon Keinan, Alastair Kent, Angeliki Kerasidou, Yingrui Li, Rasika Mathias, Gil A. McVean, Andres Moreno-Estrada, Pilar N. Ossorio, Michael Parker, Alissa M. Resch, Charles N. Rotimi, Charmaine D. Royal, Karla Sandoval, Yeyang Su, Ralf Sudbrak, Zhongming Tian, Sarah Tishkoff, Lorraine H. Toji, Chris Tyler-Smith, Marc Via, Yuhong Wang, Huanming Yang, Ling Yang, Jiayong Zhu, Walter Bodmer, Gabriel Bedoya, Andres Ruiz-Linares, Zhiming Cai, Yang Gao, Jiayou Chu, Leena Peltonen, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Juan C. Martinez-Cruzado, Taras K. Oleksyk, Kathleen C. Barnes, Rasika A. Mathias, Anselm Hennis, Harold Watson, Colin McKenzie, Firdausi Qadri, Regina LaRocque, Pardis C. Sabeti, Jiayong Zhu, Xiaoyan Deng, Pardis C. Sabeti, Danny Asogun, Onikepe Folarin, Christian Happi, Omonwunmi Omoniwa, Matt Stremlau, Ridhi Tariyal, Muminatou Jallow, Fatoumatta Sisay Joof, Tumani Corrah, Kirk Rockett, Dominic Kwiatkowski, Jaspal Kooner, Trâ‘n T nh Hiê‘n, Sarah J. Dunstan, Nguyen Thuy Hang, Richard Fonnie, Robert Garry, Lansana Kanneh, Lina Moses, Pardis C. Sabeti, John Schieffelin, Donald S. Grant, Carla Gallo, Giovanni Poletti, Danish Saleheen, Asif Rasheed, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Yekaterina Vaydylevich, Eric D. Green, Audrey Duncanson, Michael Dunn, Jeffery A. Schloss, Jun Wang, Huanming Yang, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[5] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10, 03 2015.

[6] Noah A. Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman. Genetic structure of human populations. *Science*, 298(5602):2381–2385, 2002.

[7] John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649, 2008.

[8] Brian Chisholm, L Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza. The History and Geography of Human Genes. *The Journal of Asian Studies*, 54(2):490, 1995.

[9] Felicia Gomez, Jibril Hirbo, and Sarah A. Tishkoff. Genetic variation and adaptation in africa: implications for human evolution and disease. *Cold Spring Harbor perspectives in biology*, 6(7):a008524–a008524, 2014. 24984772[pmid].

[10] Graham Coop, Joseph K. Pickrell, John Novembre, Sridhar Kudaravalli, Jun Li, Devin Absher, Richard M. Myers, Luigi Luca Cavalli-Sforza, Marcus W. Feldman, and Jonathan K. Pritchard. The role of geography in human adaptation. *PLOS Genetics*, 5(6):1–16, 06 2009.

[11] L. Luca Cavalli-Sforza. Genes, peoples, and languages. *Proceedings of the National Academy of Sciences*, 94(15):7719–7724, 1997.

[12] P Menozzi, A Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies in europeans. *Science*, 201(4358):786–792, 1978.

[13] Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2010.

[14] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[15] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[16] John FC Kingman. Origins of the coalescent: 1974-1982. *Genetics*, 156(4):1461–1463, 2000.

[17] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

[18] John Wakeley. The limits of theoretical population genetics. *Genetics*, 169(1):1–7, 2005.

[19] Jonathan Marchini, Lon R. Cardon, Michael S. Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517, 2004.

[20] Hua Tang, Tom Quertermous, Beatriz Rodriguez, Sharon L. R. Kardia, Xiaofeng Zhu, Andrew Brown, James S. Pankow, Michael A. Province, Steven C. Hunt, Eric Boerwinkle, Nicholas J. Schork, and Neil J. Risch. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *American journal of human genetics*, 76(2):268–275, Feb 2005. 15625622[pmid].

[21] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5 – 22, 2017.

[22] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, et al. Fto obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine*, 373(10):895–907, 2015.

[23] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41, 2016.

[24] Katrina M. de Lange, Loukas Moutsianas, James C. Lee, Christopher A. Lamb, Yang Luo, Nicholas A. Kennedy, Luke Jostins, Daniel L. Rice, Javier Gutierrez-Achury, Sun-Gou Ji, Graham Heap, Elaine R. Nimmo, Cathryn Edwards, Paul Henderson, Craig Mowat, Jeremy Sanderson, Jack Satsangi, Alison Simmons, David C. Wilson, Mark Tremelling, Ailsa Hart, Christopher G. Mathew, William G. Newman, Miles Parkes, Charlie W. Lees, Holm Uhlig, Chris Hawkey, Natalie J. Prescott, Tariq Ahmad, John C. Mansfield, Carl A. Anderson, and Jeffrey C. Barrett. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, 49(2):256–261, Feb 2017. 28067908[pmid].

[25] Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421, 2014.

[26] Patrick F Sullivan. The psychiatric gwas consortium: big science comes to psychiatry. *Neuron*, 68(2):182–186, 2010.

[27] Mashaal Sohail, Robert M Maier, Andrea Ganna, Alex Bloemendal, Alicia R Martin, Michael C Turchin, Charleston WK Chiang, Joel Hirschhorn, Mark J Daly, Nick Patterson, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife*, 8:e39702, 2019.

[28] Jeremy J Berg, Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan August Boyle, Xinjun Zhang, Fernando Racimo, Jonathan K Pritchard, et al. Reduced signal for polygenic adaptation of height in uk biobank. *eLife*, 8:e39725, 2019.

[29] Michael C Turchin, Charleston WK Chiang, Cameron D Palmer, Sriram Sankararaman, David Reich, Joel N Hirschhorn, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, et al. Evidence of widespread selection on standing variation in europe at height-associated snps. *Nature genetics*, 44(9):1015, 2012.

[30] Jeremy J Berg and Graham Coop. A population genetic signal of polygenic adaptation. *PLoS genetics*, 10(8):e1004412, 2014.

[31] Matthew R Robinson, Gibran Hemani, Carolina Medina-Gomez, Massimo Mezzavilla, Tonu Esko, Konstantin Shakhbazov, Joseph E Powell, Anna Vinkhuyzen, Sonja I Berndt, Stefan Gustafsson, et al. Population genetic differentiation of height and body mass index across europe. *Nature genetics*, 47(11):1357, 2015.

[32] Anne M Bowcock, Andres Ruiz-Linares, James Tomfohrde, Eric Minch, Judith R Kidd, and L Luca Cavalli-Sforza. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368(6470):455, 1994.

[33] William Astle, David J Balding, et al. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 2009.

[34] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, Jun 2000. 10835412[pmid].

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

[36] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[37] 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, 2006.

[38] Luigi Luca Cavalli-Sforza, Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza. *The history and geography of human genes*. Princeton university press, 1994.

[39] David Reich, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price, and Lalji Singh. Reconstructing Indian population history. *Nature*, 461(7263):489–494, 2009.

[40] Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

[41] Benjamin M. Peter. Admixture, population structure, and f-statistics. *Genetics*, 202(4):1485–1501, 2016.

[42] Iosif Lazaridis, Dani Nadel, Gary Rollefson, Deborah C. Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes, Mario Novak, Beatriz Gamarra, Kendra Sirak, Sarah Connell, Kristin Stewardson, Eadaoin Harney, Qiaomei Fu, Gloria Gonzalez-Fortes, Eppie R. Jones, Songül Alpaslan Roodenberg, György Lengyel, Fanny Bocquentin, Boris Gasparian, Janet M. Monge, Michael Gregg, Vered Eshed, Ahuva-Sivan Mizrahi, Christopher Meiklejohn, Fokke Gerritsen, Luminita Bejenaru, Matthias Blüher, Archie Campbell, Gianpiero Cavalleri, David Comas, Philippe Froguel, Edmund Gilbert, Shona M. Kerr, Peter Kovacs, Johannes Krause, Darren McGettigan, Michael Merrigan, D. Andrew Merriwether, Seamus O'Reilly, Martin B. Richards, Ornella Semino, Michel Shamoon-Pour, Gheorghe Stefanescu, Michael Stumvoll, Anke Tönjes, Antonio Torroni, James F. Wilson, Loic Yengo, Nelli A. Hovhannisyan, Nick Patterson, Ron Pinhasi, and David Reich. Genomic insights into the origin of farming in the ancient near east. *Nature*, 536:419 EP –, Jul 2016. Article.

[43] Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, Qiaomei Fu, Alissa Mittnik, Eszter Banffy, Christos Economou, Michael Francken, Susanne Friederich, Rafael Garrido Pena, Fredrik Hallgren, Valery Khartanovich, Aleksandr Khokhlov, Michael Kunst, Pavel Kuznetsov, Harald Meller, Oleg Mochalov, Vayacheslav Moiseyev, Nicole Nicklisch, Sandra L. Pichler, Roberto Risch, Manuel A. Rojo Guerra, Christina Roth, Anna Szecsenyi-Nagy, Joachim Wahl, Matthias Meyer, Johannes Krause, Dorcas Brown, David Anthony, Alan Cooper, Kurt Werner Alt, and David Reich. Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, 522(7555):207–211, Jun 2015. Letter.

[44] 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, 2012.

[45] Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477, 2008.

[46] Jonathan K Pritchard and Molly Przeworski. Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14, 2001.

[47] WG Hill and Alan Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231, 1968.

[48] J. Hein, M. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory*. Oxford University Press, USA, 2004.

[49] Richard R. Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[50] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[51] Heather J Cordell and David G Clayton. Genetic association studies. *The Lancet*, 366(9491):1121–1131, 2005.

[52] Minsun Song, Wei Hao, and John D Storey. Testing for genetic associations in arbitrarily structured populations. *Nature genetics*, 47(5):550, 2015.

[53] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348, 2010.

[54] Shizhong Xu and William R Atchley. Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics*, 143(3):1417–1424, 1996.

[55] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[56] George Stamatoyannopoulos, Aritra Bose, Athanasios Teodosiadis, Fotis Tsetsos, Anna Plantinga, Nikoletta Psatha, Nikos Zogas, Evangelia Yannaki, Pierre Zalloua, Kenneth K Kidd, Brian L Browning, John Stamatoyannopoulos, Peristera Paschou, and Petros Drineas. Genetics of the peloponnesean populations and the theory of extinction of the medieval peloponnesean Greeks. *European Journal of Human Genetics*, 25(5):637–645, 2017.

[57] Colin Renfrew. *Archaeology and language: the puzzle of Indo-European origins*. CUP Archive, 1990.

[58] Catherine Perlès. *The early Neolithic in Greece: the first farming communities in Europe*. Cambridge University Press, 2001.

[59] J. Chadwick and E.R.C.C.U.H.F.J. Chadwick. *The Mycenaean World*. Cambridge University Press, 1976.

[60] J.T. Hooker. *The coming of the Greeks*. Regina Books, 1999.

[61] Marija Gimbutas. *Proto-Indo-European Culture: The Kurgan Culture During the Fifth, Fourth, and Thord Millennia BC*. Verlag nicht ermittelbar, 1970.

[62] A.D. Godley. *Herodotus*. Number v. 2 in Half-title: The Loeb classical library. W. Heinemann, 1921.

[63] Florin Curta, Paul Stephenson, et al. *Southeastern Europe in the Middle Ages, 500-1250*. Cambridge University Press, 2006.

[64] John VA Fine and John Van Antwerp Fine. *The early medieval Balkans: a critical survey from the sixth to the late twelfth century*. University of Michigan Press, 1991.

[65] J.P. Fallmerayer. *Geschichte der Halbinsel Morea während des Mittelalters: ein historischer Versuch*. Number v. 1 in Geschichte der Halbinsel Morea während des Mittelalters: ein historischer Versuch. In der J.G. Cotta'schen Buchhandlung, 1830.

[66] F. Curta. *The Edinburgh History of the Greeks, C. 500 to 1050: The Early Middle Ages*. Edinburgh University Press Series. Edinburgh University Press, 2011.

[67] Peristera Paschou, Michael W. Mahoney, Asif Javed, Judith R. Kidd, Andrew J. Pakstis, Sheng Gu, Kenneth K. Kidd, and Petros Drineas. Intra- and interpopulation genotype reconstruction from tagging snps. *Genome research*, 17(1):96–107, Jan 2007. 17151345[pmid].

[68] Aritra Bose, Vassilis Kalantzis, Eugenia-Maria Kontopoulou, Mai Elkady, Peristera Paschou, and Petros Drineas. Terapca: a fast and scalable software package to study genetic variation in tera-scale genotypes. *Bioinformatics*, 2019.

[69] Peristera Paschou, Michael W. Mahoney, Asif Javed, Judith R. Kidd, Andrew J. Pakstis, Sheng Gu, Kenneth K. Kidd, and Petros Drineas. Intra- and interpopulation genotype reconstruction from tagging SNPs. *Genome Research*, 17(1):96–107, 2007.

[70] Peristera Paschou, Petros Drineas, Evangelia Yannaki, Anna Razou, Katerina Kanaki, Fotis Tsetsos, Shanmukha Sampath Padmanabhuni, Manolis Michalodimitrakis, Maria C Renda, Sonja Pavlovic, et al. Maritime route of colonization of europe. *Proceedings of the National Academy of Sciences*, 111(25):9211–9216, 2014.

[71] Cornelia Di Gaetano, Nicoletta Cerutti, Francesca Crobu, Carlo Robino, Serena Inturri, Sarah Gino, Simonetta Guarrera, Peter A Underhill, Roy J King, Valentino Romano, et al. Differential greek and northern african migrations to sicily are supported by genetic evidence from the y chromosome. *European Journal of Human Genetics*, 17(1):91, 2009.

[72] Giovanni Fiorito, Cornelia Di Gaetano, Simonetta Guarrera, Fabio Rosa, Marcus W Feldman, Alberto Piazza, and Giuseppe Matullo. The italian genome reflects the history of europe and the mediterranean basin. *European Journal of Human Genetics*, 24(7):1056, 2016.

[73] Peter Charanis. The transfer of population as a policy in the byzantine empire. *Comparative Studies in Society and History*, 3(2):140–154, 1961.

[74] Constantine VII Porphyrogenitus (Emperor of the East) and G. Moravcsik. *Constantine Porphyrogenitus de Administrando Imperio*. Corpus fontium historiae Byzantinae. Dumbarton Oaks Center for Byzantine Studies, 1967.

[75] M. Moosa. *The Maronites in history*. Syracuse University Press, 1986.

[76] Nick Nicholas. A history of the greek colony of corsica. 2005.

[77] Michael Bamshad, Toomas Kivisild, W. Scott Watkins, Mary E. Dixon, Chris E. Ricker, Baskara B. Rao, J. Mastan Naidu, B. V Ravi Prasad, P. Govinda Reddy, Arani Rasanayagam, Surinder S. Papiha, Richard Villems, Alan J. Redd, Michael F. Hammer, Son V. Nguyen, Marion L. Carroll, Mark A. Batzer, and Lynn B. Jorde. Genetic evidence on the origins of Indian caste populations. *Genome Research*, 11(6):994–1004, 2001.

[78] P. P. Majumder. Indian caste origins: Genomic insights and future outlook. *Genome Research*, 11(6):931–932, 2001.

[79] Analabha Basu, Namita Mukherjee, Sangita Roy, Sanghamitra Sengupta, Sanat Banerjee, Madan Chakraborty, Badal Dey, Monami Roy, Bidyut Roy, Nitai P. Bhattacharyya, Susanta Roychoudhury, and Partha P. Majumder. Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Research*, 13(10):2277–2290, 2003.

[80] Analabha Basu, Neeta Sarkar-Roy, and Partha P. Majumder. Genomic reconstruction of the history of extant populations of india reveals five distinct ancestral components and a complex structure. *Proceedings of the National Academy of Sciences*, 113(6):1594–1599, 2016.

[81] Priya Moorjani, Kumarasamy Thangaraj, Nick Patterson, Mark Lipson, Po Ru Loh, Periyasamy Govindaraj, Bonnie Berger, David Reich, and Lalji Singh. Genetic evidence for recent population mixture in India. *American Journal of Human Genetics*, 93(3):422–438, 2013.

[82] Marina Silva, Marisa Oliveira, Daniel Vieira, Andreia Brandão, Teresa Rito, Joana B. Pereira, Ross M. Fraser, Bob Hudson, Francesca Gandini, Ceiridwen Edwards, Maria Pala, John Koch, James F. Wilson, Luísa Pereira, Martin B. Richards, and Pedro Soares. A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. *BMC Evolutionary Biology*, 17(1):88, 2017.

[83] Samir K. Brahmachari, Lalji Singh, Abhay Sharma, Mitali Mukerji, Kunal Ray, Susanta Roychoudhury, G. R. Chandak, K. Thangaraj, Saman Habib, D. Parmar, Partha P. Majumder, Shantanu Sengupta, Dwaipayan Bharadwaj, Debasis Dash, Srikanta K. Rath, R. Shankar, Jagmohan Singh, Komal Virdi, Samira Bahl, V. R. Rao, Swapnil Sinha, Ashok Singh, Amit Mitra, Shrawan K. Mishra, B. R K Shukla, Qadar Pasha, Souvik Maiti, Amitabh Sharma, Jitender Kumar, Aarif Ahsan, Tsering Stobdan, Chitra Chauhan, Saurabh Malhotra, Ajay Vidhani, S. Siva, Aradhita Baral, Rajesh Pandey, Ravishankar Roy, Mridula Singh, S. P. Singh, Nitin Maurya, Arun Bandyopadhyay, Ganga Nath Jha, Somnath Dutta, Gautam Ghosh, Tufan Naiya, Manoj Jain, J. P. Srivatava, J. R. Gupta, Vinay Khanna, Alok Dhawan, Mohini Anand, R. S. Bharti, Madhu Singh, Arvind P. Singh, Anwar J. Khan, Kamlesh Kumar Bisht, Ashok Kumar, Balaram Ghosh, Swapan Kumar Das, Taruna Madan, Chitra Chauhan, Ranjana Verma, Uma Mittal, Anubha Mahajan, Sreenivas Chavali, Rubina Tabassum, Vijaya Banerjee, Jyotsna Batra, Rana Nagarkatti, Shilpy Sharma, Mamta Sharma, Rajshekhar Chatterjee, Jinny A. Paul, Pragya Srivastava, Rupali Chopra, Ankur Saxena, Charu Rajput, Prashant Kumar Singh, Mudit Vaid, Sumantra Das, Keya Chaudhuri, Rukhsana Chowdhury, Arijit Mukhopadhyay, Moulinath Acharya, Ashima Bhattacharyya, Atreyee Saha, Arindam Biswas, Moumita Chaki, Arnab Gupta, Saibal Mukherjee, Suddhasil Mookherjee, Ishita Chattopadhyay, Taraswi Banerjee, Meenakshi Chakravorty, Chaitali Misra, Gourish Monadal, Shiladitya Sengupta, Ishani Deb, and Arunava Banerjee. The Indian Genome Variation database (IGVdb): A project overview. *Human Genetics*, 118(1):1–11, 2005.

[84] Mait Metspalu, Irene Gallego Romero, Bayazit Yunusbayev, Gyaneshwer Chaubey, Chandana Basu Mallick, Georgi Hudjashov, Mari Nelis, Reedik Mägi, Ene Metspalu, Maido Remm, Ramasamy Pitchappan, Lalji Singh, Kumarasamy Thangaraj, Richard Villems, and Toomas Kivisild. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *American Journal of Human Genetics*, 89(6):731–744, 2011.

[85] GaneshPrasad ArunKumar, David F. Soria-Hernanz, Valampuri John Kavitha, Varatharajan Santhakumari Arun, Adhikarla Syama, Kumaran Samy Ashokan, Kavandanpatti Thangaraj Gandhirajan, Koothapuli Vijayakumar, Muthuswamy Narayanan, Mariakuttikan Jayalakshmi, Janet S. Ziegle, Ajay K. Royyuru, Laxmi Parida, R. Spencer Wells, Colin Renfrew, Theodore G. Schurr, Chris Tyler Smith, Daniel E. Platt, and Ramasamy Pitchappan. Population Differentiation of Southern Indian Male Lineages Correlates with Agricultural Expansions Predating the Caste System. *PLoS One*, 7(11), 2012.

[86] S. Roychoudhury, S. Roy, A. Basu, R. Banerjee, H. Vishwanathan, M. V. Usha Rani, S. K. Sil, M. Mitra, and P. P. Majumder. Genomic structures and population histories of linguistically distinct tribal groups of India. *Human Genetics*, 109(3):339–350, 2001.

[87] Vagheesh M Narasimhan, Nick J Patterson, Priya Moorjani, Iosif Lazaridis, Lipson Mark, Swapan Mallick, Nadin Rohland, Rebecca Bernardos, Alexander M Kim, Nathan Nakatsuka, Inigo Olalde, Alfredo Coppa, James Mallory, Vyacheslav Moiseyev, Janet Monge, Luca M Olivieri, Nicole Adamski, Nasreen Broomandkhoshbacht, Francesca Candilio, Olivia Cheronet, Brendan J Culleton, Matthew Ferry, Daniel Fernandes, Beatriz Gamarra, Daniel Gaudio, Mateja Hajdinjak, Eadaoin Harney, Thomas K Harper, Denise Keating, Ann-Marie Lawson, Megan Michel, Mario Novak, Jonas Oppenheimer, Niraj Rai, Kendra Sirak, Viviane Slon, Kristin Stewardson, Zhao Zhang, Gaziz Akhatov, Anatoly N Bagashev, Baurzhan Baitanayev, Gian Luca Bonora, Tatiana Chikisheva, Anatoly Derevianko, Enshin Dmitry, Katerina Douka, Nadezhda Dubova, Andrey Epimakhov, Suzanne Freilich, Dorian Fuller, Alexander Goryachev, Andrey Gromov, Bryan Hanks, Margaret Judd, Erlan Kazizov, Aleksander Khokhlov, Egor Kitov, Elena Kupriyanova, Pavel Kuznetsov, Donata Luiselli, Farhad Maksudov, Chris Meiklejohn, Deborah C Merrett, Roberto Micheli, Oleg Mochalov, Zahir Muhammed, Samridin Mustafakulov, Ayushi Nayak, Rykun M Petrovna, Davide Pettner, Richard Potts, Dmitry Razhev, Stefania Sarno, Kulyan Sikhymbaevae, Sergey M Slepchenko, Nadezhda Stepanova, Svetlana Svyatko, Sergey Vasilyev, Massimo Vidale, Dima Voyakin, Antonina Yermolayeva, Alisa Zubova, Vasant S Shinde, Carles Lalueza-Fox, Matthias Meyer, David Anthony, Nicole Boivin, Kumarasmy Thangaraj, Douglas Kennett, Michael Frachetti, Ron Pinhasi, and David Reich. The genomic formation of south and central asia. *bioRxiv*, 2018.

[88] Ajai K. Pathak, Anurag Kadian, Alena Kushniarevich, Francesco Montinaro, Mayukh Mondal, Linda Ongaro, Manvendra Singh, Pramod Kumar, Niraj Rai, JÃ¼ri Parik, Ene Metspalu, Siiri Rootsi, Luca Pagani, Toomas Kivisild, Mait Metspalu, Gyaneshwer Chaubey, and Richard Villems. The genetic ancestry of modern indus valley populations from northwest india. *The American Journal of Human Genetics*, 103(6):918 – 929, 2018.

[89] Partha P. Majumder. The Human Genetic History of South Asia. *Current Biology*, 20(4):R184–R187, 2010.

[90] R. Thapar. *Early India: From the Origins to AD 1300*. University of California Press, 2004.

[91] Stephen Wooding, Christopher Ostler, B. V. Ravi Prasad, W. Scott Watkins, Sandy Sung, Mike Bamshad, and Lynn B. Jorde. Directional migration in the hindu castes: inferences from mitochondrial, autosomal and y-chromosomal data. *Human Genetics*, 115(3):221–229, Aug 2004.

[92] Gideon S. Bradburd, Peter L. Ralph, and Graham M. Coop. Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, 67(11):3258–3273, 2013.

[93] Carina M. Schlebusch, Pontus Skoglund, Per Sjödin, Lucie M. Gattepaille, Dena Hernandez, Flora Jay, Sen Li, Michael De Jongh, Andrew Singleton, Michael G. B. Blum, Himla Soodyall, and Mattias Jakobsson. Genomic variation in seven khoe-san groups reveals adaptation and complex african history. *Science*, 338(6105):374–379, 2012.

[94] Christopher C. Chang, Carson C. Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, 2015.

[95] Gyaneshwer Chaubey, Mait Metspalu, Ying Choi, Reedik Mägi, Irene Gallego Romero, Pedro Soares, Mannis Van Oven, Doron M. Behar, Siiri Rootsi, Georgi Hudjashov, Chandana Basu Mallick, Monika Karmin, Mari Nelis, Jüri Parik, Alla Goverdhana Reddy, Ene Metspalu, George Van Driem, Yali Xue, Chris Tyler-Smith, Kumarasamy Thangaraj, Lalji Singh, Maido Remm, Martin B. Richards, Marta Mirazon Lahr, Manfred Kayser, Richard Villems, and Toomas Kivisild. Population genetic structure in indian austroasiatic speakers: The role of landscape barriers and sex-specific admixture. *Molecular Biology and Evolution*, 28(2):1013–1024, 2011.

[96] A. S. Hinrichs. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(90001):D590–D598, 2006.

[97] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, 81(3):559–575, 2007.

[98] Julie Di Cristofaro, Erwan Pennarun, Stéphane Mazières, Natalie M. Myres, Alice a. Lin, Shah Aga Temori, Mait Metspalu, Ene Metspalu, Michael Witzel, Roy J. King, Peter a. Underhill, Richard Villems, and Jacques Chiaroni. Afghan Hindu Kush: Where Eurasian Sub-Continent Gene Flows Converge. *PLoS One*, 8(10), 2013.

[99] Bayazit Yunusbayev, Mait Metspalu, Mari Järve, Ildus Kutuev, Siiri Rootsi, Ene Metspalu, Doron M. Behar, Kärt Varendi, Hovhannes Sahakyan, Rita Khusainova, Levon Yepiskoposyan, Elza K. Khusnutdinova, Peter A. Underhill, Toomas Kivisild, and Richard Villems. The caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Molecular Biology and Evolution*, 29(1):359–365, 2012.

[100] Bayazit Yunusbayev, Mait Metspalu, Ene Metspalu, Albert Valeev, Sergei Litvinov, Ruslan Valiev, Vita Akhmetova, Elena Balanovska, Oleg Balanovsky, Shahlo Turdikulova, Dilbar Dalimova, Pagbajabyn Nymadawa, Ardeshir Bahmanimehr, Hovhannes Sahakyan, Kristiina Tambets, Sardana Fedorova, Nikolay Barashkov, Irina Khidiyatova, Evelin Mihailov, Rita Khusainova, Larisa Damba, Miroslava Derenko, Boris Malyarchuk, Ludmila Osipova, Mikhail Voevoda, Levon Yepiskoposyan, Toomas Kivisild, Elza Khusnutdinova, and Richard Villems. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *PLoS Genetics*, 11(4):1–24, 2015.

[101] Lejla Kovacevic, Kristiina Tambets, Anne-Mai Ilumäe, Alena Kushniarevich, Bayazit Yunusbayev, Anu Solnik, Tamer Bego, Dragan Primorac, Vedrana Skaro, Andreja Leskovac, Zlatko Jakovski, Katja Drobnic, Helle-Viivi Tolk, Sandra Kovacevic, Pavao Rudan, Ene Metspalu, and Damir Marjanovic. Standing at the Gateway to Europe - The Genetic Structure of Western Balkan Populations Based on Autosomal and Haploid Markers. *PLoS One*, 9(8):e105090, 2014.

[102] Doron M. Behar, Bayazit Yunusbayev, Mait Metspalu, Ene Metspalu, Saharon Rosset, Juri Parik, Siiri Rootsi, Gyaneshwer Chaubey, Ildus Kutuev, Guennady Yudkovsky, Elza K. Khusnutdinova, Oleg Balanovsky, Ornella Semino, Luisa Pereira, David Comas, David Gurwitz, Batsheva Bonne-Tamir, Tudor Parfitt, Michael F. Hammer, Karl Skorecki, and Richard Villems. The genome-wide structure of the jewish people. *Nature*, 466(7303):238–242, Jul 2010.

[103] S.A. Fedorova et al. Autosomal and uniparental portraits of the native populations of sakha (yakutia): implications for the peopling of northeast eurasia. *BMC Evol. Biol.*, 13(1):127, 2013.

[104] M. Raghavan et al. Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature*, 505(7481):87–91, Jan 2014. Letter.

[105] H. Rajeevan, M. V. Osier, K. H. Cheung, H. Deng, L. Druskin, R. Heinzen, J. R. Kidd, S. Stein, A. J. Pakstis, N. P. Tosches, C. C. Yeh, P. L. Miller, and K. K. Kidd. ALFRED: The ALelle FREquency Database. Update. *Nucleic Acids Research*, 31(1):270–271, 2003.

[106] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936.

[107] Haim Avron, Christos Boutsidis, Sivan Toledo, and Anastasios Zouzias. Efficient dimensionality reduction for canonical correlation analysis. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 347–355, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[108] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[109] Joseph K. Pickrell and Jonathan K. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, 8(11):1–17, 11 2012.

[110] Petros Drineas, Jamey Lewis, and Peristera Paschou. Inferring geographic coordinates of origin for europeans using small panels of ancestry informative markers. *PLOS ONE*, 5(8):1–6, 08 2010.

[111] Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 775–783, Cambridge, MA, USA, 2015. MIT Press.

[112] Kai Tätte, Luca Pagani, Ajai Kumar Pathak, Sulev Kõks, Binh Ho Duy, Xuan Dung Ho, Gazi Nurun Nahar Sultana, Mohd Istiaq Sharif, Md Asaduzzaman, Doron M. Behar, Yarin Hadid, Richard Villems, Gyaneshwer Chaubey, Toomas Kivisild, and Mait Metspalu. The genetic legacy of continental scale admixture in indian austroasiatic speakers. *bioRxiv*, 2018.

[113] D.D. Kosambi. *The culture and civilisation of Ancient India in Historical Outline.* Vikas Publishing House Pvt. Ltd., 1964.

[114] Gyaneshwer Chaubey, Rakesh Tamang, Erwan Pennarun, Pavan Dubey, Niraj Rai, Rakesh Kumar Upadhyay, Rajendra Prasad Meena, Jayanti R. Patel, George van Driem, Kumarasamy Thangaraj, Mait Metspalu, and Richard Villems. Reconstructing the population history of the largest tribe of india: the dravidian speaking gond. *Eur J Hum Genet*, 25(4):493–498, Apr 2017.

[115] Gyaneshwer Chaubey, Manvendra Singh, Federica Crivellaro, Rakesh Tamang, Amrita Nandan, Kamayani Singh, Varun Kumar Sharma, Ajai Kumar Pathak, Anish M. Shah, Vishwas Sharma, Vipin Kumar Singh, Deepa Selvi Rani, Niraj Rai, Alena Kushniarevich, Anne-Mai Ilumae, Monika Karmin, Anand Phillip, Abhilasha Verma,

Erik Prank, Vijay Kumar Singh, Blaise Li, Periyasamy Govindaraj, Akhilesh Kumar Chaubey, Pavan Kumar Dubey, Alla G. Reddy, Kumpati Premkumar, Satti Vishnupriya, Veena Pande, Juri Parik, Siiri Rootsi, Phillip Endicott, Mait Metspalu, Marta Mirazon Lahr, George van Driem, Richard Villems, Toomas Kivisild, Lalji Singh, and Kumarasamy Thangaraj. Unravelling the distinct strains of tharu ancestry. *Eur J Hum Genet*, 22(12):1404–1412, Dec 2014. Article.

[116] Dorian Q Fuller. An agricultural perspective on dravidian historical linguistics: archaeological crop packages, livestock and dravidian crop vocabulary. McDonald Institute for Archaeological Research, 2003.

[117] Morten Rasmussen, Yingrui Li, Stinus Lindgreen, Jakob Skou Pedersen, Anders Albrechtsen, Ida Moltke, Mait Metspalu, Ene Metspalu, Toomas Kivisild, Ramneek Gupta, Marcelo Bertalan, Kasper Nielsen, M. Thomas P. Gilbert, Yong Wang, Maanasa Raghavan, Paula F. Campos, Hanne Munkholm Kamp, Andrew S. Wilson, Andrew Gledhill, Silvana Tridico, Michael Bunce, Eline D. Lorenzen, Jonas Binladen, Xiaosen Guo, Jing Zhao, Xiuqing Zhang, Hao Zhang, Zhuo Li, Minfeng Chen, Ludovic Orlando, Karsten Kristiansen, Mads Bak, Niels Tommerup, Christian Bendixen, Tracey L. Pierre, Bjarne Grønnow, Morten Meldgaard, Claus Andreasen, Sardana A. Fedorova, Ludmila P. Osipova, Thomas F. G. Higham, Christopher Bronk Ramsey, Thomas v. O. Hansen, Finn C. Nielsen, Michael H. Crawford, Søren Brunak, Thomas Sicheritz-Pontén, Richard Villems, Rasmus Nielsen, Anders Krogh, Jun Wang, and Eske Willerslev. Ancient human genome sequence of an extinct palaeo-eskimo. *Nature*, 463(7282):757–762, Feb 2010.

[118] Elena Arciero, Thirsa Kraaijenbrink, Asan, Marc Haber, Massimo Mezzavilla, Qasim Ayub, Wei Wang, Zhaxi Pingcuo, Huanming Yang, Jian Wang, Mark A. Jobling, George van Driem, Yali Xue, Peter de Knijff, and Chris Tyler-Smith. Demographic history and genetic adaptation in the himalayan region inferred from genome-wide snp genotypes of 49 populations. *Mol Biol Evol*, 35(8):1916–1933, Aug 2018. 29796643[pmid].

[119] Bing Su, Chunjie Xiao, Ranjan Deka, Mark T. Seielstad, Daoroong Kangwanpong, Junhua Xiao, Daru Lu, Peter Underhill, Luca Cavalli-Sforza, Ranajit Chakraborty, and Li Jin. Y chromosome haplotypes reveal prehistorical migrations to the himalayas. *Human Genetics*, 107(6):582–590, Dec 2000.

[120] Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, Mark Lipson, Bonnie Berger, Christos Economou, Ruth Bollongino, Qiaomei Fu, Kirsten I. Bos, Susanne Nordenfelt, Heng Li, Cesare de Filippo, Kay Prüfer, Susanna Sawyer, Cosimo Posth, Wolfgang Haak, Fredrik Hallgren, Elin Fornander, Nadin Rohland, Dominique Delsate, Michael Francken, Jean-Michel Guinet, Joachim Wahl, George Ayodo, Hamza A. Babiker, Graciela Bailliet, Elena Balanovska, Oleg Balanovsky, Ramiro Barrantes, Gabriel Bedoya, Haim Ben-Ami, Judit Bene, Fouad Berrada, Claudio M. Bravi, Francesca Brisighelli, George B. J. Busby, Francesco Cali, Mikhail Churnosov, David E. C. Cole, Daniel Corach, Larissa Damba, George van Driem, Stanislav Dryomov, Jean-Michel Dugoujon, Sardana A. Fedorova, Irene Gallego Romero, Marina Gubina, Michael Hammer, Brenna M. Henn, Tor Hervig, Ugur Hodoglugil, Aashish R. Jha, Sena Karachanak-Yankova, Rita Khusainova, Elza Khusnutdinova, Rick Kittles, Toomas Kivisild, William Klitz, Vaidutis Kucinskas, Alena Kushniarevich, Leila Laredj, Sergey Litvinov, Theologos Loukidis, Robert W. Mahley, Béla Melegh, Ene Metspalu, Julio Molina, Joanna Mountain, Klemetti

Näkkäläjärvi, Desislava Nesheva, Thomas Nyambo, Ludmila Osipova, Jüri Parik, Fedor Platonov, Olga Posukh, Valentino Romano, Francisco Rothhammer, Igor Rudan, Ruslan Ruizbakiev, Hovhannes Sahakyan, Antti Sajantila, Antonio Salas, Elena B. Starikovskaya, Ayele Tarekegn, Draga Toncheva, Shahlo Turdikulova, Ingrida Uktveryte, Olga Utevska, René Vasquez, Mercedes Villena, Mikhail Voevoda, Cheryl A. Winkler, Levon Yepiskoposyan, Pierre Zalloua, Tatijana Zemunik, Alan Cooper, Cristian Capelli, Mark G. Thomas, Andres Ruiz-Linares, Sarah A. Tishkoff, Lalji Singh, Kumarasamy Thangaraj, Richard Villems, David Comas, Rem Sukernik, Mait Metspalu, Matthias Meyer, Evan E. Eichler, Joachim Burger, Montgomery Slatkin, Svante Pääbo, Janet Kelso, David Reich, and Johannes Krause. Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature*, 513:409 EP –, Sep 2014.

[121] Michael Witzel. Substrate languages in old indo-aryan. (gvedic, middle and late vedic). *Electronic Journal of Vedic Studies*, 5(1):1–67, 2016.

[122] J.P. Mallory and D.Q. Adams. *Encyclopedia of Indo-European Culture*. Fitzroy Dearborn, 1997.

[123] Rebecca L. Cann. Genetic clues to dispersal in human populations: Retracing the past from the present. *Science*, 291(5509):1742–1748, 2001.

[124] Paul Mellars. Going east: New genetic and archaeological perspectives on the modern human colonization of eurasia. *Science*, 313(5788):796–800, 2006.

[125] Vincent Macaulay, Catherine Hill, Alessandro Achilli, Chiara Rengo, Douglas Clarke, William Meehan, James Blackburn, Ornella Semino, Rosaria Scozzari, Fulvio Cruciani, Adi Taha, Norazila Kassim Shaari, Joseph Maripa Raja, Patimah Ismail, Zafarina Zainuddin, William Goodwin, David Bulbeck, Hans-Jürgen Bandelt, Stephen Oppenheimer, Antonio Torroni, and Martin Richards. Single, rapid coastal settlement of asia revealed by analysis of complete mitochondrial genomes. *Science*, 308(5724):1034–1036, 2005.

[126] Lluís Quintana-Murci, Ornella Semino, Hans-J Bandelt, Giuseppe Passarino, Ken McElreavey, and A. Silvana Santachiara-Benerecetti. Genetic evidence of an early exit of homo sapiens sapiens from africa through eastern africa. *Nature Genetics*, 23:437, Dec 1999.

[127] Sanghamitra Sengupta, Lev A. Zhivotovsky, Roy King, S. Q. Mehdi, Christopher A. Edmonds, Cheryl-Emiliane T. Chow, Alice A. Lin, Mitashree Mitra, Samir K. Sil, A. Ramesh, M. V. Usha Rani, Chitra M. Thakur, L. Luca Cavalli-Sforza, Partha P. Majumder, and Peter A. Underhill. Polarity and temporality of high-resolution y-chromosome distributions in india identify both indigenous and exogenous expansions and reveal minor genetic influence of central asian pastoralists. *American Journal of Human Genetics*, 78(2):202–221, 2017/06/01 2006.

[128] Richard Cordaux, Robert Aunger, Gillian Bentley, Ivane Nasidze, S.M. Sirajuddin, and Mark Stoneking. Independent origins of indian caste and tribal paternal lineages. *Current Biology*, 14(3):231 – 235, 2004.

[129] David W. Anthony. *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World*. Princeton University Press, 2007.

[130] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[131] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

[132] Chaolong Wang, Zachary A Szpiech, James Degnan, Mattias Jakobsson, Trevor J Pemberton, John Hardy, Andrew B Singleton, and Noah A Rosenberg. *Comparing Spatial Maps of Human Population-Genetic Variation Using Procrustes Analysis*, volume 9. Statistical applications in genetics and molecular biology, 2010.

[133] Alkes L. Price, Noah A. Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*, 11(7):459–463, Jul 2010. 20548291[pmid].

[134] Peristera Paschou, Petros Drineas, Jamey Lewis, Caroline M. Nievergelt, Deborah A. Nickerson, Joshua D. Smith, Paul M. Ridker, Daniel I. Chasman, Ronald M. Krauss, and Elad Ziv. Tracing sub-structure in the european american population with pca-informative markers. *PLOS Genetics*, 4(7):1–13, 07 2008.

[135] Aritra Bose, Daniel E. Platt, Laxmi Parida, Peristera Paschou, and Petros Drineas. Dissecting population substructure in india via correlation optimization of genetics and geodemographics. *bioRxiv*, 2017.

[136] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[137] B. Parlett. *The Symmetric Eigenvalue Problem*. Society for Industrial and Applied Mathematics, 1998.

[138] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Society for Industrial and Applied Mathematics, 2011.

[139] Gad Abraham, Yixuan Qiu, and Michael Inouye. Flashpca2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, 2017.

[140] P. Drineas and M. W. Mahoney. RandNLA: Randomized Numerical Linear Algebra. *Communications of the ACM*, 59(6):80–90, 2016.

[141] Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1396–1404. Curran Associates, Inc., 2015.

[142] P. Drineas and M. W. Mahoney. *Lectures on Randomized Numerical Linear Algebra, The Mathematics of Data*, volume 25, pages 1–45. Amer. Math. Soc., Providence, RI, 2018.

[143] P. Drineas, I. C. F Ipsen, E. Kontopoulou, and M. Magdon-Ismail. Structural convergence results for low-rank approximations from block krylov spaces. *SIAM Journal of Matrix Analysis and Applications, to appear*, 2018.

[144] Kevin J. Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J. Patterson, and Alkes L. Price. Fast principal-component analysis reveals convergent evolution of ¡em¿adh1b¡/em¿ in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, 2018/03/24 2016.

[145] Prem Gopalan, Wei Hao, David M. Blei, and John D. Storey. Scaling probabilistic models of genetic variation to millions of humans. *Nat Genet*, 48(12):1587–1590, Dec 2016. 27819665[pmid].

[146] B. S. Weir and C. Clark Cockerham. Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370, 1984.

[147] Gad Abraham and Michael Inouye. Fast principal component analysis of large-scale genome-wide data. *PLOS ONE*, 9(4):1–5, 04 2014.

[148] L. S. Blackford, J. Demmel, J. Dongarra, I. Duff, S. Hammarling, G. Henry, M. Heroux, L. Kaufman, A. Lumsdaine, A. Petitet, R. Pozo, K. Remington, and R. C. Whaley. An updated set of basic linear algebra subprograms (blas). *ACM Transactions on Mathematical Software*, 28:135–151, 2001.

[149] Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

[150] Richard R Hudson. Estimating the recombination parameter of a finite population model without selection. *Genetics Research*, 50(3):245–250, 1987.

[151] Gregory Ewing and Joachim Hermisson. Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065, 2010.

[152] R. Leblois, A. Estoup, and F. Rousset. IBDSim: a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources*, 9(1):107–109, 2009.

[153] Liming Liang, Sebastian Zöllner, and Gonçalo R. Abecasis. Genome: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–1567, 2007.

[154] Thomas Mailund, Mikkel H. Schierup, Christian NS Pedersen, Peter JM Mechlenborg, Jesper N. Madsen, and Leif Schauser. Coasim: A flexible environment for simulating genetic data under coalescent models. *BMC Bioinformatics*, 6(1):252, Oct 2005.

[155] Anna Paola Carrieri, Filippo Utro, and Laxmi Parida. Sampling arg of multiple populations under complex configurations of subdivision and admixture. *Bioinformatics*, 32(7):1048–1056, 2016.

[156] R.C. Griffiths and P. Marjoram. An ancestral recombination graph. *In: Donnely, P. and Tavare, S (eds.) Progress in Population Genetics and Human Evolution, IMA Vols in Mathematics and its Applications, Springer, New York, USA*, 87:257–270, 1997.

[157] J. F. C. Kingman. On the geneaology of large populations. *Journal of Applied Probability*, 19:27–43, 1982.

[158] Laurent Excoffier and Matthieu Foll. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, 27(9):1332–1334, 2011.

[159] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

[160] Gary K. Chen, Paul Marjoram, and Jeffrey D. Wall. Fast and flexible simulation of dna sequence data. *Genome Res*, 19(1):136–142, Jan 2009. 19029539[pmid].

[161] Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

[162] Gregory Ewing and Joachim Hermisson. Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065, 2010.

[163] Andrew D. Kern and Daniel R. Schrider. Discoal: flexible coalescent simulations with selection. *Bioinformatics*, 32(24):3839–3841, 2016.

[164] Ilya Shlyakhter, Pardis C. Sabeti, and Stephen F. Schaffner. Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427–3429, 2014.

[165] Kosuke M. Teshima and Hideki Innan. mbs: modifying hudson's ms software to generate samples of dna sequences with a biallelic site under selection. *BMC Bioinformatics*, 10(1):166, 2009.

[166] N. H. Barton. How does epistasis influence the response to selection? *Heredity*, 118:96 EP –, Nov 2016.

[167] Russell Corbett-Detig and Matt Jones. Selam: simulation of epistasis and local adaptation during admixture with mate choice. *Bioinformatics*, 32(19):3035–3037, 2016.

[168] Philipp W. Messer. Slim: Simulating evolution with selection and linkage. *Genetics*, 194(4):1037–1039, 2013.

[169] Benjamin C Haller and Philipp W Messer. Slim 3: Forward genetic simulations beyond the wright–fisher model. *Molecular biology and evolution*, 36(3):632–637, 2019.

[170] Badri Padhukasahasram, Paul Marjoram, Jeffrey D. Wall, Carlos D. Bustamante, and Magnus Nordborg. Exploring population genetic models with recombination using efficient forward-time simulations. *Genetics*, 178(4):2417–2427, Apr 2008.

[171] Bo Peng and Christopher I. Amos. Forward-time simulations of non-random mating populations using simupop. *Bioinformatics*, 24(11):1408–1409, 2008.

[172] Darren Kessner and John Novembre. forqs: forward-in-time simulation of recombination, quantitative traits and selection. *Bioinformatics*, 30(4):576–577, 2014.

[173] Chris C. A. Spencer and Graham Coop. Selsim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, 20(18):3673–3675, 2004.

[174] N. L. Kaplan, T. Darden, and R. R. Hudson. The coalescent proess in models with selection. *genetics*, 120:819–829, 1988.

[175] Claudia Neuhauser and Stephen M. Krone. The genealogy of samples in models with selection. *Genetics*, 145:519–534, 1997.

[176] Asif Javed, Marc Pybus, Marta Melé, Filippo Utro, Jaume Bertranpetit, Francesc Calafell, and Laxmi Parida. Iris: Construction of arg networks at genomic scales. *Bioinformatics*, 27(17):2448–2450, 2011.

[177] Marta Melé, Asif Javed, Marc Pybus, Francesc Calafell, Laxmi Parida, Jaume Bertranpetit, and The Genographic Consortium. A new method to reconstruct recombination events at a genomic scale. *PLOS Computational Biology*, 6(11):1–13, 11 2010.

[178] Laxmi Parida, Marta Melé, Francesc Calafell, and Jaume Bertranpetit. Estimating the ancestral recombinations graph (arg) as compatible networks of snp patterns. *Journal of Computational Biology*, 15(9):1133–1153, Oct 2008.

[179] Jerome Kelleher, Kevin R Thornton, Jaime Ashander, and Peter L Ralph. Efficient pedigree recording for fast population genetics simulation. *PLoS computational biology*, 14(11):e1006581, 2018.

[180] Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, 1983.

[181] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2013.

[182] Jian Zeng, Ronald De Vlaming, Yang Wu, Matthew R Robinson, Luke R Lloyd-Jones, Loic Yengo, Chloe X Yap, Angli Xue, Julia Sidorenko, Allan F McRae, et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nature genetics*, 50(5):746, 2018.

[183] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[184] Brendan Maher. Personal genomes: The case of the missing heritability. *Nature News*, 456(7218):18–21, 2008.

[185] Nan M Laird and Christoph Lange. *The fundamentals of modern statistical genetics*. Springer Science & Business Media, 2010.

[186] S Hong Lee, Teresa R DeCandia, Stephan Ripke, Jian Yang, Patrick F Sullivan, Michael E Goddard, Matthew C Keller, Peter M Visscher, Naomi R Wray, Schizophrenia Psychiatric Genome-Wide Association Study Consortium, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. *Nature genetics*, 44(3):247, 2012.

[187] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552, 2014.

[188] Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[189] Ronald Aylmer Fisher. Studies in crop variation. i. an examination of the yield of dressed grain from broadbalk. *The Journal of Agricultural Science*, 11(2):107–135, 1921.

[190] Charles R Henderson et al. *Applications of linear models in animal breeding*, volume 462. University of Guelph Guelph, 1984.

[191] Warren J Ewens and Richard S Spielman. The transmission/disequilibrium test: history, subdivision, and admixture. *American journal of human genetics*, 57(2):455, 1995.

[192] Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[193] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173, 2014.

[194] Jing Guo, Yang Wu, Zhihong Zhu, Zhili Zheng, Maciej Trzaskowski, Jian Zeng, Matthew R Robinson, Peter M Visscher, and Jian Yang. Global genetic differentiation of complex traits shaped by natural selection in humans. *Nature communications*, 9(1):1865, 2018.

[195] Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak, et al. Genome-wide patterns of selection in 230 ancient eurasians. *Nature*, 528(7583):499, 2015.

[196] Lawrence H Uricchio, Hugo C Kitano, Alexander Gusev, and Noah A Zaitlen. An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evolution letters*, 3(1):69–79, 2019.

[197] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018.

[198] Wei Hao, Minsun Song, and John D Storey. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721, 2015.

[199] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010.

[200] Peter Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[201] Boby Mathew, Jens Léon, and Mikko J Sillanpää. A novel linkage-disequilibrium corrected genomic relationship matrix for snp-heritability estimation and genomic prediction. *Heredity*, 120(4):356, 2018.

[202] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

[203] Ann FS Mitchell and Wojtek J Krzanowski. The mahalanobis distance and elliptic distributions. *Biometrika*, 72(2):464–467, 1985.

[204] Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

[205] Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012.

[206] Petros Drineas, Michael W Mahoney, Shan Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische mathematik*, 117(2):219–249, 2011.

[207] Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

[208] David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[209] Agniva Chowdhury, Jiasen Yang, and Petros Drineas. An iterative, sketching-based framework for ridge regression. In *International Conference on Machine Learning*, pages 988–997, 2018.

[210] Irving B Weiner. *Handbook of psychology, history of psychology*, volume 1. John Wiley & Sons, 2003.

[211] Shing Wan Choi and Paul F O'Reilly. Prsice-2: Polygenic risk score software for biobank-scale data. *GigaScience*, 8(7):giz082, 2019.

[212] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4):576–592, 2015.

[213] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821, 2012.

APPENDICES

Reconstructing Genetic Population History

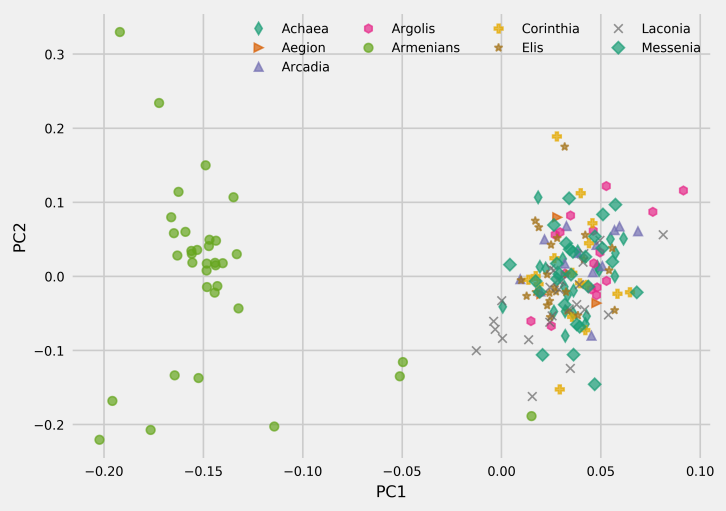
A.1 Genetics of the Peloponnesean Populations

A.1.1 Supplementary Information

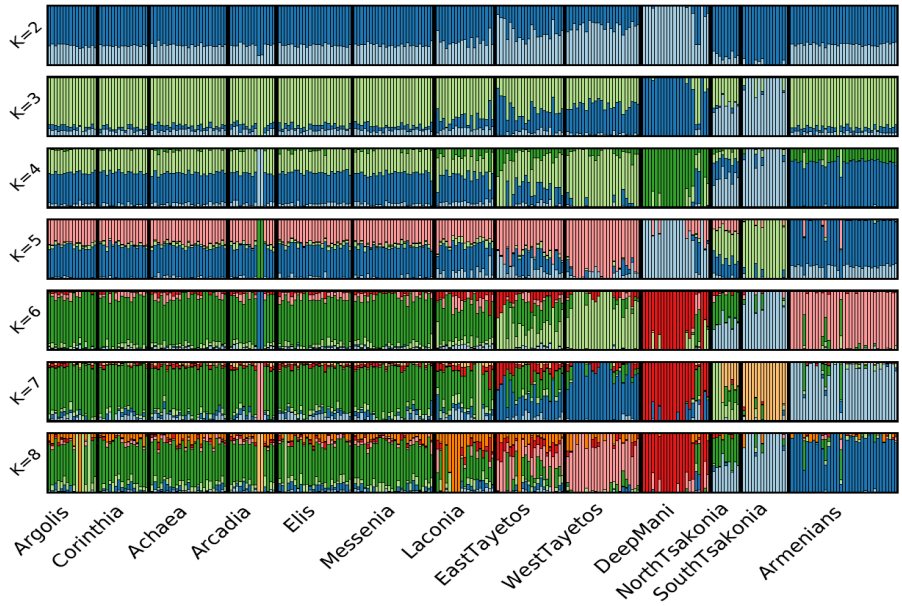
Table A.1.: Districts of origin of the subjects

Districts	Subjects	Used in analysis
Achaea	25	25
Argolis	16	16
Arcadia	15	13
Elis	24	24
Corinthia	16	16
Laconia	26	25
Messenia	26	26
East Tayetos	23	23
West Tayateos	23	24
Deep Mani	22	22
Tsakonia	24	24
TOTAL	241	238

Figure A.1.: Locations of the populations listed in Supplementary Table 1

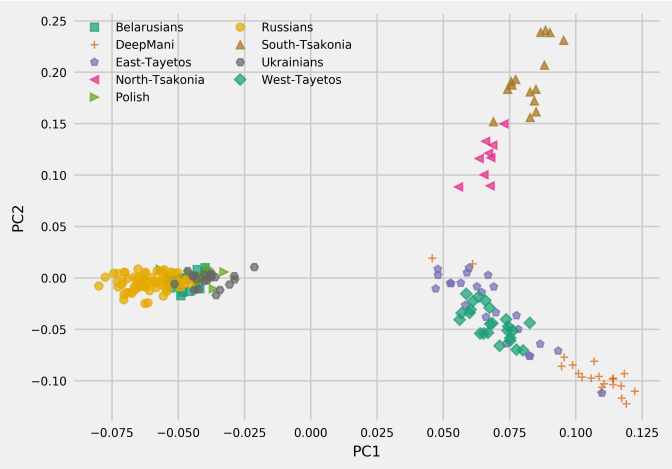


(a) Peloponnesean samples with Armenians projected on the top two PCs

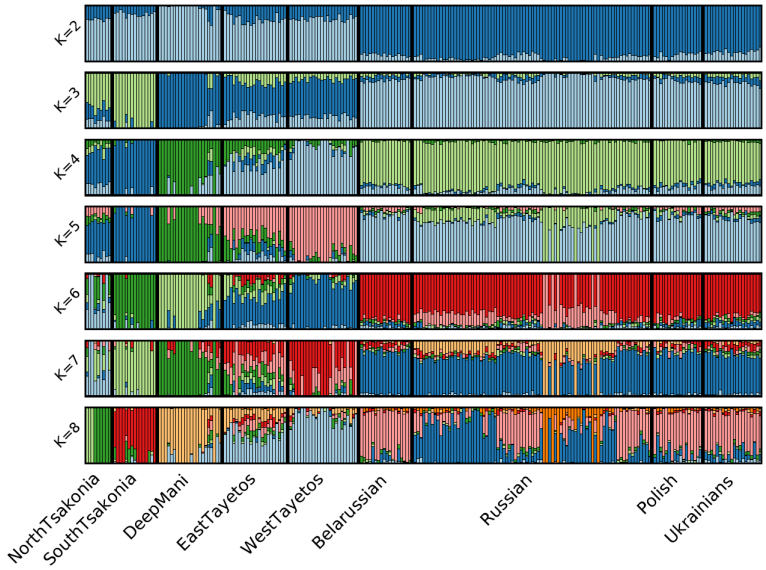


(b) ADMIXTURE plot (K = 2:8) for Peloponnesean and Armenian populations.

Figure A.2.: Testing the hypothesis of Armenian ancestry of Peloponneseans. Fallmerayer proposed that Armenians were among the medieval populations moved to Peloponnese by the Byzantines. Comparison of Peloponneseans with the Armenians by, (a) PCA analysis (b) ADMIXTURE analysis, makes this hypothesis unlikely.

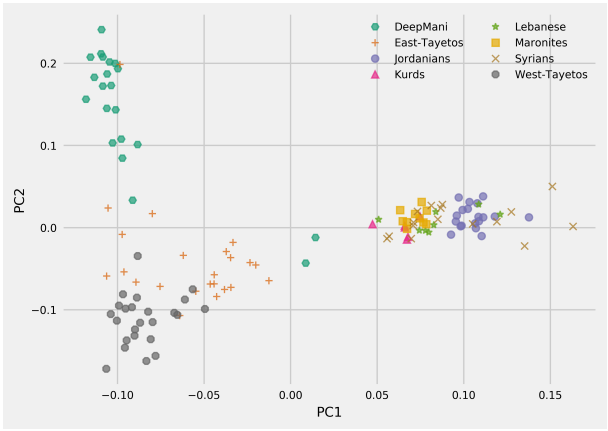


(a) Maniots and Slavic populations projected on the top two PCs.

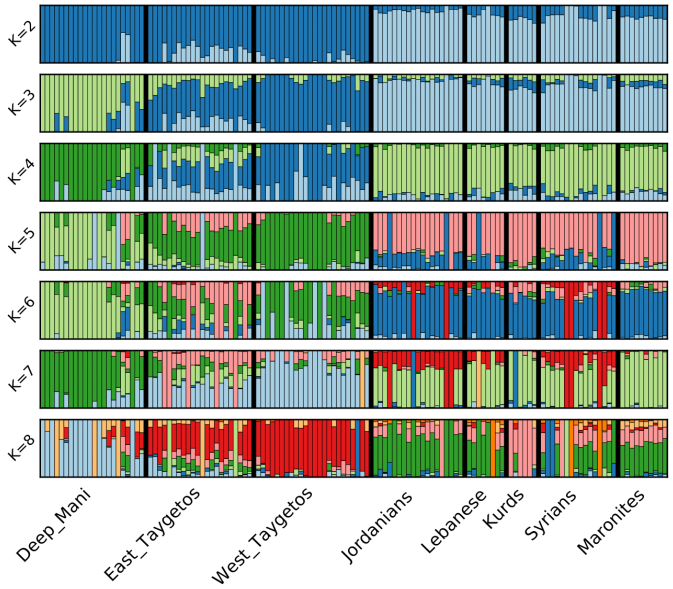


(b) ADMIXTURE plot (K = 2:8) for Maniots and Slavic populations.

Figure A.3.: Testing the hypothesis of Slavic origin of culturally distinct Peloponnesean populations. PCA comparisons of (a) The Maniots of Deep Mani, Tayetos and Tsakones, with populations of the Slavic homeland (Ukrainians, Polish, Russians and Belarusians). Notice the broad separation between the Slavs and the Peloponnesean populations. (b) ADMIXTURE analysis shows the complete separation of Maniots and Tsakones from the Slavs in all K values.



(a) Maniots and populations from Middle East projected on the top two PCs.



(b) ADMIXTURE plot (K = 2:8) for Maniots and Middle Eastern populations.

Figure A.4.: Testing the hypothesis of Mardaitic origin of Maniots. The Mardaites were a medieval Middle Eastern population considered by some historians to be the ancestors of the Maronites of Lebanon. Comparison of Maniots with Maronites and other Middle Eastern populations by (a) PCA and, (b) ADMIXTURE analysis makes this hypothesis unlikely.

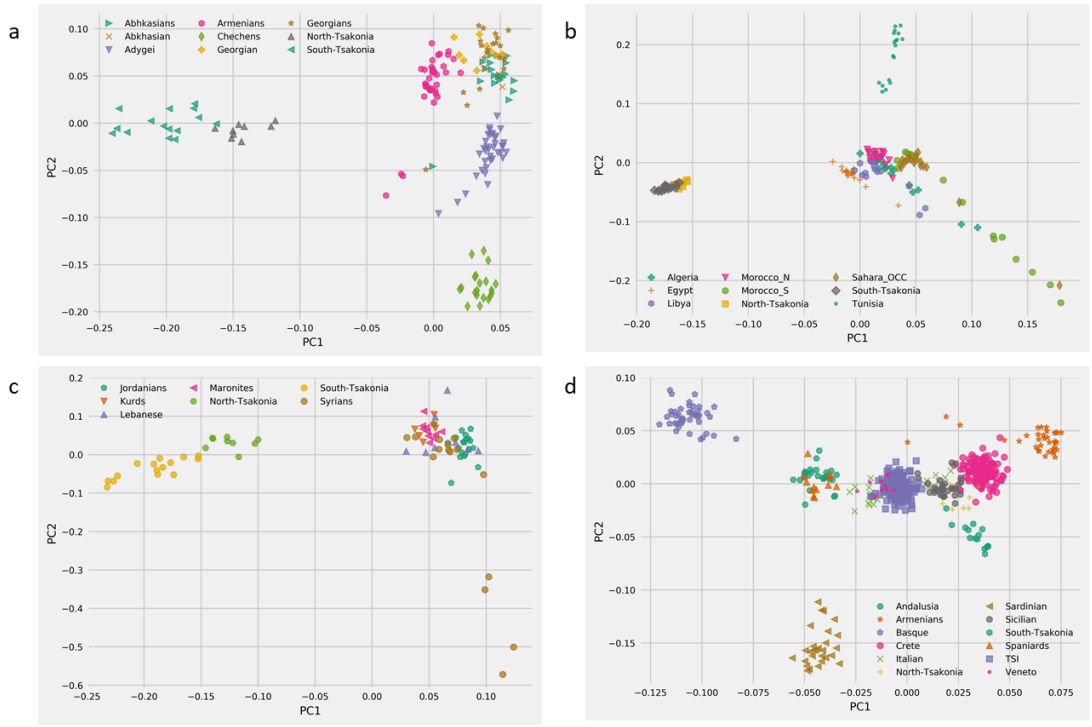


Figure A.5.: Unique genetic structure of the population of Tsakonia. PCA comparisons of Tsakones with A. the Eastern Europeans. B. North Africans C. Near Eastern populations D. Southern Europeans.

A.2 Integrating linguistics, social structure and geography to model gene flow in India

A.2.1 Supplementary Information

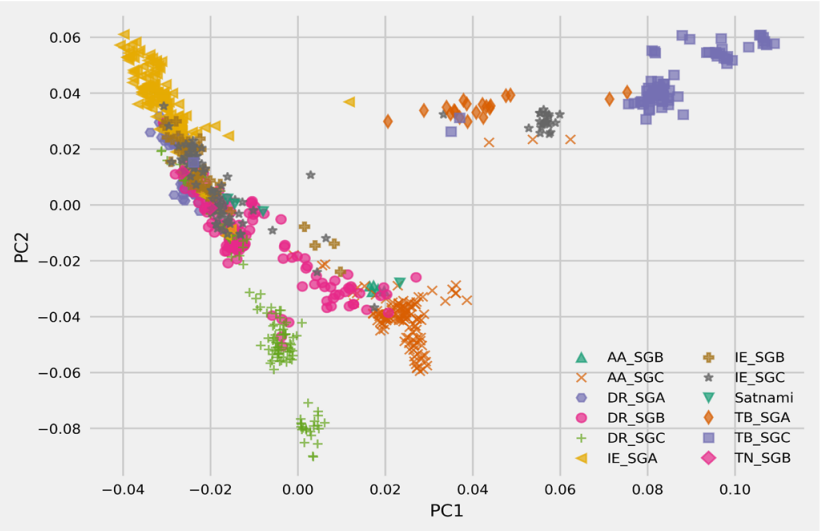


Figure A.6.: PCA plot of all Indian samples. We note that the formation of the clusters is primarily dominated by language groups, with some populations (Gond, Manipuri Brahmins, Dusadh) showing a certain amount of admixture between the language groups. A few tribal populations across IE and DR languages (Vedda, Madiga, Kol, Bhil, Chamar, Kuruchiyan) cluster together. We also observe that the Irulas, Paniyas, Kurumba and Kadars show divergence from other DR_SGC populations.

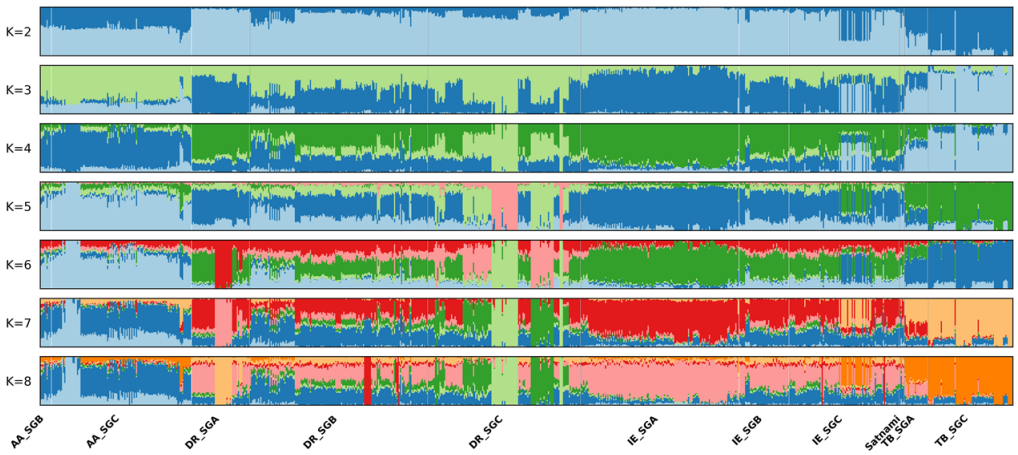
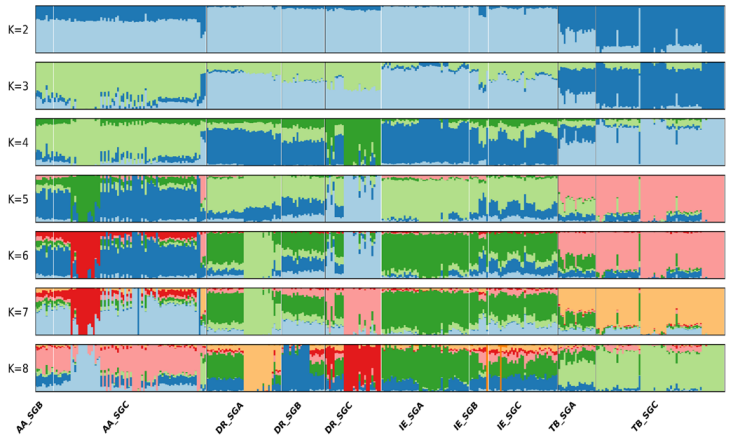
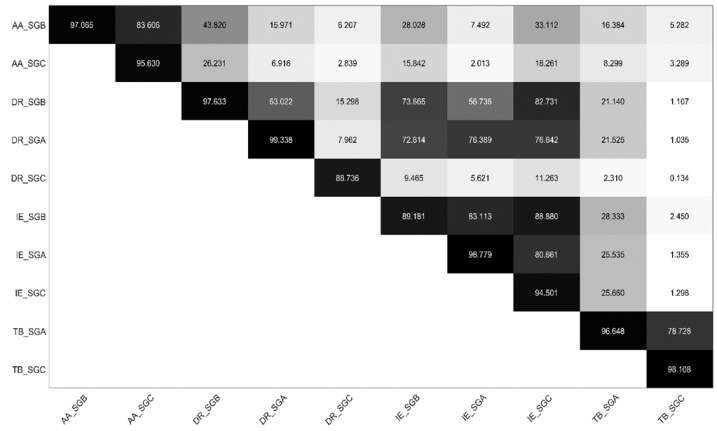


Figure A.7.: ADMIXTURE plot of all Indian populations for values of K between two and eights. Our findings are very similar to the observations in Supplementary Figure 1. The main observation is (again) that the formation of the clusters is primarily dominated by language groups, especially for larger values of K.

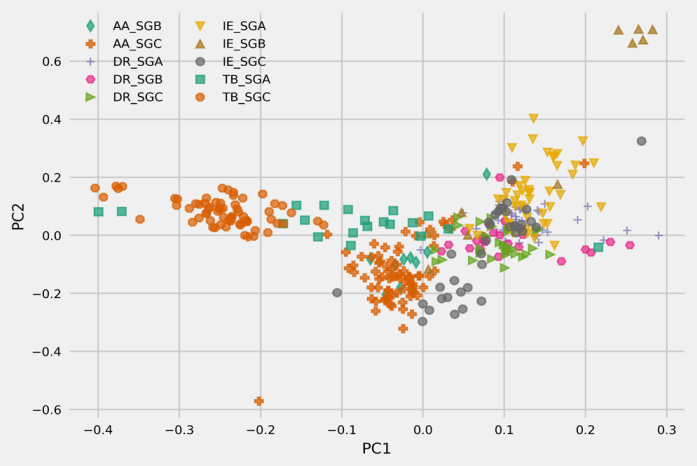


(a) ADMIXTURE plot of normalized Indian populations

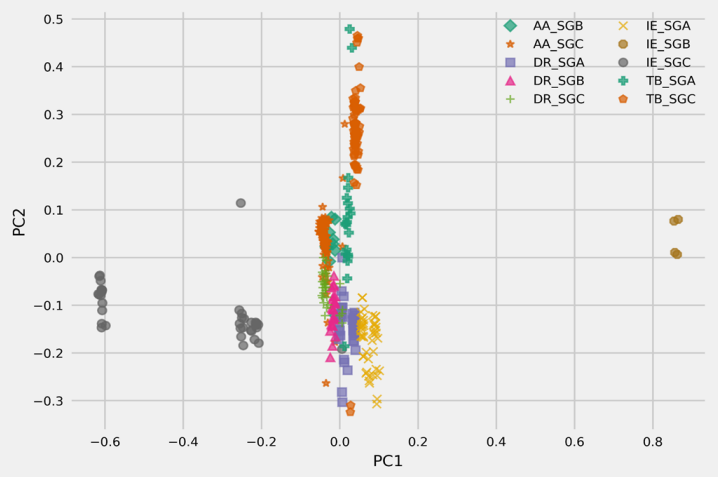


(b) Shared ancestry of ADMIXTURE plot (a) for K between 4 and 8

Figure A.8.: An ADMIXTURE plot (for values of K between two and eight) of the normalized data set (368 individuals 48,373 SNPs) clearly shows the four main components related to language groups (Dravidian, Indo-European, Tibeto-Burman, and Austro-Asiatic); see, for example, the plot for K equal to five or six. The plot also shows the divergence of the DR_SGC. We perfomed a meta-analysis of the results of the ADMIXTURE plot (see 3.1.2 for details) to visually and numerically quantify the amount of shared ancestry (as revealed by ADMIXTURE) between any pair of populations. Darker colors indicate larger amounts of shared ancestry; we observe a higher amount of shared ancestry between the IE and DR populations, across all social groups, indicating the existence of significant admixture between the two linguistic groups. The isolation of the DR_SGC samples is primarily due to the isolation of hill SGCs (such as Irula, Kadar, Paniyas, etc.)

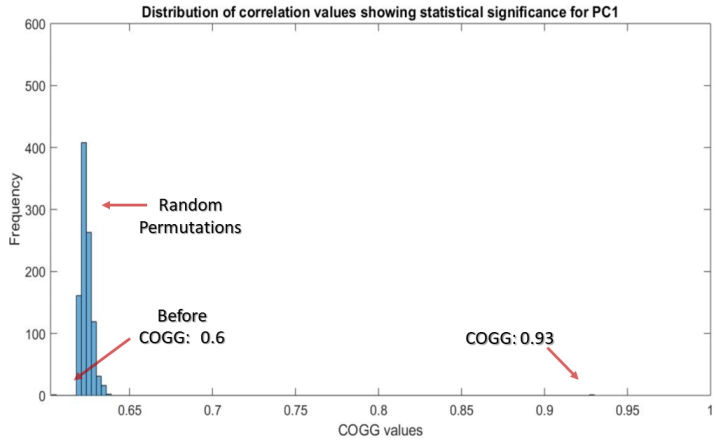


(a) Discriminated by regions

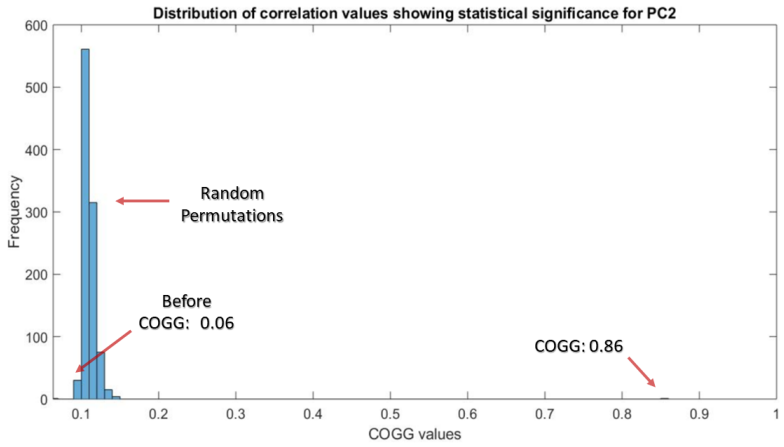


(b)Discriminated by language groups

Figure A.9.: Plotting the top two discriminants by (a) region and (b) language groups. Clearly, this follows much what we saw in Figure A.6. However, looking closely we see the following: (a) we see a geographical gradient, starting from IE_SGA and IE_SGB in Northwestern India to the other Indo-European and Dravidian SGA. We also see that the IE_SGC sit closer to the Austro-Asiatic speakers, justifying their geographical location in Central India. This is followed by the Tibeto-Burman speakers forming another cluster, concluding the other spectrum of the gradient. (b) Layers of stratification appears, from right to left. Although the LDA was performed by language groups, we see a two-layer stratification, first by castes and then by languages. The IE_SGA form a separate cline, followed by DR_SGA; then, the IE and DR SGBs follow. Then some DR and AA tribal populations cluster together, followed by a separate cluster of IE tribal populations.

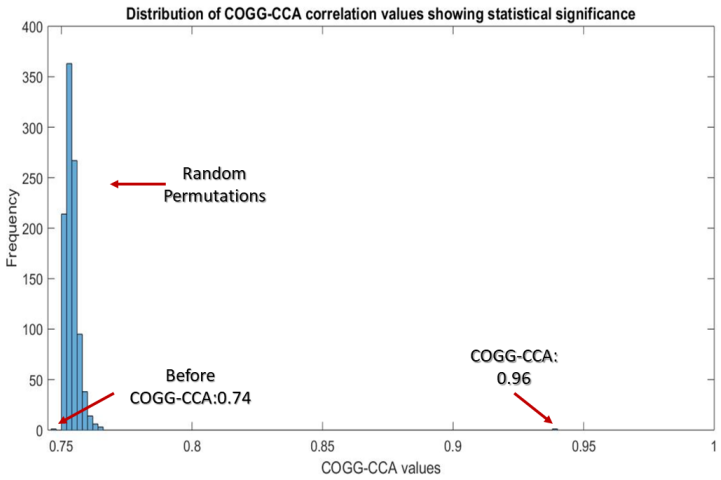


(a) histogram of permutations with first PC

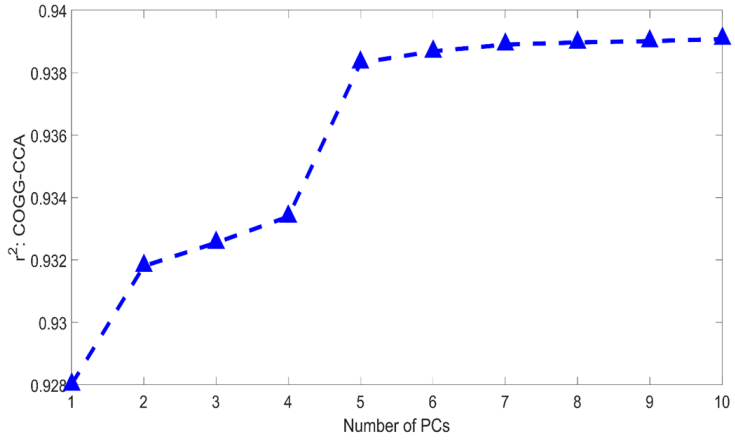


(b) histogram of permutations with second PC

Figure A.10.: Statistical significance of the COGG output (using random permutations of the features) Clearly, COGG is statistically significant for both the first and the second principal components



(a) histogram of permutations running COGG-CCA



(b) r^2 obtained from COGG-CCA varying with number of PCs

Figure A.11.: (a) COGG-CCA, when run with top 8 PCs, shows statistical significance with $r^2 = 0.94$ when compared against random permutations of the variables with average $r^2 = 0.75$. (b) Varying number of PCs to perform COGG-CCA results in the maximum r^2 when top 6 to 8 PCs are used.

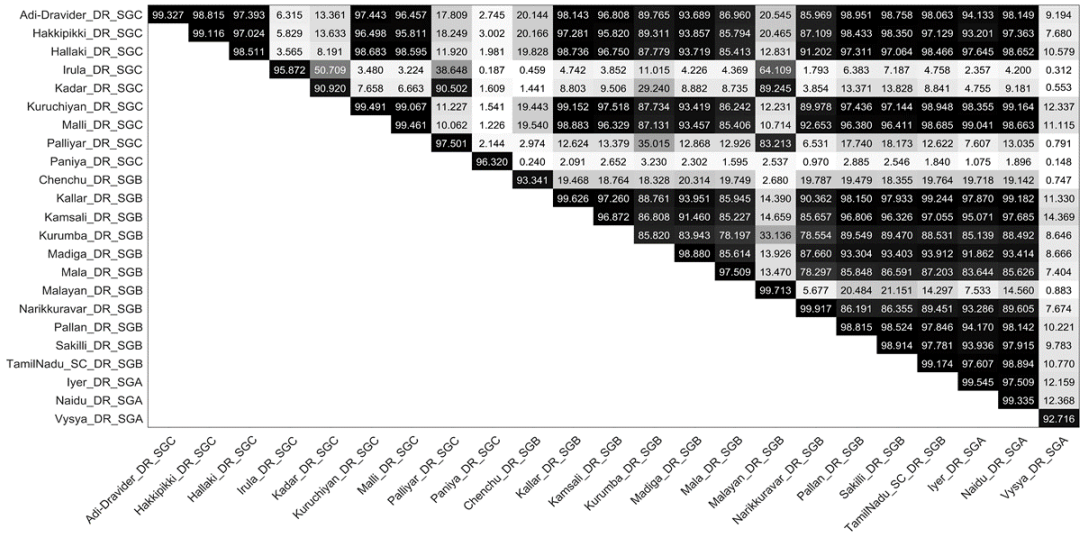


Figure A.12.: The pairwise shared ancestry matrix of relatedness within DR show high relatedness among a large portion of DR speakers across caste affiliations. The Tribes such as Irula, Kadar, Palliyar, Paniya and Malayan show significant divergence from the others. Among them the Paniyas show absolute divergence, with very less amount of ancestry with all DR speakers, whereas the others tend to form a cluster and show that although they share significant amount of ancestry with each other, than the DR_SGA. The SGB and SGAs tend to cluster together showing high relatedness with some SGCs such as Adi-Dravider, Hakkipikki, Hallaki, Kuruchiyan, etc.

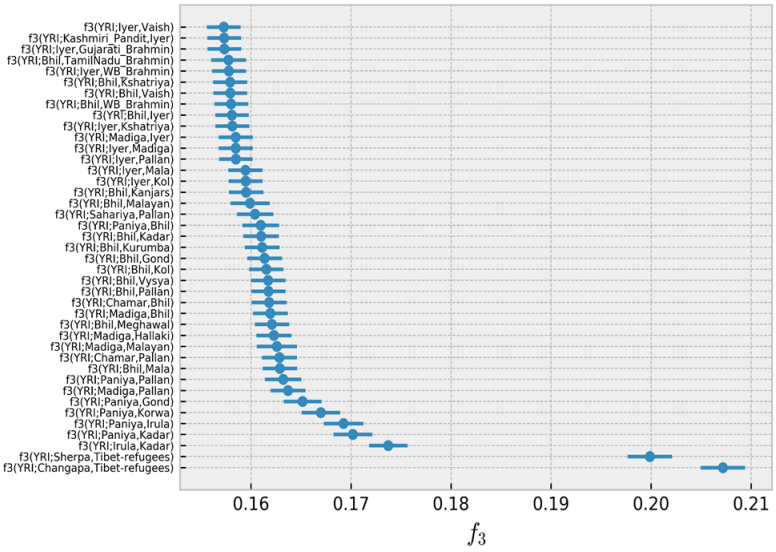


Figure A.13.: Most significant (Z-score higher than 85) outgroup f_3 statistics of the form $f_3(YRI; A, B)$ where YRI is the outgroup, A are the groups from Table 3.4 and B are all the pan-Indian populations in our data spanning across social groups and language families.

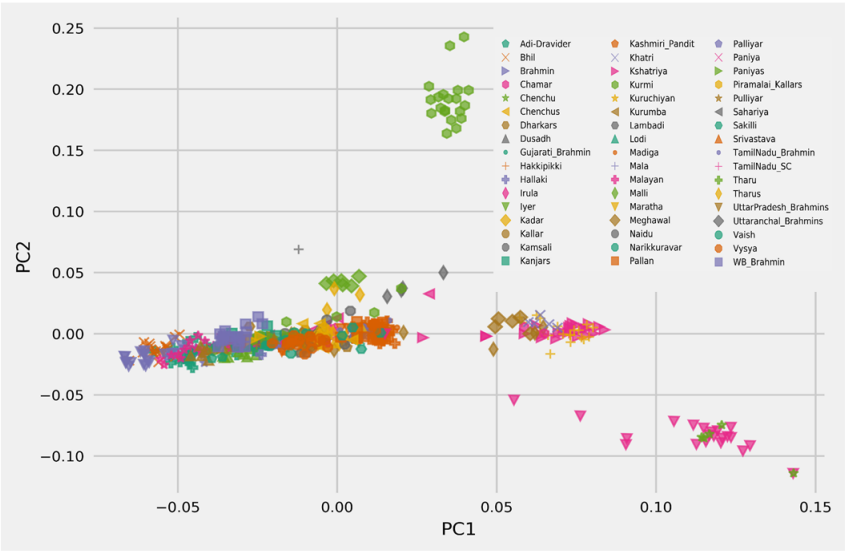


Figure A.14.: The top two principal components show a long cline of IE and DR speakers with some divergence by few SGCs, such as Tharu, Irula, Palliyar, Paniyas, etc.

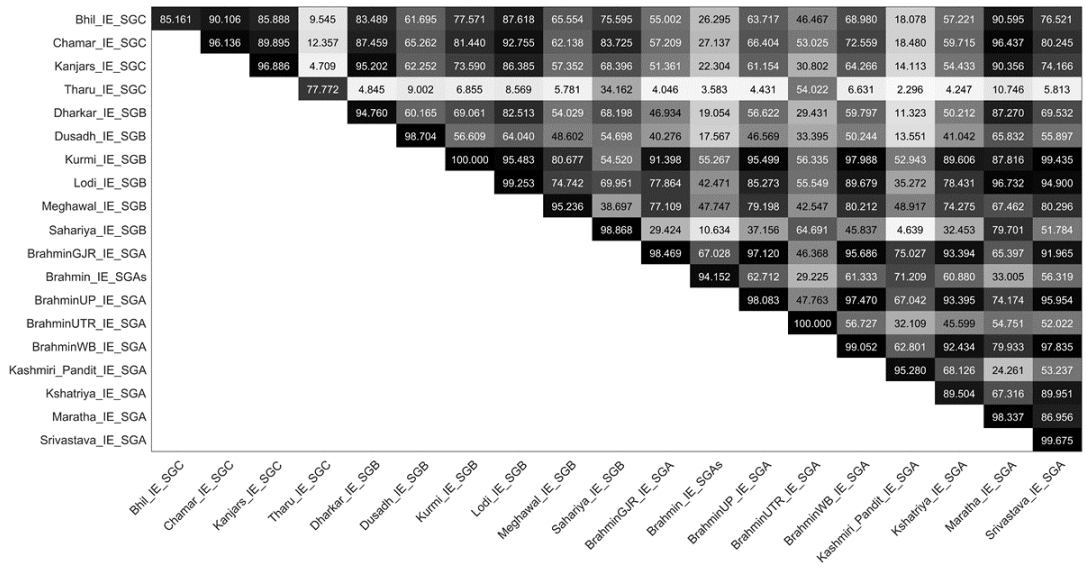


Figure A.15.: The pairwise shared ancestry matrix of relatedness within IE show high relatedness among most of the IE speakers across caste affiliations. The Tharus show divergence from rest of the IE speakers except the Uttaranchal Brahmins, who share close relatedness with the East Asian component in their gene pool. The Brahmin groups (GJR – Gujarati; UP – Uttar Pradesh; UTR – Uttaranchal; WB – West Bengal) show high values of shared ancestry within each other and rest of the IE speakers. Only UTR Brahmins show some divergence. The tribes such as Sahariya, Bhil and Chamar are more closely related to the fellow SGCs than the SGA, but still show around 70% of relatedness with them.

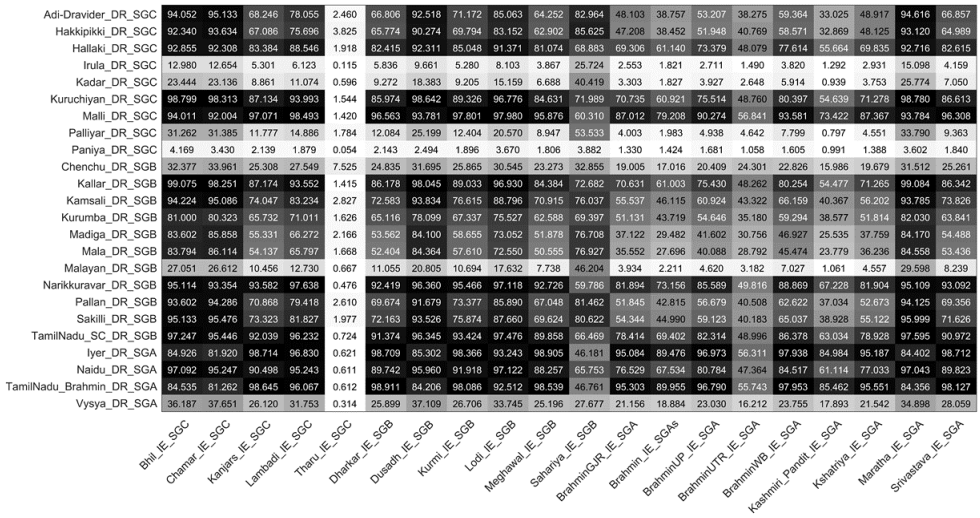


Figure A.16.: The shared ancestry matrix of relatedness between IE and DR speakers show that high relatedness with some divergent groups, following from the PC plot in Figure A.15. The DR_SGA share very high ancestry with IE SGA and SGC, showing that there was high admixture and contact between these groups prior to endogamy.

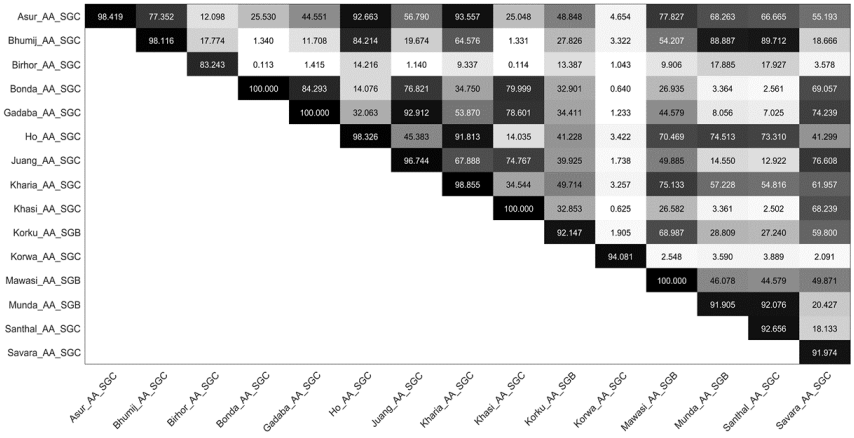


Figure A.17.: The pairwise shared ancestry matrix of relatedness within AA show very high relatedness among almost all AA speakers. Birhors, who are nomadic hunter-gatherer people dwelling in forests share less ancestry than others, probably because of their subsistence nature, where they roam around the forests of eastern and central India. The Khasis also show divergence from the AA speakers because of their location in northeastern India near TB_SGC and presence of admixture from TB speakers.

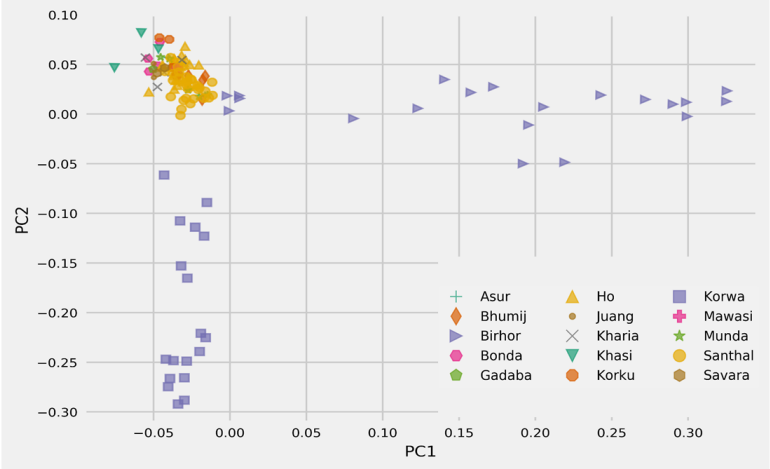


Figure A.18.: The top two PCs of AA speakers in India show most of the groups form a cluster with Birhor and Korwa showing divergence from the main cluster.

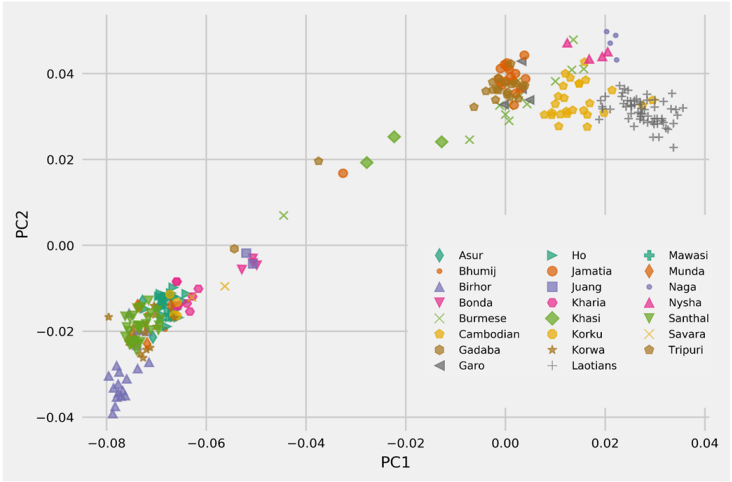
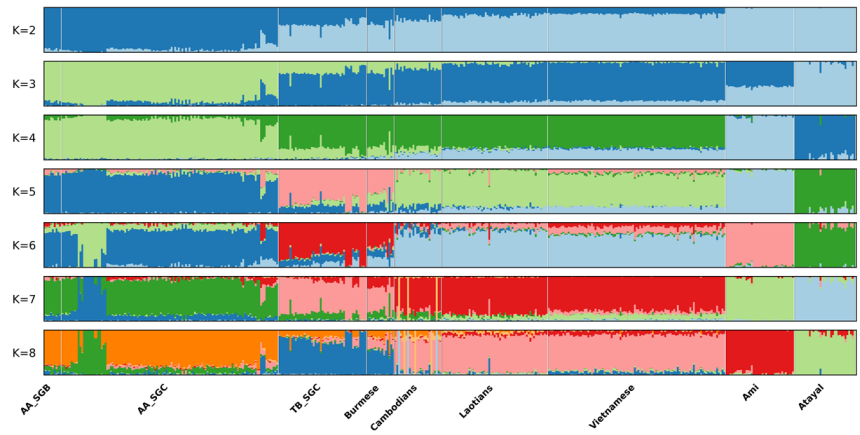


Figure A.19.: PCA plot of the first two PCs reveals the Austronesians (Ami and Atayal) and the IE and DR speakers to be distinct from the rest of the southeast Asians along with the Indian AA speakers.

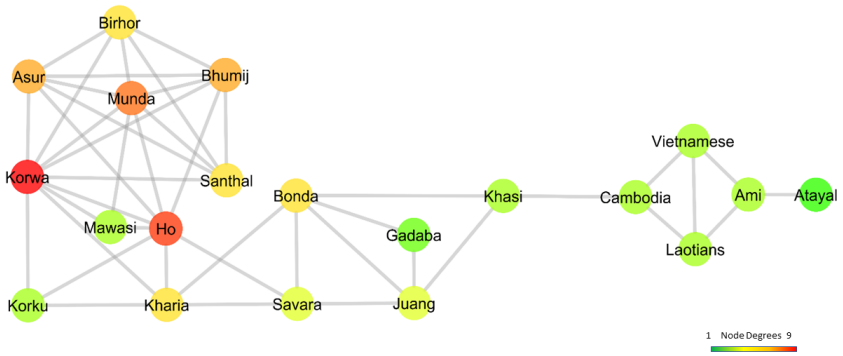


(a) ADMIXTURE plot for K (two to eight)

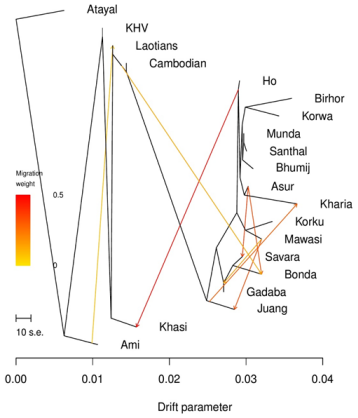
	Tripuri_TB_SGC	Garo_TB_SGC	Jamatia_TB_SGC	Nagal_TB_SGC	Nyisha_TB_SGC	Cambodian	Vietnamese	Laotians	Ami	Atayal
Asur_AA_SGC	13.614	11.801	9.511	1.519	1.922	3.660	0.295	0.491	0.003	0.067
Bhumij_AA_SGC	10.319	8.863	6.706	0.514	0.799	3.559	0.217	0.431	0.125	0.024
Birhor_AA_SGC	5.054	3.936	3.422	0.119	0.271	1.597	0.040	0.129	0.010	0.022
Bonda_AA_SGC	16.428	16.776	11.192	2.732	3.438	17.738	7.634	8.891	0.150	0.048
Gadaba_AA_SGC	21.109	20.854	15.131	4.256	5.072	14.599	5.608	6.463	0.003	0.001
Ho_AA_SGC	11.540	10.462	7.553	0.812	1.205	6.156	1.023	1.499	0.049	0.016
Juang_AA_SGC	18.666	18.893	12.947	3.280	4.036	16.621	6.991	8.023	0.090	0.023
Kharia_AA_SGC	11.059	10.351	7.116	0.821	1.225	7.719	1.715	2.351	0.095	0.031
Khasi_AA_SGC	76.951	78.779	67.668	50.277	51.676	32.186	23.869	22.254	0.030	0.031
Korku_AA_SGB	15.201	13.292	10.781	1.885	2.362	4.112	0.456	0.684	0.053	0.062
Korwa_AA_SGC	8.586	7.379	5.285	0.217	0.429	3.656	0.226	0.466	0.067	0.025
Mawasi_AA_SGB	15.423	13.519	10.971	1.950	2.445	4.266	0.481	0.723	0.009	
Munda_AA_SGB	10.939	9.555	7.204	0.740	1.064	4.326	0.444	0.737	0.024	0.023
Santhal_AA_SGC	10.497	9.052	6.865	0.593	0.889	3.710	0.240	0.468	0.020	0.030
Savara_AA_SGC	17.088	16.132	11.984	2.518	3.150	9.518	2.751	3.368	0.001	0.086

(b) Shared ancestry matrix for ADMIXTURE (K between four to eight)

Figure A.20.: (a) ADMIXTURE plot (for values of K between two and eight) of the Indian dataset merged with Southeast Asian populations shows that the AA and TB speakers do not share a lot of admixture with other Austric speakers from Southeast Asia; (b) The pairwise shared ancestry matrix of AA and TB speakers highlighting that the Khasis share very high amount of ancestry with TB tribals, unlike other AA groups.

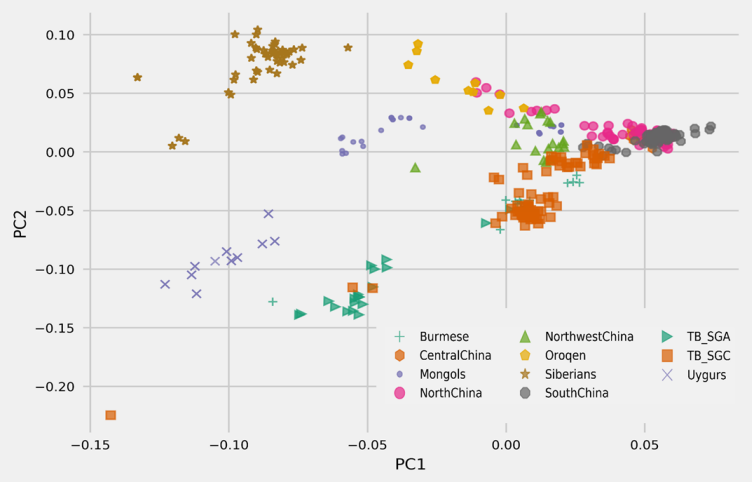


(a) Network analysis based on shared ancestry

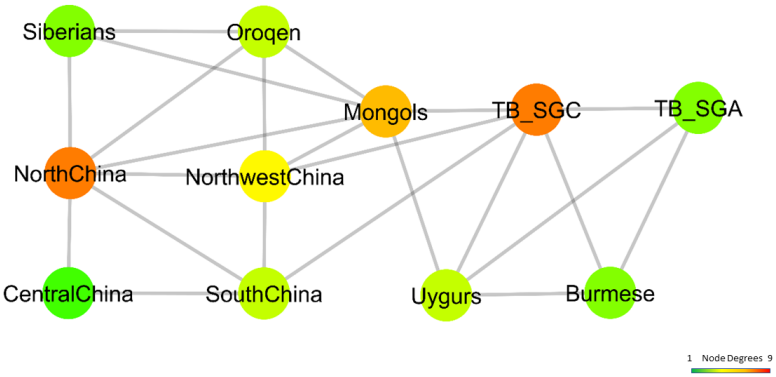


(b) Treemix plot with six migration edges

Figure A.21.: (a) Network analysis for top 2 PCs and 5 nearest neighbors show that the Khasis forming a bridge between Indian AA speakers and southeast Asia; (b) TreeMix plot of Indian and Southeast Asian AA speakers with 8 migration edges reveal that there is a migration edge from Cambodian to Bonda, who are Indian AA speakers attributed to southeastern Asian admixture.

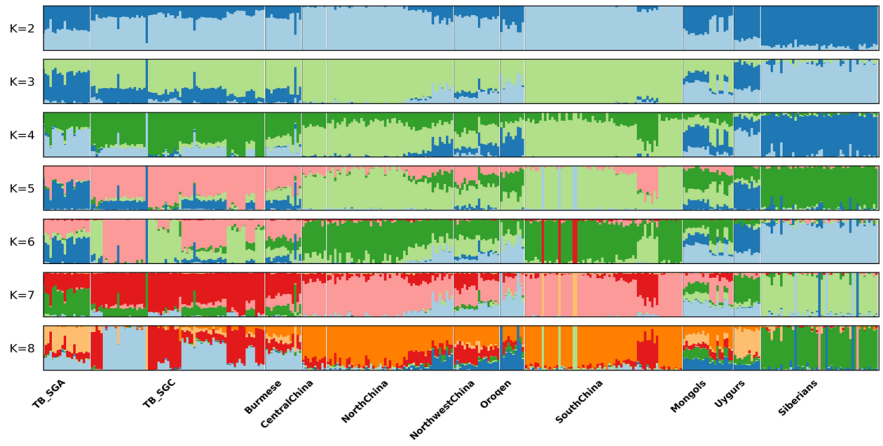


(a) Plotting top two PCs

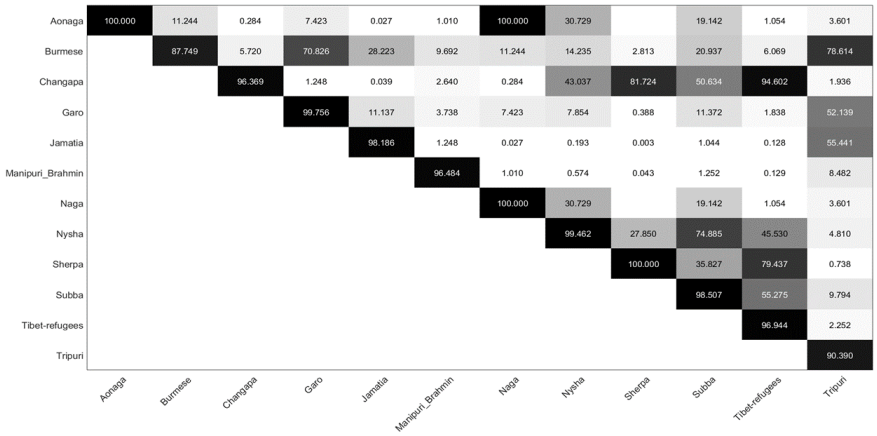


(b) Network analysis based on shared ancestry

Figure A.22.: (a) Network analysis for top 2 PCs and 5 nearest neighbors show that the Khasis forming a bridge between Indian AA speakers and southeast Asia; (b) TreeMix plot of Indian and Southeast Asian AA speakers with 8 migration edges reveal that there is a migration edge from Cambodian to Bonda, who are Indian AA speakers attributed to southeastern Asian admixture.



(a) ADMIXTURE plot for K (two to eight)



(b) Shared ancestry matrix for ADMIXTURE (K between four to eight)

Figure A.23.: (a) PCA plot of the top two principal components of Indian TB speakers and mainland Chinese populations show that the TB_SGC are closer to the southern Chinese; (b) Network analysis show that TB_SGC are closer to Central and Southern China who are geographically closer to northeast India.

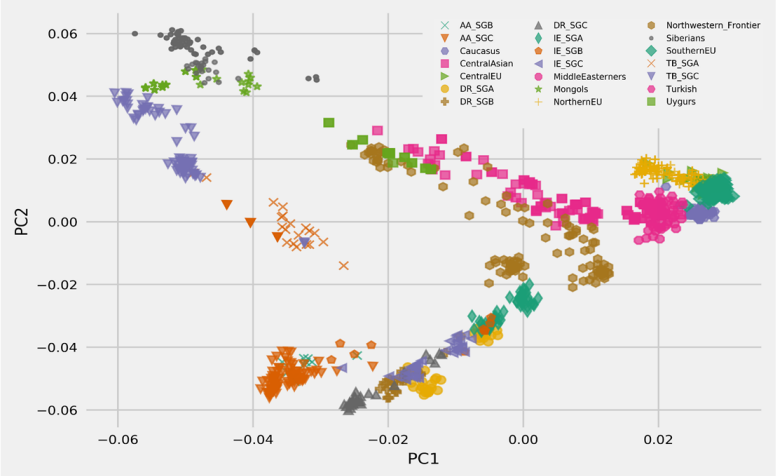


Figure A.24.: Plotting of Indian and Eurasian populations projected on the top two PCs, mirror the geography of Eurasia uncovering a triangular structure with Europeans residing in one corner, the Chinese on another corner and the DR and AA speaking tribal populations of India occupying the third corner.

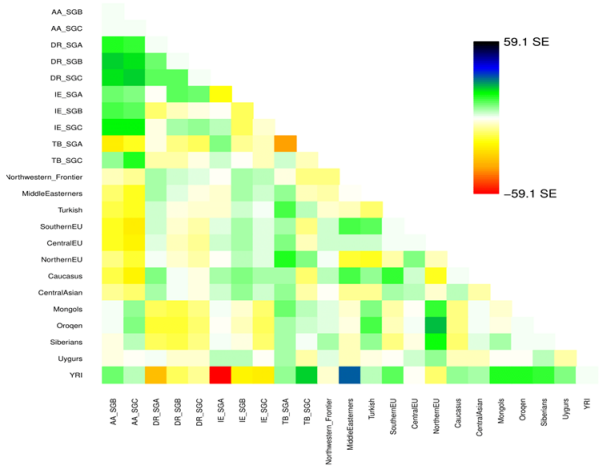


Figure A.25.: Residual fit from the maximum likelihood tree in Fig 3. The residuals are normalized over the residual covariance between each pair i and j. Residuals above zero represent populations that are more closely related to each other and are candidates for admixture events.

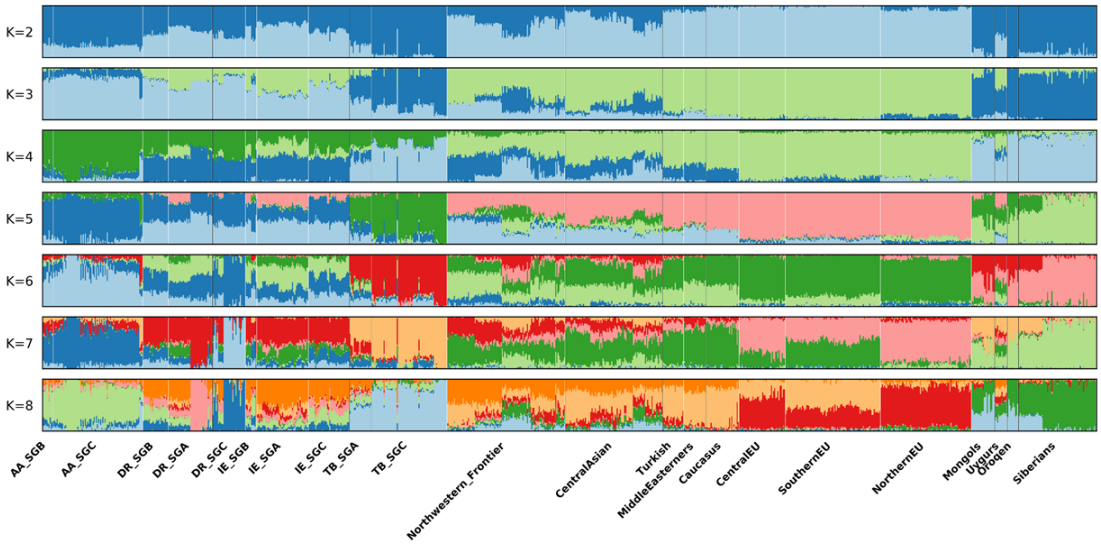


Figure A.26.: ADMIXTURE plot (for values of K between two and eight) of the Indian dataset merged with Eurasian populations (1,332 individuals, 42,973 SNPs). Meta-Analysis of this plot in Fig 4a, quantifies the relationship between populations. The IE and DR Forward and Backward Castes share significant amount of ancestry with the Northwestern Frontier populations of Afghanistan and Pakistan, followed by ancestry from Central Asia, Turkey and Caucasia. The TB tribals belong to the same cluster as the Chinese populations along with, Mongolia and Uygurs.

Table A.2.: Top 10% of the significant f_3 statistics ($f_3(C; A, B)$) highlighting the most admixed populations in India. Gounders, Manipuri Brahmins, Tharus and Gonds are the most admixed among all tribes in India.

A	B	C	F3	Err	Z
DR_SGB	IE_SGA	Gounder	-0.02328	0.000644	-36.114
IE_SGA	TB_SGC	Manipuri_Brahmin	-0.01583	0.000452	-35.019
DR_SGB	IE_SGC	Gounder	-0.02188	0.000657	-33.315
IE_SGA	TB_SGC	Tharu	-0.01364	0.000447	-30.518
DR_SGA	TB_SGC	Tharu	-0.01292	0.000429	-30.084
IE_SGC	TB_SGC	Tharu	-0.00843	0.000389	-21.647
DR_SGC	TB_SGC	Tharu	-0.00913	0.000436	-20.922
DR_SGC	TB_SGC	Manipuri_Brahmin	-0.0094	0.000484	-19.415
IE_SGA	AA_SGC	Iyer	-0.00343	0.000241	-14.211
IE_SGA	AA_SGC	Gond	-0.00449	0.000321	-13.989
DR_SGC	AA_SGC	Gond	-0.00419	0.000305	-13.722
IE_SGC	AA_SGC	Gond	-0.00226	0.000171	-13.245
IE_SGA	AA_SGC	Kol	-0.00347	0.000266	-13.002
IE_SGA	AA_SGC	Pallan	-0.00411	0.000325	-12.638
IE_SGA	AA_SGC	Bhil	-0.00326	0.000277	-11.758
IE_SGA	DR_SGC	Bhil	-0.0036	0.000343	-10.489
IE_SGC	TB_SGC	Khasi	-0.01008	0.001166	-8.648
IE_SGB	TB_SGC	Khasi	-0.00981	0.001155	-8.49
IE_SGA	AA_SGC	Chamar	-0.00292	0.000358	-8.152
IE_SGA	AA_SGC	Satnami	-0.00503	0.000853	-5.898

Appendix B. TeraPCA: a fast and scalable software package to study genetic variation in tera-scale genotypes

A.3 Supplementary Information

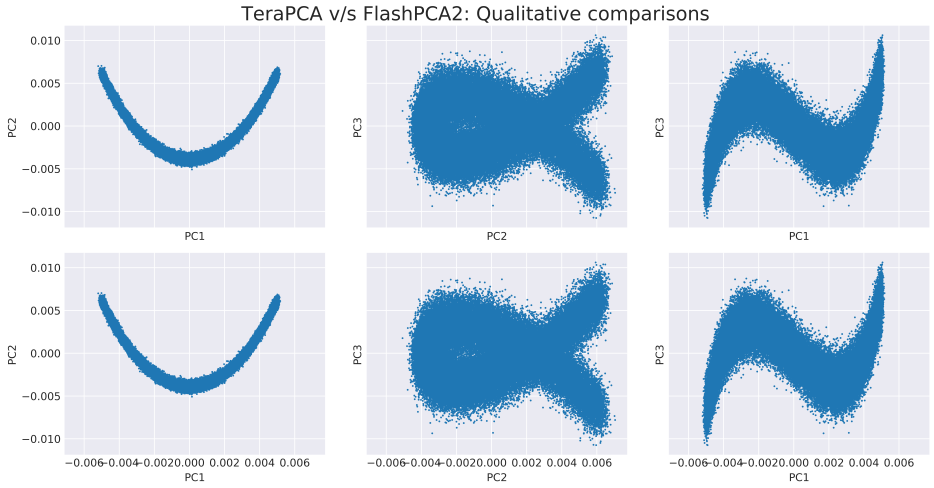


Figure A.27.: Plots of the three leading eigenvectors returned by TeraPCA and FlashPCA2 for the simulated dataset S_6.

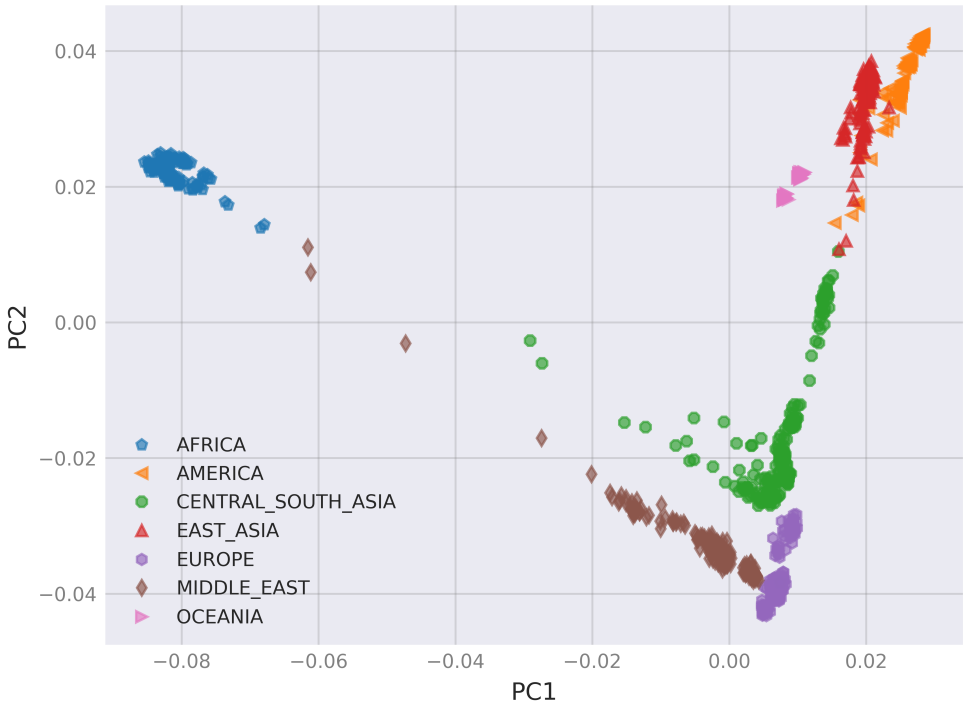


Figure A.28.: The projection of the HGDP dataset along the two leading eigenvectors computed by TeraPCA.

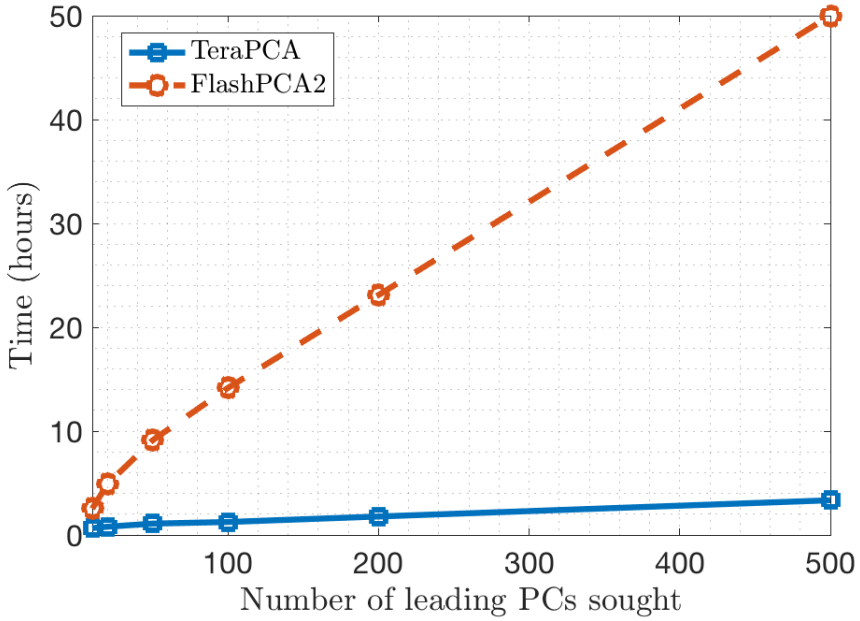


Figure A.29.: The wall-clock times achieved by TeraPCA and FlashPCA2 when the number of eigenvectors that we seek to extract (k) ranges from 10 to 500 for the dataset S_6.

Table A.3.: Accuracy of the ten leading eigenvalues computed by TeraPCA and FlashPCA2.

eigenvalue index	# correct digits		relative error	
	TeraPCA	FlashPCA2	TeraPCA	FlashPCA2
1	15	3	9.91E-15	0.00174228
2	14	4	1.02E-13	0.00129037
3	11	4	5.65E-11	0.00148699
4	9	4	2.18E-08	0.00130829
5	6	3	2.65E-06	0.00110305
6	6	4	3.01E-06	0.00076299
7	6	4	3.36E-06	0.00146959
8	6	4	1.04E-05	0.00068089
9	5	4	7.11E-05	0.00127518
10	4	4	1.74E-04	0.00074424

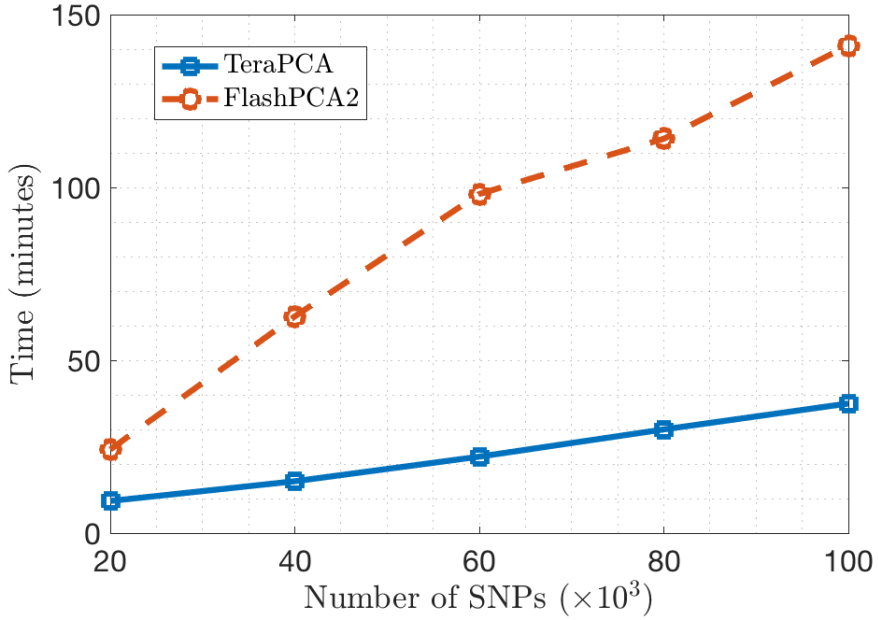


Figure A.30.: The wall-clock times achieved TeraPCA and FlashPCA2 when the number of SNPs ranges from 20K to 100K on for the dataset S_6.

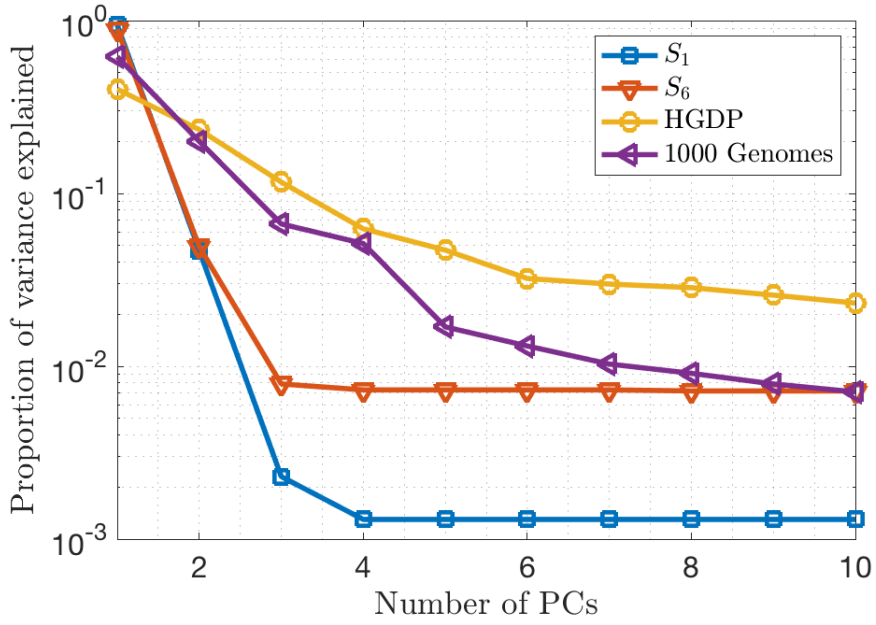


Figure A.31.: Proportion of variance captured by the ten leading eigenvectors returned by TeraPCA when applied on the 1000 Genomes dataset (FlashPCA2 returns essentially the same values for the proportion of variance captured by the top ten eigenvectors).

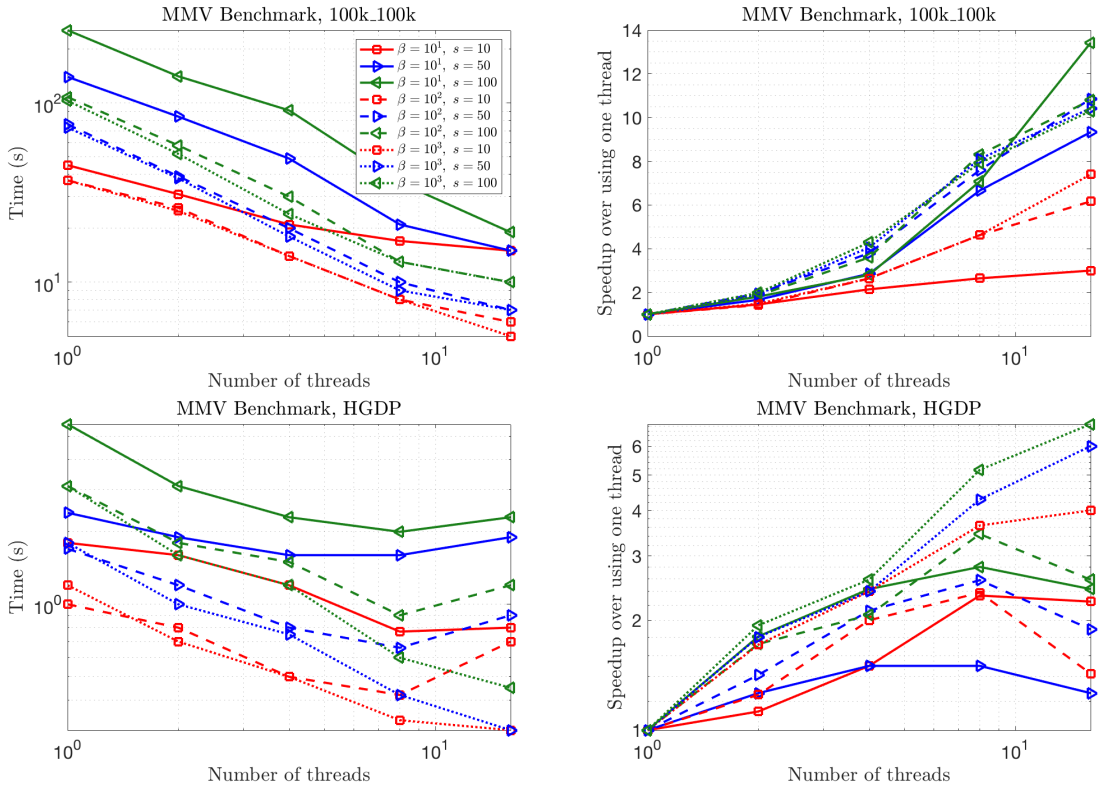


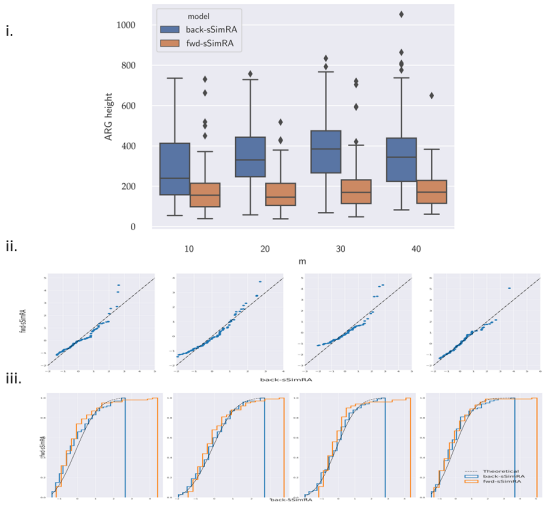
Figure A.32.: Amount of time required to multiply the (normalized) covariance matrix by a set of s vectors using the DGEMM BLAS routine of MKL for different values of s, β and threads, for the datasets S_6 and HGDP.

Appendix C. sSimRA: Multiple Loci Selection with Multiway Epistasis in Coalescence with Recombinations

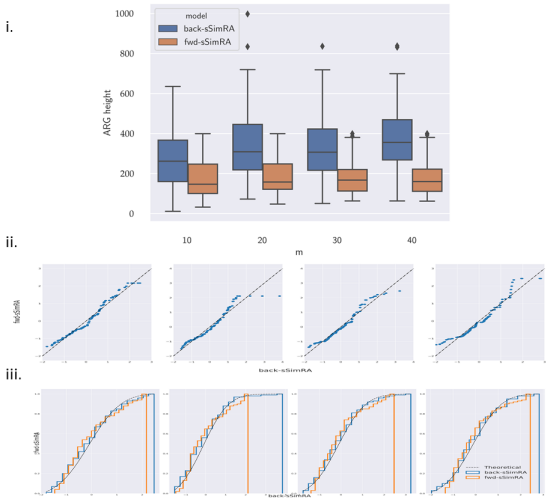
A.4 Supplementary Information

3 interacting loci			e_s	m	p-value	Test statistic
s_1	s_2	s_3				
				10	0.1400	0.16
				20	0.4431	0.12
×	×	×	×	30	0.3439	0.13
				40	0.9995	0.05
				10	0.6766	0.08
				20	0.7942	0.08
s_1	×	×	×	30	0.6766	0.10
				40	0.5750	0.11
				10	0.9921	0.06
				20	0.5560	0.11
s_1	s_2	×	×	30	0.7942	0.09
				40	0.8938	0.08
				10	0.8938	0.08
				20	0.9995	0.05
s_1	s_2	×	0.1	30	0.9710	0.06
				40	0.7942	0.09
				10	0.3439	0.13
				20	0.7942	0.08
s_1	s_2	s_3	×	30	0.6766	0.10
				40	0.5576	0.11
				10	0.9610	0.07
				20	0.9610	0.07
s_1	s_2	s_3	0.1	30	0.3556	0.13
				40	0.6766	0.10

Table A.4.: K-S test statistics with corresponding p-values showing that the probability distributions of H as returned by *fwd-sSimRA* and *back-sSimRA* abstracts each other very closely.

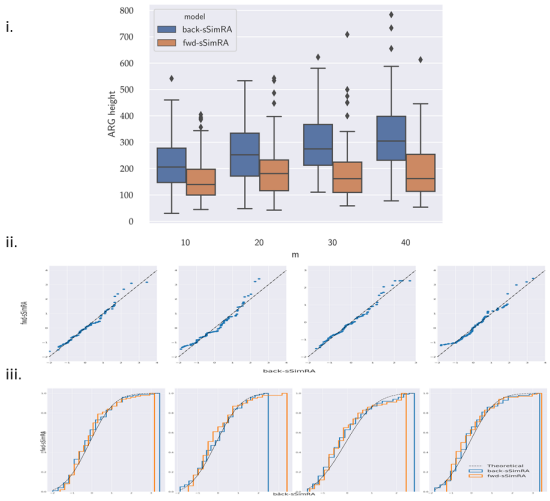


(a) Neutral model

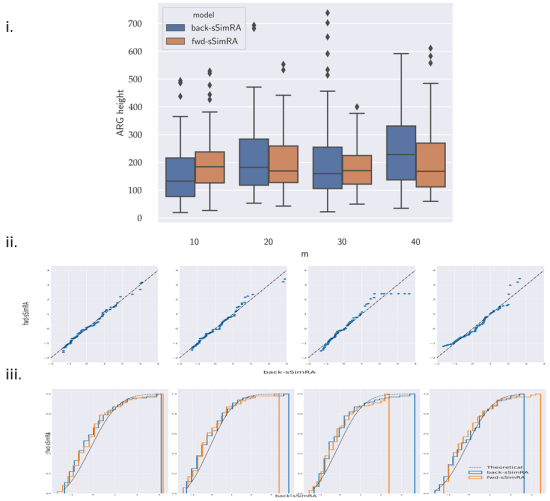


(b) Selection at single locus

Figure A.33.: Comparing the height of the ARG (H) between the *fwd-sSimRa* and *back-sSimRA* for selection at two-loci with and without epistasis, respectively. We set $g = 25K$, $r = 1.0 \times 10^{-8}$ $N = 100$, $s = 0.3$, $e_s = \{0, 0.1\}$ and $m = \{10, 20, 30, 40\}$. (i) The box-and-whisker plot summarizes the result for each m. On each box, the central mark is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. (ii) Q-Q plots for each m showing that the distributions of H from *fwd-sSimRa* and *back-sSimRA* agrees (iii) Plot showing the CDFs of *fwd-sSimRa* and *back-sSimRA* reconfirming the agreement between them.

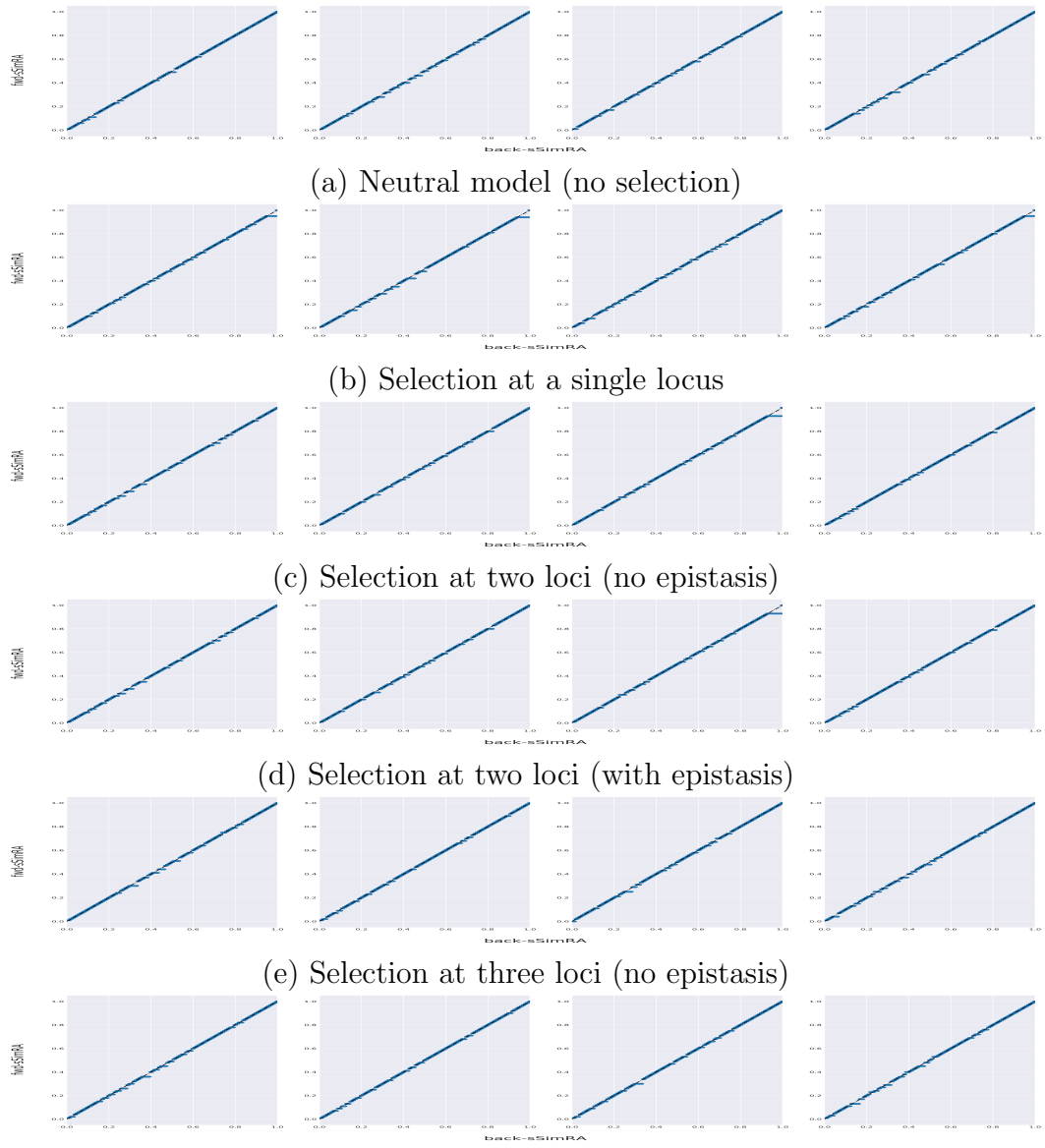


(a) Without epistasis



(b) With epistasis

Figure A.34.: Comparing the height of the ARG (H) between the *fwd-sSimRa* and *back-sSimRA* for selection at two-loci with and without epistasis, respectively. We set $g = 25K$, $r = 1.0 \times 10^{-8}$ $N = 100$, $s = \{0.3, 0.3\}$, $e_s = \{0, 0.1\}$ and $m = \{10, 20, 30, 40\}$. (i) The box-and-whisker plot summarizes the result for each m. On each box, the central mark is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. (ii) Q-Q plots for each m showing that the distributions of H from *fwd-sSimRa* and *back-sSimRA* agrees (iii) Plot showing the CDFs of *fwd-sSimRa* and *back-sSimRA* reconfirming the agreement between them.



(a) Neutral model (no selection)

(b) Selection at a single locus

(c) Selection at two loci (no epistasis)

(d) Selection at two loci (with epistasis)

(e) Selection at three loci (no epistasis)

(f) Selection at three loci (with epistasis)

Figure A.35.: P-P plots of distributions of the height of the ARG (H) between *fwd-sSimRa* and *back-sSimRA* for the neutral model with epistasis and no epistasis at two loci respectively, by setting $g = 25K$, $r = 1.0 \times 10^{-8}$ $N = 100$, $s = 0.3$, $e_s = \{0, 0.1\}$ and $m = \{10, 20, 30, 40\}$.

VITA

VITA

Aritra Bose was born in Kolkata, India on August 8^{th}, 1990. He received a Bachelor of Technology degree in Information Technology from West Bengal University of Technology, Kolkata, India. During his undergraduate he gained his first research experience in Indian Statistical Institute, Kolkata working as an intern advised by Dr. Pabitra Pal Choudhury. Thereafter he held research intern positions in Indian Institute of Technology, Guwahati and Bose Institute, Kolkata working under Dr. Ashish Anand and Dr. Zhumur Ghosh, respectively. After graduating in 2013, he worked in Teradata Corporation as an Analyst for 9 months in Hyderabad, India. He started graduate school at Rensselaer Polytechnic Institute, Troy, NY in the Fall of 2014 and obtained a Master of Science degree in Computer Science in the Summer of 2016. During this time he worked as a teaching assitant as well as a research assistant. He spent the summer of 2016 in Computational Biology Center, IBM T.J. Watson Research Center working with Dr. Laxmi Parida as a research intern. Aritra joined Purdue University in the Fall of 2016 continuing his work on Computational Genetics advised by Dr. Petros Drineas in the Computer Science department and Dr. Peristera Paschou in Biological Sciences department. He spent the summers of 2017 and 2018 working as a research intern in IBM T.J. Watson Research Center, Yorktown Heights, NY. His academic interets include population genetics, statistical genetics, data mining and computational biology.