Powerful New Methods for Decomposing Genome-wide Ancestry and Performing Trait Association

Hrushikesh Loya

Balliol College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas 2024

Abstract

The past decade has revealed numerous admixture events involving populations well-characterized by modern and ancient genomic data. However, current methods often struggle to detect events involving unknown populations, populations with poorly defined genomic composition, or those leaving only subtle traces in our genomes. To overcome these challenges, we introduce GhostBuster, a novel method that leverages genome-wide genealogies to robustly identify such hidden events. We validate GhostBuster against state-of-the-art approaches, demonstrating its accuracy in detecting known admixture events, and apply it to investigate African population history. Our analysis uncovers two key findings: first, a Holocene-era back-to-Africa migration introducing detectable levels of Eurasian and Neanderthal ancestry across Africa; and second, an ancient admixture event with a population closely related to the ancestors of non-African modern humans. These findings are further supported by mutational signatures and affinities with archaic populations. Our results highlight the power of genome-wide genealogical approaches in illuminating humanity's shared past.

The rapid growth of modern biobanks is creating new opportunities for large-scale genome-wide association studies (GWAS) and the analysis of complex traits. However, performing GWAS on millions of samples often leads to trade-offs between computational efficiency and statistical power, reducing the benefits of large-scale data collection efforts. We developed Quickdraws, a method that increases association power in quantitative and binary traits without sacrificing computational efficiency, leveraging a spike-and-slab prior on variant effects, stochastic variational inference and GPU acceleration. We applied Quickdraws to 79 quantitative and 50 binary traits in 405,000 UK Biobank samples, identifying 4.97% and 3.25% more associations than REGENIE, and 22.71% and 7.07% more than FastGWA. Quickdraws had costs comparable to REGENIE, FastGWA, and SAIGE on the UK Biobank RAP service, while being substantially faster than BOLT-LMM. These results highlight the promise of leveraging machine learning techniques for scalable GWAS without sacrificing power or robustness.

Powerful New Methods for Decomposing Genome-wide Ancestry and Performing Trait Association



Hrushikesh Loya

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2024

Acknowledgements

Personal

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Professor Simon Myers and Professor Pier Palamara, for their invaluable guidance, patience, and unwavering support throughout my DPhil journey. Their enthusiasm for science and problem-solving has been both infectious and inspiring, profoundly shaping this work and my growth as a researcher. I am also deeply thankful to my co-authors and collaborators — Leo Speidel, Georgios Kalantzis, and Fergus Cooper — for generously sharing their expertise and insights, which were instrumental in bringing this thesis to fruition.

I would like to express my heartfelt gratitude to my family - my dad, my mom, and my brother - for their unwavering support and belief in me throughout this journey. My parents have made countless sacrifices and difficult decisions, always prioritizing my future and well-being over their own. Their selflessness and love have been a constant source of strength and motivation, and I know that no words can ever fully capture the depth of my gratitude. I am profoundly blessed to have them in my life. My brother, Gaurish, has always been a source of care and positivity, offering his unwavering support through challenging times and I am deeply thankful for his presence in my life. I am also incredibly grateful to Pita, who has been my rock and a constant source of strength over the past six years. She has shared in my joys and triumphs, and has stood by me through the inevitable challenges and setbacks. Her unwavering support, patience, and love have been invaluable, and her presence in my life is something I cherish beyond words.

My experience at Oxford would be incomplete without the incredible friends I made along the way. I am deeply grateful to my housemates — Aston, Javier, Jiarui, Justin, Kristine, Marcel, Rahul, Robbie, Sophie, and Xinran (alphabetically ordered) — whose company made every shared meal, walk, restaurant outing, road trip, and, most importantly, the endless banter in the house unforgettable. Whether it was mocking each other or being there for one another, you all made this journey brighter. I would also like to thank my office and climbing friends — Jiazheng, Lino, Marius, Miriam, Riddhi, Romain, Sinan, and Zoi (alphabetically ordered) — for making my time in the office more enjoyable and eventful. The outdoor climbing trips and sessions at Brookes are unforgettable highlights of my Oxford experience, and they will always hold a special place in my memories.

Institutional

I would like to extend my sincere gratitude to the Clarendon Fund at the University of Oxford, Balliol College at the University of Oxford, the Medical Sciences Doctoral Training Centre (MSDTC) studentship, and the Wellcome Trust studentship for their generous financial support throughout my DPhil.

Previously published material

I am deeply grateful for the opportunity to collaborate with talented and dedicated colleagues throughout my DPhil journey. Their invaluable input has shaped much of the work in this thesis, contributing to peer-reviewed publications. Since scientific ideas and outputs are often developed collaboratively, I use "we" throughout this thesis to reflect the collective nature of these efforts. Where data processing, quality control, or scripts were adapted from others, I have acknowledged their contributions. Unless otherwise stated, all the work presented here, including the code, analyses, figures, and tables, is my own.

Chapters 4 and 5 are adapted from my first-authored paper, "A Scalable Variational Inference Approach for Increased Mixed-Model Association Power," published in Nature Genetics. I have obtained written permission from my supervisor, Prof. Pier Palamara, to include the content of this paper in my thesis.

Abstract

The past decade has revealed numerous admixture events involving populations well-characterized by modern and ancient genomic data. However, current methods often struggle to detect events involving unknown populations, populations with poorly defined genomic composition, or those leaving only subtle traces in our genomes. To overcome these challenges, we introduce GhostBuster, a novel method that leverages genome-wide genealogies to robustly identify such hidden events. We validate GhostBuster against state-of-the-art approaches, demonstrating its accuracy in detecting known admixture events, and apply it to investigate African population history. Our analysis uncovers two key findings: first, a Holocene-era back-to-Africa migration introducing detectable levels of Eurasian and Neanderthal ancestry across Africa; and second, an ancient admixture event with a population closely related to the ancestors of non-African modern humans. These findings are further supported by mutational signatures and affinities with archaic populations. Our results highlight the power of genome-wide genealogical approaches in illuminating humanity's shared past.

The rapid growth of modern biobanks is creating new opportunities for large-scale genome-wide association studies (GWAS) and the analysis of complex traits. However, performing GWAS on millions of samples often leads to trade-offs between computational efficiency and statistical power, reducing the benefits of large-scale data collection efforts. We developed Quickdraws, a method that increases association power in quantitative and binary traits without sacrificing computational efficiency, leveraging a spike-and-slab prior on variant effects, stochastic variational inference and GPU acceleration. We applied Quickdraws to 79 quantitative and 50 binary traits in 405,000 UK Biobank samples, identifying 4.97% and 3.25% more associations than REGENIE, and 22.71% and 7.07% more than FastGWA. Quickdraws had costs comparable to REGENIE, FastGWA, and SAIGE on the UK Biobank RAP service, while being substantially faster than BOLT-LMM. These results highlight the promise of leveraging machine learning techniques for scalable GWAS without sacrificing power or robustness.

Contents

List of Figures **xi**

List of Tables **xxxvii**

List of Abbreviations **xli**

1 Introduction **1**
- 1.1 Setting the scene . . . 1
- 1.2 Understanding population history through genetic data . . . 3
  - 1.2.1 Basic definitions . . . 3
  - 1.2.2 Theoretical framework to understand population history from genetic data . . . 6
  - 1.2.3 Survey of methods to infer population structure and admixture history . . . 11
  - 1.2.4 Large-scale inference of genome-wide genealogies . . . 16
  - 1.2.5 Genome-wide genealogies to understand evolutionary history 19
  - 1.2.6 Motivation for GhostBuster . . . 22
- 1.3 Understanding complex traits through genetic data . . . 26
  - 1.3.1 Basic definitions . . . 26
  - 1.3.2 Genome-wide association studies: usage and applications . . 28
  - 1.3.3 Theoretical framework to perform GWAS . . . 29
  - 1.3.4 Survey of methods to perform GWAS . . . 34
  - 1.3.5 Motivation for Quickdraws . . . 36

2 A method to find ghost populations using genome-wide genealogies 41
- 2.1 Chapter overview . . . 42
- 2.2 Theoretical framework for our method . . . 44
  - 2.2.1 Introduction to mixture models . . . 44
  - 2.2.2 Mixture model for the coalescent . . . 47
  - 2.2.3 EM derivation: terminology . . . 52
  - 2.2.4 EM derivation: the likelihood . . . 55
  - 2.2.5 EM derivation: inference of coalescence rates . . . 59
  - 2.2.6 EM derivation: inference for admixture time and proportions 63

2.2.7 Summary of the algorithm 68
2.3 Method description 69
2.3.1 Graphic overview of the method 69
2.3.2 PCA visualization to visualize ancestral components 69
2.3.3 Coancestry curves to date admixture event 70
2.3.4 Joint analysis of multiple targets 72
2.3.5 Tree Filtering 73
2.3.6 Choosing the number of components 73
2.3.7 Calculation of node persistence 75
2.3.8 Handling uncertain coalescence events 76
2.3.9 Initialization 77
2.4 Data and quality control 79
2.4.1 Modern human datasets 79
2.4.2 High quality ancient DNA resources 80
2.4.3 Building genome-wide genealogies for moderns and ancients 81
2.5 Accurate local ancestry inference in simulations 82
2.5.1 Simulation design 82
2.5.2 Four-way recent admixture 83
2.5.3 Denisovan-like Introgression into Papuans 89
2.6 Identifying several recent admixtures accurately 95
2.7 Identifying multi-way admixture in holocene Europe 100

3 Understanding population structure in Africa using GhostBuster 103
3.1 Chapter overview 104
3.2 Overview of African population history 104
3.2.1 Colonization, maritime trade and Bantu expansion 105
3.2.2 Back-to-Africa migration 107
3.2.3 Deep population structure within Africa 109
3.3 Evidence for widespread back-to-Africa migration 111
3.3.1 Finding Eurasian-like ancestry in Africans 111
3.3.2 Dating the admixture event 116
3.3.3 Eurasian ancestry explains the Neanderthal ancestry in Africans 119
3.3.4 Mutational profiles validate Eurasian segments in Africans . 126
3.3.5 Identifying source populations for Eurasian ancestry 131
3.4 Evidence of a deep admixture event in Africans 134
3.4.1 Ancient admixture across sub-Saharan African populations . 134
3.4.2 Dating the admixture event 137
3.4.3 Mutational profiles validate the admixture event 139
3.4.4 Variation in PRDM9 activity across populations 141

3.4.5 Characterizing the admixture event using coalescence rates . 146
3.4.6 Impact of the admixture event on polygenic score portability in Africa . . . 149
3.5 Discussion . . . 154
3.5.1 Limitations and future work . . . 156
3.5.2 Understanding admixtures in non human organisms . . . 157

4 A scalable variational inference approach for increased mixed-model association power 159
4.1 Chapter overview . . . 160
4.2 Theoretical framework for our method . . . 161
4.2.1 Bayesian inference . . . 161
4.2.2 Variational inference . . . 162
4.2.3 Evidence lower bound and mean-field variational inference . 162
4.2.4 Coordinate ascent variational inference . . . 164
4.2.5 Stochastic variational inference . . . 166
4.3 Method description . . . 169
4.3.1 Overview of the method . . . 169
4.3.2 Bayesian regression with spike-and-slab prior . . . 171
4.3.3 Heritability estimation . . . 176
4.3.4 Test statistics calculation . . . 177
4.3.5 Performance optimization . . . 181
4.3.6 Hyperparameter tuning . . . 183
4.4 Summary of the Quickdraws algorithm . . . 185

5 Application to simulated and real phenotypes in UK Biobank 187
5.1 Chapter overview . . . 188
5.2 Data and quality control . . . 188
5.2.1 UK Biobank data . . . 188
5.2.2 Quality control and preprocessing for GWAS . . . 189
5.3 Performance in simulated data . . . 190
5.3.1 Simulation design . . . 190
5.3.2 False positive rate evaluation . . . 193
5.3.3 Statistical power evaluation . . . 199
5.4 UK Biobank analysis . . . 204
5.4.1 Analysis of false positive rates . . . 204
5.4.2 Number of associations . . . 207
5.4.3 Replication analysis with Biobank Japan, Finngen and other large-scale GWAS . . . 209
5.4.4 Predictive power in out-of-sample data . . . 212

5.4.5 Correction of participation bias 216
5.5 Cost analysis on UK Biobank RAP 216
5.6 Discussion 218
5.6.1 Sharing and combining posterior effect estimates 220
5.6.2 Limitations and future work 220

Appendices

A Additional proofs **225**
A.1 KL-divergence for non-overlapping mixture distributions 225

B Additional figures and tables **227**
B.1 Additional figures 227
B.2 Additional tables 261

References **277**

List of Figures

1.1 **Example of three lineages coalescing backwards in time in the Wright Fisher model.** Each individual chooses its ancestors from the set of ancestors with replacement until all the lineages coalesce into a single common ancestor. The three samples are colored blue and the common ancestor for the three samples red . . 7

1.2 **Example of an ancestral recombination graph (ARG) and the corresponding sequence of trees.** A. In the coalescent with recombination model the Lineages going backward in time can either coalesce or recombine, which means different portions of the genome have slightly different history. B. Corresponding sequence of trees representation of the ARG, **X** denotes the recombination position. . 10

1.3 **PCA visualization of** $500,000$ **individual genotypes from the UK Biobank.** PC1 and PC2 separate the African individuals from non-African individuals, whereas PC3 and PC4 separate individuals in Eurasia. The figure was taken from [3] (Creative Commons License). 14

1.4 **Evolutionary forces impact genotypes through genealogies.** Evolutionary forces like mutations, recombination, population bottlenecks, admixture, migration and selection affect the observed genotypes through the unobserved genealogies 17

1.5 **Some examples of migration and admixture events in human history.** A. Archaeological and genetic evidence suggests modern humans originated in Africa and migrated out between 50,000 and 100,000 years ago [18, 77]. They interbred with Neanderthals, who contributed up to 3% of non-African DNA. After leaving Africa, the population split into west and east Eurasian groups, creating genetic differences between Europeans and Asians. East Eurasians also admixed with Denisovans, though the timing and location are debated [25, 78]. B. Europe has experienced significant migrations, starting with Anatolian farmers 7,000-9,000 years ago, followed by Steppe pastoralists around 5,000 years ago, which heavily influenced present-day European ancestry. There is also evidence of recent gene flow from North African groups into southern Europeans, such as Italians and Spanish, and ancient Siberian-related ancestry in northeastern Europeans like Finns and Russians [21, 58]. C. The peopling of America likely began during the last glacial period, which created a land bridge to Eurasia [79], ending around 15,000 years ago. Ancient DNA in America shows genetic links to ancient Siberians (MA1), Han Chinese, and the Onge tribe [80]. Present-day American genetics are influenced by historical events like colonization and the trans-Atlantic slave trade, resulting in admixture from Native American, African, and European sources [22, 23]. D. Previously identified human admixtures were determined using various methods, including haplotype-based approaches [51, 53–55], ancient DNA, and archaic DNA analyses. The histogram shows the ages of ancient DNA samples sequenced to date, sourced from the Allen Ancient DNA Resource [81]. The illustrations were downloaded from Flaticon. 25

1.6 **Overview of a typical GWAS pipeline.** The GWAS pipeline typically begins with genotyping and phenotyping a large cohort of individuals, including both cases and controls for disease traits. Genotyping can be conducted using either low-cost SNP arrays or whole-genome sequencing. Quality control steps are crucial and involve filtering variants and samples based on various criteria (see Section 5.2 for detailed quality control procedures). If genotyping was performed instead of whole-genome sequencing, genotype imputation follows as the next step. Subsequently, genome-wide association testing is conducted for each variant and phenotype of interest to generate p-values and effect estimates, indicating the association between each variant and the phenotype. The resulting GWAS summary statistics have several downstream applications. These include replication and meta-analyses to increase statistical power, fine-mapping and functional analyses to identify causal variants among the associated variants, and the calculation of polygenic scores. This pipeline structure was adapted from [101], with the fine-mapping figure from [102] and the polygenic scores figure from [103]. 30

2.1 **A simplified overview of GhostBuster.** GhostBuster implements an iterative expectation-maximization (EM) algorithm which cycles between two states A and B. (A) inferring local ancestry of a target individual given multiple coalescence rates of the target individuals with a set of reference populations. (B) using the local ancestry to refine the estimates of coalescence rates of target individual and the reference populations. 43

2.2 **Gaussian mixture model.** The observed data and inferred clusters using the Gaussian mixture model after 100 EM iterations, the Gaussian mixtures inferred are represented by the red ellipse. 46

2.3 **Overview of the GhostBuster algorithm.** The EM algorithm computes a local maximum likelihood estimate (MLE) for the mixture model, clustering genomic regions based on variations in coalescence histories for a given target individual. In the E-step, the expected log-likelihood of the data is calculated, conditioned on the local tree topology and the current estimates of model parameters (proportions, admixture time, and coalescence rates). An HMM is then used to infer the local ancestry of the target individual, with emissions modeled using the likelihood of coalescence times and transitions modeled as a Poisson process dependent on recombination rates. In the M-step, the expected log-likelihood obtained from the E-step is maximized to update the model parameters. The EM algorithm ensures non-decreasing likelihood updates at each iteration, leading to convergence at a local optimum [136]. The likelihood is conditioned on the tree topology by carefully accounting for lineage contributions based on their descendants. This approach provides higher statistical power to detect admixture events by avoiding assumptions of independence among reference individuals or reliance on composite likelihoods. Detailed derivations of the likelihood and EM algorithm updates are presented in Section 2.2. 69

2.4 **Comparison of Approximate and True Node Persistence in Simulated Trees.** True node persistence is determined by identifying the exact coalescence times across neighboring trees, while approximate node persistence is estimated based on the correlation of descendant sets between adjacent trees. In this case, the correlation cutoff for approximate node persistence is set to 0.5, with the number of overlaps indicating the number of 10kb intervals for which the node persists. 76

2.5 **Handling uncertainty in coalescence events.** We calculate the expected number of coalescence events for subsets of nodes jointly supported by a mutation. In this example, there are three such subsets on the target lineage, with weights for each coalescence event calculated based on the number of nodes in the subset. 78

2.6 **Simulation design for performance evaluation.** A. Four-way recent admixture simulation, B. Simulated Denisovan introgression into Papuans. The simulation parameters including the population size, mutation rate and recombination map are described in Section 2.5.1. 84

2.7 **Decomposing focal individuals in the four-way admix simulation with presence of all source populations.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Precision-recall curves, (f) Calibration curves, and (g) Prediction R^2 for inferred local ancestry compared to the simulated ground truth, where inference was done using true trees or Relate trees. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. 87

2.8 **Decomposing focal individuals in the four-way admix simulation with absence of one source population.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Precision-recall curves, (f) Calibration curves, and (g) Prediction R^2 for inferred local ancestry compared to the simulated ground truth, where inference was done using true trees or Relate trees. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. 88

2.9 **Decomposing Papuan individuals in simulations with Mbuti and Denisovans as additional references.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Precision-recall curves, (f) Calibration curves, and (g) Prediction R^2 for inferred local ancestry compared to the simulated ground truth, where inference was done using true trees or Relate trees. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. 92

2.10 **Decomposing Papuan individuals in simulations with only Mbuti as additional references.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Precision-recall curves, (f) Calibration curves, and (g) Prediction R^2 for inferred local ancestry compared to the simulated ground truth, where inference was done using true trees or Relate trees. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. 93

2.11 **Decomposing Mbuti individuals in simulations with Papuan as additional references.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Histogram of the inferred local ancestry posteriors. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. 94

2.12 **Decomposing Hazara individuals in Human Genome Diversity Project.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Local ancestry inference for GhostBuster vs. Mosaic. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. The diploid local ancestry in (e) is based on chromosome 1 for HGDP samples HGDP00102 and HGDP00103. 97

2.13 **Decomposing Bedouin individuals in Human Genome Diversity Project.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Local ancestry inference for GhostBuster vs. Mosaic. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. The diploid local ancestry in (e) is based on chromosome 1 for HGDP samples HGDP00607 and HGDP00608. 98

2.14 **Decomposing Maya individuals in Human Genome Diversity Project.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Local ancestry inference for GhostBuster vs. Mosaic. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. The diploid local ancestry in (e) is based on chromosome 1 for HGDP samples HGDP00856 and HGDP00860. 99

2.15 **Decomposing British individuals in** 1,000**GP with ancient farmers, hunter-gatherers and Yamnaya.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Histogram of the inferred local ancestry posteriors. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. 102

3.1 **Finding Eurasian-like ancestry in sub-Saharan Africans.** (a) Inferred inverse coalescence rates and proportions when decomposing Yorubans: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Map of the African samples analyzed with red inverted triangles corresponding to modern samples in HGDP, 1,000GP and SGDP, and green squares for 5 ancient African samples (with their estimated sampling ages in brackets). (d) Proportion of Eurasian-like ancestry and (e) Expected coefficient of determination across all African populations analyzed. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. Individual data-points overlaid on the violin plots. 115

3.2 **Dating the Eurasian-like ancestry in sub-Saharan Africans.** Coancestry curves and (a) one-date admixture model fit, (b) two-date admixture model fit and (c) constant migration rate continuous admixture model fit for Yorubans. (d) Per-population least squared log-likelihood corresponding to various admixture models along with the admixture dates and weights inferred using two-date admixture model. The log-likelihood values are aggregated across 20 jackknife runs. Populations are sorted based on highest to lowest proportion of Eurasian-like ancestry, pink represents the younger admixture date and purple represents the older admixture date in (d). Individual data-points overlaid on the violin plots. Admixture dates in the plot are in generations. 117

3.3 **Eurasian ancestry explains the Neanderthal ancestry in Africans.** (a) Inferred inverse coalescence rates and proportions for Neanderthal ancestry in Yorubans: the line represents the inverse coalescence rate profiles with sequenced Neanderthals. (b) Sample HGDP01405 showing a 20 Mb region enriched for Neanderthal ancestry, overlapping Eurasian-like segments. The average enrichment is calculated as the ratio of the mean posterior probability of Neanderthal ancestry in Eurasian segments, $P(\text{Nea} \mid \text{Eurasian})$, to the mean posterior probability in African segments, $P(\text{Nea} \mid \text{African})$. (c) Coancestry curve showing the normalized joint probability and single admixture date fit for Eurasian segments carrying Neanderthal ancestry, averaged across all African samples analyzed. (d) Admixture dates derived from coancestry curves generated by Eurasian segments carrying Neanderthal ancestry for each population. For Bantu Kenya a two-date admixture model fit is shown, as it better explains the coancestry patterns. More detailed results are provided in Figure B.12. Grey lines in (c) and error bars in (d) represent jackknife estimates and 95% confidence intervals around the mean. For panel (d) single date admixtures shown in golden, whereas two date admixtures shown in pink and purple. Results derived from HGDP+1,000GP+aDNA+archaic genealogies. 125

3.4 **Enrichment of TCC to TTC mutation rates in African components.** (a) Enrichment of TCC to TTC mutation rates across various continental groups, including NFE (non-Finnish Europeans), SAS (South Asians), EAS (East Asians), and MID (Middle Eastern and North African populations). (b) Enrichment of TCC to TTC mutation rates in Eurasian and non-Eurasian (African) segments in African populations. (c) TCC to TTC mutation rate enrichment in five ancient African samples, including new001, ela001, baa001, I5950, and I10871 (see population descriptions in Table 3.1b). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. In panel (b), the Eurasian segments in Africans are further categorized based on their lengths: segments shorter than 0.1 cM, those between 0.1 and 0.5 cM, and segments longer than 0.5 cM. Results derived from HGDP+1,000GP and HGDP+1,000GP+aDNA+archaic genealogies. . . 128

3.5 **Enrichment of weak to strong mutation rates in African components.** (a) Enrichment of weak to strong mutation rates across various continental groups, including NFE (non-Finnish Europeans), SAS (South Asians), EAS (East Asians), and MID (Middle Eastern and North African populations). (b) Enrichment of weak to strong mutation rates in Eurasian and non-Eurasian (African) segments in African populations. (c) weak to strong mutation rate enrichment in five ancient African samples, including new001, ela001, baa001, I5950, and I10871 (see population descriptions in Table 3.1b). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. In panel (b), the Eurasian segments in Africans are further categorized based on their lengths: segments shorter than 0.1 cM, those between 0.1 and 0.5 cM, and segments longer than 0.5 cM. Results derived from HGDP+1,000GP and HGDP+1,000GP+aDNA+archaic genealogies. . . 130

3.6 **Coalescence rates with modern and ancient samples.** (a) Matrix of pairwise coalescence rates for all modern and ancient groups in the epoch spanning 10,000–15,848 years. (b) Dendrogram corresponding to the UPGMA hierarchical clustering of the pairwise coalescence rate matrix in (a). (c) PCA visualization of coalescence rates for the epoch spanning 10,000–15,848 years, with PC1 and PC2 representing the eigenvectors capturing the greatest variance among modern samples. (d-e) Top 25 groups based on coalescence rates with (d) younger Eurasian-like segments in the San, and (e) younger Eurasian-like segments in Yoruba during the epoch spanning 10,000–15,848 years. Row and column colors in (a), dendrogram leaves in (b), and dots and bars in (c-e) are colored according to the common legend. NFE = non-Finnish Europeans, EAS = East Asians, MID = Middle Easterners and North Africans, SAS = South Asians, OCE = Oceanians. The numbers in brackets in (c) indicate the proportion of variance explained by each principal component. The error bars in (d-e) represent 95% confidence intervals calculated using 100 bootstraps. Results derived from HGDP+1,000GP+aDNA+archaic genealogies. 133

3.7 **Decomposing sub-Saharan Africans to find a deep admixture event.** (a) Inferred inverse coalescence rates and proportions when decomposing Yorubans: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Proportion of the out of Africa-like ancestry and (e) Expected coefficient of determination across all African populations analyzed. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. Individual data-points overlaid on the violin plots. 136

3.8 **Dating the ancestral components corresponding to deep admixture event.** Coancestry curves and (a) one-date admixture model fit, (b) two-date admixture model fit and (c) constant migration rate continuous admixture model fit for Yorubans. (d) Per-population least squared log-likelihood corresponding to various admixture models along with the admixture dates and weights inferred using two-date admixture model. The log-likelihood values are aggregated across 20 jackknife runs. Populations are sorted based on highest to lowest proportion of Eurasian-like ancestry, pink represents the younger admixture date and purple represents the older admixture date in (d). Individual data-points overlaid on the violin plots. Admixture dates in the plot are in generations. 138

3.9 **Enrichment of weak to strong mutation rates in the deep admixture African components.** (a) Enrichment of weak to strong mutation rates across various continental groups, including NFE (non-Finnish Europeans), SAS (South Asians), EAS (East Asians), and MID (Middle Eastern and North African populations). (b) Enrichment of weak to strong mutation rates in OOA-like and non-OOA-like (African) segments in African populations. (c) Weak to strong mutation rate enrichment in five ancient African samples, including new001, ela001, baa001, I5950, and I10871 (see population descriptions in Table 3.1b). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. Results derived from HGDP+1,000GP and HGDP+1,000GP+aDNA+archaic genealogies. 140

3.10 **Weak to strong mutation rates at recombination hotspots across populations.** Weak to strong mutation counts across populations and deep admixture component in (a-b) PRDM9-A hotspots, (c-d) PRDM9-C hotspots. Panel (a) and (c) measure the number of individuals carrying weak to strong mutations around the PRDM9 hotspots, normalized by the total number of individuals and the AT content of the region. Whereas, Panel (b) and (d) measure the time-stratified normalized mutation rate enrichment with respect to weak to strong mutation rates in East Asians. In panel (b) and (d) the shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. Results derived from HGDP+1,000GP and HGDP+1,000GP+aDNA+archaic genealogies. . . 145

3.11 **Coalescence rates with modern and ancient samples.** (a) Matrix of pairwise coalescence rates for all modern and ancient groups in the epoch spanning 64,000–100,000 years. (b) Dendrogram corresponding to the UPGMA hierarchical clustering of the pairwise coalescence rate matrix in (a). (c) PCA visualization of coalescence rates for the epoch spanning 64,000–100,000 years, with PC1 and PC2 representing the eigenvectors capturing the greatest variance among modern samples. (d-e) Top 25 groups based on coalescence rates with (d) Denisovan, and (e) Neanderthals during the epoch spanning 64,000–100,000 years. (f-g) Inverse coalescence rates (ICR) of (f) Denisovan and (g) Neanderthals with other modern human groups from 64,000-1.6million years. Darker lines represent population-averaged ICRs. Row and column colors in (a), dendrogram leaves in (b), dots and bars in (c-e), and lines in (f-g) are colored according to the common legend. NFE = non-Finnish Europeans, EAS = East Asians, MID = Middle Easterners and North Africans, SAS = South Asians, OCE = Oceanians. The numbers in brackets in (c) indicate the proportion of variance explained by each principal component. The error bars in (d-e) represent 95% confidence intervals calculated using 100 bootstraps. Results derived from HGDP+1,000GP+aDNA+archaic genealogies. 147

3.12 **Inverse coalescence rates (ICRs) of OOA-like and Non-OOA-like component in Yoruba.** ICRs of OOA-like (YOROBA_MINOR) and Non-OOA-like (YORUBA_MAJOR) components in Yoruba with other modern human and archaic groups from 64,000-1.6million years ago. Darker lines represent population-averaged ICRs and fainter lines represent within population ICRs. NFE = non-Finnish Europeans, EAS = East Asians, MID = Middle Easterners and North Africans, SAS = South Asians, OCE = Oceanians. 148

3.13 **Differences in PGS predictive power across populations and deep admixture components.** (a) PGS prediction R^2 across 32 quantitative phenotypes and four population groups: self-identified non-British Europeans (EUR), South Asians (SAS), East Asians (EAS), and Africans (AFR). Error bars represent the 95% confidence intervals for the prediction R^2. (b) Normalized predictive ability of the PGS based on the OOA-like and non-OOA-like ancestry components across the same 32 quantitative phenotypes. Error bars represent the 95% confidence intervals calculated using 1,000 bootstrap replicates. Traits are sorted based on their prediction R^2 in Europeans. PGS was calculated using posterior mean effect estimates obtained during Quickdraws' model fitting step. A meta-analysis using inverse variance-weighted average across traits is presented in (a) and (b). 153

3.14 **Decomposing American wolves with grey wolves and coyote as reference.** (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Histogram of the inferred local ancestry posteriors. (e) Coancestry curves and model fit for one-date, two-date, and continuous migration scenarios (admixture dates in generations). The figure title includes 'logl', which refers to the log-likelihood of the fit (higher values indicate a better fit). The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. 158

4.1 **A simplified overview of Quickdraws.** Quickdraws is a scalable mixed model designed for genome-wide association studies (GWAS). It improves the power of association testing through a two-step process. First, it estimates the genomic prediction for a trait by excluding the testing chromosome, employing a leave-one-chromosome-out method. The prediction models sparsity of effects and is optimized using advanced variational inference techniques ensuring scalability for large sample sizes. Second, the model leverages this genomic prediction to improve the signal-to-noise ratio for finding association between a trait and a variant. 160

4.2 **Overview of the Quickdraws algorithm.** The algorithm is divided into two main steps, first step deals with estimating genetic effects and the second step with calculating and calibrating per-variant test statistics. The first step is further divided into heritability estimation using moment-based methods [111] and Bayesian regression for phenotype prediction which uses stochastic variational inference and GPU-based matrix multiplication for increased scalability. In the test statistics calculation step we use a novel calibration strategy which estimates and matches effective sample size with linear regression model run on unrelated and homogeneous subset of the data, along with approximate Firth logistic regression to correct for biases in binary traits. The illustrations were downloaded from Flaticon. . . 171

4.3 **Transfer-learning in Bayesian regression.** We measure the causal χ^2 and false positive rate at 5×10^{-4}, with and without pretraining (for 90 epochs), and varying the number of LOCO training epochs. Error bars represent the standard error of the mean χ^2 or false positive rate across 50 independent quantitative traits with GB-unrel structure and 2% polygenicity. 182

5.1 **Phenotype simulation design.** We simulate causal effects using a spike-and-slab, mixture of Gaussian, Laplace or Gaussian distribution only on odd chromosomes, whereas we treat even chromosomes as null controls to test for false positive rate and calibration of test statistics. 191

5.2 **Summary of calibration in simulations for quantitative traits.** False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the number of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, GB-rel+, EUR) is provided in Section 5.3.1. 194

5.3 **Summary of calibration in simulations for binary traits and common variants (MAF $\geq 1\%$).** False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the number of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, EUR) is provided in Section 5.3.1. 196

5.4 **Summary of calibration in simulations for binary traits and rare variants (MAF $< 1\%$).** False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the number of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, EUR) is provided in Section 5.3.1. 197

5.5 **Statistical power in simulated quantitative and binary traits for unrelated British samples.** (a) % increase in average χ^2 test statistics at causal variants with respect to linear regression (Lin. Reg.) for quantitative traits, varying the number of simulated causal variants from 5,000 to 50,000 variants. (b) % increase in average χ^2 test-statistics at causal variants with respect to logistic regression (Log. Reg.) for binary traits, varying the number of simulated causal variants from 1,250 to 10,000 variants. Traits in panel (a) and (b) are simulated for 50,000 samples with $h^2 = 0.4$; the prevalence is fixed to 30% for binary traits in panel (b). Error bars represent the standard error in the percentage improvement, measured using 50 independent traits. The causal χ^2 is normalized by the mean χ^2 at null variants for each trait. 200

5.6 **Power and FPR in N= 405k simulation.** (a) The mean χ^2 at causal variants across different methods and polygenicities for Quantitative traits, and prevalence for binary traits. The χ^2 at causal variants is not normalized by the χ^2 at null variants in this plot. (b) The mean false positive rates (FPR) at 0.005 calculated at null variants across different methods and polygenicities for Quantitative traits, and prevalence for binary traits. The simulation was performed using the N= 405k set of white British individuals. The error bars represent 95% confidence intervals. The red dashed line corresponds to false positive rate = 0.005. 202

5.7 **Power and FPR in N= 460k simulation.** (a) The mean χ^2 at causal variants across different methods and polygenicities for quantitative traits, and prevalence for binary traits. The χ^2 at causal variants is not normalized by the χ^2 at null variants in this plot. (b) The mean false positive rates (FPR) at 0.005 calculated at null variants across different methods and polygenicities for Quantitative traits, and prevalence for binary traits. The simulation was performed using the N= 460k set of UK Biobank self-identified Europeans [3]. The error bars represent 95% confidence intervals. The red dashed line corresponds to false positive rate = 0.005. 203

5.8 **Calibration in real data analysis.** Attenuation ratio of Quickdraws (N=405k) vs. linear regression in unrelated samples (N=337k); vertical and horizontal lines represent mean ± standard errors in the attenuation ratio estimate for each method. 205

5.9 **Functional enrichment at associated variants for quantitative traits.** Functional enrichment profile of variants associated across 79 quantitative traits. We considered either (a) the set of variants significantly associated ($p \leq 5 \times 10^{-8}$) using both Quickdraws and REGENIE, matching the p-value distribution (see Methods), or (b) the set of variants associated using Quickdraws but not REGENIE. Error bars represent jackknife standard errors; the red dashed line corresponds to no enrichment. 206

5.10 **Functional enrichment at associated variants for binary traits.** Functional enrichment profile of variants associated across 50 binary traits. We considered either (a) the set of variants significantly associated ($p \leq 5 \times 10^{-8}$) using both Quickdraws and REGENIE, matching the p-value distribution (see Methods), or (b) the set of variants associated using Quickdraws but not REGENIE. Error bars represent jackknife standard errors; the red dashed line corresponds to no enrichment. 206

5.11 **Approximately independent loci and effective sample size in UK Biobank analysis.** (a) Number of approximately independent loci after plink clumping in FastGWA (x-axis) vs. REGENIE, BOLT-LMM, and Quickdraws for 79 quantitative traits. (b) Number of approximately independent loci after Plink clumping in REGENIE (x-axis) vs. Quickdraws for 250 randomly sampled plasma protein traits. (c) The χ^2 for 79 quantitative traits and N= 405k UK Biobank set conditioned on genome-wide significance ($p < 5 \times 10^{-8}$) in linear regression run on unrelated subset of the data (N= 337k). Median χ^2: FastGWA = 58.57, REGENIE = 68.08, BOLT-LMM = 70.16, Quickdraws = 70.64. (d) Histogram of the effective-sample size increase compared to FastGWA for 79 quantitative traits, measured as the mean χ^2 minus 1 at genome-wide significant variants inferred though linear regression run on unrelated subset of the data (N= 337k). For (a-c) The gray line represents the y=x line and the dashed lines represent a linear regression fit for each method, for (d) the dashed line represents the median improvement in effective sample-size across traits. In (c), the slope refers to the linear regression slope of the χ^2 values between each pair of methods. 208

5.12 **Replication analysis in Biobank Japan and Finngen.** (a) The χ^2 in UK Biobank conditioned on genome-wide significance ($p < 5 \times 10^{-8}$) in replicating biobank across GWAS methods. The variants are aggregated across traits, quantitative and binary traits for REGENIE and FastGWA, and only quantitative traits for BOLT-LMM. The gray line represents the y=x line, and the dashed line the linear regression fit without intercept. Median χ^2 for quantitative traits: FastGWA = 28.36, REGENIE = 33.44, BOLT-LMM = 34.47, Quickdraws = 36.32; binary traits: FastGWA = 14.08, REGENIE = 14.34, Quickdraws = 14.51. (b-c) Replication venn diagram for (b) 30 overlapping quantitative traits in Biobank Japan and (b) 23 overlapping binary traits in Biobank Japan and Finngen. (d-e) Number of replicated variants vs replication rate for (d) quantitative and (e) binary traits. The discovery threshold was fixed to 5×10^{-9} and the replication threshold was varied from 5×10^{-2} to 5×10^{-8}. The error bars represent standard errors calculated using block jack-knife across chromosomes. The dashed line represents a cubic spline fit to the datapoints. In (a), the slope represents the linear regression slope of the χ^2 values between methods, and in (b-c), the percentages indicate the proportion of associations found with respect to the union of all methods. 211

5.13 **Phenotype prediction analyses in the UK Biobank.** (a) Held-out mean phenotype prediction R^2 comparing step 1 posterior estimates from Quickdraws, BOLT-LMM and BOLT-LMM-Inf. (b) Comparing Quickdraws' step 1 posterior estimates with PGS calculated using Quickdraws' association statistics and P+T (pruning and thresholding as implemented in PRSice) or PRS-CS. (c-d) Comparing predictive power for PGS calculated using association statistics from different GWAS methods and different PGS methods, (c) P+T as implemented in PRSice and (d) PRS-CS. All analyses was performed on 27,683 held-out non-British Europeans, 9,044 self-identified south Asians, 1,457 self identified east Asians and 7,204 self identified African of African American samples. Results are aggregated across the 79 quantitative traits we analyzed, and the error bars represent 95% confidence interval of the mean prediction R^2 for each method in each population subgroup. 213

5.14 **Effects of PGS-adjustment on association power.** (a) Total number of independent loci found on chromosomes 1-5, using different GWAS methods with and without PGS adjustment in 79 UK Biobank quantitative traits. (b) Mean χ^2 - 1 at significant variants with and without PGS adjustment for the same traits and regions. Significant variants were defined as those with LR-unrel p-value $\leq 5 \times 10^{-8}$. FastGWA with PGS adjustment yielded an increase in the number of independent loci (with PGS adjustment, 159.1; without, 141.6; paired t-test $p = 1.9 \times 10^{-9}$) and higher mean χ^2-1 at significant variants (with PGS adjustment, 96.0; without, 89.8; $p = 3.3 \times 10^{-11}$). REGENIE with PGS adjustment yielded fewer independent loci (with PGS adjustment, 144.1; without, 150.1; $p = 7.4 \times 10^{-6}$), and similar mean χ^2 - 1 values (with PGS adjustment, 124.6; without, 124.1; $p > 0.05$). 215

5.15 **Computational performance of Quickdraws on different computing architectures.** We assessed the running time and cost associated with Quickdraws' model fitting step (Step 1) across varying numbers of samples (N) and phenotypes (P). Performance was compared for three architectures: an Nvidia A10G GPU node on the UK Biobank Research Analysis Platform (RAP) (`mem2_ssd2_gpu1_x8`), a 14-core MacBook Pro equipped with the Apple M3 chip, and an 8-core CPU node on the UK Biobank RAP (`mem2_ssd1_v2_x8`). Running costs were calculated only for the GPU and CPU nodes on the RAP. 222

B.1 **Comparison of allele frequencies between HGDP+1,000GP and ancient samples.** We analyze biallelic SNPs on chromosome 20 within the masked regions. The comparison includes union ("all sites") and intersection ("overlapping sites") of SNPs in the two datasets seperately. 228
B.2 **Pairwise inverse coalescence rates estimated from Relate trees for the HGDP+1,000GP dataset.** Genome-wide average inverse cross-coalescence rates between afr (Sub-Saharan Africans), eas (East Asians), nfe (Non-Finnish Europeans), and sas (South Asians). 228
B.3 **Pairwise inverse coalescence rates estimated from Relate trees for the HGDP+1,000GP+aDNA dataset.** Genome-wide average inverse cross-coalescence rates between I9590 (Mota individual, Ethiopia), afr (modern Sub-Saharan Africans), baa001 (Ballito Bay individual, South Africa), and nfe (modern Non-Finnish Europeans). 229
B.4 **Pairwise inverse coalescence rates estimated from Relate trees for the HGDP+1,000GP+archaics dataset.** Genome-wide average inverse cross-coalescence rates between Denisovan, Neanderthals, afr (modern Sub-Saharan Africans), and nfe (modern Non-Finnish Europeans). 229
B.5 **Pairwise inverse coalescence rates estimated from Relate trees for the SGDP dataset.** Genome-wide average inverse cross-coalescence rates between Africans, East Asians, South Asians, and West Eurasians. 230
B.6 **Pairwise inverse coalescence rates estimated from Relate trees for the European subset of the HGDP+1,000GP+aDNA dataset.** Genome-wide average inverse cross-coalescence rates between Farmers, WHG (Western Hunter-Gatherers), Yamnaya, and nfe (modern Non-Finnish Europeans). 230
B.7 **Co-ancestry curves for 3-way admixture in Europe.** We refit each co-ancestry curve with an exponential fit to estimate the date of a multi-way admixture event. In this case component 1 represent WHG ancestry, component 2 farmer and component 3 Yamnaya-like ancestry. Inferred admixture dates in the plot are in generations. . . 231
B.8 **Mean Eurasian-like ancestry proportion stratified by background selection intensity.** The mean Eurasian-like ancestry proportion is calculated across African samples from the HGDP and 1,000GP. The background selection B-map was obtained from [192]. Error bars represent standard errors around the mean. 232

B.9 **Eurasian-like ancestry segment length distribution across African populations.** Panels (a-v) show histograms of Eurasian-like ancestry segment lengths for each African population analyzed in the HGDP, 1,000GP, and SGDP datasets. Each panel title includes the population label, the reciprocal of the African proportion, and the mean segment length (an estimate of single-event admixture date in generations). The x-axis is truncated at 5 cM. 235

B.10 **Detecting Neanderthal ancestry in Eurasian populations with GhostBuster.** (a) Proportion of Neanderthal ancestry estimated for two East Asian (Han, Japanese), two European (French, Russian), and two South Asian (Pathan, Brahui) populations from the HGDP+1,000GP dataset. (b) Chromosome-wide average proportion of Neanderthal ancestry. (c) Genome-wide average proportion of Neanderthal ancestry along the genome. (d) Average Neanderthal ancestry across five previously identified archaic deserts [173, 199]. In panel (c), the green dashed line represents the genome-wide average Neanderthal ancestry. In panel (d), the dashed lines mark the start and end positions of the deserts. Regions within and 100 kb around the ENCODE blacklist [287] have been excluded from the analysis. Panel (a) overlays individual genome-wide Neanderthal ancestry data points for each sample on the violin plots. The positions shown in panel (d) are based on the hg38 genome build. 236

B.11 **Joint-regression of Neanderthal and Eurasian-like ancestry in Africans.** (a) Mean coefficient estimates and confidence intervals from the joint regression analysis of genome-wide Neanderthal ancestry proportions against Eurasian-like ancestry proportions, divided into younger and older Eurasian-like segments. (b) Genome-wide residual Neanderthal ancestry vs. Eurasian-like ancestry proportions across all African samples. The residual Neanderthal proportion is defined as the genome-wide Neanderthal ancestry proportion minus the estimated proportion attributed to other Eurasian-like ancestry, as inferred from the joint regression. Error bars in (a) represent 95% confidence intervals around the mean estimated coefficients using least squares. Error bars in (b) represent the population-specific standard deviation of Eurasian and Neanderthal proportions, while the dashed line in (b) indicating the least squares fit. 237

B.12 **Dating Eurasian-like segments carrying Neanderthal ancestry in Africans.** Coancestry curves were generated by considering only Eurasian-like segments that carry Neanderthal ancestry (with posterior probability ≥ 0.1) at at least one position within the segment. Eurasian-like segments that did not carry any Neanderthal ancestry were treated as African (non-Eurasian-like) ancestry. The figure displays the log-likelihoods corresponding to various admixture models, including one-date, two-date, and constant continuous migration models. Single admixture model fits and two-date admixture model fits are shown, with populations sorted from highest to lowest proportion of Eurasian-like ancestry. The log-likelihood values are aggregated across 20 jackknife runs. For the two-date admixture model, pink represents the younger admixture date, and purple represents the older admixture date. Admixture dates in the plot are in generations. 238

B.13 **Comparison of observed Neanderthal proportion with theoretical expectations under the assumption of no Neanderthal ancestry in older Eurasian admixture.** The posterior probability of being Neanderthal is measured for Eurasian-like segments in Africa with segment sizes less than or equal to the specified value (x-axis). Observed values are compared to theoretical Neanderthal proportions assuming two admixture events, where only the younger event contributes to Neanderthal ancestry. Shaded regions represent 95% confidence intervals for the observed Neanderthal proportion and theoretical estimates, accounting for variability in admixture dating. The x-axis is limited to 5 cM. 239

B.14 **Evolution of number of weak to strong mutations relative to strong to weak mutations across populations.** Ratio of number of weak to strong mutations divided by strong to weak mutations across various continental groups, including NFE (non-Finnish Europeans), SAS (South Asians), EAS (East Asians), and MID (Middle Eastern and North African populations), AFR (sub-Saharan Africans). CpG sites are excluded from calling both weak and strong mutations. Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. 240

B.15 **Enrichment of TCC to TTC mutation rates across African ancestral components in diverse African populations.** Enrichment of TCC to TTC mutation rates in younger and older Eurasian-like segments compared to non-Eurasian-like (African) segments across 11 modern African populations (see population descriptions in 3.1a). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. 241

B.16 **Enrichment of weak to strong mutation rates across African ancestral components in diverse African populations.** Enrichment of weak to strong mutation rates in younger and older Eurasian-like segments compared to non-Eurasian-like (African) segments across 11 modern African populations (see population descriptions in 3.1a). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. 242

B.17 **Mutation count enrichment across all 96 trimer types in Eurasian-like segments within African individuals from the HGDP+1,000GP dataset.** (a) Younger Eurasian-like segments and (b) older Eurasian-like segments are shown. The y-axis represents the enrichment, calculated relative to the mutation count in non-Eurasian-like segments of African individuals. Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. 243

B.18 **Coalescence rates between Eurasian-like and non-Eurasian-like segments in Africans.** Pairwise coalescence rate matrices and corresponding UPGMA dendrograms for (a-b) Younger Eurasian-like segments and (c-d) Older Eurasian-like segments. The coalescence rates correspond to the epoch spanning 10,000 to 15,848 years. . . . 244

B.19 **Coalescence rates with modern and ancient samples.** Each bar plot shows the coalescence rate for the epoch spanning 10,000–15,848 years between Eurasian-like segments from the younger admixture date and various modern and ancient samples. Only the top 25 groups are displayed. Error bars represent 95% confidence intervals, calculated from 100 bootstrap samples. 245

B.20 **Held-out likelihood and expected R^2 for the deep admixture event across African populations.** The held-out likelihood was calculated by fitting the EM algorithm on chromosome 1 and testing on chromosome 2. The number of components varied between 1 and 4 or 5. 246

B.21 **Dating the ancestral components of the deep admixture event.** This analysis is similar to Figure 3.8, but with local ancestry segments identified as Eurasian-like in the back-to-Africa analysis masked. Additionally, regions within 5 cM of these segments were also excluded. The plot displays the log-likelihoods for one-date, two-date, and continuous admixture models across populations, as well as the inferred admixture date for the single-date admixture model. The log-likelihood values are aggregated across 20 jackknife runs. The legend remains identical to that in Figure 3.8. Admixture dates in the plot are in generations. 247

B.22 **Comparison of GhostBuster runs with and without the HMM.** Chromosome 1 of San individuals from the HGDP+1,000GP dataset was decomposed, fixing the number of components to 2. The HMM was turned off, treating each site independently for local ancestry calculation. (a) shows the inverse coalescence rates when the HMM is turned off, and (b) compares the correlation in local ancestry inferred with and without the HMM. Plot (b) is based on binning the posterior probabilities into 500 equally spaced bins and calculating the average posterior within each bin. 248

B.23 **Comparison of GhostBuster runs with and without Eurasian reference samples.** Chromosome 1 of San individuals from the HGDP+1,000GP dataset was analyzed, fixing the number of components to 2. The Eurasian samples were excluded from the reference panel prior to ancestry decomposition. (a) shows the inverse coalescence rates when only African populations were used as the reference, and (b) compares the correlation in local ancestry inferred with and without Eurasian samples in the reference panel. Plot (b) is based on binning the posterior probabilities into 500 equally spaced bins and calculating the average posterior within each bin. 248

B.24 **Mutation count enrichment across all 96 trimer types in OOA-like segments within African individuals from the HGDP+1,000GP dataset.** The y-axis represents the enrichment, calculated relative to the mutation count in non-OOA-like segments of African individuals. Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. 249

B.25 **Relate-inferred tree at the PRDM9 gene.** Relate-inferred tree at chromosome 5:23,532,534 (hg38 build), highlighting the PRDM9-C allele (rs6889665) as a star on the lineage along with the descendant population composition. The genealogy was subset to include Africans (OOA-like and Non-OOA-like), East Asians (eas), South Asians (sas), and non-Finnish Europeans (nfe). 249

B.26 **Manhattan plots for some blood traits.** (a) Apolipoprotein B, (b) Gamma Glutamyl Transferase (GGT), (c) Total Bilirubin. The GWAS summary statistics was calculated using Quickdraws (more details in chapter 5). The green line indicates $P = 1 \times 10^{-20}$. . . . 250

B.27 **Summary of calibration in simulations with varying levels of population structure.** False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the fraction of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, EUR, PAS) is provided in section 5.3.1. 251

B.28 **Summary of calibration in simulations with varying levels of relatedness.** False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the number of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. GB-unrel refers to simulations including only unrelated British individuals, GB-rel-ukb refers to randomly sampling from the related white British subset, GB-rel refers to the default relatedness setting with 3.4× more relative pairs compared to the UK Biobank, and GB-rel+ refers to the extreme relatedness case of 7.3× and 4.8× more first and second degree relative pairs compared to the UK Biobank. The prevalence of binary traits was fixed at 10%. 252

B.29 **Summary of calibration in simulations with varying causal effect distributions.** The first two rows correspond to causal effects simulated from a mixture of two Gaussian. Polygenicity refers to the proportion corresponding to the Gaussian with higher variance, and f refers to the fraction of total variance explained by the Gaussian with lower variance. False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the fraction of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile (IQR) range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. 253

B.30 **Summary of statistical power in simulations for quantitative traits.** We measure the normalized causal χ^2 at top 1,000 and top 10,000 causal variants. To correct for confounding, the causal χ^2 is normalized by the average χ^2 at null variants on even chromosomes. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, GB-rel+, EUR) is provided in section 5.3.1. 254

B.31 **Summary of statistical power in simulations for binary traits.** We measure the normalized causal χ^2 at top 1,000 and top 10,000 causal variants. To correct for confounding, the causal χ^2 is normalized by the average χ^2 at null variants on even chromosomes. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, EUR) is provided in the section 5.3.1. 255

B.32 **Summary of statistical power in simulations with varying causal effect distributions.** The first two rows correspond to causal effects simulated from a mixture of two Gaussian. Polygenicity refers the proportion corresponding to the Gaussian with higher variance, and f refers to the fraction of total variance explained by the Gaussian with lower variance. We measure the normalized causal χ^2 at top 1,000 and top 10,000 causal variants. To correct for confounding, the causal χ^2 is normalized by the average χ^2 at null variants on even chromosomes. The line inside each box indicates the median value, the central box indicates the interquartile range (IQR), whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. 256

B.33 **Manhattan plots for two binary traits.** (a) Asthma (prevalence = 0.119) and (b) Basal cell carcinoma (prevalence = 0.012). ∼13.3 million variants were tested. Red lines indicate a genome-wide significance threshold of $P = 5 \times 10^{-8}$ and the green line indicates $P = 1 \times 10^{-20}$. 257

B.34 **Manhattan plots for two binary traits.** (a) Celiac disease (prevalence = 0.0047) and (b) Vitiligo (prevalence = 0.0005). ∼13.3 million variants were tested. Red lines indicate a genome-wide significance threshold of $P = 5 \times 10^{-8}$ and the green line indicates $P = 1 \times 10^{-20}$. 258

B.35 **Manhattan plots for two quantitative traits.** (a) Eosinophil count (RHE-me $h^2 = 0.213$) and (b) Haemoglobin concentration (RHE-mc $h^2 = 0.158$). ∼13.3 million variants were tested. Red lines indicate a genome-wide significance threshold of $P = 5 \times 10^{-8}$ and the green line indicates $P = 1 \times 10^{-20}$. 259

B.36 **Manhattan plots for two quantitative traits.** (a) Lymphocyte percentage (RHE-mc $h^2 = 0.196$) and (b) Mean platelet volume (RHE-mc $h^2 = 0.448$). ∼13.3 million variants were tested. Red lines indicate a genome-wide significance threshold of $P = 5 \times 10^{-8}$ and the green line indicates $P = 1 \times 10^{-20}$. 260

List of Tables

1.1 **Comparison of methods to perform GWAS.** We compare five most popular GWAS softwares namely, Plink [112], BOLT-LMM[106, 107], SAIGE[119], REGENIE[123] and FastGWA[108, 122] comparing their overall scalability, ability to model non-infinitesimal genetic architecture for higher statistical power, multi-trait vectorization for better scalability, support for binary traits and ability to control FPR in presence of population structure and relatedness. 39

3.1 **Sub-Saharan African populations analyzed.** (a) Modern samples and, (b) Ancient samples 112

4.1 **Genetic relationship estimates based on KING values.** . . . 180

5.1 **Summary of calibration in simulations involving (a)** $N = 405$**k white British individuals and (b)** $N = 460$**k European individuals [3].** We report the false positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the fraction of variants on even chromosomes with p-value lower than α. We varied the polygenicity from 1% to 10% in quantitative traits. For binary traits, we changed the prevalence from 0.3 to 0.001, fixing the polygenicity to 2%. Mean FPR refers to the mean across 50 independent simulated traits, SE refers to standard errors, and p-values are computed for the significance of a deviation from the FPR threshold. 198

5.2 **Computational efficiency of Quickdraws for (a) quantitative and (b) binary trait association.** We compare the computational requirements for recent GWAS algorithms with Quickdraws to generate summary statistics for 13.3 million variants and 50 phenotypes with either $N = 50,000$ or $N = 405,088$ and $458,464$ genotyped markers for model fitting ($89,177$ genotyped markers for SAIGE, see Methods); *Total memory includes CPU RAM memory and GPU memory, only Quickdraws requires GPU memory which is reported separately in brackets; **A more detailed cost analysis can be found in table B.12 and B.13 Running times are computed using the same hardware for all methods (mem2_ssd1_v2_x8, with 32 GB of RAM, 8-core processor for $N = 50k$ and mem1_ssd1_v2_x36, 72 GB of RAM, 36-core processor for $N = 405k$ datasets). 218

B.1 **Ancient samples built in the genealogy.** Ancient sample IDs, sampling age (in years), country of origin and genomic coverage . . 261

B.1 *(Continued)* 262

B.2 **Admixture date estimates for the back to Africa admixture event in Africans.** Estimated dates for the younger admixture event based on two-date co-ancestry curve fitting, measured in generations. The lower and upper bounds represent the 95% confidence intervals derived from 20 jackknife replicates during the co-ancestry curve fitting process. 262

B.3 **Admixture date estimates for the deep admixture event in Africans.** Estimated dates for the deep admixture event based on two-date co-ancestry curve fitting, measured in generations. The lower and upper bounds represent the 95% confidence intervals derived from 20 jackknife replicates during the co-ancestry curve fitting process. 263

B.4 **False positive rates (FPR) for FastGWA in quantitative trait simulations with up to 460,000 self-identified Europeans.** The table presents the FPR for FastGWA under various sample sizes, controlling for residual population stratification by adjusting the number of included principal components (PCs) or by omitting them entirely. The polygenicity of the simulated trait is fixed at 1%. S.E. refers to the standard error, and * indicates a Bonferroni-significant t-test p-value. The significance threshold for the FPR was set at 0.005. Each simulation condition varied one of the following baseline parameters: sample size (N = 460k), ancestry differences explaining 5% of phenotypic variance, and the use of 10 PCs to correct for stratification. 263

B.5 **Quantitative traits analyzed.** Quantitative traits analyzed, including RHE-MC h^2 estimates and sample sizes (N). 264
B.5 *(Continued)* 265
B.6 **Self-reported disease traits analyzed.** Self-reported disease traits analyzed, including their disease code and prevalence. 266
B.6 *(Continued)* 267
B.7 **Number of independent associated loci for quantitative traits.** Number of independent associated loci for quantitative traits after Plink clumping using summary statistics from FastGWA, Regenie, BOLT-LMM, and Quickdraws 267
B.7 *(Continued)* 268
B.8 **Number of independent associated loci for binary traits.** Number of independent associated loci for binary traits after Plink clumping using summary statistics from FastGWA, SAIGE, Regenie, and Quickdraws. We excluded traits (enlarged prostate, prostate cancer, vaginal prolapse, and uterine fibroids) which led to matrix inversion errors or non-convergence of Firth logistic regression, or for which no associations were found using any method. 269
B.9 **Number of individual variants replicated using trait-specific summary statistics.** Number of individual variants replicated using summary statistics for Crohn's disease [263], Type 2 Diabetes [264], Celiac disease [265], Depression [266], and Ulcerative Colitis [267]. We use a discovery threshold of 5×10^{-8} and vary the replication threshold from 5×10^{-2} to 5×10^{-6}, and compare summary statistics from FastGWA, Regenie, and Quickdraws. 270
B.10 **Number of loci replicated in Biobank Japan.** Number of loci replicated in Biobank Japan using summary statistics from FastGWA, Regenie, and Quickdraws. We exclude traits that did not yield any associations using any method and which do not have a matching phenotype in Biobank Japan. 271
B.11 **Number of independent associated loci for quantitative traits after applying the correction for participation bias.** Number of independent associated loci for quantitative traits after PLINK clumping using summary statistics from weighted GWAS in Quickdraws ($N \approx 338,738$) compared to weighted GWAS on unrelated samples in LDAK ($N \approx 286,432$). Note that the adjustment for participation bias was only applied to step 2 of the Quickdraws algorithm. 272
B.11 *(Continued)* 273

B.12 **Computational efficiency of Quickdraws for quantitative trait association.** We compare the computational requirements of several GWAS algorithms and Quickdraws across 3 to 4 RAP cloud instances for ∼13.3 million tested variants and 50 phenotypes with either $N = 50{,}000$ or $N = 405{,}088$ samples and 458,464 genotyped markers for model fitting. * The step 1 for Quickdraws is run using `mem3_ssd1_gpu_x8`; ** extrapolated value; "-" indicates that BOLT-LMM was not run in this setting. 274

B.13 **Computational efficiency of Quickdraws for binary trait association.** We compare the computational requirements of several GWAS algorithms and Quickdraws across 3 to 4 RAP cloud instances for ∼13.3 million tested variants and 50 phenotypes with either $N = 50{,}000$ or $N = 405{,}088$ samples and 458,464 genotyped markers for model fitting. * The step 1 for Quickdraws is run using `mem3_ssd1_gpu_x8`. All methods except FastGWA were run on 1 million test variants and extrapolated. 275

B.14 **Computational efficiency of Quickdraws for a single quantitative or binary trait.** We compare the computational requirements of Regenie and Quickdraws, which support the parallal analysis of multiple traits, with those of FastGWA, BOLT-LMM, and SAIGE. All methods are used to compute summary statistics for ∼13.3 million variants and one phenotype with $N = 405{,}088$ and 458,464 genotyped markers for model fitting ($89,177$ genotyped markers for SAIGE, see Methods). Running times and costs are computed using the same hardware for all methods (mem1_ssd1_v2_x36, 72 GB of RAM, 36-core processor). 275

B.15 **Computational efficiency of Quickdraws for $N = 1{,}000{,}000$ samples.** We compare the computational requirements of Quickdraws and Regenie when computing summary statistics for 458k, 13.3 million and 600 million variants, 50 quantitative, and 50 binary traits, using $N = 1{,}000{,}000$ and 458,464 genotyped markers for model fitting. The data set was generated by duplicating samples from the $N = 405k$ subset of white British individuals. * indicates that the cost of imputed variant association testing is extrapolated from the testing time for $458,464$ genotyped markers. Quickdraws was run using the low-memory option. Running times and costs were computed using the same hardware for both methods (mem1_ssd1_v2_x36, 72 GB of RAM, 36-core processor). 276

List of Abbreviations

AF . . . . . Allele frequency

ARG . . . . . Ancestral recombination graph

GRM . . . . . Genomic relationship matrix

GWAS . . . . . Genome-wide association study

HLA . . . . . Human leukocyte antigen

HMM . . . . . Hidden Markov model

HRC+UK10k Haplotype reference panel consortium + UK10k

IBD . . . . . Identity by descent

ICR . . . . . Inverse coalescence rates

LMM . . . . . Linear mixed model

LD . . . . . Linkage disequilibrium

LOCO . . . . . Leave-one-chromosome-out

MAF . . . . . Minor allele frequency

PCA . . . . . Principal component analysis

PCs . . . . . Principal components

PGS . . . . . Polygenic score

RHE . . . . . Randomized Haseman-Elston

SNP . . . . . Single nucleotide polymorphism

SGDP . . . . . Simons genome diversity project

UKB . . . . . UK biobank

WGS . . . . . Whole genome sequencing

1,000**GP** . . . . 1,000 genomes project

1 Introduction

Contents

1.1 Setting the scene 1
1.2 Understanding population history through genetic data 3
1.2.1 Basic definitions 3
1.2.2 Theoretical framework to understand population history from genetic data 6
1.2.3 Survey of methods to infer population structure and admixture history 11
1.2.4 Large-scale inference of genome-wide genealogies 16
1.2.5 Genome-wide genealogies to understand evolutionary history 19
1.2.6 Motivation for GhostBuster 22
1.3 Understanding complex traits through genetic data . . 26
1.3.1 Basic definitions 26
1.3.2 Genome-wide association studies: usage and applications 28
1.3.3 Theoretical framework to perform GWAS 29
1.3.4 Survey of methods to perform GWAS 34
1.3.5 Motivation for Quickdraws 36

1.1 Setting the scene

The costs of DNA sequencing have dropped significantly over the last couple of decades, from $2.7 billion to sequence one human genome in the Human Genome Project [1] to around $600 [2]. This dramatic reduction in cost has facilitated

the generation of vast amounts of genetic data, enabling researchers to draw numerous inferences about the complexity of human genetics. With this influx of data, the field of statistical genetics has emerged as a critical area of study, focusing on the development of sophisticated statistical methods to manage and interpret this big data. Statistical genetics can be broadly divided into two main directions. First, quantitative genetics aims to understand how genetic variation affects observable phenotypes, providing valuable insights into the relationship between genes and traits. On the other hand, population genetics uses genomic data to look backward in time and aims to reconstruct the evolutionary forces that have shaped present-day genetic diversity.

In this thesis, we harness the wealth of genomic data now available from comprehensive datasets such as the UK Biobank [3], the Human Genome Diversity Project [4], and the 1000 Genomes Project [5] to investigate both key areas of statistical genetics. First, in the realm of population genetics, we introduce a method called GhostBuster, designed to identify mixtures between ancestral human groups, including those involving unsampled populations, shedding light on the evolutionary forces shaping human diversity. Second, in the field of quantitative genetics, we present Quickdraws, a tool that utilizes present-day genetic variation to conduct genome-wide association studies, enabling us to understand how these variations contribute to complex traits.

In this introductory chapter, we discuss the motivation, research aims, and literature review for analyses in Chapters 2 to 5. We structure the rest of the chapter to discuss the population genetics and quantitative genetics aspects separately. In Chapter 2, we introduce a novel method called GhostBuster, which relies on recent advances in building genome-wide genealogies at scale [6–8]. In Chapter 3, we apply this method to understand population structure in Africa. In Chapter 4, we shift to the quantitative genetics angle and introduce Quickdraws, a scalable variational inference method for performing genome-wide associations. In Chapter 5, we assess Quickdraws' efficiency, power, and robustness in simulations and real-data

applications. We discuss the key takeaways, limitations and future directions for both GhostBuster and Quickdraws separately in Chapter 3 and 5 respectively.

1.2 Understanding population history through genetic data

The study of how evolutionary forces impact genetics, and therefore phenotypes, has been conducted for more than a century. Even with limited data, mathematicians have theorized frameworks to understand the relationship between evolutionary forces and present-day genetics. With the availability of genetic data, it has now become possible to construct a quantitative understanding of how evolutionary forces such as mutations, recombination, selection, and demographic changes like population structure and admixture have impacted our present-day genetic makeup. In this thesis, we focus on two such important forces that have shaped the genetics of present-day individuals: admixture and population structure. In this section, we will discuss various established theoretical foundations to understand population history through genetic data, particularly genome-wide genealogies, and their application to recently available genomic data.

1.2.1 Basic definitions

Deoxyribonucleic acid (DNA) encodes the hereditary information of an individual. It consists of four nucleotide bases: adenine (A), cytosine (C), guanine (G), and thymine (T). Adenine pairs with thymine, and cytosine pairs with guanine, forming base pairs. Each nucleotide also includes a sugar molecule and a phosphate group. DNA sequences are arranged in a double helix structure, which is tightly coiled around proteins called histones to form chromosomes. Humans are diploid organisms, meaning they have two copies of each chromosome, one inherited from each parent. The human genome is compacted into 23 pairs of chromosomes, of which 22 pairs are autosomal chromosomes shared by both sexes, while one pair is sex-specific. DNA varies between individuals due to evolutionary processes including mutations and recombination.

Mutations are random changes in the DNA sequence that can happen during DNA replication or due to external factors. While mutations can occur in any cell, those occurring in germ cells are passed down to offspring and are called germline or de novo mutations. Humans have an estimated average mutation rate of 1.2×10^{-8} per nucleotide per generation [9], with one generation approximately corresponding to 28 years in humans [10, 11]. Males contribute more de novo mutations than females [12], and, on average, humans inherit 70 de novo mutations compared to their parents. These mutations can be of various types, including single nucleotide variants (SNVs), which are the most common and involve the alteration of a single nucleotide base. Insertions and deletions (indels) involve the addition or removal of one or more nucleotide bases, copy number variants (CNVs) result in the duplication or deletion of large sections of DNA, and structural variants encompass even larger alterations to DNA.

Recombination is another key mechanism that ensures genetic diversity in several organisms. In humans, recombination mainly occurs during meiosis, a type of cell division responsible for the production of germ cells. During meiosis, the chromosomes from each parent align and cross over, creating a mosaic of both parental chromosomes. Consequently, offspring inherit a proportion of their genetic material from each parent's maternal chromosomes and a portion from that parent's paternal chromosomes. In most mammals, including humans, the locations of these crossover events are concentrated in specific regions of the chromosomes called recombination hotspots. It has been shown that the location of these hotspots is determined by where a crucial gene called PRDM9 [13] binds to DNA.

Apart from mutations and recombination, which are almost stochastic in nature, various demographic factors like population structure, migration and admixture also influence the genetic makeup of present-day individuals. Population structure is defined as a systematic difference in allele frequencies between subpopulations. This occurs due to non-random mating within a species, either due to geography or other factors. For example, the geographic distance between continents has caused human populations within a certain continent to mix more frequently among

themselves than with populations from other continents, resulting in systematic genetic differences between individuals from different continents [14]. It has also been shown recently that population structure within humans exists not only on a continental scale but also within smaller geographical regions, termed fine-scale population structure [15–17].

Another key demographic factor influencing present-day genetics is migration, which refers to the systematic movement of individuals and their genetic material between populations. With the availability of genetic data and archaeological records we have now established impacts of mass migrations in human history. For example it is now well established there was a huge wave of migration of anatomically modern humans from Africa to Eurasia around 50,000-100,000 years ago which gave rise to present-day non African populations [18]. More recently, there have been migrations from East Asia to America [19], spread of Anatolian or farming ancestry in Europe [20] and a rapid migration from the steppe into Europe and south Asia [21].

Migrations involve geographic movement of individuals over an extended duration of time, but two populations can mix more abruptly or over a shorter time duration. This is referred to as an admixture event. Admixture affects genetics by producing offspring that are a mixture of two or more distinct subpopulations. Some examples of previously documented admixtures in humans include present-day African American individuals in America, whose genomes are admixed from African, Native American, and European sources [22, 23]. Introgression is similar to admixture except it involves small amount of gene-flow from one population to another usually involving populations or species which are very diverged. For example, the Neanderthal introgression into humans, which has resulted in 1-3% of modern day non-African DNA coming from Neanderthals [24]. Similarly, Denisovan introgression into South Asia and Oceania, contributes up to 5% of Denisovan ancestry in Papuans [25].

When two subpopulations admix, we often do not know these source groups. It might also be possible that, due to extensive admixture, we no longer have source groups that are 100% comprised of one subpopulation or another. We define a “ghost”

population as a population (or a group of related individuals) that contributed genetic material to present-day individuals but have not yet been sequenced or identified in modern genetic studies. These populations may no longer exist or remain unsampled, making their detection reliant on indirect inference from present-day genetic variation. By analyzing patterns of population structure and admixture, we can infer their historical presence even in the absence of direct genomic data.

1.2.2 Theoretical framework to understand population history from genetic data

The Wright-Fisher model [26, 27] is one of the most popular population genetics models, which formalizes how populations trace their ancestry backward in time. The simplest version of the Wright-Fisher model simulates the genetic history of samples as follows: (i) Starting with N haploid individuals in the population, each individual, going backward in time, chooses its immediate ancestor uniformly and independently at random from a set of N ancestors with replacement. (ii) Generations are discrete and do not overlap, so (i) can be repeated backward in time until there is only one common ancestor of the population or a sample. Furthermore, the Wright-Fisher model assumes tracing a single-site, equality of fitness and lack of any population structure. A simple depiction of how the common ancestor is chosen backward in time for three lineages is shown in Figure 1.1.

It is possible to derive the probability distribution of the time to the most recent common ancestor (TMRCA) under the Wright-Fisher model assuming a constant effective population size N_e. The TMRCA for a pair of samples is geometrically distributed and takes the form:

$$P(\tau_2 = t) = \left(1 - \frac{1}{N_e}\right)^{t-1} \frac{1}{N_e} \tag{1.1}$$

where τ_2 is the time to most recent common ancestor, and N_e is the effective population size. The cumulative distribution function (cdf) of τ_2 looks like:

$$P(\tau_2 \leq t) = 1 - \left(1 - \frac{1}{N_e}\right)^{t} \tag{1.2}$$

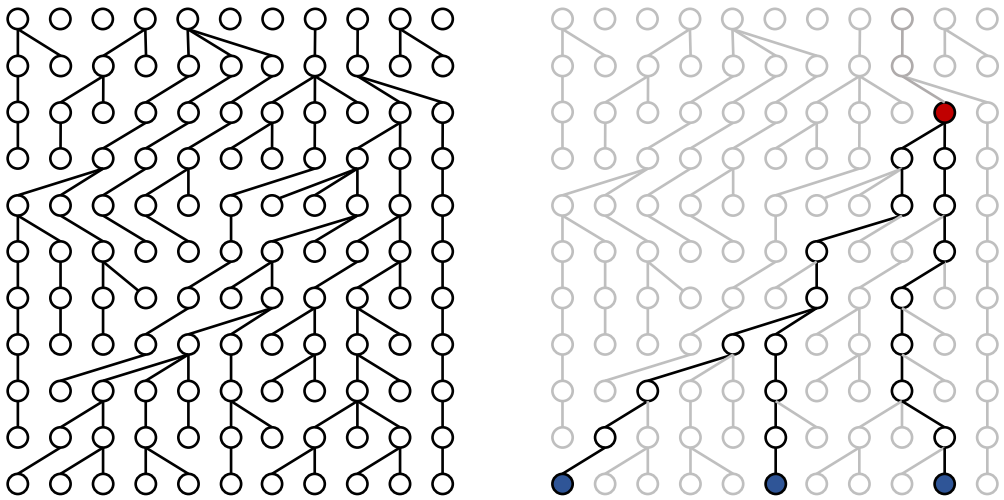


Figure 1.1: Example of three lineages coalescing backwards in time in the Wright Fisher model. Each individual chooses its ancestors from the set of ancestors with replacement until all the lineages coalesce into a single common ancestor. The three samples are colored blue and the common ancestor for the three samples red

Assuming a large and constant population size, $N_e \to \infty$ and setting $T_2 = \frac{\tau_2}{N_e}$ we can approximate the cdf as follows:

$$\begin{aligned} P(T_2 \leq t) &= P(\tau_2 \leq N_e t) = P(\tau_2 \leq \lfloor N_e t \rfloor) \\ &= 1 - \left(1 - \frac{1}{N_e}\right)^{\lfloor N_e t \rfloor} \\ &\to 1 - e^{-t} \text{ as } N_e \to \infty \end{aligned} \tag{1.3}$$

Note that the TMRCA takes the form of a standard exponential distribution. This is the coalescent model, introduced by three scientists independently in 1982-1983: Kingman, Hudson and Tajima [28–32]. As we will see the coalescent is a simplification of the Wright-Fisher model [26, 27] for large populations which generalizes to various features in human genetics like mutations, recombinations and population structure.

It is possible to generalize the coalescent to n lineages instead of a pair. Let us define $\xi(\tau)$ as the number of lineages remaining at time τ (measured backwards, in generations). Going backward in time with the initialization $\xi(0) = n$, $\xi(\tau)$ behaves as a Markov process. That is the distribution of $\xi(\tau + 1)$ only depends

on the previous one, $\xi(\tau)$. We can define the transition probability of going from i lineages to j lineages in one generation as:

$$p_{i,j} = P(\xi(\tau+1) = j \mid \xi(\tau) = i) \tag{1.4}$$

It can be shown under the standard coalescent model, setting $N_e \to \infty$ and ignoring second order terms, the transition matrix simplifies as follows:

$$\begin{aligned} p_{i,i} &\approx 1 - \frac{\binom{i}{2}}{N_e} \\ p_{i,i-1} &\approx \frac{\binom{i}{2}}{N_e} \\ p_{i,j} &\approx 0 \text{ for any } j < i-1 \end{aligned} \tag{1.5}$$

This implies the coalescent for n samples only allows up to one pair of lineages to coalesce per generation, making a binary "coalescent" tree. Additionally, the rate of coalescence depends on the number of lineages present at a given time. Let $T_j = \frac{\tau_j}{N_e}$ be the times to go from j to $j-1$ ancestors (scaled by the population size N_e) then according to the coalescent T_j is independent of T_i $(i \neq j)$ and is distributed as follows:

$$T_j \sim exp\left(\binom{j}{2}\right),\ f_j(t) = \binom{j}{2} e^{-\binom{j}{2}t} \tag{1.6}$$

It is possible to generalize the above model for time-varying population sizes. Let us define the effective population size at time t as $N_e(t)$. We can write the effective population size as a product of population size at $t = 0$ and relative population size $\nu(t)$. Specifically, $N_e(t) = N_e(0)\nu(t)$. Scaling the coalescence times based on population size at $t = 0$, $N_e(0)$, the probability distribution when j ancestors remain is given as follows:

$$P(T_j > t \mid T_n + T_{n-1} + ... + T_{j+1} = s) = exp\left(-\binom{j}{2}\int_s^{s+t} \frac{1}{\nu(u)} du\right) \tag{1.7}$$

The above equation simplifies to the standard coalescent with constant population size if we set $\nu(t) = 1$.

Coalescent with mutations: To simulate mutations on the coalescent, we make an additional assumption called the infinite sites assumption. This assumption

states that whenever a mutation occurs, it will always result in a new and unique mutated site, thereby disallowing any site from having more than one mutation. In humans, where the mutation rate is on the order of 10^{-8} per base per generation, it is realistic to assume that a mutation does not hit the same site more than once. We can model mutations as a homogeneous Poisson process on the coalescent. This means the total number of mutation events will follow a Poisson distribution with mean μtL, where μ is the mutation rate per base pair per generation, t is the time to the most recent common ancestor in generations, and L is the length of the genomic span being considered in base pairs. This is the generative model. During inference, where we already have information about mutations, we aim to reconstruct population characteristics backward in time.

Coalescent with recombination: In addition to modeling mutation, we also need to account for recombination to fully infer human population characteristics. We can modify the coalescent model to account for recombination as follows: Each pair of lineages can not only coalesce but also undergo recombination events, which split the ancestry of a single lineage into two parts. Going backward in time, at each generation, each segment of DNA can trace its ancestry independently, potentially to different ancestors. This process creates a branching structure where different segments of an individual's genome may trace back to different ancestors, creating a more complex ancestral structure than a simple coalescent tree. This model is called the coalescent with recombination [33]. This graph structure was first introduced in [34], and is called the ancestral recombination graph (ARG). An illustrative example of an ARG is depicted in Figure 1.2. ARGs induce a sequence of trees along the genome (see Figure 1.2b). Genome-wide inference of ARGs is also loosely referred to as inferring genome-wide genealogies. In this thesis, we use the terms genealogies, tree sequences, and ARGs interchangeably, but a more descriptive difference can be found in [35].

The generality and simplicity of the coalescent limiting model enables extending it to account for many factors, including non-uniform recombination rates, time-varying population sizes and multiple samples. There is a suite of methods to

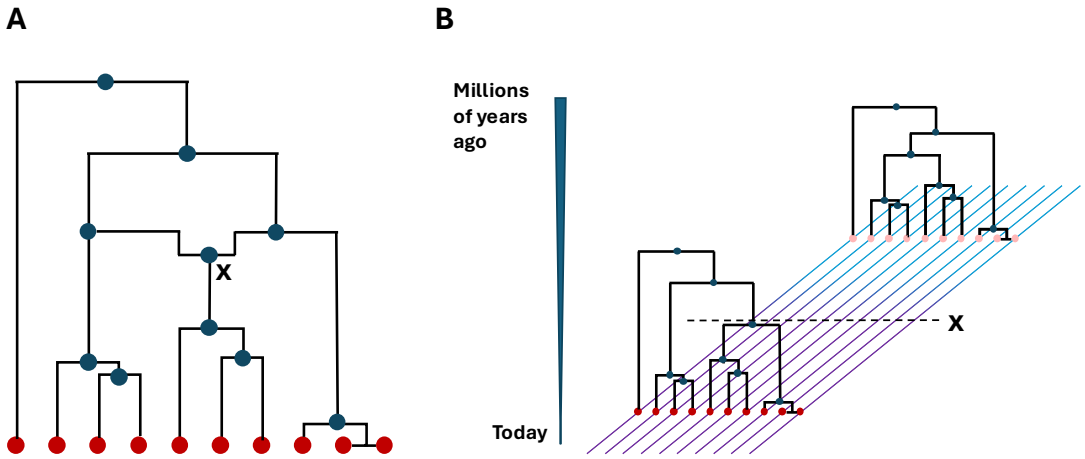


Figure 1.2: Example of an ancestral recombination graph (ARG) and the corresponding sequence of trees. A. In the coalescent with recombination model the Lineages going backward in time can either coalesce or recombine, which means different portions of the genome have slightly different history. B. Corresponding sequence of trees representation of the ARG, **X** denotes the recombination position.

infer these ARGs which rely on the coalescent theory. We will introduce some of these methods in Section 1.2.4.

Estimation of coalescence rates for a pair of lineages: We can also perform inference under the coalescent. Let us simplify the coalescent for a pair of haplotypes and infer time-varying coalescence rates assuming piece-wise constant coalescence rates. We assume step-wise constant coalescence rates defined by fixed time intervals or epochs, labeled as I_1, I_2, ... I_m. Let us say the coalescence rate in each epoch $e \in 1, \ldots, m$ is $\gamma(e)$ and is the parameter to be inferred. The epoch index when the coalescence happens in the l-th tree is given by e_l and the coalescence time is denoted by t_l ($I_{e_l} \leq t_l \leq I_{e_l+1}$). Extending the coalescent from Equation 1.7 for piece-wise constant coalescence rates, we can write the likelihood of the coalescence time (for a pair of lineages) as:

$$P(t_l) = \gamma(e_l) e^{-\gamma(e_l)(t_l - I_{e_l})} \prod_{e=1}^{e_l} e^{-\gamma(e-1)(I_e - I_{e-1})} \tag{1.8}$$

For maximum likelihood estimation, we assume the **L** trees (or data points) are

independent, so we sum the log-likelihood across trees and estimate the coalescence rates by maximizing the following:

$$\log P(t_l) = \log \gamma(e_l) - \gamma(e_l)(t_l - I_{e_l}) - \sum_{e=1}^{e_l} \gamma(e-1)(I_e - I_{e-1}),$$
$$\log P(t) = \sum_{l=1}^{\mathbf{L}} \left(\log \gamma(e_l) - \gamma(e_l)(t_l - I_{e_l}) - \sum_{e=1}^{e_l} \gamma(e-1)(I_e - I_{e-1}) \right). \quad (1.9)$$

The MLE estimate for the coalescence rates are given by:

$$\hat{\gamma}(e) = \frac{n_e}{\sum_{l:e=e_l}(t_l - I_e) + \sum_{l:e<e_l}(I_e - I_{e-1})} \quad (1.10)$$

where, n_e is the number of trees which coalesce in the e-th epoch and is referred to as "coalescence count" and $\sum_{l:e=e_l}(t_l - I_e) + \sum_{l:e<e_l}(I_e - I_{e-1})$ is the sum of branch length until the haplotype coalesces and is referred as the "opportunity".

1.2.3 Survey of methods to infer population structure and admixture history

Current methods to infer population structure can be classified into two main categories. The first set of approaches are model-based approaches, where the methods rely on a predefined generative model of how the population structure would impact the genetics and use this model to solve the reverse problem of inferring the population structure given the genetics. Popular model-based methods to identify population structure include Structure [36], Admixture [37], and FineStructure [38]. The other category of methods are model-free approaches, which do not rely on modeling assumptions and are more exploratory in nature. Dimensionality reduction methods like Principal Component Analysis (PCA) are a popular example of such methods. The choice of which method to use depends on the type of genetic data being analyzed, computational resources, data sizes, and modeling assumptions.

Some of the earliest methods to understand population structure relied on quantifying genetic distances between individuals within and across predefined reference populations. One such measure, the "fixation index" also referred to as F_{st} [39, 40], measures the normalized allele frequency difference between two populations and within a population. Under several simplifying assumptions,

including constant population sizes, the fixation index equals $1 - T_0/T$, where T_0 is the average coalescence time for individuals within the population and T is the average coalescence time of two lineages drawn at random from the collection of the two populations. The index is bounded between 0 to 1, where 0 refers to two populations being indistinguishable from each other, and 1 refers to when they are highly diverged. Since F_{st} measures genetic differences between populations due to genetic drift, population size plays a vital role in dictating how much drift is present. Smaller population sizes experience stronger drift, making methods that rely solely on F_{st} sensitive to assumptions about population sizes.

Early work to infer population structure also relied on analyzing variation in the Y-chromosome and mitochondrial DNA (mtDNA), as they do not recombine [41, 42]. This allows for inferring a single genealogical tree that relates various populations, instead of an ARG or an average genealogical tree. In particular, the analysis of population structure using Y-chromosome or mitochondrial DNA involves considering prespecified "haplogroups", which are specific haplotypes thought to have arisen early in human history and are present in differing amounts in various human populations. By analyzing the number of specific haplogroups among populations, one can infer the coalescence times and topology of the genealogical tree explaining this pattern. Additionally, because Y-chromosomes are only inherited from the paternal lineage and mtDNA from the maternal lineage, this allows inference of population history that might be sex-biased. Although its importance in identifying sex-biased population structure persists even today, the number of haplogroups is limited, making it not always possible to discern subtle differences in populations by only looking at the Y-chromosome or mtDNA.

A popular model-based method to infer population structure is "STRUCTURE" [36], which clusters individuals into a fixed number of distinct populations based on their genotypic similarity. STRUCTURE assumes a similar generative model as latent Dirichlet allocation (LDA) in machine learning [43], which is used for topic modeling (though STRUCTURE predates LDA). The method works by partitioning the individuals into K populations, each characterized by allele frequency spectrum.

It explicitly models the probability of observing the genotypes given the data and the assignment of the individual and proceeds with a Bayesian algorithm to infer the assignment of the individual and the corresponding allele frequencies that define the clusters. Several developments have improved the scalability of STRUCTURE, including methods like ADMIXTURE [37] and fastSTRUCTURE [44].

PCA is a model-free method to infer and visualize population structure in the data [45–47]. It proceeds by eigen-decomposing the genotype matrix for all individuals and projecting the samples onto the top K eigenvectors. Due to eigen-decomposition, the top K eigenvectors maximize the amount of variance explained while being orthogonal to each other. It has been demonstrated that the top eigenvectors or principal components of the genotype matrix can capture the demographic history of samples without any specific modeling assumptions about the demography [46, 48]. For example, the first two principal components (PCs) of a widely studied human genetic dataset, the UK Biobank, clearly separate populations from sub-Saharan Africa from the rest of the samples. Furthermore, PC3 and PC4 start separating populations outside Africa. An accompanying plot which includes the principal components for the UK Biobank data is presented in Figure 1.3. This plot is taken from [3].

Both STRUCTURE and PCA work well when the population structure is strong, making differences in allele frequencies or genotypes among individuals identifiable from noise [46]. There are a suite of methods to infer more subtle differences in populations, also referred to as fine-scale population structure. Identity-by-descent (IBD) methods use shared genetic segments between individuals to identify recent genetic relatedness and infer fine-scale population structure [49]. Other Methods like ChromoPainter and FineStructure [38] are model-based and use the Li and Stephens [50] algorithm to model an individual's haplotype as a mosaic of other individuals in a reference panel. Once the chromosome is painted, follow-up methods like FineStructure summarize the painting. The Li and Stephens algorithm is also used extensively to infer genealogies from sequencing data, as we will see in Section 1.2.4.

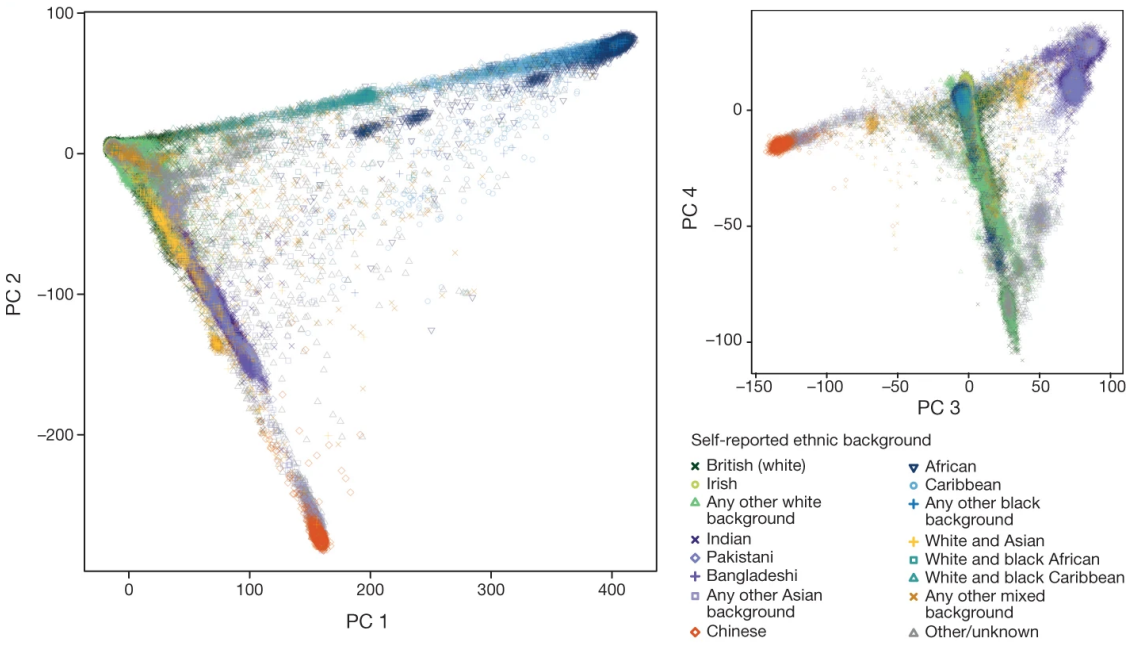


Figure 1.3: PCA visualization of $500,000$ **individual genotypes from the UK Biobank.** PC1 and PC2 separate the African individuals from non-African individuals, whereas PC3 and PC4 separate individuals in Eurasia. The figure was taken from [3] (Creative Commons License).

Methods to identify population structure can also identify signatures of admixture in the data. For example, if there is recent admixture between two populations, the individuals who are admixed would have their STRUCTURE assignment as a mixture of two or more clusters, or would show a "cline" in the PCA, representing a continuum of ancestry between source groups. The American individuals in the UK Biobank dataset are admixed with varying admixture proportions from Native Americans, Europeans, and Africans, and this can be visualized in the PC plots in Figure 1.3.

Other methods to infer admixture can characterize the admixture event by inferring an admixture date. Methods like Globetrotter [51], TRACTS [52], ALDER [53], and Rolloff [54, 55] use haplotype information to infer admixture proportions, sources, and admixture times. The idea behind these methods is based on the observation that if admixture is recent, there will be long segments of DNA from each ancestry because there has been less time for recombination to mix the ancestries thoroughly. In a simple 2-way admixture, the length of the DNA segment from a

particular ancestry is exponentially distributed with the rate proportional to the admixture time; this is called a co-ancestry curve. More formally, the probability of no recombination between two points a distance g (in genetic distance units) apart since admixture is $e^{-g\lambda}$, where λ is the admixture time. For a two-way admixture model with a single pulse of gene flow, the joint probability of ancestries at loci separated by a genetic distance g, conditional on recombination, is given by:

$$\begin{aligned} p_{AA}(g) &= \alpha(e^{-g\lambda} + \alpha(1 - e^{-g\lambda})), \\ p_{AB}(g) &= p_{BA}(g) = \alpha(1-\alpha)(1 - e^{-g\lambda}), \\ p_{BB}(g) &= (1-\alpha)(e^{-g\lambda} + (1-\alpha)(1 - e^{-g\lambda})) \end{aligned} \tag{1.11}$$

where, α is the genome-wide proportion of ancestry 'A' and $1 - \alpha$ of ancestry 'B'. In the case of multi-way admixture (more than two sources), the distribution becomes a mixture of exponential [51]. Therefore, using ChromoPainter to infer the painting of a target individual, methods like Globetrotter can accurately infer the admixture sources and admixture time. It is important to note that inferring admixture becomes challenging for old admixture events or admixture involving similar groups as recombination breaks up the long ancestry segments or ancestry segment inference becomes noisy when groups involved are very similar.

Once we identify the presence of an admixture event, various supervised learning methods can be employed to infer the local ancestry (ancestry along the genome) of a target. RFmix [56] uses support vector machines (SVMs) to classify regions of the sequence that belong to a particular reference population and infers the local ancestry quite accurately for recent admixtures between deeply diverged populations, such as the 3-way admixture in Americans. FLARE [57] is designed to scale efficiently to large sample sizes, while Mosaic [58] corrects for phasing errors during local ancestry inference and does not require prior knowledge of source groups, making it particularly useful for more complex scenarios.

F-statistics are another suite of methods, first proposed by Nick Patterson [19, 21, 55, 59, 60], which are useful for detecting population structure and admixture. These statistics, like the fixation index (F_{st}), rely on allele frequencies in different populations to infer genetic distances between individuals, the presence of admixture,

or population splits in the dataset. There are three main kinds of f-statistics. The f_2 statistic is simply the average squared allele frequency difference between two populations. Similar to F_{st}, it measures the amount of genetic drift (and therefore genetic difference) between two populations. The f_3 statistic is a three-population test measuring the amount of shared genetic drift between two populations from a common ancestor. The f_3 statistic is also used as a test to detect admixture. The f_4 statistic is a four-population test that measures how clade-like the four populations are. The f_2, f_3, and f_4 statistics are defined as follows:

$$\begin{aligned} f_2(A,B) &= \frac{1}{L}\sum_{i=1}^{L}(p_{A,i}-p_{B,i})^2, \\ f_3(A;B,C) &= \frac{1}{L}\sum_{i=1}^{L}(p_{A,i}-p_{B,i})(p_{A,i}-p_{C,i}), \\ f_4(A,B;C,D) &= \frac{1}{L}\sum_{i=1}^{L}(p_{A,i}-p_{B,i})(p_{C,i}-p_{D,i}) \end{aligned} \tag{1.12}$$

where, $p_{X,i}$ is the allele frequency in population X at locus i and L is the number of loci considered. As f-statistics only take allele frequencies into account, they have become a widely used method to understand population structure, especially in ancient DNA studies where there is scarcity of high-coverage sequencing data but an abundance of low-coverage ancient samples. A more detailed explanation of f-statistics and their geometric interpretation can be found in [61].

1.2.4 Large-scale inference of genome-wide genealogies

Many model-based methods to detect population structure and admixture rely on coalescent theory to model the likelihood of the observed genotype given the population sizes or admixtures. It is reasonable to question if we can infer the ARGs directly from the genetic variation data. Assuming the validity of assumptions under the coalescent, inferring the ARG directly would make understanding the evolutionary forces responsible for observed genetic variation much easier, as under the coalescent model the evolutionary forces impact the genetic variation through the ARGs (see Figure 1.4).

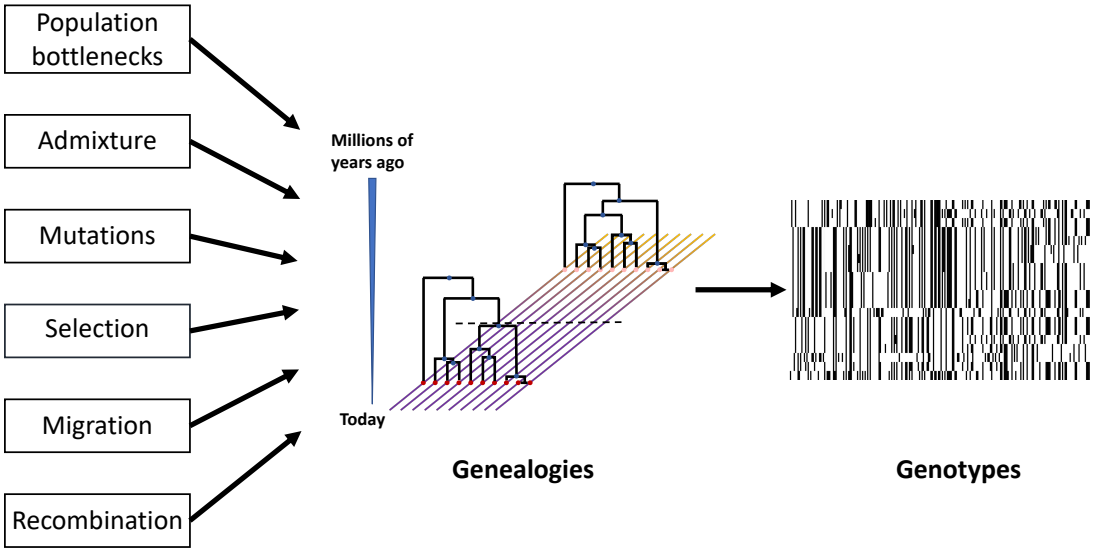


Figure 1.4: Evolutionary forces impact genotypes through genealogies. Evolutionary forces like mutations, recombination, population bottlenecks, admixture, migration and selection affect the observed genotypes through the unobserved genealogies

Inferring Ancestral Recombination Graphs (ARGs) directly from genetic variation data is challenging because each recombination event splits the genealogical tree into two trees. The fact that mutation rates and recombination rates in humans (and many other species) are roughly similar means that, on average, there is only one mutation per branch on the tree, making it difficult to infer the genealogy. However, ARG inference methods leverage the fact that recombinations are unevenly distributed along the genome and still result in correlated adjacent trees. Over the past decade, new statistical methods have been developed to utilize these insights for ARG inference.

ARG-weaver [62] is one of the earliest scalable genealogy inference methods. ARG-weaver uses a Markovian approximation to the coalescent with recombination, the Sequentially Markovian Coalescent (SMC) [63], to constrain the space of possible ARGs. The SMC model reduces the search space of possible ARGs and makes it Markovian by ignoring coalescent events between lineages that do not share overlapping ancestral material. It essentially moves along the chromosome and updates the coalescent tree instead of the full ancestral recombination graph

whenever a recombination occurs. ARG-weaver simplifies the inference of Ancestral Recombination Graphs (ARGs) by discretizing time, which further reduces the space of possible ARGs. It employs a Gibbs sampling approach, where a sample is "threaded[1]" into the ARG based on the conditional probability of the rest of the ARG. Despite these approximations, ARG-weaver is limited in its scalability and can handle at most a few hundred samples.

Tsinfer [7] and tsdate [64] are recent algorithms for ARG inference that are able to scale to hundreds of thousands of samples. Tsinfer uses a variation of the Li and Stephens algorithm to infer the tree sequence. It first estimates various possibilities of ancestral haplotypes at a given site and then uses the Li and Stephens copying process to paint the younger ancestors using older ancestors, thereby generating an ARG structure around the variant. Tsdate uses the topologies inferred using tsinfer to date the coalescence events using a conditional coalescent prior, conditioning on the number of descendants of each node on a local tree. Tsinfer and tsdate also demonstrated that storing genotype information in the tree sequence format saves a significant amount of disk space, leveraging the tree structure to save the variants instead of treating all variants independently.

Relate [6], like tsinfer and tsdate, uses a variation of the Li and Stephens algorithm to infer genealogies and can scale to thousands of samples. It proceeds by using the Li and Stephens algorithm to calculate pairwise distances between samples at a given site, followed by hierarchical clustering to build a tree that best represents the observed variation around this site. Once the local trees are inferred and heuristically matched with adjacent trees, Relate estimates the coalescence times using an MCMC routine under the coalescent prior, conditioning on the tree topology. Relate also supports an iterative method to jointly estimate the population sizes along with estimating the coalescence times, making it more flexible for use in human datasets.

ARG-Needle [8] is another recent algorithm that is able to scale to hundreds of thousands of samples. Like ARGweaver, it threads a sample into the genealogy

[1] essentially integrating a sample into an existing ARG structure

one at a time, but uses heuristics to speed up the process. It essentially proceeds as follows: for a new sample to be threaded, it detects K closely related samples around the variant, uses ASMC [65] to infer the coalescence times with those K possible samples, and then uses the sample corresponding to the minimum coalescence time to thread the new sample into the ARG. This process is repeated until all samples are threaded into the genealogy. Apart from genealogy inference, ARG-Needle also demonstrated the promise of using genealogies for phenotype prediction and association testing.

Singer [66] is a Bayesian method for ARG inference that scales to hundreds of samples. Like ARGweaver and ARG-Needle, Singer uses threading to infer the genealogies in a Bayesian way. Specifically, Singer employs a two-step threading algorithm where the first step samples the possible branch to thread and the second step samples the coalescence time for this threading. This is done under an MCMC routine, which enables the sampling of not only coalescence times but also tree topologies.

Overall, there has been recent development in accurate and scalable inference of genome-wide genealogies from genetic variation data. A recent evaluation of the performance of some of these ARG inference methods under some accuracy metrics can be found in [67]. This opens a new suite of directions to understand the evolutionary forces responsible for our genetic variation using these genealogies. In the next section, we survey some novel methods that rely on genealogies and help us understand our human past more accurately.

1.2.5 Genome-wide genealogies to understand evolutionary history

By inferring genome-wide genealogies, we hope to be one step closer to understanding evolutionary history, as the evolutionary history has impacted genetic variation through these genealogies. Let us explore the various evolutionary forces we can infer once we know the genealogies.

Inferring population sizes using genealogies: Under the coalescent, the population sizes are defined as the reciprocal of the coalescent rates between individuals in the same population. Therefore, given the genealogy, it is straightforward to infer the population sizes. A common approach to infer these time-varying population sizes is to discretize the time into several epochs and count the number of coalescence events between individuals in each epoch, divided by the opportunity for the coalescence event to happen, that is, the sum of branch lengths in the epoch. In a simplistic case, this can be done for all pairs of individuals in the population, which results in an unbiased estimate of the population size.

Inferring population structure using genealogies: Following the argument for inferring population sizes, one can also infer cross-population coalescence rates, that is, the chance of coalescence between two individuals from different populations. These rates are a common way to identify population structure or migration [68], as populations with higher cross-coalescence rates would be more similar to each other either due to a recent population split or recent migration between the two populations.

Inferring presence of admixture using genealogies: A simple way to identify recent admixture in a genealogy is by looking at genealogical nearest neighbors. If there is a recent admixture event, different regions of the genome would have different affinities to reference populations. A straightforward method to infer regions of the genome where a target individual is closer to one population than another is to look at its first coalescence event with the reference panel. A population description of the sister individuals at the tree would roughly indicate which population the target individual most closely relates to in that section of the genome. This is also loosely how admixture inference methods like ChromoPainter work. A more careful analysis is needed if there are multiple admixture events or if the admixture event is old and can not be summarized by just the first coalescence event. Our method, GhostBuster (discussed in Chapters 2-3), addresses this problem.

Inferring positive selection: Relate [69] devised a simple statistic to detect the presence of positive selection in the data. Specifically, the statistic compares

the speed of spread of a certain lineage relative to other lineages in the tree. Relate's selection test uses results from the standard coalescent to model the joint distribution of the number of descendants given k lineages. Relate's selection test produced controlled false positive rates while outperforming other selection inference methods, such as singleton density scores [70], which did not leverage the genealogy to infer selection. A similar test was devised in ASMC [65], where they considered pairwise genealogies and regions of significantly lower TMRCA to detect positive selection. Additionally, CLUES [71] and CLUES2 [72] use inferred genealogical information in an importance sampling scheme to directly estimate the likelihood of alleles under selection.

Among other applications, twigstats [69], used inferred genealogies to more accurately estimate f-statistics. By effectively truncating trees and focusing only on mutations in the recent part of the tree, twigstats improved f-statistics standard errors to detect fine-scale structure. Another approach, introduced glike [73], involves estimating the likelihood of the observed genealogy given the demographic parameters, thereby accurately estimating parameters like population sizes, admixtures, and migrations from the observed genealogies. This method surpasses techniques that only use genetic data for inference. ARG-based methods also have their limitations. Firstly, most methods rely on the inferred ARG to be an accurate estimate of the true underlying genealogy. Second, ARG-based methods are not necessarily computationally faster than methods which directly operate on genotype data. Third, accurate estimation of ARGs currently requires high-quality genetic data, which may not always be available. However, despite these limitations, ARG-based methods provide powerful insights into population history and evolutionary processes that are difficult to achieve with other approaches. A comprehensive review comparing different ARG inference methods and discussing their application in understanding human history can be found in [74, 75].

1.2.6 Motivation for GhostBuster

Population movements and mixtures have been a constant feature of human history across various time scales. Notable examples include the Out of Africa migration, Neanderthal and Denisovan introgression over 40,000 years ago, multiple admixtures in Holocene Europe, and recent migrations and multi-way admixtures in the Americas within the last few hundred years (see Figure 1.5a-c). Despite this, the genetic legacy of some events remains elusive due to the limited resolution of statistical methods and the scarcity of ancient DNA samples (see Figure 1.5d).

Many previous methods for deciphering admixtures have relied on detecting long segments of DNA corresponding to recent admixture events. These methods often use local ancestry correlation to date admixtures, which have proven effective for identifying events within the last few thousand years. However, as recombination events shorten these segments, the statistical power of these methods diminishes, making it challenging to infer older admixtures [51, 53–55]. Another widely used approach involves sequencing or genotyping ancient DNA samples, providing direct evidence of the genetic makeup of our ancestors. Ancient DNA studies, employing methods such as PCA [46], STRUCTURE [36], and f-statistics [19, 21, 55, 59, 60], have significantly extended our understanding of human history up to approximately 10,000 years ago—beyond the temporal reach of haplotype-based methods. However, the deeper human past remains difficult to uncover due to the scarcity of ancient DNA samples from earlier periods, which are challenging to preserve over such long timescales. While the limited number of archaic DNA samples has offered some insights, they remain insufficient to fully reconstruct the complex genetic history of early human populations.

Furthermore, previous methods typically rely on deriving summary statistics from genetic data to infer the presence of admixture and population structure. Identifying subtle changes in population history directly from genetic data is particularly challenging when there is minimal evidence or when source groups remain unsampled. The large-scale inference of genome-wide genealogies provides a more direct approach to examining the demographic factors influencing genetic variation. By modeling

these demographic changes, genealogies offer a more interpretable and detailed view of human history, allowing researchers to trace population dynamics with greater accuracy and uncover previously hidden aspects of our ancestral past.

In this thesis, we introduce a novel method called GhostBuster, which leverages coalescent theory and inferred genealogies to analyze past mixture events with greater statistical power and interpretability. GhostBuster is a mixture model that clusters the genealogical history of a target individual based on its coalescence rates with a set of reference populations. This approach not only characterizes mixture events but also decomposes the local ancestry of target individuals. Unlike traditional methods, GhostBuster does not require sampled source populations to identify admixture events, making it capable of detecting ghost admixtures. Through simulations and real-data examples, we demonstrate that GhostBuster can successfully recover signatures of both ancient and recent admixtures, accurately identifying source groups, admixture times, and the local ancestry of target individuals.

We apply GhostBuster to both modern and ancient samples from sub-Saharan Africa, a region where ancient DNA is particularly scarce due to environmental conditions that hinder DNA preservation. Our analysis uncovered two significant events that occurred in Africa. First, we identified a Holocene-era back-to-Africa migration that introduced detectable levels of Eurasian ancestry into all sub-Saharan African populations we analyzed. This event, which occurred between the last millennium and approximately 10,000 years ago, provides a very likely explanation for the presence of small amounts of Neanderthal ancestry in modern Africans. Leveraging inferred local ancestry, we detected differential enrichment for specific trimer mutations, including TCC to TTC substitutions, which have been previously linked to West Eurasian populations [76]. Further characterization of this event using coalescence rate affinities with various ancient populations revealed that the Eurasian ancestry in Africans may trace back to early European farmers from Anatolia, suggesting a potential spread of farming practices from Europe to the Middle East and North Africa, eventually reaching sub-Saharan Africa. The second event represents a more ancient admixture between two African population groups,

one of which is closely related to the out-of-Africa population that gave rise to present-day Eurasians. Our findings suggest that these populations diverged at least 300,000 years ago and subsequently admixed around or after the major out-of-Africa migration. This admixture is associated with differences in population sizes and PRDM9 hotspot activity between the two groups. Furthermore, we observed distinct affinities of these ancient populations with other archaic humans, such as Neanderthals and Denisovans, from that era. Finally, our analysis revealed that the two ancient populations exhibit differing levels of polygenic score (PGS) portability when genome-wide association studies (GWAS) are conducted in European cohort, highlighting the importance of ancestry decomposition in understanding genotypic variation affecting phenotypes.

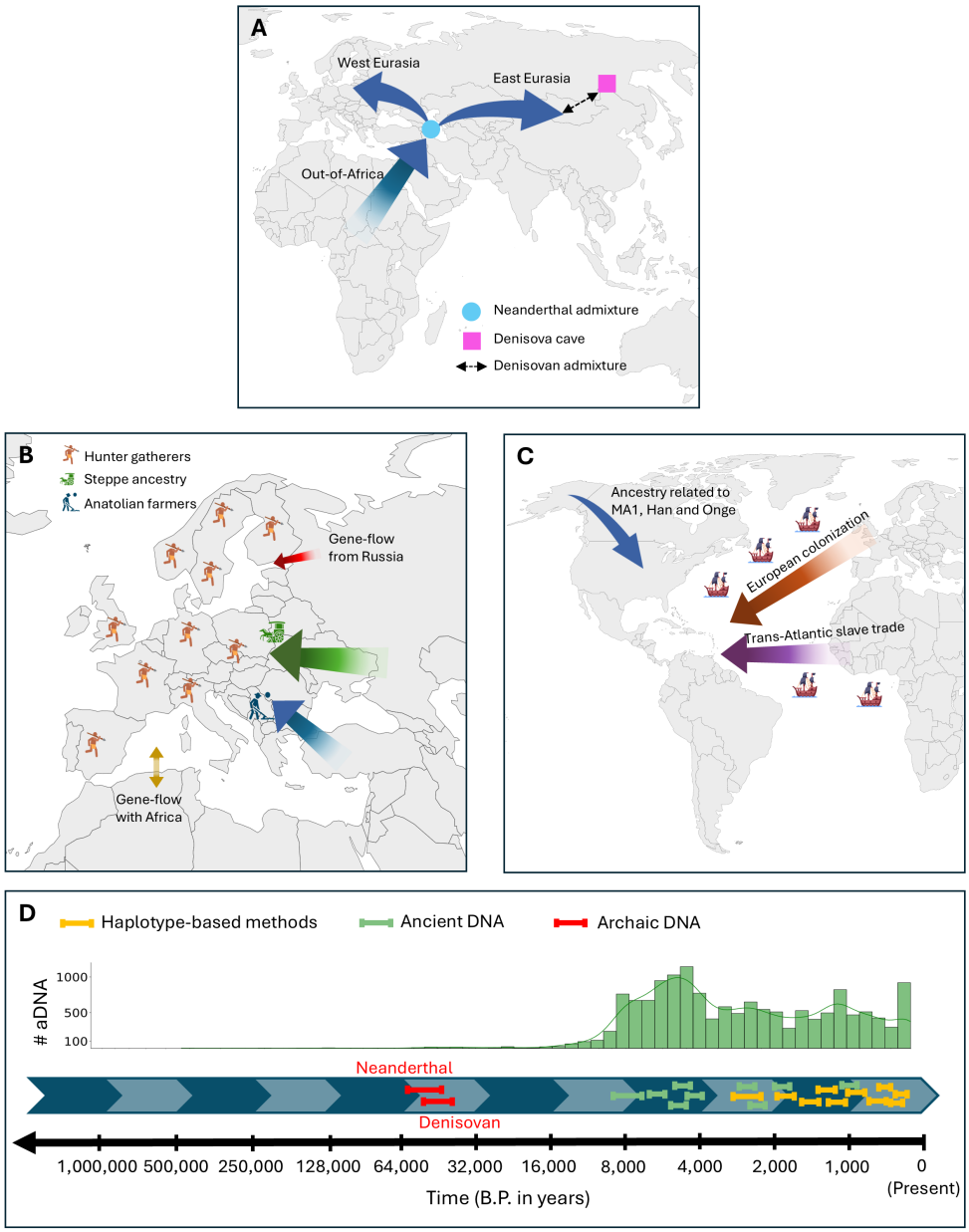


Figure 1.5: Some examples of migration and admixture events in human history. A. Archaeological and genetic evidence suggests modern humans originated in Africa and migrated out between 50,000 and 100,000 years ago [18, 77]. They interbred with Neanderthals, who contributed up to 3% of non-African DNA. After leaving Africa, the population split into west and east Eurasian groups, creating genetic differences between Europeans and Asians. East Eurasians also admixed with Denisovans, though the timing and location are debated [25, 78]. B. Europe has experienced significant migrations, starting with Anatolian farmers 7,000-9,000 years ago, followed by Steppe pastoralists around 5,000 years ago, which heavily influenced present-day European ancestry. There is also evidence of recent gene flow from North African groups into southern Europeans, such as Italians and Spanish, and ancient Siberian-related ancestry in northeastern Europeans like Finns and Russians [21, 58]. C. The peopling of America likely began during the last glacial period, which created a land bridge to Eurasia [79], ending around 15,000 years ago. Ancient DNA in America shows genetic links to ancient Siberians (MA1), Han Chinese, and the Onge tribe [80]. Present-day American genetics are influenced by historical events like colonization and the trans-Atlantic slave trade, resulting in admixture from Native American, African, and European sources [22, 23]. D. Previously identified human admixtures were determined using various methods, including haplotype-based approaches [51, 53–55], ancient DNA, and archaic DNA analyses. The histogram shows the ages of ancient DNA samples sequenced to date, sourced from the Allen Ancient DNA Resource [81]. The illustrations were downloaded from Flaticon.

1.3 Understanding complex traits through genetic data

An important line of work in statistical genetics focuses on understanding how the patterns of genetic variation in present-day individuals affect phenotypes. The first disease gene was mapped in 1983, identifying a short tandem repeat on chromosome 4 responsible for Huntington's disease [82]. This study involved analyzing pedigrees with fewer than 100 samples to pinpoint the genetic cause of Huntington's disease. Huntington's disease is an example of a monogenic disorder, where alterations in a single gene (in this case, Huntington gene) explains the entire genetic basis of the disease. With the growing size of genetic datasets, we now know that most complex traits and common diseases in humans are not monogenic but polygenic, where hundreds or thousands of small alterations in the genome explain the genetic basis of a trait. Finding these small polygenic effects spread across the genome requires examining genetic variation data genome-wide in a larger number of samples.

The Wellcome Trust Case Control Consortium (WTCCC) [83] was one of the first efforts to perform a large-scale genetic association study. It genotyped 17,000 individuals, comprising 2,000 cases and 3,000 common controls across seven common diseases (type 1 diabetes, type 2 diabetes, coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis, and Crohn's disease). The WTCCC performed a genome-wide association study to look at polymorphisms in the genome that correlate more than usual with the case-control status and found 90 new variants across all the diseases analyzed, confirming the presence of 28 previously established variants. Since then, Genome-wide association studies have become a cornerstone in current-day quantitative genetics to understand how genetic variation affects phenotypes.

1.3.1 Basic definitions

The heritability of a trait, roughly speaking, is the proportion of total trait variance explained by genetic variation. It measures how important genetics are in predicting a given trait, where a heritability of 1 means genetics explain all the variance in the

trait between individuals, and a heritability of 0 means genetics does not explain any variance in the trait. It is important to note that heritability for a trait depends on the quality of phenotypic measurements and the environment itself. Heritability should not be confused with the variance associated with the direct biological mechanisms affecting a trait. In many complex traits, the path from genes to traits is complex and indiscernible [84]. A more detailed understanding of heritability, its implications, and common misunderstandings can be found in [85, 86].

Linkage disequilibrium (LD) refers to the correlation patterns in genetic variation data due to the fact that variants close to each other are transmitted together more frequently than variants further away from each other. LD is further exacerbated because most recombination in humans happens at recombination hotspots, creating distinct LD patterns along the genome. LD can be both a boon and a bane in analyzing human variation data. It is a boon because common variants used in low-cost genotyping arrays, can capture the signal from nearby unobserved causal variants, thus providing substantial power in predicting the phenotype using low-cost common variation data. However, it is a bane because the same LD makes it challenging to distinguish causal variants from associated variants, as tightly correlated variants obscure the identification of the true causal variant.

Two major challenges in performing genome-wide association studies (GWAS) are the presence of population structure and relatedness in the data, both of which can confound genetic signals and inflate test statistics. We have already discussed population structure in the context of population genetics in Section 1.2.1. In quantitative genetics, population structure can lead to allele frequency differences between cases and controls, resulting in spurious associations between genotypes and traits — a phenomenon known as population stratification. Relatedness between samples is another significant confounder in association testing and is typically estimated through the genetic relatedness matrix (GRM). The (GRM) is computed as $\frac{XX^T}{M}$, where X is an $N \times M$ mean-centered and standardized genotype matrix, N is the sample size, and M is the number of variants. The GRM plays a crucial role

in linear mixed model (LMM) methods, which are widely used in GWAS to account for these confounders. We will discuss LMM methods in more detail in Section 1.3.3.

1.3.2 Genome-wide association studies: usage and applications

After it became clear that many human traits are polygenic in nature, an unbiased genome-wide scan for variants associated with these traits was warranted. Following the first large-scale genome-wide association study published by the Wellcome Trust Case Control Consortium (WTCCC) in 2007, a huge wave of large-scale population-wide studies aimed at understanding the genetic architectures of various complex traits ensued. Today, 17 years after WTCCC, GWAS has become a standard protocol for understanding the genetic basis of most complex traits. With sequencing and phenotyping efforts from pharmaceutical industries, academia, and independent nation-wide "biobanks", GWASs have been performed on thousands of human traits, using millions of individuals, and have identified hundreds of thousands of genetic associations [87].

GWAS has become a vital element in drug discovery, with two-thirds of the new drugs approved by the FDA in 2021 supported by human genetic evidence [88–90]. Some examples include Finerenone, an NR3C2 antagonist approved for the treatment of chronic kidney disease [91, 92]. Baricitinib, identified through GWAS studies for COVID-19 severity, was shown to reduce the COVID-19 mortality rate by approximately 20% [93, 94]. It is important to note that GWASs are not enough to support drug target discovery on their own. Further downstream analysis, such as replication, expression QTL studies in relevant tissues and cell types, and understanding off-target effects, are necessary and are part of the GWAS pipeline to obtain suitable drug targets.

Apart from identifying genetic associations with traits, GWAS is also utilized to build polygenic scores for a trait or disease susceptibility. Polygenic scores, or PGS, use the estimated effects for a trait to build a predictor of the genetic likelihood of a trait. PGS have found clinical potential for personalized medicine and treatment,

with current utility for various genetic disorders including cardiovascular diseases [95]. GWAS has also enabled the identification of genetic similarity between traits and diseases, helping to identify biological pathways for gene functions [96]. On a more general level, GWAS hits have also helped us understand the evolution of certain human traits, where selection mechanisms like directional, stabilizing, and pleiotropic are among the most likely explanations of how trait-associated variants spread forward in time [97]. Additionally, GWAS has been useful in causal inference studies, as a genetic instrument in Mendelian randomization, helping to identify disease-causing traits [98, 99]. An overview of the GWAS pipeline and its application can be found in Figure 1.6 and following reviews [90, 100].

1.3.3 Theoretical framework to perform GWAS

There are two main categories of methods to perform genome-wide association studies. The first set of methods involves fitting a linear model to estimate the effect of a variant on the phenotype, taking into account various covariates. The linear models fit the following model:

$$y = x_{test}\beta_{test} + \alpha C + \epsilon. \tag{1.13}$$

In the above, y is an $N \times 1$ vector of mean-centered and standardized phenotype values, x_{test} is an $N \times 1$ vector of mean-centered and standardized bi-allelic variant values (that is, the original $\tilde{x}_{test}$ takes values in $\{0, 1, 2\}^N$ for diploid samples), C is an $N \times C$ matrix of covariates of interest, such as principal components, age, and sex, and ϵ is Gaussian noise.

Linear models fit the model in Equation 1.13 using a simple OLS estimate for linear regression or an iterative scheme for logistic regression. Test statistics are calculated from standard closed-form solutions for the standard error of the estimates. However, simple linear models fail to correct for the presence of population structure and cryptic relatedness in the data, making the resulting test statistic estimates inflated. Therefore, methods that rely on linear models require working with a subset of unrelated and homogeneous individuals to ensure the false positive rates

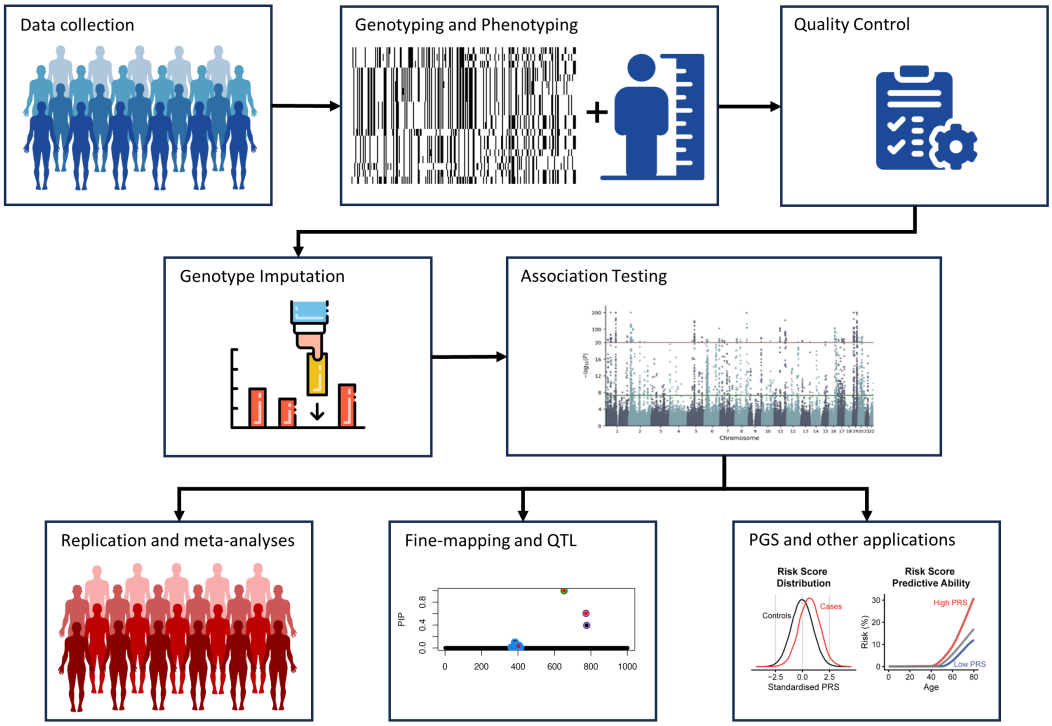


Figure 1.6: Overview of a typical GWAS pipeline. The GWAS pipeline typically begins with genotyping and phenotyping a large cohort of individuals, including both cases and controls for disease traits. Genotyping can be conducted using either low-cost SNP arrays or whole-genome sequencing. Quality control steps are crucial and involve filtering variants and samples based on various criteria (see Section 5.2 for detailed quality control procedures). If genotyping was performed instead of whole-genome sequencing, genotype imputation follows as the next step. Subsequently, genome-wide association testing is conducted for each variant and phenotype of interest to generate p-values and effect estimates, indicating the association between each variant and the phenotype. The resulting GWAS summary statistics have several downstream applications. These include replication and meta-analyses to increase statistical power, fine-mapping and functional analyses to identify causal variants among the associated variants, and the calculation of polygenic scores. This pipeline structure was adapted from [101], with the fine-mapping figure from [102] and the polygenic scores figure from [103].

are not too high. This, in turn, reduces the sample size and power to detect genetic associations in the dataset.

The second class of methods are Linear mixed models or LMMs. LMMs are

an extension of linear models that facilitate the analysis of non-independent data. They are commonly adopted in the analysis of genomic data involving population structure or cryptic relatedness. LMMs handle non-independence by modeling random effects that capture variability within different subpopulations or phenotypic correlation across close relatives. In an LMM for genome-wide association analysis, the phenotype of interest is often modeled as a combination of fixed effects, which include the effects of the test variant and covariates, and random effects, which include genetic effects and environmental effects:

$$y = x_{test}\beta_{test} + \alpha C + g + \epsilon. \tag{1.14}$$

Similar to the linear model above, y is an $N \times 1$ vector of mean-centered and standardized phenotype values, x_{test} is an $N \times 1$ vector of mean-centered and standardized bi-allelic variant values, C is an $N \times C$ matrix of covariates. We are mainly interested in inferring the fixed effect of a particular variant, β_{test}, given the random genetic effects, which help capture the effects of population structure and relatedness. We model the genetic and environmental effects using

$$\begin{aligned} g &= \beta_{GRM} X_{GRM} \\ \epsilon &\sim \mathcal{N}(0, \sigma_e^2), \end{aligned} \tag{1.15}$$

where X_{GRM} is an $N \times M_{GRM}$ mean-centered and standardized genotype matrix, independent of x_{test}, and σ_e^2 is the environmental variance for a given phenotype. For simplicity, the environmental random effect is assumed to be independent and identically distributed (i.i.d.) from a normal distribution, and the random genetic effect is modeled as a weighted sum of standardized genotypes. The random genetic effect g can also be thought of as a sample from $\mathcal{N}(0, \sigma_g^2 K)$, where σ_g^2 is the genetic variance for a given phenotype and $K = \frac{1}{M} X_{GRM} X_{GRM}^T$ is the genetic relatedness matrix.

Leave-one-chromosome-out: It has been observed that in order to avoid reducing the power of association, the matrix X_{GRM} should not contain the test variant x_{test}, as well as any other variants in linkage disequilibrium (LD) with

x_{test}. This problem has been referred to as "proximal contamination" [104]. Fitting an LMM with genetic effects from variants not in LD with the testing variant, however, would require re-training the model to infer genetic effects for every tested variant, which would cause major computational overheads. [105] proposed a simple solution to this problem, referred to as the leave-one-chromosome-out (LOCO) scheme, where the X_{GRM} is constructed using all variants except those on the same chromosome as the test variant. For example, assuming 22 autosomal chromosomes and a test variant on chromosome 7, X_{GRM} is built to contain all variants from chromosome 1-6 and 8-22. This approach, which requires training a separate model for each chromosome, allows circumventing proximal contamination while increasing association power, and has been widely adopted by recent LMM association algorithms [106–108].

Calculation of the test statistic: To infer the fixed effect β_{test} in Equation 1.14, one can marginalize the random effects $\beta_{GRM}X_{GRM}$ and ϵ to get the marginalized log-likelihood:

$$\begin{aligned} y \mid x_{test}, \beta_{test}, \sigma_e^2, \sigma_g^2 &\sim \mathcal{N}(x_{test}\beta_{test}, \frac{\sigma_g^2}{M} X_{GRM} X_{GRM}^T + \sigma_e^2 I_N) \\ \log P(y \mid x_{test}, \beta_{test}, \sigma_e^2, \sigma_g^2) &\propto -\frac{1}{2}(y - x_{test}\beta_{test})^T V^{-1}(y - x_{test}\beta_{test}) - \frac{1}{2}|\log(V)|, \end{aligned} \tag{1.16}$$

where σ_g^2 and σ_e^2 are genetic and environmental variance components as described above, and V is a covariance matrix, defined as $V = \frac{\sigma_g^2}{M} X_{GRM} X_{GRM}^T + \sigma_e^2 I_N$.

The maximum likelihood estimate (MLE) for β_{test} is obtained by setting $\frac{\partial \log P(y|x_{test},\beta_{test})}{\partial \beta_{test}} = 0$; the variance of the estimate is obtained using the inverse Fisher information $-\mathbf{E}\left[\frac{\partial^2 \log P(y|x_{test},\beta_{test})}{\partial \beta_{test}^2} \mid \hat{\beta}_{test}\right]^{-1}$ and takes the form:

$$\begin{aligned} \hat{\beta}_{test} &= (x_{test}^T V^{-1} x_{test})^{-1}(x_{test}^T V^{-1} y) \\ var(\hat{\beta}_{test}) &= (x_{test}^T V^{-1} x_{test})^{-1}. \end{aligned} \tag{1.17}$$

The variance components, σ_g^2 and σ_e^2, are estimated using restricted likelihood maximization (REML) [106] or moment-based methods [109, 110] rather than standard maximum likelihood estimation. In this work, we estimate the variance components $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ using the RHE-MC algorithm [109, 110], a highly

scalable moment-based approach that was recently extended to allow processing multiple traits in parallel [111]. We then estimate the covariance matrix using $\hat{V} = \frac{\hat{\sigma}_g^2}{M} X_{GRM} X_{GRM}^T + \hat{\sigma}_e^2 I_N$, and the effects using

$$\begin{aligned} \hat{\beta}_{test} &= (x_{test}^T \hat{V}^{-1} x_{test})^{-1} (x_{test}^T \hat{V}^{-1} y), \\ var(\hat{\beta}_{test}) &= (x_{test}^T \hat{V}^{-1} x_{test})^{-1}. \end{aligned} \tag{1.18}$$

Using the asymptotic property of the MLE,

$$\frac{\hat{\beta}_{test} - \beta_{test}}{\sqrt{var(\hat{\beta}_{test})}} \sim \mathcal{N}(0, 1). \tag{1.19}$$

To test the null hypothesis $\beta_{test} = 0$, we therefore substitute the MLE estimate and the variance of the estimate from Equation 1.18, and square both sides to get:

$$\frac{(x_{test}^T \hat{V}^{-1} y)^2}{x_{test}^T \hat{V}^{-1} x_{test}} \sim \chi_1^2. \tag{1.20}$$

There are two common modeling assumptions for the distribution of genetic effects along the genome, leading to different LMM foundations - Infinitesimal and non-infinitesimal models.

Infinitesimal model: The infinitesimal model places a normal prior on the effect estimates β_{GRM} from Equation 1.15, and assumes the total genetic effect explained by all the genetic markers to be σ_g^2:

$$\beta_{GRM} \sim \mathcal{N}(0, \sigma_g^2 I_M / M) \tag{1.21}$$

As the prior and likelihood are conjugate, we can infer the posterior distribution $P(\beta_{GRM} \mid X_{GRM}, y)$:

$$\begin{aligned} P(\beta_{GRM} \mid X_{GRM}, y) = \mathcal{N}\Bigg(& \left(\frac{I_M M}{\sigma_g^2} + \frac{X_{GRM}^T X_{GRM}}{\sigma_e^2} \right)^{-1} \frac{X_{GRM}^T y}{\sigma_e^2}, \\ & \left(\frac{I_M M}{\sigma_g^2} + \frac{X_{GRM}^T X_{GRM}}{\sigma_e^2} \right)^{-1} \Bigg) \end{aligned} \tag{1.22}$$

The posterior mean estimate of β_{GRM} in Equation 1.22, also referred to as the best linear unbiased predictor (BLUP), can be written as:

$$\hat{\beta}_{GRM} = \left(\frac{I_M M}{\sigma_g^2} + \frac{X_{GRM}^T X_{GRM}}{\sigma_e^2} \right)^{-1} \frac{X_{GRM}^T y}{\sigma_e^2} \tag{1.23}$$

The residual phenotype, formed by regressing out the effect of other variants on the phenotype, is given by $\tilde{y} = y - X_{GRM}\hat{\beta}_{GRM}$ and simplifies to

$$\tilde{y} = y - X_{GRM}\hat{\beta}_{GRM} = \sigma_e^2 V^{-1} y \tag{1.24}$$

This provides a link between the BLUP and test statistics for x_{test} in Equation 1.20, which is exploited to compute test statistics for non-infinitesimal models. Under the infinitesimal model, the test statistic can be rewritten as:

$$\frac{(x_{test}^T \tilde{y}/\sigma_e^2)^2}{x_{test}^T \hat{V}^{-1} x_{test}} \sim \chi_1^2 \tag{1.25}$$

Non-infinitesimal model: It is often the case that a large fraction of variants are not linked to the phenotype, so the above model can be improved by adopting a more appropriate prior, such as a spike-and-slab prior, or a mixture-of-Gaussian prior. The spike-and-slab prior can be modeled as follows:

$$P(\beta_{GRM}) \sim p_0 \mathcal{N}(0, \sigma_g^2/Mp_0) + (1-p_0)\delta(0), \tag{1.26}$$

where $\delta(0)$ is the Dirac-delta function at 0. A mixture-of-Gaussian prior replaces the Dirac-delta function using a low-variance normally distributed component, and can be written as:

$$P(\beta_{GRM}) \sim p_0 \mathcal{N}(0, \sigma_{g,1}^2/M) + (1-p_0)\mathcal{N}(0, \sigma_{g,2}^2/M). \tag{1.27}$$

Equation 1.25 can still be used to compute test statistics, by replacing $\tilde{y}$ with the phenotype residuals calculated using BLUP under the non-infinitesimal model. Note that it is often not possible to obtain a closed-form solution for the BLUP under an non-infinitesimal model, but can be approximated through MCMC or variational inference algorithms.

1.3.4 Survey of methods to perform GWAS

Plink [112] is one of the most widely used linear model-based GWAS algorithms. Its subsequent versions [113] improve its scalability by implementing vectorized operations over phenotypes, working with the disk-space efficient pgen format, and

supporting sparse matrix multiplication [114]. Despite its efficiency, Plink does not implement a linear mixed model to correct for population structure and relatedness, and therefore is only used with unrelated homogeneous subsets of the data.

EMMA [115] is among the earliest methods to implement a linear mixed model for association testing in human datasets. It optimized the fixed effects by maximizing the full likelihood in Equation 1.16 relying on spectral decomposition. The overall algorithm was still cubic in sample size and required constructing the $N \times N$ kinship matrix. EMMAX, or EMMA expedited [116], was a follow-up to EMMA which improved its scalability by more than two orders of magnitude. Instead of recomputing the variance parameters for each variant separately, EMMAX observed that a single variant does not have a huge effect on the variance components and therefore estimated it once and reused it for association testing.

Fast-LMM [117] used the spectral decomposition of the kinship matrix and its low-rank representation to speed up association testing further compared to EMMAX. Fast-LMM-Select [118] followed as a subsequent method, which subsets the markers for kinship calculation based on their predictability for the trait using a linear model. It was later shown that this strategy of reducing the number of markers in kinship calculation leads to inflation in the test statistics, as background polygenic effects are not fully captured [105].

BOLT-LMM [106, 107] was one of the first methods to scale linearly with the number of markers and sub-quadratically with the number of samples. The overall complexity of BOLT-LMM is $O(MN^{1.5})$. BOLT-LMM achieved this by using conjugate gradient techniques for matrix inversions (thus never explicitly computing the kinship matrix), an efficient grid search algorithm for estimating variance parameters, and a variational inference routine to predict the polygenic effects. Furthermore, BOLT-LMM's variational inference routine imposed a mixture of Gaussian priors on the effect estimates, allowing it to model the non-infinitesimal genetic architecture, which led to improved statistical power in GWAS.

The assumption of normality about the test statistic distribution is often violated in binary traits, especially when the ratio of cases to controls is imbalanced or when

testing with a rare variant. SAIGE [119] implements saddle point approximation (SPA) [120, 121] to correct the test statistics and ensure calibration in binary traits. SAIGE also employs conjugate gradient tricks similar to BOLT-LMM, which speeds up the calculation of test statistics while avoiding explicit computation of the kinship matrix.

FastGWA [108] and FastGWA-GLMM [122] are among the fastest methods to perform linear mixed model association testing for quantitative and binary traits. FastGWA-GLMM is the generalized mixed model extension of FastGWA, aimed at analyzing binary traits. These methods approximate the GRM as a sparse matrix, thresholding it to only keep close relatives. The sparse GRM is generated once and used for all subsequent analyses, making matrix multiplications in test statistics calculation extremely fast. However, approximating the kinship with a sparse matrix loses some statistical power compared to methods that use the dense matrix.

REGENIE [123] is a recent method for association testing that uses block-wise ridge regression to estimate the polygenic effect and then vectorized test statistics calculation routines, allowing for the analysis of multiple phenotypes at scale. REGENIE also implements an approximate version of Firth logistic regression and saddle point approximation to correct test statistics for binary traits. REGENIE scales linearly with sample size and number of markers, making it approximately two orders of magnitude faster than BOLT-LMM in the analysis of 50 quantitative traits in the UK Biobank. However, it still does not match BOLT-LMM's statistical power, as it does not account for the non-infinitesimal genetic architecture.

A table comparing modern GWAS methods can be found in table 1.1.

1.3.5 Motivation for Quickdraws

As we saw in Section 1.3.2, genome-wide association studies (GWAS) are a key step in modern analyses of complex traits and diseases. These association studies are being enabled by the rapid growth of modern biobanks [3, 124–128], which contain vast amounts of genomic, phenotypic, and environmental information for increasingly diverse groups [129]. The scale of these data sets, however,

creates several computational obstacles, which have led to significant advances in statistical methodology and the development of highly scalable linear-mixed-model GWAS algorithms [104, 106–108, 115–118, 130–133]. Among these, the BOLT-LMM algorithm [106, 107] achieves state-of-the-art association power by adopting a Bayesian mixture prior to model non-infinitesimal trait architectures, but is computationally intensive, particularly when applied to multiple traits, and has limited applicability to binary traits. More recent methods, including SAIGE [119], FastGWA [108, 122], and REGENIE [123], are highly scalable and may be used for the analysis of both quantitative and binary traits. These algorithms have enabled performing GWAS for millions of genomic variants and individuals [134] across thousands of traits and diseases [135], but rely on modeling approximations, such as the use of block-wise ridge regression or sparse genetic relationship matrices, that may reduce statistical power. As a result, current GWAS algorithms are either highly scalable and cost effective or highly powered to detect association, but not both.

In this thesis we describe a method, called Quickdraws, that uses machine learning strategies to simultaneously achieve state-of-the-art association power and computational efficiency, while being applicable to both quantitative and binary traits. Quickdraws uses a Bayesian regression model with a spike-and-slab prior on variant effects, which is efficiently trained by leveraging recent advances in stochastic variational inference, transfer learning, and graphics processing units (GPUs) matrix operations. We use simulations to show that, applied to the analysis of quantitative traits, Quickdraws matches the association power of BOLT-LMM, the most powerful method, while requiring orders of magnitude less computation. For binary traits, Quickdraws achieves higher statistical power compared to existing methods such as SAIGE, REGENIE, and FastGWA, equivalent to analyzing up to 13.4% more samples, while having a comparable computational speed. We apply Quickdraws and other methods to test 13.3 million variants for 79 quantitative traits and 50 disease traits in up to $\sim 405k$ individuals from the UK Biobank data set [3]. For quantitative traits, Quickdraws identifies 4.97% more independent associations than REGENIE and 22.71% more than FastGWA; for disease traits, it

detects 3.25% more associations than REGENIE and 7.07% more than FastGWA. These associations result in similar gains in the number of replicated signals in Biobank Japan [124] and Finngen [128]. We estimate computational costs using the UK Biobank RAP platform, observing that despite these gains in statistical power, the cost for running these analyses using Quickdraws is comparable to that of other highly scalable methods.

GWAS Method	Main innovation	Scalability	Uses non-infinitesimal prior	Multi-trait vectorization	Binary trait correction	Corrects population structure & relatedness
Plink v2	Implements a very fast linear model, uses efficient data storage and access	$\mathcal{O}(MN)$	✗	✓	✗	✗
BOLT-LMM	Implements a linear mixed model, uses variational inference to fit a mixture of normal prior on the effects	$\mathcal{O}(MN^{1.5})$	✓	✗	✗	✓
SAIGE v2	Implements a logistic mixed model, uses sparse GRM and SPA to correct for binary traits	$\mathcal{O}(MN)$	✗	✗	✓	✓
REGENIE	Implements block-wise ridge regression with SPA and Firth logistic regression to correct for binary traits	$\mathcal{O}(MN)$	✗	✓	✓	✓
FastGWA	Implements linear and logistic mixed model, uses sparse GRMs for quick test statistics calculation	$\mathcal{O}(MN)$	✗	✗	✓	✓

Table 1.1: Comparison of methods to perform GWAS. We compare five most popular GWAS softwares namely, Plink [112], BOLT-LMM[106, 107], SAIGE[119], REGENIE[123] and FastGWA[108, 122] comparing their overall scalability, ability to model non-infinitesimal genetic architecture for higher statistical power, multi-trait vectorization for better scalability, support for binary traits and ability to control FPR in presence of population structure and relatedness.

2

A method to find ghost populations using genome-wide genealogies

Contents

2.1 Chapter overview 42
2.2 Theoretical framework for our method 44
2.2.1 Introduction to mixture models 44
2.2.2 Mixture model for the coalescent 47
2.2.3 EM derivation: terminology 52
2.2.4 EM derivation: the likelihood 55
2.2.5 EM derivation: inference of coalescence rates 59
2.2.6 EM derivation: inference for admixture time and proportions 63
2.2.7 Summary of the algorithm 68
2.3 Method description 69
2.3.1 Graphic overview of the method 69
2.3.2 PCA visualization to visualize ancestral components . . 69
2.3.3 Coancestry curves to date admixture event 70
2.3.4 Joint analysis of multiple targets 72
2.3.5 Tree Filtering 73
2.3.6 Choosing the number of components 73
2.3.7 Calculation of node persistence 75
2.3.8 Handling uncertain coalescence events 76
2.3.9 Initialization 77
2.4 Data and quality control 79
2.4.1 Modern human datasets 79
2.4.2 High quality ancient DNA resources 80
2.4.3 Building genome-wide genealogies for moderns and ancients 81
2.5 Accurate local ancestry inference in simulations 82

2.5.1 Simulation design 82
2.5.2 Four-way recent admixture 83
2.5.3 Denisovan-like Introgression into Papuans 89
2.6 Identifying several recent admixtures accurately 95
2.7 Identifying multi-way admixture in holocene Europe . 100

2.1 Chapter overview

In this chapter, we introduce GhostBuster, a novel method for detecting admixture events by analyzing genome-wide genealogies inferred from genomic variation data. GhostBuster employs a mixture model to probabilistically cluster regions of the genome, identifying distinct "types" of coalescence histories for a target individual, with each ancestry type characterized by its coalescence rates to reference populations. Using an efficient expectation-maximization (EM) algorithm, the method iteratively updates these coalescence rates and the local ancestry of the target individual. The algorithm operates in two iterative steps. In the first step, GhostBuster assigns each genomic region of the target individual to an ancestry type, effectively decoding local ancestry (see Figure 2.1A). This is done by comparing who the target individual coalesces with back in time, with coalescence rates describing each ancestry type. In the second step, GhostBuster refines the coalescence rates that define each ancestry by leveraging the updated local ancestry information, analogous to updating cluster means and covariances in Gaussian mixture models (see Figure 2.1B).

We begin by presenting the theoretical framework underlying GhostBuster in Section 2.2, where we employ coalescent theory to model the observed genealogies. Next, in Section 2.3, we delve into the algorithmic and methodological details of the method. In Section 2.4, we discuss the various modern and ancient DNA resources used in the analyses and the construction of genealogies. In Section 2.5, we apply GhostBuster to two simulated scenarios, demonstrating its ability to recover ghost admixture events and exploring different aspects of the method. In Section 2.6, we apply GhostBuster to real data, replicating several well-established

admixture signals and comparing local ancestry estimates to prior studies. Finally, in Section 2.7, we apply GhostBuster to decompose and interpret admixture events over the last 10,000 years in Europe, a period that has gained new clarity through the availability of ancient DNA.

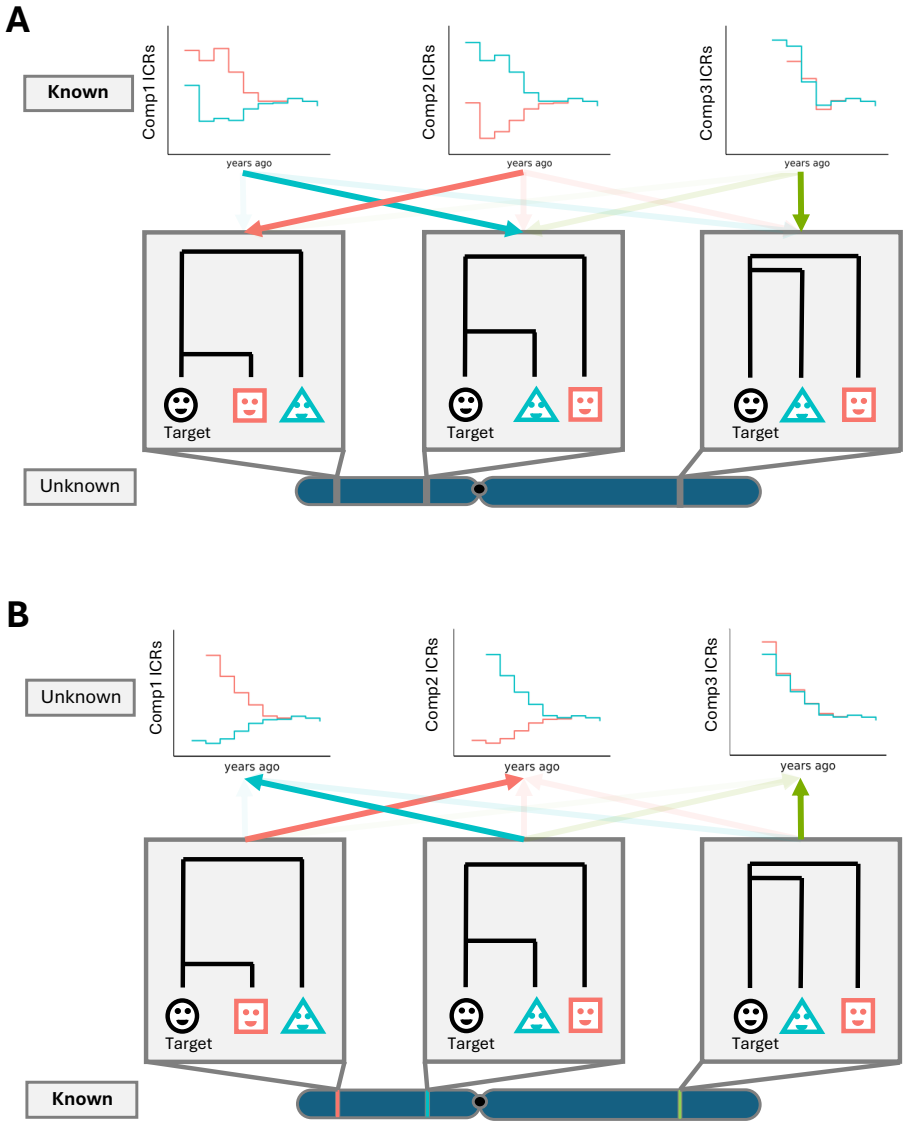


Figure 2.1: A simplified overview of GhostBuster. GhostBuster implements an iterative expectation-maximization (EM) algorithm which cycles between two states A and B. (A) inferring local ancestry of a target individual given multiple coalescence rates of the target individuals with a set of reference populations. (B) using the local ancestry to refine the estimates of coalescence rates of target individual and the reference populations.

2.2 Theoretical framework for our method

2.2.1 Introduction to mixture models

Mixture models are probabilistic models that represent subpopulations within a larger population by assuming the data is generated from a mixture of distributions, each corresponding to a subpopulation or cluster.

Gaussian mixture models

One of the simplest and most widely used mixture models is the Gaussian Mixture Model (GMM). The GMM assumes that the data is generated from a mixture of multiple Gaussian distributions, where each observation is drawn from one of the Gaussian components.

Consider a dataset $x = (x_1, x_2, \ldots, x_n)$, where x is a vector of n independent observations. For simplicity, assume that these observations are sampled from a mixture of two Gaussian distributions. To model this mixture, we introduce a set of hidden variables $z = (z_1, z_2, \ldots, z_n)$, where each z_i is a categorical variable indicating which Gaussian component generated the corresponding observation x_i. Specifically, $z_i = 0$ if x_i is drawn from the first Gaussian component, and $z_i = 1$ if x_i is drawn from the second. The conditional distribution of the random variable X_i, given its corresponding hidden variable Z_i, is defined as follows:

$$X_i|(Z_i = 0) \sim \mathcal{N}(\mu_1, \Sigma_1),$$
$$X_i|(Z_i = 1) \sim \mathcal{N}(\mu_2, \Sigma_2),$$

where, $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ denote Gaussian distributions with means μ_1 and μ_2, and covariance matrices Σ_1 and Σ_2, respectively. Additionally, let us assume the prior probabilities of the components are given by $P(Z_i = 0) = \pi_1$ and $P(Z_i = 1) = \pi_2$, where $\pi_2 = 1 - \pi_1$.

The aim of the GMM is to estimate the parameters corresponding to the mixture of two Gaussians, namely $\theta = (\pi_1, \mu_1, \Sigma_1, \mu_2, \Sigma_2)$, by maximizing the likelihood of the observed data. However, it is important to note that the vector z is a hidden variable, meaning we do not know which observation corresponds to which

cluster. Consequently, we need to maximize the "incomplete-data likelihood," which integrates out the hidden variable z:

$$L(\theta; x) = \prod_{i=1}^{n} \Big[\pi_1 \mathcal{N}(x_i|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x_i|\mu_2, \Sigma_2)\Big]. \tag{2.1}$$

where, $\mathcal{N}(x_i|\mu_j, \Sigma_j)$ represents the Gaussian probability density function for the j-th component, evaluated at x_i. Since the likelihood involves a summation inside the product, direct maximization is challenging. To address this, the Expectation-Maximization (EM) algorithm is employed [136]. Instead of directly maximizing the incomplete-data likelihood, the EM algorithm operates on the "complete-data likelihood," which assumes knowledge of the hidden variables z:

$$L(\theta; x, z) = \prod_{i=1}^{n} \prod_{j=1}^{2} \Big[\pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)\Big]^{\mathbb{1}(z_i=j)} \tag{2.2}$$

where, $\mathbb{1}(z_i = j)$ is an indicator variable that equals 1 if the observation x_i belongs to component j, and 0 otherwise. This formulation represents the likelihood as a product and the log-likelihood as a sum over components and observations, which eventually simplifies the optimization process.

$$\begin{aligned}\log L(\theta; x, z) &= \sum_{i=1}^{n} \sum_{j=1}^{2} \mathbb{1}(z_i = j)\Big[\log \pi_j + \log \mathcal{N}(x_i|\mu_j, \Sigma_j)\Big] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{2} \mathbb{1}(z_i = j)\Big[\log \pi_j - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1}(x_i - \mu_j) - \frac{d}{2}\log(2\pi)\Big].\end{aligned} \tag{2.3}$$

The EM algorithm alternates between the E-step, where the expected values of the hidden variables z_i are computed given the current parameter estimates, and the M-step, which maximizes the expected complete-data log-likelihood with respect to the parameters θ. Through these iterative updates, the EM algorithm refines the parameter estimates, ultimately converging to a local maximum of the incomplete-data likelihood [136] (see Figure 2.2). The EM algorithm for GMM operates as follows:

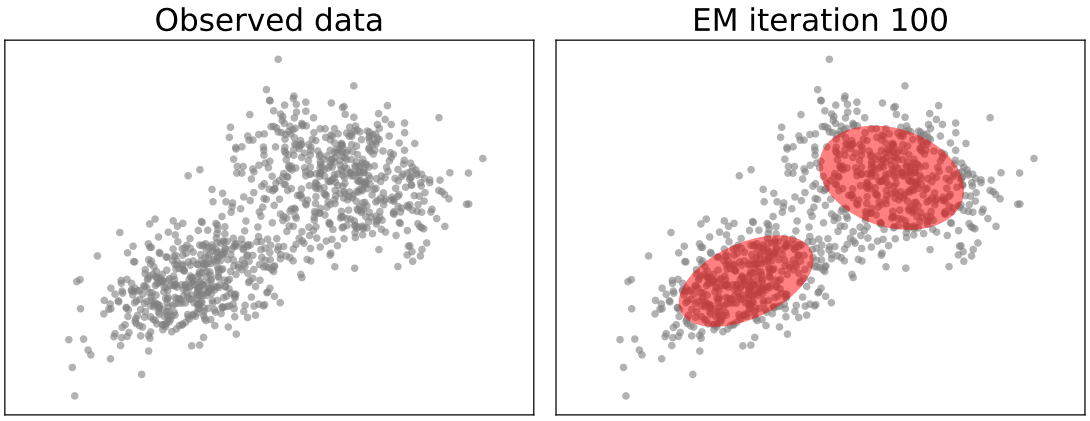


Figure 2.2: Gaussian mixture model. The observed data and inferred clusters using the Gaussian mixture model after 100 EM iterations, the Gaussian mixtures inferred are represented by the red ellipse.

1. **E-step**: Calculates the expected log-likelihood of the complete data, given the current parameter estimates.

$$
\begin{aligned}
&\mathbb{E}_{Z|X=x;\theta^{(t)}}\left[\log L(\theta; X, Z)\right] \\
&= \mathbb{E}_{Z|X=x;\theta^{(t)}}\left[\log \prod_{i=1}^{n} L(\theta; x_i, Z_i)\right] \\
&= \mathbb{E}_{Z|X=x;\theta^{(t)}}\left[\sum_{i=1}^{n} \log L(\theta; x_i, Z_i)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{Z_i|X_i=x_i;\theta^{(t)}}\left[\log L(\theta; x_i, Z_i)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{2} P(Z_i = j|X_i = x_i; \theta^{(t)}) \log L(\theta_j; x_i, j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{2} \gamma_j^{(t)}(x_i)\left[\log \tau_j - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x_i-\mu_j)^\top \Sigma_j^{-1}(x_i-\mu_j) - \frac{d}{2}\log(2\pi)\right]
\end{aligned}
\tag{2.4}
$$

   where, the log-likelihood of the complete data comes from Equation 2.3; $\gamma_j(x_i)$ is the posterior probability that the data point x_i was generated by the j-th Gaussian distribution:

$$
\gamma_j^{(t)}(x_i) = \frac{\pi_j^{(t)}\mathcal{N}(x_i|\mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{j=1}^{2}\pi_j^{(t)}\mathcal{N}(x_i|\mu_j^{(t)}, \Sigma_j^{(t)})} \tag{2.5}
$$

2. **M-step**: Maximizes the expected log-likelihood, computed in the E-step, with respect to the model parameters θ. For a GMM, this results in straightforward

analytical update equations, as shown below:

$$\begin{aligned}
\pi_j^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\gamma_j^{(t)}(x_i) \\
\mu_j^{(t+1)} &= \frac{\sum_{i=1}^{n}\gamma_j^{(t)}(x_i)x_i}{\sum_{j=1}^{n}\gamma_j^{(t)}(x_i)} \\
\Sigma_j^{(t+1)} &= \frac{\sum_{i=1}^{n}\gamma_j^{(t)}(x_i)(x_i-\mu_j^{t+1})(x_i-\mu_j^{t+1})^\top}{\sum_{j=1}^{n}\gamma_j^{(t)}(x_i)}
\end{aligned} \tag{2.6}$$

2.2.2 Mixture model for the coalescent

We aim to extend the basic example of mixture models to genealogies by specifically modeling coalescence events on a given target lineage as a mixture distribution. The input to our method includes, for each site, information about which individuals the target sample coalesces with and the corresponding coalescence times as we trace backwards in time. Our goal is to model this as a mixture distribution, clustering locations based on differences in the coalescence histories of the target lineage. The expected output of our model is a per-component coalescence rate matrix, which specifies how the target individual coalesces with each reference population across the components of the mixture.

Our method is based on the assumption that admixture can be characterized by coalescence rates with a set of reference populations. To illustrate this, consider the example of Denisovan admixture in Papuans. It is well known that Papuans carry up to 5% Denisovan ancestry in their genomes contributed around 40,000 years ago [25]. This archaic ancestry has different coalescence rates to reference populations compared to the majority of Papuan ancestry: regions with Denisovan ancestry exhibit more frequent coalescence events with Denisovans recently (last $40{,}000 - 700{,}000$ years), but do not tend to coalesce with modern human lineages until the time of the Denisovan-human split, estimated around 700,000 years ago [25]. Therefore, there is a clear difference in the coalescence history between archaic and non-archaic genomic regions in Papuans, which we leverage to infer admixture. Importantly, our method does not require explicit source groups or unadmixed outgroups to infer an admixture event. In the Papuan example,

Denisovan ancestry not only exhibits quicker coalescence with Denisovans but also shows slower coalescence with other Papuans and modern human groups. We can leverage these differences in coalescence rates with modern human groups alone to characterize the admixture, enabling us to infer admixture without relying on explicit source groups or additional assumptions about outgroups.

The generative model

To formalize the model, we first define the probabilistic process that gives rise to the observed data, which in our case are coalescent trees along the genome. Our approach is based on a structured generative model where we fix the model parameters - namely, the global ancestry proportions (π), the admixture time (λ), and the coalescence rates with reference groups (γ) - and then generate the genealogies at each position under these assumptions.

1. Global ancestry proportions (π): The target lineage is assumed to originate from a mixture of ancestries, where the global ancestry proportions π describe the expected fraction of the genome inherited from each ancestral source. Therefore, π is a vector of length equal to the number of ancestries.

2. Admixture time (λ): The admixture event is assumed to have occurred λ generations in the past. Therefore, λ is a scalar.

3. Coalescence rates (γ): For each ancestry, the coalescence rates with reference populations are determined by the matrix γ which encodes the instantaneous rate at which the target lineage coalesces with lineages belonging to each reference group. Therefore, γ is a matrix of dimension equal to the number of ancestries $\times$ number of reference groups $\times$ number of time epochs. The ancestry the target individual belongs to can change along the genome, and we will therefore refer to the local ancestry of a target individual.

The observed data under the model consist of the coalescence genealogies of the target individual with reference populations. The generative process for producing the observed genealogies is as follows:

1. **Sample local ancestry**: The local ancestry states along the genome are sampled according to a Hidden Markov Model (HMM) with global ancestry proportions π and transition probabilities determined by the admixture time λ. More details about the HMM can be found in section 2.2.6.

2. **Thread the target sample into the genealogy**: The target sample is integrated into an existing genealogy of reference samples based on the sampled local ancestry and corresponding coalescence rates. The probability of coalescing with a particular lineage is calculated based on the descendant composition of the coalescing lineage. Further details on the likelihood associated with coalescence with a lineage can be found in the next subsection.

3. **Construct the final genealogy**: The resulting genealogy consists of the original reference samples plus the target sample at each genomic position.

Coalescence rates with a 'mixed' lineage

As demonstrated in Section 1.2.2, the properties of the coalescent provide a closed-form solution for the Maximum Likelihood Estimate (MLE) of coalescence rates when only one reference population is considered. However, when dealing with lineages whose descendants originate from a mixture of populations, computing the likelihood becomes more complex.

Such mixed lineages arise primarily from two factors: (1) historical gene-flow or population structure among reference samples and (2) uncertainty in genealogy inference from genetic variation data. To illustrate this, we revisit the case of Denisovan admixture in Papuans. Suppose we analyze a Papuan individual ($\approx 5\%$ Denisovan ancestry) alongside other Papuans (who also carry some Denisovan ancestry) and archaic Denisovans as reference populations. In this scenario, the target Papuan may coalesce with lineages containing both Papuan and Denisovan samples. This could be due to the merging of Denisovan and human populations over 700,000 years ago, Denisovan admixture in Papuan reference samples more recently, or simply uncertainty in tree topology and dating estimates in genealogy inference methods, leading to noisier split times.

To compute the likelihood of a lineage coalescing with a mixed lineage, we leverage the descendant composition of lineages - i.e., the reference population assignments at the leaves of the tree - to inform the coalescence rate with the target. This likelihood is calculated in two steps: (1) determining the population of the lineage and (2) sampling the coalescence time conditioned on this information. Specifically, we assume that the population label of a lineage is drawn uniformly from the set of its descendants, using the local tree to establish this distribution. For instance, if a lineage has five Papuan and one Denisovan descendant, it is assigned to the Papuan population with probability $\frac{5}{6}$ and to the Denisovan population with probability $\frac{1}{6}$. Given the population label, the coalescence time with the target follows an exponential distribution, with a rate determined by the coalescence rate between the target and the sampled population. This approach integrates naturally with the EM algorithm introduced later. By treating the lineage label as a latent variable, we effectively "complete" the data, enabling analytical updates at each iteration of the algorithm.

Beyond its computational efficiency, we believe this approach is also one of the simplest ways to model 'mixed' lineages. Instead of relying on knowledge of the population history in the reference population, our method samples descendants uniformly. This increases robustness to errors in genealogy inference by preventing abrupt lineage label switches. Additionally, in a clean split model, when populations merge, their coalescence rates become equal, rendering the choice of averaging inconsequential. Finally, under the assumption of approximately equal sample sizes and genetic drift across reference populations — and without prior knowledge of the genetic history of reference samples - this weighting provides an unbiased likelihood of coalescence with a given lineage. Future work can relax some of these assumptions with improved weighting schemes.

This approach of randomly sampling a descendant and computing the pairwise likelihood leads to a mixture of exponential weighted by the descendants. It is important to note that this differs from a simpler composite likelihood approach, which assumes pairwise independence between lineages and, by not conditioning

on the inferred tree topology, overcounts non-independent events, leading to overconfident estimates and reduced statistical power.

Additionally, the maximum likelihood estimation (MLE) of coalescence rates derived in Section 1.2.2 assumes independence of trees along the genome, which is an unrealistic assumption. To address this, we employ a Hidden Markov Model (HMM) to account for ancestry switches along the genome, modeling each location as Markovian. This approach relaxes the independence assumption and improves the robustness and accuracy of our likelihood estimation.

Likelihood of a single coalescence event

Under our model, we derive the likelihood of a single coalescence event involving a lineage whose descendants are drawn from a mixture of populations. Let the coalescing clade or lineage have n_g descendants from population g, with the total number of descendants across R reference populations given by $\sum_{g=1}^{R} n_g$. The probability density for observing a coalescence event at time t, given the matrix of coalescence rates per group per epoch γ, is expressed as:

$$P(t|\gamma) = \frac{\sum_{g=1}^{R} \gamma_g(m) n_g}{\sum_{g=1}^{R} n_g} e^{-\sum_{g=1}^{R} \gamma_g \cdot O_g(t)}, \tag{2.7}$$

where, n_g is the number of descendants in group g, $\gamma_g(m)$ is the coalescence rate of group g during epoch m, and $\gamma_g \cdot O_g(t)$ is the dot product (over time epochs) of the coalescence rates γ_g for group g and opportunity $O_g(t)$, defined as the sum of branch lengths from the previous coalescence time to t for all group g individuals.

Lemma. This weighting is equivalent to randomly sampling a descendant from the coalescing lineage and computing the pairwise coalescence likelihoods.

Proof. Let q be a categorical random variable that indicates which individual coalesces with the target. The probability of selecting the k-th descendant is given by:

$$P(q = k \mid \gamma) = \frac{1}{\sum_{g=1}^{R} n_g}, \tag{2.8}$$

where n_g is the number of descendants in group g. The joint probability of observing a coalescence event at time t and the descendant $q = k$, assuming conditional independence between t and q, is expressed as:

$$\begin{aligned} P(t, q = k \mid \gamma) &= P(t \mid q = k, \gamma) P(q = k \mid \gamma) \\ &= \gamma_{g(k)}(m) e^{-\sum_{g=1}^{R} \gamma_g \cdot O_g(t)} \frac{1}{\sum_{g=1}^{R} n_g}, \end{aligned} \tag{2.9}$$

where $P(t \mid q = k, \gamma)$ is derived from the pairwise coalescence likelihood. Summing over all possible descendants in the coalescing clade, we obtain the following expression for the likelihood of observing a coalescence event at time t:

$$\begin{aligned} P(t \mid \gamma) &= \sum_{k \in \text{Clade}} P(t, q = k \mid \gamma) \\ &= \frac{\sum_{g=1}^{R} \gamma_g(m) n_g}{\sum_{g=1}^{R} n_g} e^{-\sum_{g=1}^{R} \gamma_g \cdot O_g(t)}. \end{aligned} \tag{2.10}$$

This matches the expression in Equation 2.7.

2.2.3 EM derivation: terminology

We aim to fit a mixture model to cluster the genome based on variations in coalescence rates. To reduce computational costs, we divide the genome into T small grids and focus on inferring local ancestry at these grid points, rather than at every individual location on the genome (by default, grid points are equally spaced 10kb apart). Additionally, we approximate coalescence rates as piecewise constant functions over time, defined by E time epochs that are equally spaced on a logarithmic scale [68]. For the derivation of the EM algorithm, we use the following terminology:

Fixed constants used in derivation

1. C: Total number of coalescence events between the target lineage and reference samples across the entire genome.
2. T: Total number of grid points.
3. A: Number of clusters assumed for the target lineage.

4. E: Total number of time epochs in which the coalescence rates are assumed to be piecewise constant (by default, $E = 20$).
5. N: Total number of reference individuals in the genealogy.
6. R: Total number of reference groups in the genealogy.
7. N_{li}: Number of reference individuals subtending the lineage, which coalesces with the target lineage in the i-th coalescence event at the l-th grid position.
8. N_c: Number of reference individuals subtending the lineage corresponding to the c-th coalescence event.

Subscripts/Superscripts used in derivation

1. l: Index of the grid position.
2. i: Index of the coalescence event involving the target in the local tree.
3. g: Predefined reference group (e.g., Papuans, Denisovans, etc.).
4. j: Index of the cluster or component in the mixture model.
5. m: Index of the time epoch, measured backwards in time.
6. c: Index of the coalescence event involving the target in the entire genealogy.
7. υ: υ-th EM iteration.

Input data

1. $t_{li}, t_c \in \mathbb{R}^+$: Coalescence time, where t_{li} denotes the time (measured relative to the previous coalescence event) for the i-th coalescence event at the l-th grid position, and t_c represents the time for the c-th coalescence event.
2. $n_{lig}, n_{cg} \in \mathbb{Z}^+$: Number of individuals from group g involved in a coalescence event, where n_{lig} represents the number of individuals from group g participating in the i-th coalescence event at the l-th grid position, and n_{cg} denotes the number of individuals from group g involved in the c-th coalescence event.
3. $Q_l \in \mathbb{Z}^+$: Total number of coalescence events between the target lineage and reference samples within the tree corresponding to the l-th grid position.

4. $e_{li}, e_c \in \{1, 2, 3, \ldots, E\}$: Time epoch during which a coalescence event occurs, where e_{li} denotes the epoch of the i-th coalescence event at the l-th grid position, and e_c represents the epoch of the c-th coalescence event.
5. $r_l \in \mathbb{R}^+$: Genetic distance between the $(l-1)$-th and l-th grid positions (in Morgans).

Mixture model parameters

1. $\gamma \in \mathbb{R}_+^{A \times R \times E}$: Coalescence rate matrix, where $\gamma_{jg}(m)$ represents the coalescence rate between reference group g and the target in the j-th mixture component during epoch m.
2. $\pi \in [0, 1]^A$: Cluster proportions, where $\pi_j = P(z_l = j)$ denotes the proportion of the j-th mixture component. Additionally, $0 \leq \pi_j \leq 1, \quad \sum_{j=1}^{A} \pi_j = 1$.
3. $\lambda \in \mathbb{R}^+$: Admixture time, measured in generations.

Hidden variables in the EM

1. $z \in \{1, 2, \ldots, A\}^T$: Local ancestry vector, where z_l represents the local ancestry of the target at the l-th grid position, z_c denotes the same for the c-th coalescence event.
2. $q \in \{1, 2, \ldots, N\}^C$: Reference individual indices, where q_{li} denotes the index of the reference individual that coalesces in the i-th coalescence event at the l-th grid position, q_c denotes the same for c-th coalescence event.
3. $q' \in \{1, 2, \ldots, R\}^C$: Reference individual groups, where q'_{li} denotes the population group of the reference individual coalescing in the i-th coalescence event at the l-th grid position, q'_c denotes the same for c-th coalescence event.
4. $\omega_l \in \mathbb{R}^+$: Number of recombination events since admixture, occurring between the $(l-1)$-th and l-th grid positions.

2.2.4 EM derivation: the likelihood

The "incomplete-data likelihood"

We aim to formalize the likelihood for the mixture model described in Section 2.2.2. Following the approach of the EM algorithm for GMMs outlined in Section 2.2.1, we first define the "incomplete-data likelihood," which is based solely on the directly observable quantities. These include: (1) The vector of coalescence times for each event between the target lineage and reference samples, denoted as $t = (t_1, t_2, \ldots, t_C)$. (2) The corresponding vector of time epochs, $e = (e_1, e_2, \ldots, e_C)$. (3) The clade distribution vector, $n_g = (n_{1g}, n_{2g}, \ldots, n_{Cg})$, which records the number of individuals from group g corresponding to each coalescence event. The incomplete-data likelihood, which incorporates these observable quantities, is expressed as:

$$P(t|\gamma, \pi, \lambda) = \sum_z P(t|z, \gamma, \pi, \lambda) P(z|\gamma, \pi, \lambda) \tag{2.11}$$

Here, z is a hidden categorical random variable that determines the "type" of coalescence history for the target individual. These coalescence histories are distinguished by the different coalescence rates between the target individual and reference samples. The parameters to be inferred include γ, π, and λ: γ represents the coalescence rate matrix, π denotes the mixture proportions, and λ is the admixture time. See section 2.2.3 for their definitions. To compute the incomplete-data likelihood, we sum over all possible configurations of the hidden variable z using $\sum_z$. Notably, given the value of z, the conditional probability $P(t|z, \gamma, \pi, \lambda)$ can be factorized across coalescence events.

$$P(t|\gamma, \pi, \lambda) = \sum_z \left(\prod_{c=1}^{C} P(t_c|z_c, \gamma, \pi, \lambda) \right) P(z|\gamma, \pi, \lambda) \tag{2.12}$$

We can now utilize our model as described in Section 2.2.2 and Equation 2.7 to get $P(t_c|z_c, \gamma, \pi, \lambda)$:

$$P(t|\gamma, \pi, \lambda) = \sum_{z:z_c=j} \left(\prod_{c=1}^{C} \frac{\sum_{g=1}^{R} \gamma_{jg}(e_c) n_{cg}}{\sum_{g=1}^{R} n_{cg}} e^{-\sum_{g=1}^{R} \gamma_{jg} \cdot O_g(t)} \right) P(z|\gamma, \pi, \lambda) \tag{2.13}$$

Similar to equation 2.7, $\gamma_{jg} \cdot O_g(t)$ is the dot-product (over time epochs) between the coalescence rates of the j-th component, g-th reference group and Opportunity of the g-th group.

The "complete-data likelihood"

As with the GMM, the incomplete-data likelihood in Equation 2.13 is not directly optimizable due to its complex form. To address this, we employ the EM algorithm, which utilizes the complete-data likelihood to maximize the incomplete-data likelihood. This assumes access to complete information, including both observable quantities and hidden variables. In our case, hidden variables to facilitate computations in the EM algorithm include: (1) $z = (z_1, z_2, \ldots, z_C)$, which records the mixture component assignment for the c-th coalescence event. (2) $q = (q_1, q_2, \ldots, q_C)$, which identifies the reference individual involved in the c-th coalescence event. (3) $\omega = (\omega_1, \omega_2, \ldots, \omega_T)$, which identifies number of recombination events since admixture between two genomic positions (T is the total number of genomic positions considered).

The complete-data likelihood represents the joint probability of observing the coalescence times t, along with the hidden variables z and q, given the model parameters γ, π, and λ. We express this likelihood as the product of probabilities for each coalescence event, where the product is across all coalescence events (relating the target) c, clusters j, reference individuals k, and time epochs m:

$$\begin{aligned} P(t, z, q|\gamma, \pi, \lambda) &= P(z|\pi, \gamma, \lambda) P(t, q|z, \pi, \gamma, \lambda) \\ &= P(z|\pi, \lambda) \prod_{c=1}^{C} P(t_c, q_c|z_c, \gamma) \\ &= \underbrace{P(z|\pi, \lambda)}_{\text{Term 1}} \prod_{c=1}^{C} \prod_{j=1}^{A} \prod_{k=1}^{N_c} \prod_{m=1}^{E} \underbrace{P(t_c, q_c = k|z_c = j, \gamma)^{\mathbb{1}_{\{z_c=j\}} \mathbb{1}_{\{q_c=k\}} \mathbb{1}_{\{e_c=m\}}}}_{\text{Term 2}} \end{aligned} \tag{2.14}$$

Here, C represents the total number of coalescence events involving the target individual, A is the number of assumed clusters or components, N_c denotes the number of reference individuals corresponding to the c-th coalescence event, and E is the total number of time epochs. The indicator functions $\mathbb{1}_{\{z_c=j\}}$, $\mathbb{1}_{\{q_c=k\}}$, and $\mathbb{1}_{\{e_c=m\}}$ ensure that probability contributions are assigned to the appropriate cluster, reference individual, and time epoch, respectively. The complete-data likelihood is divided into two terms. The first term depends on the cluster proportions π and the admixture time λ, while the second term depends on the coalescence rates γ. Notably,

the second term simplifies to a product over coalescence events due to the assumption of conditional independence of coalescence times, t_c given the hidden variable z_c.

Coalescence events in an ARG can span more than one tree, particularly for nodes lower in the tree, which are less likely to be affected by recombination and thus persist longer along the genome. The likelihood in Equation 2.14 can be reformulated as a product over genomic positions. To achieve this, we propose a windowing approach that partitions the genome into grids and appropriately scales the likelihood based on the persistence of a coalescence event along the genome. This approach enables us to maintain the structure of Equation 2.14 while transitioning from a formulation based on individual coalescence events to one that operates along genomic segments:

$$\begin{aligned} &P(t, z, q|\gamma, \pi, \lambda) = \\ &= P(z|\pi, \lambda) \prod_{l=1}^{T} \prod_{j=1}^{A} \prod_{c=1}^{C} \prod_{k=1}^{N_c} \prod_{m=1}^{E} \left(P(t_c, q_c = k | z_c = j, \gamma)^{\mathbb{1}_{\{z_c=j\}} \mathbb{1}_{\{q_c=k\}} \mathbb{1}_{\{e_c=m\}}} \right)^{\frac{\mathbb{1}_{\{w_c=l\}}}{K_c}} \end{aligned} \tag{2.15}$$

where, $\mathbb{1}_{\{w_c=l\}}$ is an indicator random variable that identifies whether the l-th grid point overlaps with the node corresponding to the c-th coalescence event, and K_c represents the total number of grid points spanned by the node associated with the c-th coalescence event (referred to as node persistence). The likelihood in Equation 2.15 is scaled by $\frac{1}{K_c}$ to ensure that, when multiplied over all grids, it remains equivalent to the original likelihood in Equation 2.14. The likelihood in Equation 2.15 can be simplified further by considering only the coalescence events in the tree that overlap with the l-th grid position.

$$\begin{aligned} &P(t, z, q|\gamma, \pi, \lambda) = \\ &= P(z|\pi, \lambda) \prod_{l=1}^{T} \prod_{j=1}^{A} \prod_{i=1}^{Q_l} \prod_{k=1}^{N_{li}} \prod_{m=1}^{E} \left(P(t_{li}, q_{li} = k | z_{li} = j, \gamma)^{\mathbb{1}_{\{z_{li}=j\}} \mathbb{1}_{\{q_{li}=k\}} \mathbb{1}_{\{e_{li}=m\}}} \right)^{\frac{1}{K_{li}}} \end{aligned} \tag{2.16}$$

where, instead of taking the product over all coalescence events in the ARG, we consider only the coalescence events within the local tree at the l-th grid position. Consequently, coalescence events are indexed by the grid position l and the order of the coalescence event within the tree, denoted as i. Here, Q_l represents the total number of coalescence events between the target lineage and reference samples

within the tree corresponding to the l-th grid position. Further details about the calculation of node persistence, K_{li}, can be found in Section 2.3.7.

A approximation to the "true" likelihood

The hidden variable z_{li} in Equation 2.16 is indexed over coalescence events in the ARG. This indexing adds significant complexity to the likelihood computation and reduces its interpretability, as the hidden variables do not correspond directly to the local ancestry of the target individual at specific genomic positions. Instead, they remain tied to individual coalescence events, making it challenging to efficiently model ancestry transitions along the genome. To mitigate this, we introduce a approximation to the likelihood in Equation 2.16. Specifically, we replace z_{li}, which represents the cluster assignment for the i-th coalescence event at the l-th grid position, with a single cluster assignment z_l for the entire grid position l. This approximation assumes that all coalescence events occurring at a given genomic position share the same cluster assignment. By reducing the number of hidden variables indexed over coalescence events, this approximation significantly simplifies the likelihood computation and makes it more interpretable.

$$\begin{aligned} &P(t, z, q|\gamma, \pi, \lambda) \approx \\ &\approx P(z|\pi, \lambda) \prod_{l=1}^{T} \prod_{j=1}^{A} \prod_{i=1}^{Q_l} \prod_{k=1}^{N_{li}} \prod_{m=1}^{E} \left(P(t_{li}, q_{li} = k|z_l = j, \gamma)^{\mathbb{1}_{\{z_l=j\}}\mathbb{1}_{\{q_{li}=k\}}\mathbb{1}_{\{e_{li}=m\}}}\right)^{\frac{1}{K_{li}}} \end{aligned} \tag{2.17}$$

While this approximation is computationally efficient (as we show later in section 2.2.6 due to HMMs), it assumes that all coalescence events at a given position share the same ancestry cluster. This simplification may not fully capture the complexity of human demographic history, which involves multiple admixture and migration events. Retaining coalescence event-level indexing for z could provide a richer model that better distinguishes older from recent ancestry shifts. However, we believe that this comes with significant computational challenges. Future work could explore this direction to achieve a more fine-grained representation of human ancestry, capturing the sequential impact of migration waves across time.

Accounting for uncertain coalescence events

Genealogies inferred from genetic variation data come with the added complexity of uncertainty, as they are not always entirely accurate. The support for a particular node in the tree is stronger when there is a mutation mapped just above that node. To account for this uncertainty, we scale the likelihood by the confidence in each coalescence event. Specifically, as we move backward in time along the target lineage, we count the number of coalescence events that occur before encountering a mutation. All coalescence events between successive mutations are then weighted based on the number of such events. This adjustment is ad hoc, but improves the robustness of our method to errors in inferred genealogies, as demonstrated in empirical analyses. Overall, this results in scaling the likelihood in Equation 2.17 by an additional term, H_{li}, which calculates the effective contribution of a coalescence event to the likelihood depending on its certainty. The updated likelihood is now given as:

$$\begin{aligned} &P(t, z, q|\gamma, \pi, \lambda) \approx \\ &\approx \underbrace{P(z|\pi, \lambda)}_{\text{Term 1}} \prod_{l=1}^{T} \prod_{j=1}^{A} \prod_{i=1}^{Q_l} \prod_{k=1}^{N_{li}} \prod_{m=1}^{E} \underbrace{\left(P(t_{li}, q_{li} = k|z_l = j, \gamma)^{\mathbb{1}_{\{z_l=j\}}\mathbb{1}_{\{q_{li}=k\}}\mathbb{1}_{\{e_{li}=m\}}}\right)^{\frac{1}{B_{li}}}}_{\text{Term 2}} \end{aligned} \tag{2.18}$$

Here, $B_{li} = H_{li}K_{li}$ represents the effective scaling factor that accounts for both node persistence and the certainty of coalescence events in the inferred genealogies. Further details on the calculation of this weighting can be found in Section 2.3.8.

As discussed earlier, Term 1 and Term 2 are handled separately. We use the expectation-maximization (EM) algorithm to optimize Term 1 for estimating admixture time λ and cluster proportions π and Term 2 for estimating coalescence rates γ. In the following sections, we derive the EM updates for Term 1 (see Section 2.2.6) and Term 2 (see Section 2.2.5) independently.

2.2.5 EM derivation: inference of coalescence rates

We write the likelihood Term 2 in Equation 2.18 which only depends on coalescence rates as follows:

$$P(t, q|z, \gamma) = \prod_{l=1}^{T} \prod_{j=1}^{A} \prod_{i=1}^{Q_l} \prod_{k=1}^{N_{li}} \prod_{m=1}^{E} P(t_{li}, q_{li} = k|z_l = j, \gamma)^{\frac{\mathbb{1}_{\{z_l=j\}}\mathbb{1}_{\{q_{li}=k\}}\mathbb{1}_{\{e_{li}=m\}}}{B_{li}}} \tag{2.19}$$

where, the indicator random variables $\mathbb{1}_{\{z_l=j\}}$, $\mathbb{1}_{\{e_{li}=m\}}$ and, $\mathbb{1}_{\{q_{li}=k\}}$ are present to simplify the equation when writing it down for all the trees and coalescence events in that tree. The log-likelihood is given by:

$$\log P(t,q|z,\gamma) = \sum_{l=1}^{T}\sum_{j=1}^{A}\sum_{i=1}^{Q_l}\sum_{k=1}^{N_{li}}\sum_{m=1}^{E} \frac{\mathbb{1}_{\{q_{li}=k\}}\mathbb{1}_{\{z_l=j\}}\mathbb{1}_{\{e_{li}=m\}}\log P(t_{li}, q_{li}=k|z_l=j,\gamma)}{B_{li}} \tag{2.20}$$

We can simplify the likelihood by uniformly sampling a descendant from the coalescing lineage and considering its pairwise coalescence likelihood. See Section 2.2.2 for a detailed explanation and the underlying motivation for this model.:

$$\begin{aligned} P(t_{li}, q_{li}=k \mid z_l=j,\gamma) &= P(t_{li} \mid \gamma, z_l=j, q_{li}=k)P(q_{li}=k \mid z_l=j,\gamma) \\ P(t_{li}, q_{li}=k \mid z_l=j,\gamma) &= \gamma_{jg(k)}(m)e^{-\sum_{g=1}^{R}\gamma_{jg}\cdot O_{lg}[i]}\frac{1}{\sum_{g=1}^{R} n_{lig}} \end{aligned} \tag{2.21}$$

where $g(k)$ denotes the group assignment (among R groups) of the k-th individual in the reference panel, n_{lig} represents the number of group g individuals coalescing in the i-th coalescence event at position l, and $O_{lg}[i]$ is the sum of branch lengths from the $(i-1)$-th to the i-th coalescence event corresponding to all group g individuals. Overall, $P(t_{li}, q_{li}=k \mid z_l=j,\gamma)$ can be simplified by focusing only on the terms dependent on the cluster membership j, as follows:

$$\begin{aligned} P(t_{li}, q_{li}=k|\gamma, z_l=j) &= \frac{\gamma_{jg(k)}(m)e^{-\sum_{g=1}^{R}\gamma_{jg}\cdot O_{lg}[i]}}{\sum_{g=1}^{R} n_{lig}} \\ &\propto \gamma_{jg(k)}(m) \times e^{-\sum_{g=1}^{R}\gamma_{jg}\cdot O_{lg}[i]} \end{aligned} \tag{2.22}$$

In the following sections, we derive the E-step and M-step for the likelihood presented in Equation 2.19.

E-step

In the E-step we calculate the expected log-likelihood, where the expectation is with respect to the hidden variables conditioned on the data and previous parameter estimates. We therefore proceed by stating the expected log-likelihood in our case:

$$\mathbb{E}_{z,q|\gamma^{\upsilon-1},\lambda^{\upsilon-1},\pi^{\upsilon-1},t}\Big(\sum_{l=1}^{T}\sum_{j=1}^{A}\sum_{i=1}^{Q_l}\sum_{k=1}^{N_{li}}\sum_{m=1}^{E} \frac{\mathbb{1}_{\{q_{li}=k\}}\mathbb{1}_{\{z_l=j\}}\mathbb{1}_{\{e_{li}=m\}}\log P(t_{li}, q_{li}=k|z_l=j,\gamma)}{B_{li}}\Big) \tag{2.23}$$

We start simplifying the expected log-likelihood by taking the expectation inside the summations, and transforming the expected value of indicator random variables as probabilities.

$$= \sum_{l=1}^{T}\sum_{j=1}^{A}\sum_{i=1}^{Q_l}\sum_{k=1}^{N_{li}}\sum_{m=1}^{E} \mathbb{1}_{\{e_{li}=m\}} P(q_{li}=k, z_l=j|\gamma^{v-1},\lambda^{v-1},\pi^{v-1},t)\frac{\log(P(t_{li},q_{li}=k|z_l=j,\gamma))}{B_{li}} \tag{2.24}$$

Simplifying further and substituting the value of $\log(P(t_{li}, q_{li} = k|z_l = j, \gamma))$ from Equation 2.22. Note that we omit the constants of proportionality as they do not depend on the coalescence rates and thus will not affect the maximization in the M-step.

$$\begin{aligned}
&= \sum_{l=1}^{T}\sum_{j=1}^{A}\sum_{i=1}^{Q_l}\sum_{k=1}^{N_{li}}\sum_{m=1}^{E} \Bigg(P(q_{li}=k|\gamma^{v-1},t,z_l=j)P(z_l=j|t,\gamma^{v-1},\lambda^{v-1},\pi^{v-1}) \\
&\quad \times \mathbb{1}_{\{e_{li}=m\}} \frac{\log(P(t_{li},q_{li}=k|z_l=j,\gamma))}{B_{li}} \Bigg) \\
&= \sum_{l=1}^{T}\sum_{j=1}^{A}\sum_{i=1}^{Q_l}\sum_{k=1}^{N_{li}}\sum_{m=1}^{E} \Bigg(P(q_{li}=k|\gamma^{v-1},t,z_l=j)P(z_l=j|t,\gamma^{v-1},\lambda^{v-1},\pi^{v-1}) \\
&\quad \times \mathbb{1}_{\{e_{li}=m\}} \frac{\log(\gamma_{jg(k)}(m)) - \sum_{g=1}^{R}\gamma_{jg}\cdot O_{lg}[i]}{B_{li}} \Bigg)
\end{aligned} \tag{2.25}$$

Note the probability $P(q_{li} = k|\gamma^{v-1}, t, z_l = j)$ does not depend on admixture time and proportion. We can simplify the expected log-likelihood further and switch the individual assignment hidden variable q_{li} to more rapidly computable group assignment hidden variable q'_{li}.

$$\begin{aligned}
&\sum_{l=1}^{T}\sum_{j=1}^{A}\sum_{i=1}^{Q_l}\sum_{k=1}^{N_{li}}\sum_{m=1}^{E} \Bigg(P(q_{li}=k|\gamma^{v-1},z_l=j)P(z_l=j|t,\gamma^{v-1},\lambda^{v-1},\pi^{v-1}) \\
&\quad \times \mathbb{1}_{\{e_{li}=m\}} \frac{\log(\gamma_{jg(k)}(m))}{B_{li}} \Bigg) \\
&- \sum_{l=1}^{T}\sum_{j=1}^{A}\sum_{i=1}^{Q_l}\sum_{k=1}^{N_{li}}\sum_{m=1}^{E} \Bigg(P(q_{li}=k|\gamma^{v-1},z_l=j)P(z_l=j|t,\gamma^{v-1},\lambda^{v-1},\pi^{v-1}) \\
&\quad \times \mathbb{1}_{\{e_{li}=m\}} \frac{\sum_{g=1}^{R}\gamma_{jg}\cdot O_{lg}[i]}{B_{li}} \Bigg)
\end{aligned} \tag{2.26}$$

We group the individuals based on group assignment in the first term (red term)

and use $\sum_{k=1}^{N_{li}} P(q_{li} = k|\gamma^{\upsilon-1}, z_l = j) = 1$ in the second term (blue term).

$$\sum_{l=1}^{T}\sum_{j=1}^{A}\sum_{i=1}^{Q_l}\sum_{g=1}^{R}\sum_{m=1}^{E}\left(P(q'_{li} = g|\gamma^{\upsilon-1}, z_l = j)P(z_l = j|t, \gamma^{\upsilon-1}, \lambda^{\upsilon-1}, \pi^{\upsilon-1}) \times \mathbb{1}_{\{e_{li}=m\}}\frac{\log(\gamma_{jg}(m))}{B_{li}}\right)$$
$$-\sum_{l=1}^{T}\sum_{j=1}^{A}\sum_{i=1}^{Q_l}\sum_{m=1}^{E}\left(P(z_l = j|t, \gamma^{\upsilon-1}, \lambda^{\upsilon-1}, \pi^{\upsilon-1})\mathbb{1}_{\{e_{li}=m\}}\sum_{g=1}^{R}\frac{\gamma_{jg}\cdot O_{lg}[i]}{B_{li}}\right) \tag{2.27}$$

It is important to note we use the group-wise assignment, q'_{li} moving forward instead of the individual-wise assignment, q_{li}. Let us denote $U^{\upsilon-1}_{l,i,g,j,m}$ as the first probability in the red term above, corresponding to the membership "group g individual coalesces in the i-th coalescence event". And, let us denote $T^{\upsilon-1}_{l,j}$ as the first probability in the blue term above, corresponding to the local ancestry of the target. We use the forward-backward algorithm for HMMs to evaluate $T_{l,j}$ in Section 2.2.6. Whereas, we calculate $U^{\upsilon-1}_{l,i,g,j,m}$ based on our assumed model that the probability of coalescing with reference individual is directly proportional to its coalescence rate and for a group of individual on the weighted mean of their individual coalescence rates (see Section 2.2.2):

$$U^{\upsilon-1}_{l,i,g,j,m} = P(q'_{li} = g|\gamma^{\upsilon-1}, z_l = j) = \frac{\gamma^{\upsilon-1}_{jg}(m)n_{lig}}{\sum_{g=1}^{R}\gamma^{\upsilon-1}_{jg}(m)n_{lig}} \tag{2.28}$$

M-step

Once we have the expected log-likelihood we can maximize it in the M-step. We substitute $U^{\upsilon-1}_{l,i,g,j,m}$ and $T^{\upsilon-1}_{l,j}$ as hidden variables from Equation 2.28, and the HMM forward-backward expectation into Equation 2.27. We then maximize with respect to $\gamma_{jg}(m)$ the resulting expression:

$$\sum_{l=1}^{T}\sum_{i=1}^{Q_l}\sum_{j=1}^{A}\sum_{g=1}^{R}\sum_{m=1}^{E}\left(U^{\upsilon-1}_{l,i,g,j,m}T^{\upsilon-1}_{l,j}\mathbb{1}_{\{e_{li}=m\}}\frac{\log(\gamma_{jg}(m))}{B_{li}}\right)$$
$$-\sum_{l=1}^{T}\sum_{i=1}^{Q_l}\sum_{j=1}^{A}\sum_{m=1}^{E}\left(T^{\upsilon-1}_{l,j}\mathbb{1}_{\{e_{li}=m\}}\sum_{g=1}^{R}\frac{\gamma_{jg}\cdot O_{lg}[i]}{B_{li}}\right) \tag{2.29}$$

Differentiating with respect to $\gamma_{jg}(m)$ and setting it to 0 we get:

$$\sum_{l=1}^{T}\sum_{i=1}^{Q_l}\left(\frac{\mathbb{1}_{\{e_{li}=m\}}U^{\upsilon-1}_{l,i,g,j,m}T^{\upsilon-1}_{l,j}}{\gamma_{jg}(m)B_{li}}\right) - \sum_{l=1}^{T}\sum_{i=1}^{Q_l}\left(\frac{\mathbb{1}_{\{e_{li}=m\}}T^{\upsilon-1}_{l,j}O_{lg}[i]}{B_{li}}\right) = 0 \tag{2.30}$$

$$\gamma_{jg}^{\upsilon}(m) = \frac{\sum_{l=1}^{T}\sum_{i=1}^{Q_l}\mathbb{1}_{\{e_{li}=m\}}U_{l,i,g,j,m}^{\upsilon-1}T_{l,j}^{\upsilon-1}/B_{li}}{\sum_{l=1}^{T}\sum_{i=1}^{Q_l}\mathbb{1}_{\{e_{li}=m\}}T_{l,j}^{\upsilon-1}O_{lg}[i]/B_{li}}$$
$$\gamma_{jg}^{\upsilon}(m) = \frac{\sum_{l=1}^{T}\sum_{i=1}^{Q_l}\mathbb{1}_{\{e_{li}=m\}}\frac{\gamma_{jg}^{\upsilon-1}(m)n_{lig}}{\sum_{g=1}^{R}\gamma_{jg}^{\upsilon-1}(m)n_{lig}}T_{l,j}^{\upsilon-1}/B_{li}}{\sum_{l=1}^{T}\sum_{i=1}^{Q_l}\mathbb{1}_{\{e_{li}=m\}}T_{l,j}^{\upsilon-1}O_{lg}[i]/B_{li}} \tag{2.31}$$

Notice that the coalescence rate estimates closely resemble the pairwise maximum likelihood estimate for two lineages derived in Equation 1.10 in Section 1.2.2.

2.2.6 EM derivation: inference for admixture time and proportions

We assume a Markov model to express the likelihood for Term 1 in Equation 2.18. In addition to the local ancestry hidden variable (z), we introduce another hidden variable, ω_l, which records the number of recombination events since admixture between positions $l-1$ and l. A user-provided recombination map is assumed to be available, specifying the genetic distance between two grid points, denoted as r_l. The "complete-data likelihood" for this term can then be written as:

$$P(z,\omega \mid \lambda,\pi) = \prod_{l=1}^{T}\Big(\prod_{j=1}^{A}\pi_j^{I\{z_l=j,\omega_l>0\}}\Big)P(z_{l-1})^{I\{\omega_l=0\}}\frac{(\lambda r_l)^{\omega_l}e^{-\lambda r_l}}{\omega_l!}, \tag{2.32}$$

where the likelihood for the number of recombination events follows a Poisson counting process. Notably, a local ancestry change occurs only if there is at least one recombination event. As a boundary condition, we impose $P(\omega_l > 0) = 1$ for $l = 1$. The model parameters are λ, representing the admixture time, and π, the global cluster proportions. The corresponding log-likelihood can be expressed as:

$$\log P(z,\omega_l \mid \lambda,\pi) = \sum_{l=1}^{T}\Big(\sum_{j=1}^{A} I\{z_l=j,\omega_l>0\}\log\pi_j\Big) + \omega_l\log(\lambda r_l) - \lambda r_l + \text{constants}. \tag{2.33}$$

Note in the above equation, we have removed terms (called them 'constants') which do not depend on λ or π, as they will be discarded in the M-step.

E-step

To perform the EM algorithm, we compute the expected log-likelihood of the complete data, where the expectation is taken over the hidden variables conditioned on the observed data and the model parameters from the previous iteration (including γ).

$$\mathbb{E}_{z,\omega_l|\gamma^{v-1},\lambda^{v-1},\pi^{v-1},t}\Big(\sum_{l=1}^{T}\Big(\sum_{j=1}^{A} I\{z_l=j,\omega_l>0\}\log\pi_j\Big)+\omega_l\log\lambda r_l-\lambda r_l\Big) \tag{2.34}$$

We start simplifying the expected log-likelihood by taking the expectation inside the summation, and transforming the expected value of indicator random variables as probabilities.

$$\sum_{l=1}^{T}\Big(\sum_{j=1}^{A} P(z_l=j,\omega_l>0|\gamma^{v-1},\lambda^{v-1},\pi^{v-1},t)\log\pi_j\Big)+\mathbb{E}(\omega_l)log\lambda r_l-\lambda r_l \tag{2.35}$$

where the expectation is taken over the hidden variables conditioned on the data and previous iterations' parameter updates. We are interested in inferring three terms, (1) $P(z_l=j,\omega_l>0|\gamma^{v-1},\lambda^{v-1},\pi^{v-1},t)$, (2) $\mathbb{E}(\omega_l|\gamma^{v-1},\lambda^{v-1},\pi^{v-1},t)$ and (3) $P(z_l=j|t,\gamma^{v-1},\lambda^{v-1},\pi^{v-1})$. We start with the first term:

$$\begin{aligned}
&P(z_l=j,\omega_l>0|P^{v-1},t)=P(z_l=j|P^{v-1},t)P(\omega_l>0|z_l=j,P^{v-1},t)\\
&=P(z_l=j|P^{v-1},t)\sum_{i=1}^{A}P(\omega_l>0|z_l=j,z_{l-1}=i,P^{v-1},t)P(z_{l-1}=i|z_l=j,P^{v-1},t)\\
&=\sum_{i=1}^{A}P(z_l=j|P^{v-1},t)P(\omega_l>0|z_l=j,z_{l-1}=i,P^{v-1},t)P(z_{l-1}=i|z_l=j,P^{v-1},t)\\
&=\sum_{i=1}^{A}P(\omega_l>0|z_l=j,z_{l-1}=i,P^{v-1})P(z_l=j,z_{l-1}=i|P^{v-1},t)
\end{aligned} \tag{2.36}$$

where, $P^{v-1}=(\gamma^{v-1},\lambda^{v-1},\pi^{v-1})$ is a short-hand for the tuple of all previous parameters. We denote the second term (colored in blue) as $\epsilon_{ij}(l)$ as it denotes the probability of being in state i and state j at position $l-1$ and l respectively. Similar to the Baum-Welch algorithm we can solve for $\epsilon_{ij}(l)$ through the forward-backward pass in HMMs.

$$
\begin{aligned}
&\epsilon_{ij}(l) = P(z_l = j, z_{l-1} = i | P^{\upsilon-1}, t) \\
&= \frac{P(z_l = j, z_{l-1} = i, t | P^{\upsilon-1})}{P(t|P^{\upsilon-1})} \\
&= \frac{P(t_1.., t_{l-1}, z_{l-1} = i | P^{\upsilon-1}) P(z_l = j | z_{l-1} = i, P^{\upsilon-1}) P(t_l.., t_T | z_l = j, P^{\upsilon-1}) P(t_l | z_l, P^{\upsilon-1})}{P(t|P^{\upsilon-1})} \\
&= \frac{P(t_1.., t_{l-1}, z_{l-1} = i | P^{\upsilon-1}) P(z_l = j | z_{l-1} = i, P^{\upsilon-1}) P(t_l.., t_T | z_l = j, P^{\upsilon-1}) P(t_l | z_l, P^{\upsilon-1})}{\sum_{j=1}^{A} \sum_{i=1}^{A} P(t_1.., t_{l-1}, z_{l-1} = i | P^{\upsilon-1}) P(z_l = j | z_{l-1} = i, P^{\upsilon-1}) P(t_l.., t_T | z_l = j, P^{\upsilon-1}) P(t_l | z_l, P^{\upsilon-1})}
\end{aligned}
\tag{2.37}
$$

where, $P(t_1.., t_{l-1}, z_{l-1} = i | P^{\upsilon-1})$ is called the forward-pass and $P(t_l.., t_T | z_l = j, P^{\upsilon-1})$ is called the backward-pass. The transition probability for the HMM is given as follows:

$$
P(z_l = j | z_{l-1} = i, P^{\upsilon-1}) = \begin{cases} (1 - e^{-r_l \lambda^{\upsilon-1}}) \pi_j^{\upsilon-1}, & \text{if } i \neq j \\ e^{-r_l \lambda^{\upsilon-1}} + (1 - e^{-r_l \lambda^{\upsilon-1}}) \pi_j^{\upsilon-1}, & \text{if } i = j \end{cases} \tag{2.38}
$$

whereas the emission probability for the HMM $P(t_l | z_l, P^{\upsilon-1})$ only depends on $\gamma^{\upsilon-1}$ and follows from the scaled likelihood in Equation 2.19:

$$
P(t_l | z_l = j, P^{\upsilon-1}) = \Big(\frac{\sum_{g=1}^{R} \gamma_{jg}^{\upsilon-1}(m) n_{lig}}{\sum_{g=1}^{R} n_{lig}} e^{-\sum_{g=1}^{R} \gamma_{jg}^{\upsilon-1} \cdot O_{lg}[i]} \Big)^{\frac{1}{B_{li}}} \tag{2.39}
$$

The first term (highlighted in red) in Equation 2.36 represents the probability of at least one recombination event, conditioned on the local ancestry states z_{l-1} and z_l. It is expressed as follows:

$$
P(\omega_l > 0 | z_l = j, z_{l-1} = i, P^{\upsilon-1}) = \begin{cases} 1, & \text{if } i \neq j \\ \frac{(1 - e^{-r_l \lambda^{\upsilon-1}}) \pi_j^{\upsilon-1}}{e^{-r_l \lambda^{\upsilon-1}} + (1 - e^{-r_l \lambda^{\upsilon-1}}) \pi_j^{\upsilon-1}}, & \text{if } i = j \end{cases} \tag{2.40}
$$

The next term to be computed in the E-step is $E(\omega_l)$, where the expectation is taken over the hidden variables z_l and ω_l, conditioned on the observed data and the model parameters from the previous iteration. This expectation can be simplified as follows:

$$
\begin{aligned}
\mathbb{E}(\omega_l | P^{\upsilon-1}, t) &= \mathbb{E}(\omega_l | \omega_l > 0, P^{\upsilon-1}, t) P(\omega_l > 0 | P^{\upsilon-1}, t) \\
&= \mathbb{E}(\omega_l | \omega_l > 0, P^{\upsilon-1}) P(\omega_l > 0 | P^{\upsilon-1}, t) \\
&= \frac{\lambda^{\upsilon-1} r_l}{1 - e^{-\lambda^{\upsilon-1} r_l}} P(\omega_l > 0 | P^{\upsilon-1}, t)
\end{aligned}
\tag{2.41}
$$

In order to evaluate $P(\omega_l > 0|P^{v-1}, t)$ we do the following:

$$P(\omega_l > 0|P^{v-1}, t) = \sum_{j=1}^{A}\sum_{j=1}^{A} P(\omega_l > 0|z_l = j, z_{l-1} = i, P^{v-1}, t)P(z_l = j, z_{l-1} = i|P^{v-1}, t)$$
$$= \sum_{j=1}^{A}\sum_{j=1}^{A} P(\omega_l > 0|z_l = j, z_{l-1} = i, P^{v-1})\epsilon_{ij}(l) \quad (2.42)$$

The terms required in Equation 2.42 have already been calculated in Equations 2.37 and 2.40. As a boundary condition, we impose $P(\omega_l > 0) = 1$ for $l = 1$. Finally, to compute the local ancestry estimate $P(z_l = j \mid t, \gamma^{v-1}, \lambda^{v-1}, \pi^{v-1})$ needed in Equation 2.31, we perform a forward-backward pass:

$$P(z_l = j|t, \gamma^{v-1}, \lambda^{v-1}, \pi^{v-1}) = \frac{P(t_1.., t_{l-1}, z_{l-1} = i|P^{v-1})P(t_l.., t_T|z_l = j, P^{v-1})}{\sum_{j=1}^{A} P(t_1.., t_{l-1}, z_{l-1} = i|P^{v-1})P(t_l.., t_T|z_l = j, P^{v-1})} \quad (2.43)$$

where, $P(t_1.., t_{l-1}, z_{l-1} = i|P^{v-1})$ and $P(t_l.., t_T|z_l = j, P^{v-1})$ is the forward and backward probability respectively.

M-step

Once all the necessary equations in the E-step have been determined, we proceed to the M-step, where we maximize the expected log-likelihood. Specifically, we optimize with respect to λ and the cluster proportions π by maximizing the following equation:

$$\sum_{l=1}^{T}\Big(\sum_{j=1}^{A} P(z_l = j, \omega_l > 0|\gamma^{v-1}, \lambda^{v-1}, \pi^{v-1}, t)\log\pi_j\Big) + \mathbb{E}(\omega_l)log\lambda r_l - \lambda r_l \quad (2.44)$$

1. Differentiating with respect to π_j and setting to zero we get:

$$\pi_j^v \propto \sum_{l=1}^{T} P(z_l = j, \omega_l > 0|\gamma^{v-1}, \lambda^{v-1}, \pi^{v-1}, t)$$
$$\pi_j^v = \frac{\sum_{l=1}^{T} P(z_l = j, \omega_l > 0|\gamma^{v-1}, \lambda^{v-1}, \pi^{v-1}, t)}{\sum_{j=1}^{A}\sum_{l=1}^{T} P(z_l = j, \omega_l > 0|\gamma^{v-1}, \lambda^{v-1}, \pi^{v-1}, t)} \quad (2.45)$$

We get the term $P(z_l = j, \omega_l > 0|\gamma^{v-1}, \lambda^{v-1}, \pi^{v-1}, t)$ from the E-step Equation 2.36.

2. Differentiating with respect to λ and setting to zero we get:

$$\lambda^t = \frac{\sum_{l=1}^{T} \mathbb{E}(\omega_l | \gamma^{v-1}, \lambda^{v-1}, \pi^{v-1}, t)}{\sum_{l=1}^{T} r_l} \tag{2.46}$$

Note we get the expected number of recombinations from Equation 2.41 in E-step.

2.2.7 Summary of the algorithm

We present a summary of the EM algorithm for GhostBuster here:

Algorithm 1 EM Algorithm for GhostBuster

Input: Genealogies $\mathbf{X} = \{(t_1, n_1), (t_2, n_2), \ldots, (t_C, n_C)\}$ containing coalescence times and population counts for each coalescence event, number of components A, number of time epochs E, number of grid points T, total number of coalescence events C and number of reference groups R.
Preprocessing: Calculate likelihood scaling term B_c for each coalescence event $c = 1, 2, \ldots, C$, and r_l genetic distance between two grid points
Initialize: Coalescence rates $\gamma_j^{(0)}$, mixing coefficients $\pi_j^{(0)}$ for each component $j = 1, 2, \ldots, A$ and admixture time $\lambda^{(0)}$
Set: Iteration $\upsilon = 1$
repeat
    E-step:
    for each component $j = 1, 2, \ldots, A$ **do**
        for each data point (t_{li}, n_{li}) **do**
        ▷ t_{li} represent the coalescence time in l-th grid and i-th coalescence event
                                                ▷ $P^{\upsilon-1} = \{\gamma^{\upsilon-1}, \lambda^{\upsilon-1}, \pi^{\upsilon-1}\}$
            Compute the likelihood under the component:

$$P(t_l | z_l = j, P^{\upsilon-1}) = \Big(\frac{\sum_{g=1}^{R} \gamma_{jg}^{\upsilon-1}(m) n_{lig}}{\sum_{g=1}^{R} n_{lig}} e^{-\sum_{g=1}^{R} \gamma_{jg}^{\upsilon-1} \cdot O_{lg}[i]} \Big)^{\frac{1}{B_{li}}}$$

        end for
        Perform forward-backward pass with the HMM to get $P(z_l = j | X, P^{\upsilon-1})$
          see Equation 2.43
    end for
    M-step:
    for each component $j = 1, 2, \ldots, A$ **do**
        Update the coalescence rates:
          see Equation 2.31
        Update the cluster proportions:
          see Equation 2.45
        Update the admixture time:
          see Equation 2.46
    end for
    Increment: $\upsilon = \upsilon + 1$
until convergence
Output: Final parameters, $\{\gamma^{\upsilon}, \lambda^{\upsilon}, \pi^{\upsilon}\}$ and local ancestry posteriors, $P(z_l = j | X, P^{\upsilon})$ for $l = 1, 2, \ldots, T$ and $j = 1, 2, \ldots, A$

2.3 Method description

2.3.1 Graphic overview of the method

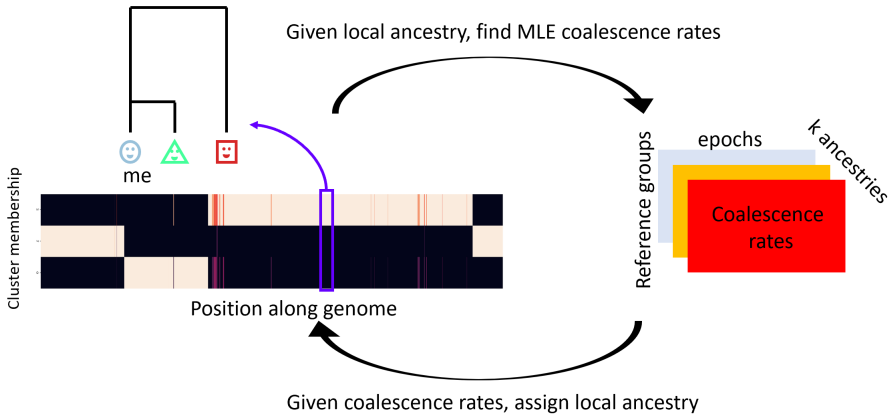


Figure 2.3: Overview of the GhostBuster algorithm. The EM algorithm computes a local maximum likelihood estimate (MLE) for the mixture model, clustering genomic regions based on variations in coalescence histories for a given target individual. In the E-step, the expected log-likelihood of the data is calculated, conditioned on the local tree topology and the current estimates of model parameters (proportions, admixture time, and coalescence rates). An HMM is then used to infer the local ancestry of the target individual, with emissions modeled using the likelihood of coalescence times and transitions modeled as a Poisson process dependent on recombination rates. In the M-step, the expected log-likelihood obtained from the E-step is maximized to update the model parameters. The EM algorithm ensures non-decreasing likelihood updates at each iteration, leading to convergence at a local optimum [136]. The likelihood is conditioned on the tree topology by carefully accounting for lineage contributions based on their descendants. This approach provides higher statistical power to detect admixture events by avoiding assumptions of independence among reference individuals or reliance on composite likelihoods. Detailed derivations of the likelihood and EM algorithm updates are presented in Section 2.2.

2.3.2 PCA visualization to visualize ancestral components

To visualize the clustering performed by GhostBuster, we perform Principal Component Analysis (PCA) on the coalescence count and opportunity matrices obtained from the genealogies. The coalescence count matrix is an $N_{\text{ref}} \times N_{\text{sites}}$ matrix, where the $(g, l)^{\text{th}}$ entry represents the total contribution of reference group g to the coalescence events at genomic position l. We aggregate the contributions from all coalescence events involving the target individual at genomic position l,

considering only those events that occur within the predefined start and end time epochs. These time epochs are aligned with the start and end epochs used in the GhostBuster analysis. These epochs control the time periods in which we expect to observe differences in coalescence patterns. We construct a similar matrix for the opportunity, also an $N_{\text{ref}} \times N_{\text{sites}}$ matrix, where each entry g, l records the total opportunity for reference group g at position l. The same time epochs are used as in the coalescence count matrix to ensure consistency.

The coalescence count and opportunity matrices are then stacked to form a $2N_{\text{ref}} \times N_{\text{sites}}$ matrix, which is mean-centered and standardized. We perform eigen-decomposition on this matrix to obtain the leading principal components. These principal components provide a visual interpretation of the clustering performed by GhostBuster, helping to better understand the patterns in coalescence across the genome.

2.3.3 Coancestry curves to date admixture event

Given the local ancestry inferred using GhostBuster, we use coancestry curves to accurately estimate the date of the admixture event. As outlined in Section 1.2.3, the joint probability of observing ancestry at a genetic distance g under a single-pulse admixture model is given by:

$$\begin{aligned} p_{AA}(g) &= \alpha^2 + \alpha(1-\alpha)\exp(-g\lambda) \\ p_{AB}(g) &= \alpha(1-\alpha) - \alpha(1-\alpha)\exp(-g\lambda) \\ p_{BB}(g) &= (1-\alpha)^2 + \alpha(1-\alpha)\exp(-g\lambda) \end{aligned} \tag{2.47}$$

where α is the genome-wide proportion of ancestry 'A', $1-\alpha$ of ancestry 'B' and λ is the admixture time in generations. We also consider admixture models with two-dates corresponding to two-pulses of gene flow. The joint probability of observing ancestry at a genetic distance g under a double admixture model is given by:

$$
\begin{aligned}
p_{AA}(g) &= \alpha_1^2(1-\alpha_3)^2 + \alpha_1^2\alpha_3(1-\alpha_3)\exp(-g\lambda_2) + \alpha_1\alpha_2(1-\alpha_3)\exp(-g\lambda_1)\\
p_{BB}(g) &= \alpha_2^2(1-\alpha_3)^2 + \alpha_2^2\alpha_3(1-\alpha_3)\exp(-g\lambda_2) + \alpha_1\alpha_2(1-\alpha_3)\exp(-g\lambda_1)\\
p_{AB}(g) &= p_{BA}(g) = \alpha_1\alpha_2(1-\alpha_3)^2 + \alpha_1\alpha_2\alpha_3(1-\alpha_3)\exp(-g\lambda_2) - \alpha_1\alpha_2(1-\alpha_3)\exp(-g\lambda_1)\\
p_{CA}(g) &= p_{AC}(g) = \alpha_3(1-\alpha_3)\alpha_1 - \alpha_3(1-\alpha_3)\alpha_1\exp(-g\lambda_2)\\
p_{CB}(g) &= p_{BC}(g) = \alpha_3(1-\alpha_3)\alpha_2 - \alpha_3(1-\alpha_3)\alpha_2\exp(-g\lambda_2)\\
p_{CC}(g) &= \alpha_3^2 + \alpha_3(1-\alpha_3)\exp(-g\lambda_2)
\end{aligned}
\tag{2.48}
$$

where, $\alpha_1(1-\alpha_3)$ is the genome-wide proportion of ancestry 'A', $\alpha_2(1-\alpha_3)$ of ancestry 'B', α_3 of ancestry 'C' and λ_1, λ_2 are the corresponding admixture times. Finally, we consider admixture models corresponding to constant migration rate continuous admixture. In this case the joint probability of observing ancestry at a genetic distance g is given by:

$$
\begin{aligned}
p_{AA}(g) &= \alpha^2 + \alpha(1-\alpha)\sum_{j=\lambda_e}^{\lambda_s} w_j \exp(-gj)\\
p_{AB}(g) &= \alpha(1-\alpha) - \alpha(1-\alpha)\sum_{j=\lambda_e}^{\lambda_s} w_j \exp(-gj)\\
p_{BB}(g) &= (1-\alpha)^2 + \alpha(1-\alpha)\sum_{j=\lambda_e}^{\lambda_s} w_j \exp(-gj)
\end{aligned}
\tag{2.49}
$$

where, $w_j = \frac{\mu(1-\mu)^{\lambda_s - j}}{1-(1-\mu)^{\lambda_s-\lambda_e+1}}$; λ_s and λ_e are the start and end time of the admixture, μ is the constant migration rate. Detailed derivation of the analytical coancestry curves can be found in the supplementary information of [51].

We fit the exponential distribution to all possible ancestry combinations and jointly maximize the log-likelihood. In the case of three-way or more complex admixture, we fit coancestry curves for each pair of ancestries separately, estimating the dates for each event independently. We use 20 block-jackknife across chromosomes to estimate standard errors when dating the admixture event. Additionally, we use diploid ancestry calls (summing the posteriors of two haplotypes) to prevent phasing errors from biasing the admixture dating.

Genetic drift can cause the fixation or loss of certain alleles, introducing a downward bias in the dating estimates derived from coancestry curves. This effect is particularly pronounced in small populations or when considering ancient events,

where the influence of drift is much greater. To mitigate this bias, we normalize the coancestry curves by the coancestry curves generated from randomly sampling haplotype pairs from different individuals, also known as pseudo-individuals. As previously done in Globetrotter [51], this approach corrects for the impact of genetic drift and provides an unbiased admixture date.

Another source of error in admixture dating can arise from inaccuracies in recombination maps. This is particularly relevant for dating ancient events, where errors on the scale of tens of kilo-bases in the recombination map can downward-bias the inferred date [137]. In our case, this issue does not affect the simulated examples, as we assume a perfect recombination map, but might have some impact on real data, especially the old admixture dates. However, due to the lack of suitable error estimates for the recombination map of the population being analyzed, we highlight this as a cautionary note for the interpretation of results from real data.

Finally, the effects of selection and the evolution of recombination maps can influence admixture dates; however, we assume these factors have a minimal impact on our real-data examples.

2.3.4 Joint analysis of multiple targets

Sometimes the reference populations can be themselves admixed. GhostBuster allows estimating the coalescence rates of the target lineage with reference populations conditioned on the local ancestry of the reference populations. It does so by counting coalescence events and opportunity separately for different ancestries in the admixed reference populations.

It is also possible that the reference populations share the same admixture event as the target individual. To account for this, we employ a two-step procedure to fit the admixture event. First, we fit a mixture model assuming unadmixed reference populations, assigning ancestry to all admixed reference populations. Next, we threshold the inferred local ancestry probabilities for the reference populations and refit the mixture model conditioned on this local ancestry. This method allows for an approximate joint fitting of multiple target individuals.

2.3.5 Tree Filtering

Relate trees have variable span along the genome. We only consider trees every 10kb and 0.05cM apart, ensuring the tree intervals overlap with the local ancestry grid at least once. We filter out trees in regions corresponding to centromeres, telomeres, and the HLA region, along with a 500kb buffer around these areas. The positions of centromeres, telomeres, and the HLA region in the appropriate genome build were obtained from UCSC genome browser [138].

2.3.6 Choosing the number of components

A critical aspect of decomposing local ancestry is selecting the optimal number of components for our method. To ensure we determine the appropriate number, we conduct several tests to evaluate and optimize this choice.

First, we assess the model's log-likelihood on an independent, held-out chromosome. This allows us to distinguish admixture events that replicate consistently across chromosomes from biases that are chromosome-specific and do not generalize well. The evaluation proceeds as follows: we estimate the parameters—such as coalescence rates and proportions—using our EM algorithm on data from all chromosomes except the held-out one (by default we use chromosomes 2-5). We then use the fitted parameters to calculate the log-likelihood on the held-out chromosome (by default we use chromosome 1), by turning off the HMM (treating grid points independently). The optimal number of components is identified by selecting the point where the total log-likelihood across held-out chromosomes is maximized.

Second, we analyze the posterior probabilities of local ancestry by plotting a histogram of the target's local ancestry posteriors and calculating the expected coefficient of determination (expected R^2), as defined in [23, 58]. Expected R^2 estimates the expected correlation between true and inferred local ancestry, assuming the true local ancestry (one-hot encoded) as a random variable following a Bernoulli distribution with the same mean as the inferred local ancestry. The calculation of expected R^2 therefore does not need the true local ancestry. Both the histogram

and expected R^2 help assess the confidence of the cluster assignments, where values greater than 0.7 generally indicate confident local ancestry inference.

The expected R^2 between true (denoted as X) and predicted local ancestry estimates (denoted as Y) can be calculated as follows:

$$\mathbb{E}(R^2_{XY}) = \mathbb{E}\left(\frac{\text{cov}(X,Y)^2}{\text{Var}(Y)\text{Var}(X)}\right) \tag{2.50}$$

$$\approx \frac{\mathbb{E}(\text{cov}(X,Y))^2}{\text{Var}(Y)\mathbb{E}(\text{Var}(X))} \tag{2.51}$$

Assuming the true local ancestry to follow a Bernoulli distribution with same empirical mean as predicted local ancestry, $P(X_g = 1) = Y_g$. Where, g indexes the local ancestry position (total = G) along the genome. We can evaluate the terms in Equation 2.51 as follows:

$$\begin{aligned}
\text{var}(Y) &= \frac{\sum_g Y_g^2}{G} - \left(\frac{\sum_g Y_g}{G}\right)^2 \\
\mathbb{E}[\text{cov}(X,Y)] &= \mathbb{E}\left[\frac{\sum_g X_g Y_g}{G} - \frac{\sum_g Y_g}{G}\frac{\sum_g X_g}{G}\right] \\
&= \frac{\sum_g Y_g \mathbb{E}[X_g]}{G} - \frac{\sum_g Y_g}{G}\frac{\sum_g \mathbb{E}[X_g]}{G} \\
&= \frac{\sum_g Y_g^2}{G} - \left(\frac{\sum_g Y_g}{G}\right)^2 \\
\mathbb{E}[\text{var}(Y)] &= \mathbb{E}\left[\frac{\sum_g X_g^2}{G} - \left(\frac{\sum_g X_g}{G}\right)^2\right] \\
&= \frac{\sum_g \mathbb{E}[X_g]^2}{G} + \text{var}(X_g) - \left(\frac{\sum_g \mathbb{E}[X_g]}{G}\right)^2 \\
&= \frac{\sum_g Y_g^2 + Y_g(1-Y_g)}{G} - \left(\frac{\sum_g Y_g}{G}\right)^2 \\
&= \frac{\sum_g Y_g}{G} - \left(\frac{\sum_g Y_g}{G}\right)^2
\end{aligned} \tag{2.52}$$

Finally the $\mathbb{E}(R^2_{XY})$ is given as:

$$\mathbb{E}(R^2_{XY}) \simeq \frac{\sum_g Y_g^2 - \left(\sum_g Y_g\right)^2/G}{\sum_g Y_g - \left(\sum_g Y_g\right)^2/G}. \tag{2.53}$$

We use a diploid version of the Equation 2.53 which sums the local ancestry for both haploids; more details and derivation for the diploid version can be found in [58].

Finally, we use the inferred local ancestry to perform coancestry curve-based dating of the admixture event. Unlike admixture tracts, which correspond to long ancestry segments, general biases tend to result in much older estimated dates. Admixture dates exceeding 2,500 generations are challenging to distinguish from spurious events and are typically regarded with caution, as they may reflect biases in genealogical inference rather than true admixture.

While this approach performs well in simulations, real data introduces additional complexities. The full admixture history of a sample is often intricate, and source populations themselves may be the result of earlier admixture. Thus, in practice, we select the number of components based on how well the components can be interpreted in the context of real data. In most cases, this results in choosing two components. Ultimately, the choice of the number of components can be data-driven, guided by the held-out log-likelihood, expected R^2, and coancestry dating, or informed by prior knowledge of the admixture event.

2.3.7 Calculation of node persistence

The majority of nodes in adjacent trees are shared. Accurate calculation of node persistence is critical in our likelihood estimation, as it prevents the over-counting of identical or similar coalescence events. In true genealogies, nodes can be distinguished by their coalescence times. These coalescence times remain consistent across adjacent trees unless a recombination event affects the corresponding node.

However, in inferred genealogies, the coalescence times are not precisely identical, even in the absence of recombination, as each tree is dated independently based on its structure. To identify equivalent nodes in inferred genealogies, we examine the descendant composition of these nodes. Specifically, we store a vector of length equal to the number of reference samples, representing the descendant composition of the lineage coalescing with the target lineage. In this vector, the mth entry is set to 1 if haplotype m descends from the coalescence event, and 0 otherwise. Similar

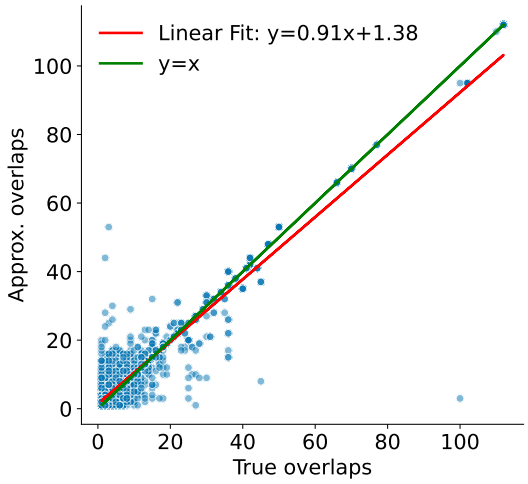


Figure 2.4: Comparison of Approximate and True Node Persistence in Simulated Trees. True node persistence is determined by identifying the exact coalescence times across neighboring trees, while approximate node persistence is estimated based on the correlation of descendant sets between adjacent trees. In this case, the correlation cutoff for approximate node persistence is set to 0.5, with the number of overlaps indicating the number of 10kb intervals for which the node persists.

to defining equivalent branches in Relate [6], we define two nodes as equivalent if the correlation coefficient between their descendant composition vectors exceeds a predefined threshold (default = 0.5). To ensure that each node is matched with at most one other node, we rank pairs of candidate nodes in descending order of their correlation coefficient. Nodes are then associated based on their position in this sorted list, with any pair involving a node already associated with another node being removed.

In true genealogies, where we have access to actual node persistence, we observed that the approximate node persistence reasonably aligns with the true values (see Figure 2.4 for comparison).

2.3.8 Handling uncertain coalescence events

Since the mutation and recombination rates in humans are approximately similar, multiple tree structures often exist to explain the genetic variation at a given site, making genealogy inference uncertain and potentially inaccurate. To address this

uncertainty, we scale the likelihood of our mixture model based on the confidence in coalescence events, which is evaluated by the presence of mutations above the node.

Specifically, we trace the coalescence events along the target lineage backward in time, identifying subsets of nodes (or coalescence events) that are jointly supported by at least one mutation (see Figure 2.5). For each subset, we calculate the expected number of coalescence events (or nodes) that are supported by mutations. To clarify, this is equivalent to asking: what is the expected number of coalescence involving the target lineage if the coalescence events between the lineages were shuffled randomly? We assume that clades coalescing with the target are resolved—that is, they have mutations supporting their tree structure. We further assume that unresolved coalescence events occur around the same time, simplifying their contribution to the likelihood. The expected number of coalescence events, conditioned on the presence of C unresolved events, can be computed as the sum of the probability of coalescing with the target for the C independent coalescence events.

$$\mathcal{E}[C] = \sum_{i=1}^{C} \frac{i}{\binom{i+1}{2}} \tag{2.54}$$

$$= \sum_{i=1}^{C} \frac{2}{i+1} \tag{2.55}$$

We scale the likelihood for each node by the expected number of coalescence events divided by the observed number of coalescence events, that is $\frac{\mathcal{E}[C]}{C}$. For cases where there are 1, 2, or 3 unresolved nodes, the corresponding expected number of coalescence events equals 1, $\frac{5}{3}$, and $\frac{13}{6}$, leading to effectively scaling the likelihood by 1, $\frac{5}{6}$, and $\frac{13}{18}$, respectively.

2.3.9 Initialization

To approximate a solution near the global maximum, we perform a random search over 20 randomly initialized values for coalescence rates, proportions, and admixture times. The coalescence rates are sampled from a log-uniform distribution within the range $[5 \times 10^{-2}, 1 \times 10^{-7}]$. The proportions are drawn from a uniform distribution between 0.01 and 0.99, while the admixture times are sampled from a log-uniform

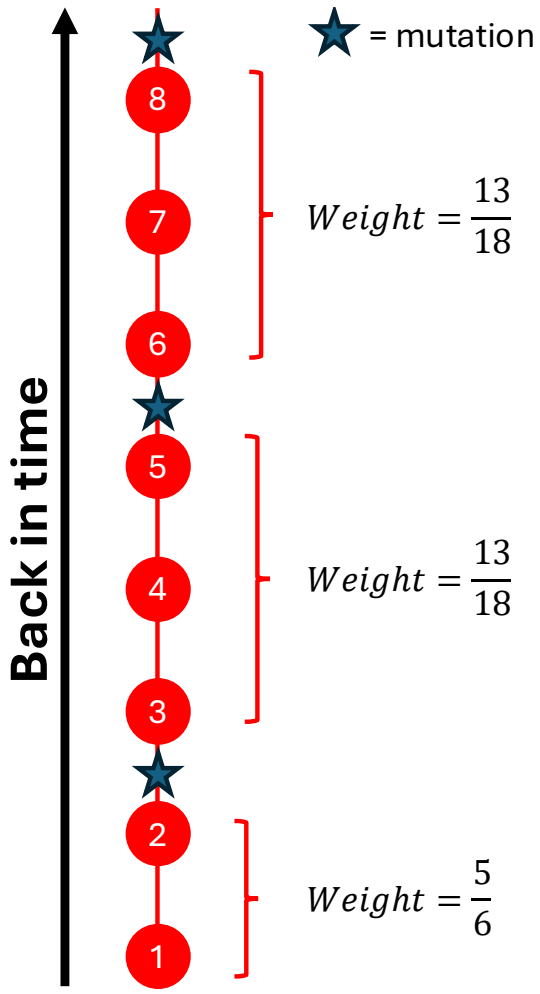


Figure 2.5: Handling uncertainty in coalescence events. We calculate the expected number of coalescence events for subsets of nodes jointly supported by a mutation. In this example, there are three such subsets on the target lineage, with weights for each coalescence event calculated based on the number of nodes in the subset.

distribution spanning 20 to 2000 generations. For each random initialization, the EM algorithm is run for 10 iterations. The initialization that achieves the highest log-likelihood after these iterations is selected as the optimal starting point for further model fitting.

2.4 Data and quality control

2.4.1 Modern human datasets

We work with multiple publicly available modern human datasets throughout our analysis. The datasets we analyze include the Human Genome Diversity Project (HGDP) [4, 139], Simons Genome Diversity Project (SGDP) [140], and 1,000 Genomes Project (1,000 GP) [5, 141]. The whole-genome sequenced version of HGDP [139] comprises 929 individuals and 51 diverse population groups sampled across the globe, with an average coverage of 34× (range: 23-75×). The 1,000 GP dataset comprises 2,504 individuals from 26 populations across the globe, with an average coverage of 32× (range: 26-66×). The SGDP dataset comprises 278 individuals from 142 diverse populations across the globe, with an average coverage of 43× (range: 34-83×).

In particular, we use the unified version of HGDP and 1,000 GP [142], where the genomes from the two projects were jointly called. We obtained the phased version of the dataset from gnomad [143]. This dataset was called with the hg38 genome build. The phased data had already undergone variant QC to filter variants with (1) HWE $\geq 1 \times 10^{-30}$, (2) F_MISSING ≤ 0.1, or (3) ExcHet ≥ 0.5 and ExcHet ≤ 1.5. We additionally excluded multi-allelic SNPs and Indels from the dataset. We also performed sample QC, retaining only a subset of African samples from the 1,000 GP (20 individuals per 6 African subgroups: 'ACB', 'ASW', 'ESN', 'LWK', 'GWD', 'MSL') and removing close relatives up to the 2nd degree. This QC step was implemented to minimize biases introduced by close relatives during genealogy inference and to prevent data imbalance between the two datasets. To remove relatives, we used the list of unrelated samples in HGDP + 1,000 GP inferred in the Allen Ancient DNA Resource [81] and retained only individuals falling within this set. Finally, we retained 45.6 million variants and 1,026 samples across 58 population groups in this dataset.

The Simons Genome Diversity Project (SGDP) was called in the hg37 genome build and was therefore not merged with the HGDP + 1,000 GP dataset. We treated

it separately as an independent validation dataset. The phased data for SGDP was obtained from Reich lab website [144]. Apart from the filtering done in [140], we additionally excluded multi-allelic SNPs and Indels from our analysis. We retained 28.3 million variants and 278 samples across 130 population groups in this dataset.

2.4.2 High quality ancient DNA resources

The assembly and quality control of the ancient DNA dataset were performed by Dr. Leo Speidel, who is one of the co-authors on the paper. Apart from the modern samples, we also consider 49 ancient DNA and 4 archaic DNA samples. These samples were all high coverage, with an average coverage of 17.3× (range: 10.3-65.1×). The ancient and archaic DNA samples used in the analysis are tabulated in B.1.

The ancient DNA BAM files were processed using `bcftools mpileup` to generate variant calls, selecting only reads with a mapping quality of 20 or higher, a base quality of 20 or higher, and a minimum depth of coverage of 5. The variant calls from the ancient DNA samples were then merged with the 1,000 GP (downloaded from [145]) and phased using `shapeit4`. The phasing process involved first phasing variants observed in 1,000 GP and then using the phasing as a scaffold to phase variants not observed in 1,000 GP within the ancient dataset. We additionally randomly phased singletons in the ancient dataset. Subsequently, the ancient DNA dataset was lifted over to the hg38 reference genome using `picard LiftoverVcf` and merged with the unified version of HGDP + 1,000 GP using the `bcftools merge -0` option. This merging resulted in less than 0.001% missingness. After merging, we filtered out multi-allelic variants, duplicate variants, and indels. As a final sanity check, we compared allele frequencies between the HGDP + 1,000 GP dataset and the ancient DNA dataset, finding a high correlation between the allele frequencies with minimal missingness (see Figure B.1).

2.4.3 Building genome-wide genealogies for moderns and ancients

We use Relate [6] to construct genealogies from phased data for both modern and ancient samples. For genealogies involving only modern samples, we apply a genomic mask provided with 1,000 Genomes Project (lifting it for hg38 datasets), retaining only regions marked as "passing" to exclude variants with high uncertainty. Additionally, we mask sites where the human ancestral allele FASTA file (accessible through Ensembl [146, 147]) shows missing data, as well as sites corresponding to multi-allelic SNPs or indels. For genealogies that include archaic samples, we further refine this mask by considering only sites that pass in all Neanderthal and Denisovan masks (accessible through [148]). Additionally, for genealogies with ancient samples, we filter out sites where ancient samples exhibit more than 10% missingness. The masks corresponding to ancient and archaic samples are lifted over to the hg38 reference genome to ensure consistency.

To identify the most likely ancestral allele for each SNP, we use an estimate of the human ancestral genome (accessible through Ensembl [146, 147]). We utilize the HapMap3 inferred recombination rate maps for builds hg37 and hg38, which can be downloaded from [149]. The mutation rate is set to 1.25×10^{-8} for building genealogies of modern samples, while for genealogies involving ancient or archaic samples, the mutation rate is adjusted to 1.2×10^{-9}, and all CpG sites are excluded from genealogy inference. To ensure consistency in the genealogies, we enforce the construction of topologies every 10kb using the `-fb` option. When working with ancient samples, we provide the estimated ages of the samples to improve the accuracy of the genealogies. The population size of the sample is estimated by iteratively fitting the (time-varying) population sizes and branch lengths for chromosome 1, over five iterations. The converged population size is then used as a prior for estimating branch lengths on the remaining chromosomes. Finally, the resulting trees are converted to tskit format, retaining only the trees to be used for model fitting in GhostBuster.

Due to differences in variant filtering and genome build, we construct multiple genealogies for distinct sets of samples. The genealogies include:

1. **HGDP + 1,000GP**: Comprises all modern samples from HGDP and a subet of African samples from 1,000 Genomes Project.

2. **HGDP + 1,000GP + aDNA**: Includes HGDP + 1,000GP samples along with 49 high-coverage ancient DNA samples.

3. **HGDP + 1,000GP + aDNA + archaic**: Extends the HGDP + 1,000GP + aDNA set by adding 4 archaic samples.

4. **SGDP**: Contains all modern samples from SGDP.

We verified our inferred genealogy for all sample sets by evaluating the inferred inverse coalescence rates between populations (see figures B.2, B.3, B.4, B.5, B.6). For real data analysis of population structure in Africa (Chapter 3), we primarily focus on using samples from HGDP and 1,000GP, while incorporating SGDP samples to augment the analyses where feasible.

2.5 Accurate local ancestry inference in simulations

2.5.1 Simulation design

To assess the power and robustness of detecting ghost populations, we simulate two distinct admixture scenarios: a four-way recent admixture and Denisovan admixture into Papuans. For both simulations, we used a constant mutation rate of 1.25×10^{-8} and the HapMap3 recombination rate map, excluding regions with zero recombination rate, centromeres, telomeres and the HLA region on chromosome 6. We simulate demography using `msprime` [150].

1. **Four-way recent admixture:** In this scenario, we assume four reference populations (A, B, C, and D), each contributing 25% to form the focal individual 5,000 years ago. Populations A and B split 50,000 years ago, as

do populations C and D, with all populations merging back together 100,000 years ago (see Figure 2.6a). We assume a constant population size of 20,000 for all populations after their split and 10,000 for the super-populations prior to 50,000 years ago. To assess the case without a ghost population, we sample 10 diploid individuals from each population and five focal (admixed) individuals. For the ghost population scenario, we sample 10 individuals each from populations A, B, and C, with none from population D.

2. **Denisovan Introgression into Papuans:** This scenario uses the simulation model and script provided by [151] to simulate Denisovan introgression into Papuans. We assume three reference populations (Papuans, Africans, and Denisovans), where Papuans and Africans split 72,036 years ago, and Denisovans split from modern humans 656,908 years ago (see Figure 2.6b). Denisovans admixed with Papuans around 43,935 years ago with a 5% admixture proportion. The population sizes are $3,899$ for modern Papuans, 27,122 for Africans, and 4,947 for Denisovans before they went extinct. We also simulate the Out-of-Africa bottleneck, where the Papuan population size drops to 136. A graphic overview of this simulation is shown in Figure 2.6b. We fit five Papuan individuals, using 25 Papuans and 25 Africans as the reference panel. Additionally, we assess the power and compare local ancestry estimates with and without an ancient Denisovan sample dated to 67,570 years ago.

2.5.2 Four-way recent admixture

We validated GhostBuster's ability to detect admixture events through a simple simulation, where we modeled a recent 4-way admixture event occurring 5,000 years ago, involving groups that had been separated for at least 50,000 years (see Figure 2.6a). The focal individual in our simulation resulted from an equal contribution of 25% ancestry from four distinct reference populations, labeled A, B, C, and D. We sampled up to 10 individuals from each reference population and used

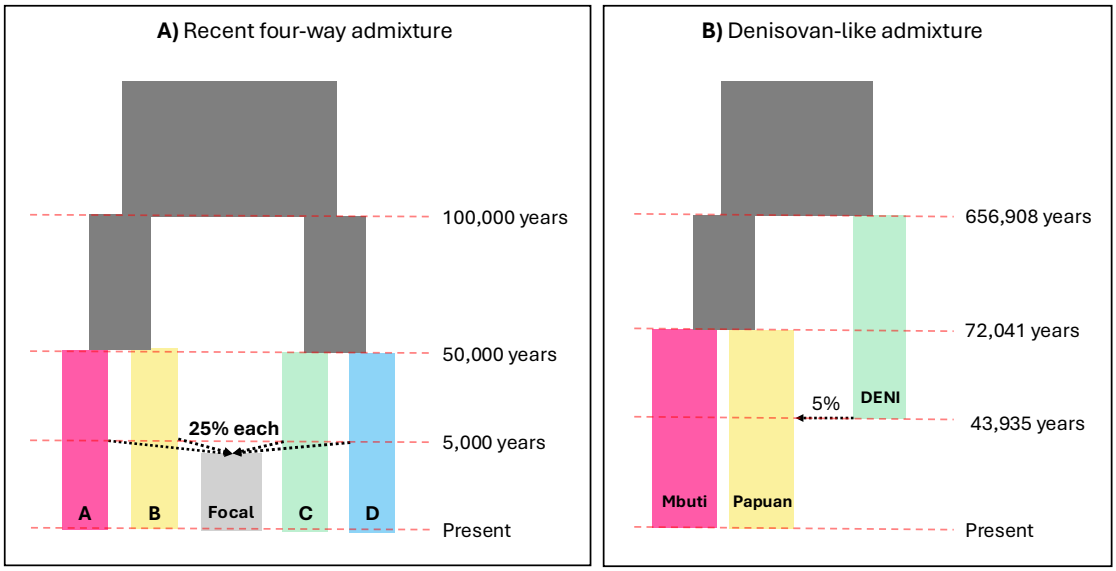


Figure 2.6: Simulation design for performance evaluation. A. Four-way recent admixture simulation, B. Simulated Denisovan introgression into Papuans. The simulation parameters including the population size, mutation rate and recombination map are described in Section 2.5.1.

chromosomes 1 through 5. More details about the simulation setup can be found in Section 2.5.1. Our objective was to decompose the focal individual's ancestry both with and without access to all reference populations, with the expectation of accurately identifying the four ancestral components. We applied our method primarily to trees inferred using genetic variation data via Relate, assuming a constant mutation rate of 1.25×10^{-8}, while results comparing with true trees simulated using msprime are also presented. Additionally, we only focused on coalescence events occurring from 0 to 50,000 years.

We first tested the method with access to all four reference populations and applied our approach to decompose the focal individual's ancestry into four distinct components or clusters. Starting with a random initialization of coalescence rates and proportions, we fit our model over 200 EM iterations until the log-likelihood converged. The resulting inverse coalescence rates represent the four clusters that best differentiated the coalescence events for the focal individual (see Figure 2.7a). The inverse coalescence rate (ICR) for each component reflects its affinity toward

one of the four reference populations—a lower ICR value indicates faster coalescence with that specific population. Our analysis revealed that the four clusters identified by GhostBuster contributed approximately 25% to the focal individual's ancestry, with each cluster more closely related to one of the four reference populations. For example, component 1 was most similar to population D, component 2 to population A, component 3 to population C, and component 4 to population B. The clusters were clearly distinguishable using the top four principal components (PCs) of the coalescence count and opportunity matrices (see Figure 2.7b). Further evaluation based on the expected coefficient of determination and held-out log-likelihood indicated that four components provided the best fit to the data, with diminishing returns in held-out log-likelihood beyond four components (see Figure 2.7c). We also accurately dated the admixture event to 185.1 generations, or 5,182.8 years, using coancestry curves (see Figure 2.7d). Finally, we assessed the accuracy and calibration of our local ancestry inference by comparing it with the simulated ground truth, demonstrating strong correlation and low calibration error (see Figure 2.7e-g).

Next, we removed population D (without loss of generality) from the genealogies and attempted to decompose the focal individual. This scenario simulates the presence of a ghost population, as population D contributed to the focal individual's genetic makeup but is not included in the reference panel. We reran our method, fixing the number of components to four. The four components still contributed roughly 25% to the total ancestry. The inverse coalescence rates (ICRs) for each component only existed for populations A, B, and C, as no individuals from population D were sampled. Component 2's ICR was closest to population B, component 3 was closest to population A, and component 4 was closest to population C (see Figure 2.8a). However, component 1 was more distantly related to all three populations, indicating slower coalescence with each. This could be the ancestry relating to population D, which may have been tagged by its slower coalescence rates with populations A-C. Similar to the run with access to population D, we performed PCA visualization (see Figure 2.8b), evaluation of the number of clusters (see Figure 2.8c), coancestry curve dating (see Figure 2.8d), and accuracy and

calibration of local ancestry inference (see Figure 2.8e-g). When comparing local ancestry with the simulated ground truth, we found a strong correlation, with an R^2 value of 0.88, indicating that component 1 effectively captured the ancestry related to population D without requiring direct samples from that population.

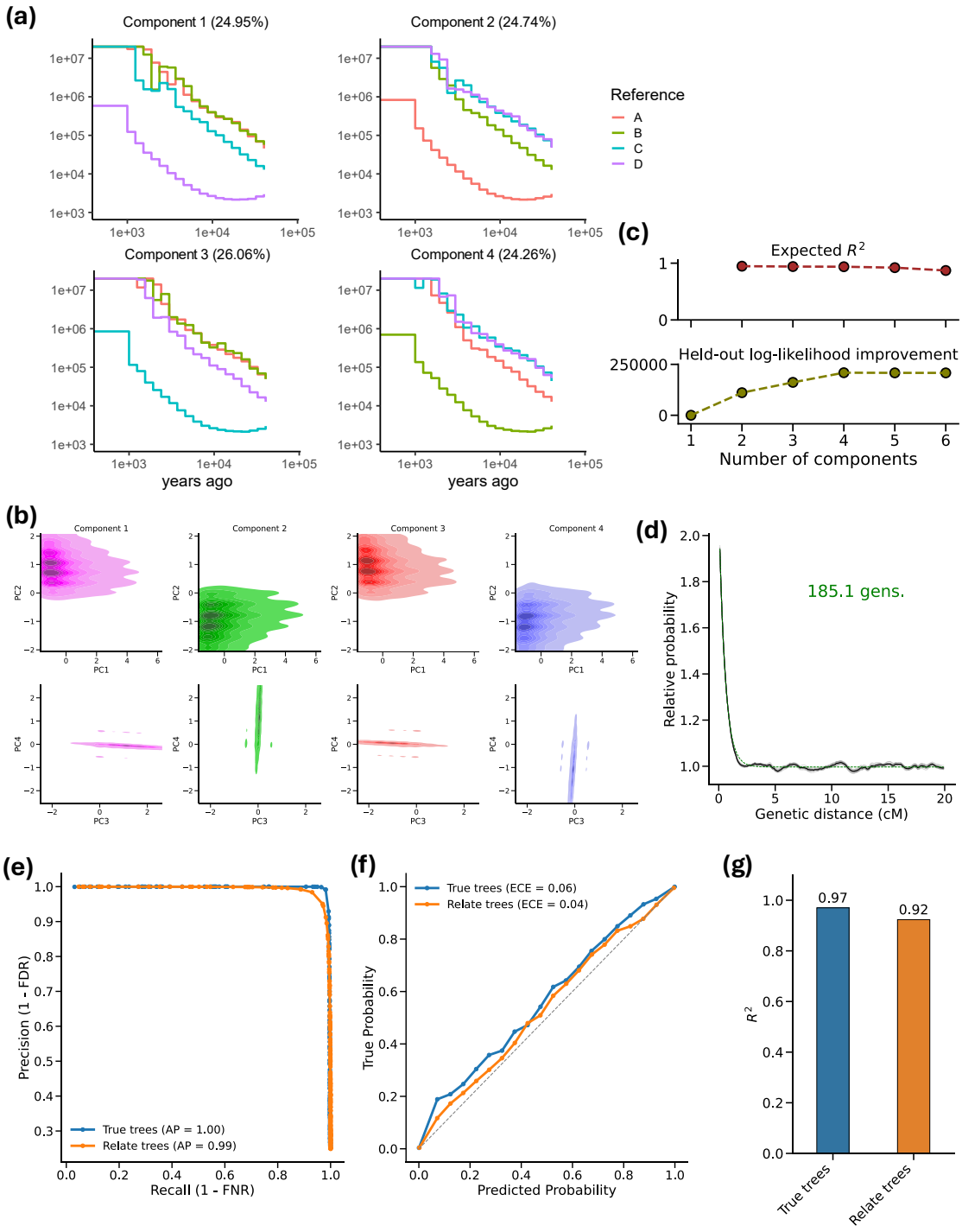


Figure 2.7: Decomposing focal individuals in the four-way admix simulation with presence of all source populations. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Precision-recall curves, (f) Calibration curves, and (g) Prediction R^2 for inferred local ancestry compared to the simulated ground truth, where inference was done using true trees or Relate trees. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5.

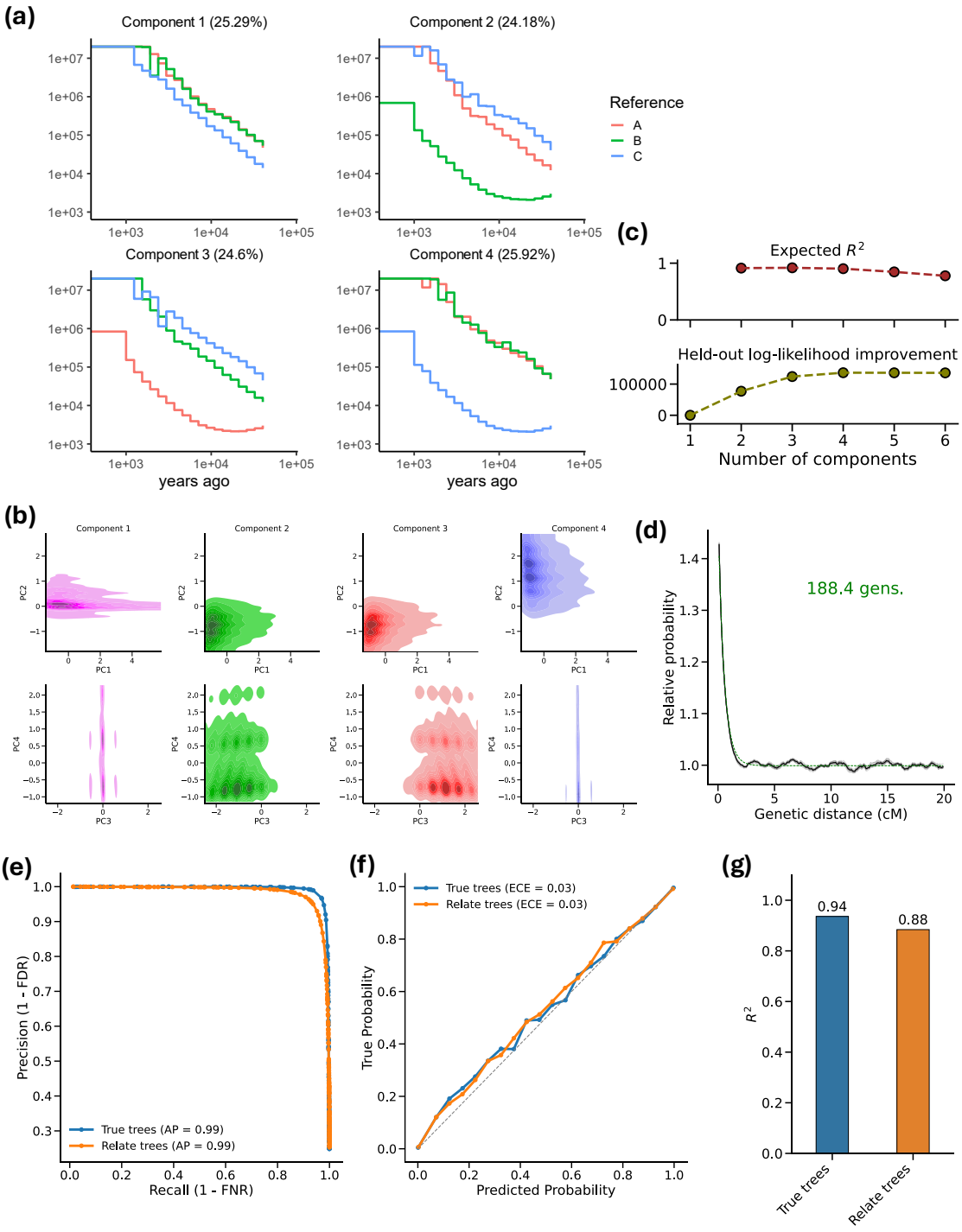


Figure 2.8: Decomposing focal individuals in the four-way admix simulation with absence of one source population. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Precision-recall curves, (f) Calibration curves, and (g) Prediction R^2 for inferred local ancestry compared to the simulated ground truth, where inference was done using true trees or Relate trees. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5.

2.5.3 Denisovan-like Introgression into Papuans

We emulated the archaic introgression of Denisovan ancestry in modern-day Papuans by simulating an admixture event that occurred approximately 44,000 years ago, with an admixture proportion of 5% (see Figure 2.6b). The simulation parameters are from [151], and further details can be found in Section 2.5.1. For our analysis, we simulated 25 Papuan-like individuals, 25 Mbuti-like individuals (representing African ancestry), and a Denisovan-like ancient sample. We focused on chromosomes 1 through 5 for our analysis. Our objective was to decompose the Papuan individuals' ancestry both with and without access to the Denisovan sample, aiming to accurately identify Denisovan segments and admixture within Papuans. Similar to our previous simulation of a four-way recent admixture, we primarily applied our method to trees inferred from genetic variation data via Relate, assuming a constant mutation rate of 1.25×10^{-8}. Finally, we focused only on coalescence events occurring from 10,000 to 1,000,000 years for this analysis.

We begin our analysis by assuming we have access to one ancient Denisovan sample. We decompose five out of the 25 Papuan individuals, assuming two distinct components. As with the previous simulation, we randomly initialize the parameters and fit the EM algorithm for 200 iterations until the log-likelihood converges. The converged components reveal a major component, contributing around 95.85% to the local ancestry, and a minor component contributing 4.15% (see Figure 2.9a). The inverse coalescence rates (ICRs) characterizing these components differ significantly. Component 1 demonstrates a slower coalescence rate with the ancient Denisovan sample and a relatively faster coalescence with the modern Mbuti samples. In contrast, component 2 exhibits a very fast coalescence rate with Denisovans but slower coalescence with modern Mbuti and Papuan samples. It is important to note that the most striking difference between the two components is the vastly accelerated coalescence with Denisovans in the minority group, up to 1,000× faster compared to the majority group. However, there is also a substantial difference in coalescence with modern groups, which we will leverage to decompose the ancestry without access to the ancient Denisovan sample.

The clusters were distinguishable using the top four principal components (PCs) of the coalescence count and opportunity matrices (see Figure 2.9b). Further evaluation based on the expected coefficient of determination and held-out log-likelihood indicated that two components provided the best fit to the data, with diminishing returns in held-out log-likelihood beyond four components (see Figure 2.9c). We also accurately dated the admixture event to 1,546.7 generations, or 44,854.3 years, using coancestry curves (see Figure 2.9d). Finally, we assessed the accuracy and calibration of our local ancestry inference by comparing it with the simulated ground truth and HMMIX, a recent HMM-based method tailored to find archaic introgression such as this one [151], demonstrating strong correlation with the ground truth, surpassing HMMIX, and showing low calibration error (see Figure 2.9e-g).

Next, we removed the Denisovan samples from our analysis and decomposed the same Papuan individuals into two components. In this scenario, the Denisovan ancestry in Papuans can be thought of as a ghost population. Our method converged on two components, contributing approximately 95.9% and 4.1% to the local ancestry, respectively (see Figure 2.10a). The ICRs of the minor component closely matched those of the minor component from the earlier analysis that included the Denisovan sample, showing slower coalescence with Mbuti and Papuans (see Figure 2.9a). Similar to the run with access to the Denisovan genome, we performed PCA visualization (see Figure 2.10b), evaluation of the number of clusters (see Figure 2.10c), coancestry curve dating (see Figure 2.10d), and accuracy and calibration of local ancestry inference (see Figure 2.10e-g). To verify that the minor component indeed represents Denisovan ancestry, we compared the local ancestry inferred using GhostBuster with the simulated ground truth and found a high degree of correlation ($R^2 = 0.63$), surpassing HMMIX [151].

In order to test the robustness and false-positive rate of our method, we attempted to decompose five Mbuti individuals in the same simulation. The Mbuti individuals are panmictic, with no simulated population structure or admixture. When we forced GhostBuster to find two components for relate trees, it identified

two components with varying coalescence rates with Papuans. However, these components were not clearly separated in the PC plots. In terms of held-out log-likelihood, the two-component solution offered only a slight improvement compared to a single-component model. Moreover, the histogram of the local ancestry posteriors and the expected $R^2 = 0.49$, both suggest that the local ancestry inferred is highly uncertain in this case. Additionally, dating this event using coancestry curves inferred an admixture date of around 2,800 generations (81,200 years), which is older than the Mbuti and Papuan split time in this simulation, raising suspicion. Thus, while it is possible to fit a two-component model, it does not necessarily provide compelling evidence for admixture. It is therefore crucial to evaluate PC plots, held-out likelihood, expected coefficient of determination, and coancestry dating along with other sources of evidence in real-data analyses to be more confident about the presence of gene flow or admixture.

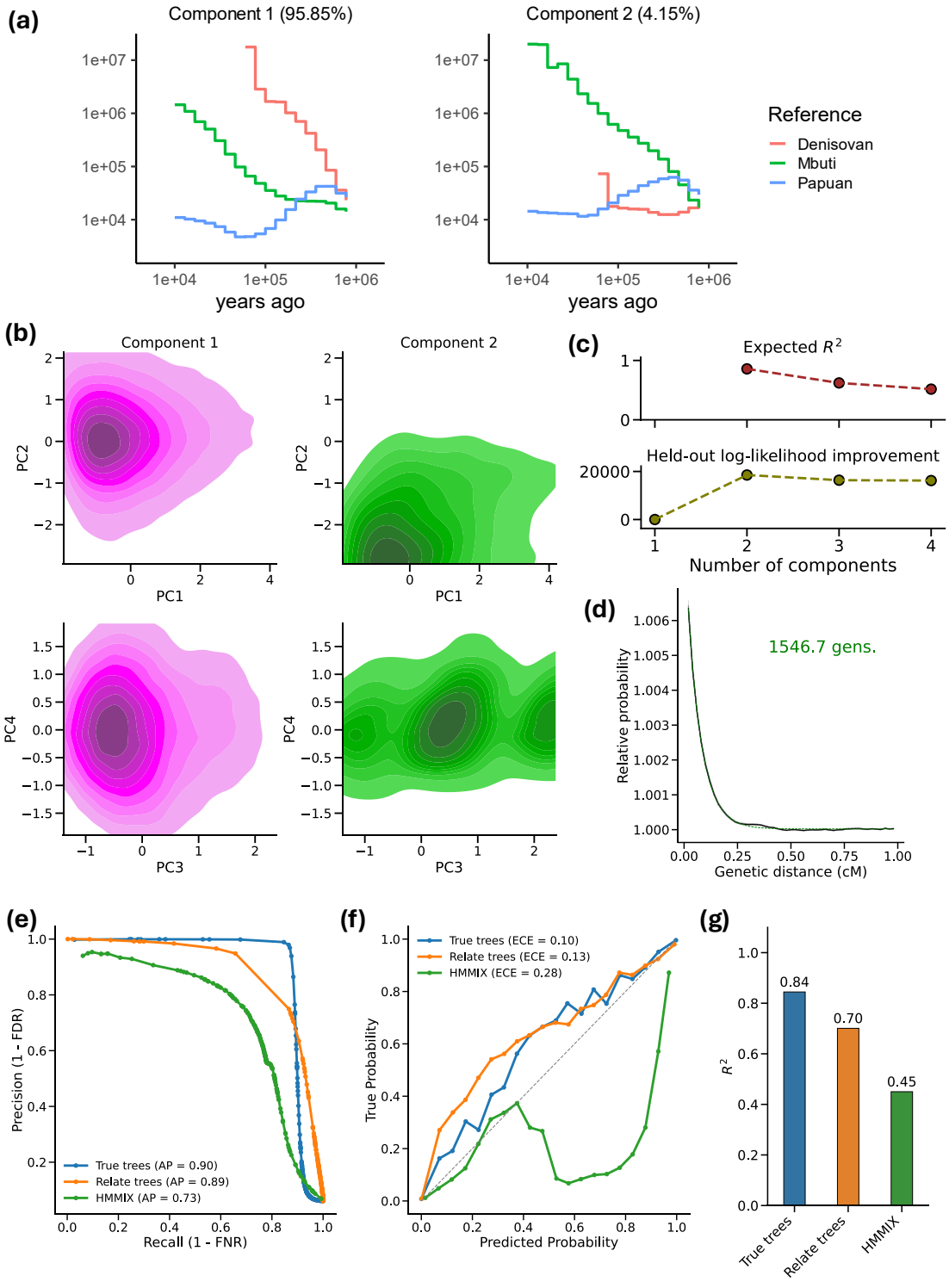


Figure 2.9: Decomposing Papuan individuals in simulations with Mbuti and Denisovans as additional references. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Precision-recall curves, (f) Calibration curves, and (g) Prediction R^2 for inferred local ancestry compared to the simulated ground truth, where inference was done using true trees or Relate trees. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5.

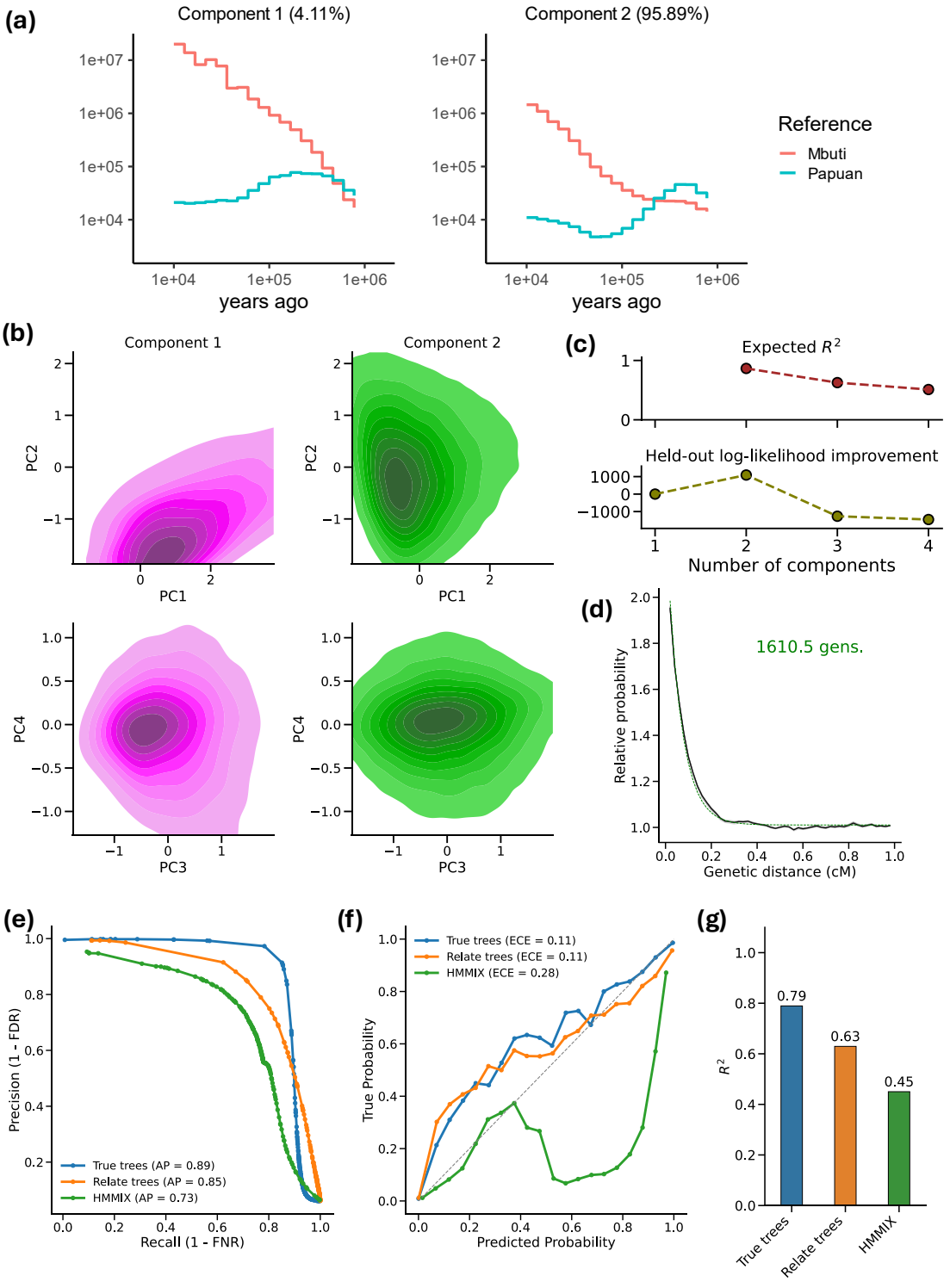


Figure 2.10: Decomposing Papuan individuals in simulations with only Mbuti as additional references. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Precision-recall curves, (f) Calibration curves, and (g) Prediction R^2 for inferred local ancestry compared to the simulated ground truth, where inference was done using true trees or Relate trees. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5.

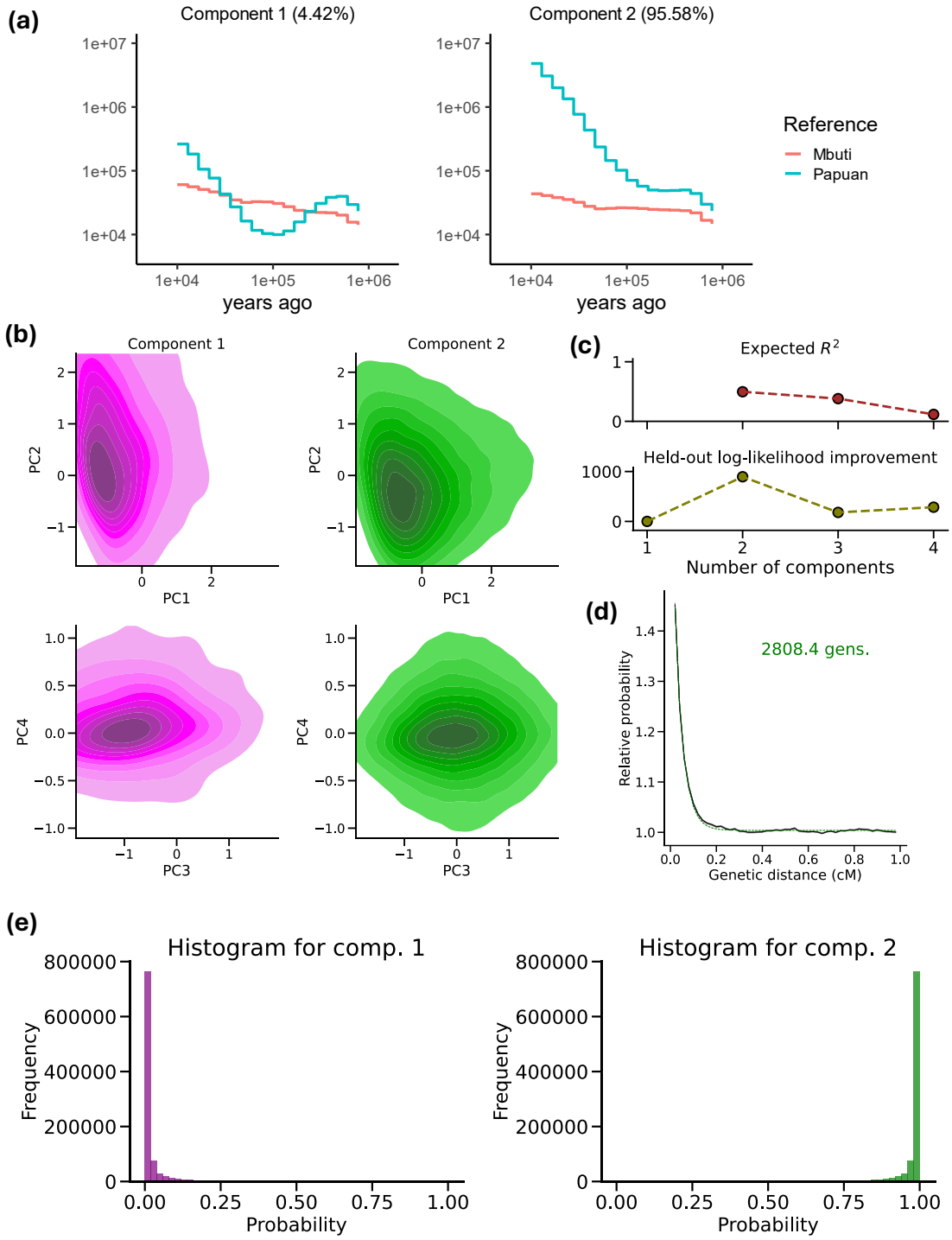


Figure 2.11: Decomposing Mbuti individuals in simulations with Papuan as additional references. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Histogram of the inferred local ancestry posteriors. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5.

2.6 Identifying several recent admixtures accurately

Haplotype-based methods like Alder [53], Globetrotter [51], and Mosaic [58] have been instrumental in inferring recent admixture events, particularly those occurring less than 2,000 years ago. We focus on several well-documented, recently admixed populations from the Human Genome Diversity Project (HGDP), namely the Hazara, Bedouin, and Maya. Each of these populations has undergone two-way admixture events within the past few thousand years, which have been reliably detected by previous methods. Our objective is to decompose the ancestry of individuals from these populations using GhostBuster and to compare the local ancestry with results from other haplotype-based methods.

We begin by decomposing the ancestry of Hazara individuals from the HGDP, who have been shown in previous studies to result from recent admixture between groups similar to modern-day Pathans in South Asia and Mongols in East Asia [51, 58]. Using all populations in our HGDP + 1,000GP dataset (as described in Section 2.4), we fit two components to decompose the ancestry of the five Hazara individuals from chromosome one to five. We focused on coalescence events within the last 10,000 years, dividing the time scale into 10 logarithmically spaced epochs. Our analysis revealed that the Hazara individuals could be decomposed into two components, each contributing roughly 50% to the overall ancestry. The inverse coalescence rates (ICRs) indicate that component 1 is closely related to East Asian groups, while component 2 is more closely aligned with South Asian groups—consistent with previous findings using Globetrotter and Mosaic (see Figure 2.12a). The two components separated clearly in the PC plots (see Figure 2.12b). Moreover, two components provided the best fit with high expected coefficient of determination and held-out likelihood improvement (see Figure 2.12c). The coancestry curves inferred admixture dates of around 26.2 generations or 733.6 years, assuming 28 years per generation (see Figure 2.12d). To further validate our results, we compared the local ancestry estimates from our method with those obtained using Mosaic, which was run using genotyped variants and all reference populations in HGDP. We

found that the local ancestry was highly correlated between the two methods, with a mean correlation of 0.87 (see Figure 2.12e). Potential differences in local ancestry estimates may be attributed to additional steps in Mosaic, such as correction for phasing, which are not present in our method.

We conducted similar analyses for other populations in the HGDP, including the Bedouin, who have a well-documented history of European and African admixture within the last 40 generations. Our findings reveal two components: a major component, comprising 90%, closely aligned with Middle Eastern and European groups, and a minor component more closely related to sub-Saharan African groups (see Figure 2.13a). The PC plots effectively distinguish between these ancestries (see Figure 2.13b), and the coancestry curves suggest an admixture date of approximately 36.6 generations or 1,024.8 years (see Figure 2.13d). Similar to our analysis of the Hazara, the local ancestry inferred with GhostBuster shows a high correlation with that inferred using Mosaic, with a mean local ancestry correlation of 0.86 (see Figure 2.13e).

Finally, we analyzed Maya individuals from Central America, who exhibit a well-documented two-way admixture between Native American and European ancestries, largely resulting from European colonization. We identified two components: the major component, comprising 86%, is closely aligned with other Native American groups, while the minor component shows a closer relationship to European groups (see Figure 2.14a). The PC plots successfully separate the ancestries (see Figure 2.14b), and the coancestry curves estimate an admixture date of approximately 6.4 generations, or 179.2 years (see Figure 2.14d). As with the Hazara and Bedouin populations, the local ancestry inferred with GhostBuster showed a strong correlation with Mosaic, with a mean local ancestry correlation of 0.92 (see Figure 2.14e).

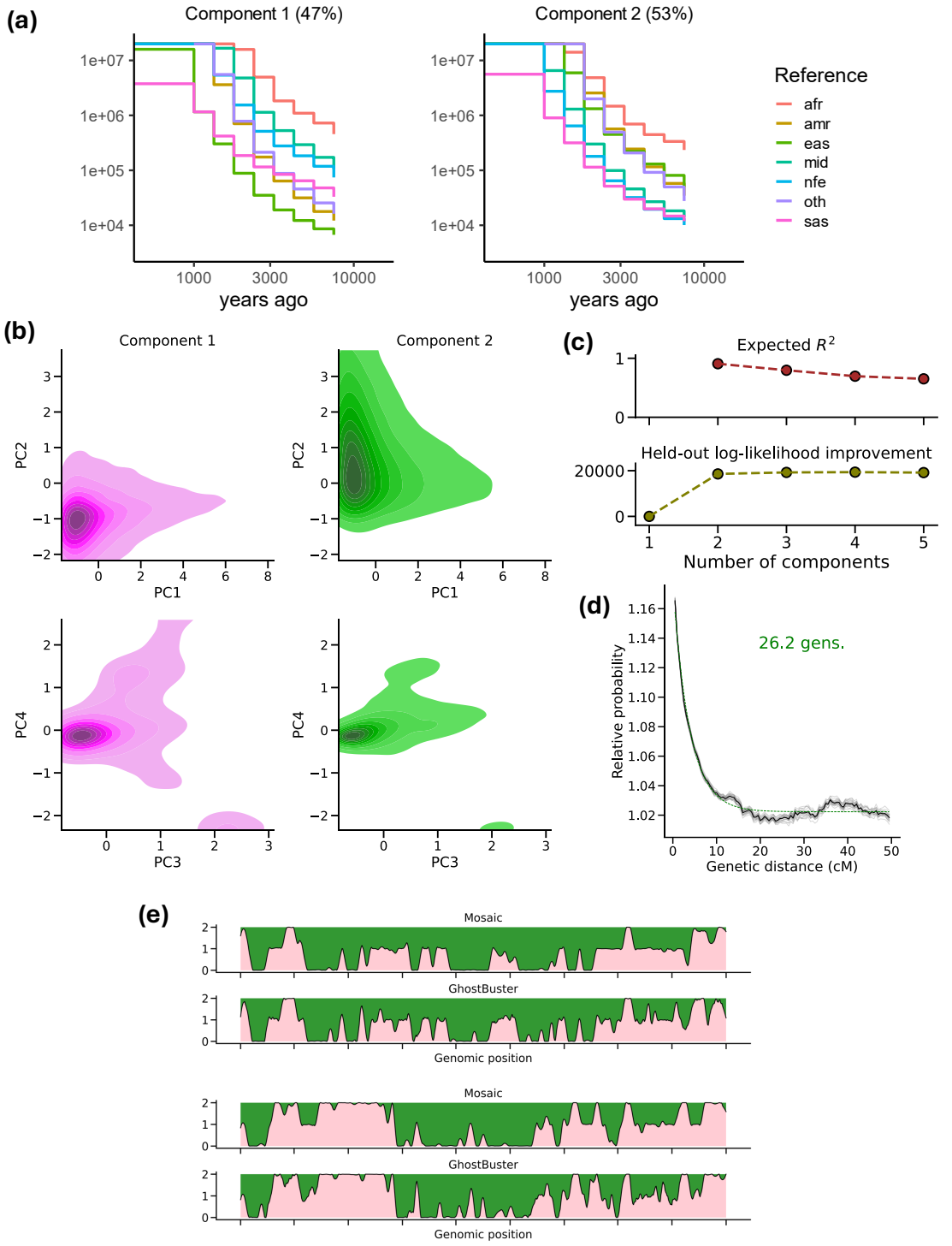


Figure 2.12: Decomposing Hazara individuals in Human Genome Diversity Project. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Local ancestry inference for GhostBuster vs. Mosaic. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. The diploid local ancestry in (e) is based on chromosome 1 for HGDP samples HGDP00102 and HGDP00103.

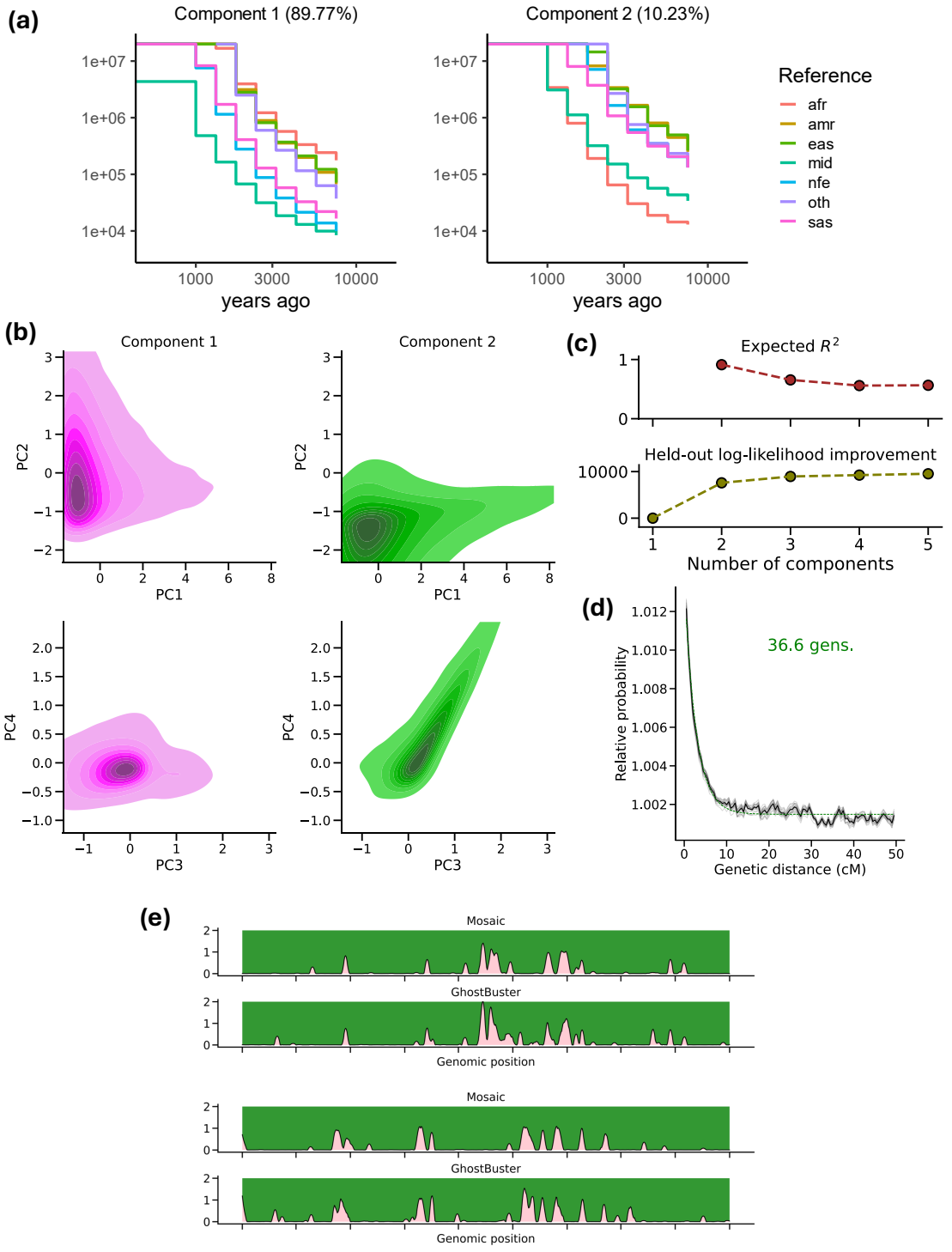


Figure 2.13: Decomposing Bedouin individuals in Human Genome Diversity Project. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Local ancestry inference for GhostBuster vs. Mosaic. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. The diploid local ancestry in (e) is based on chromosome 1 for HGDP samples HGDP00607 and HGDP00608.

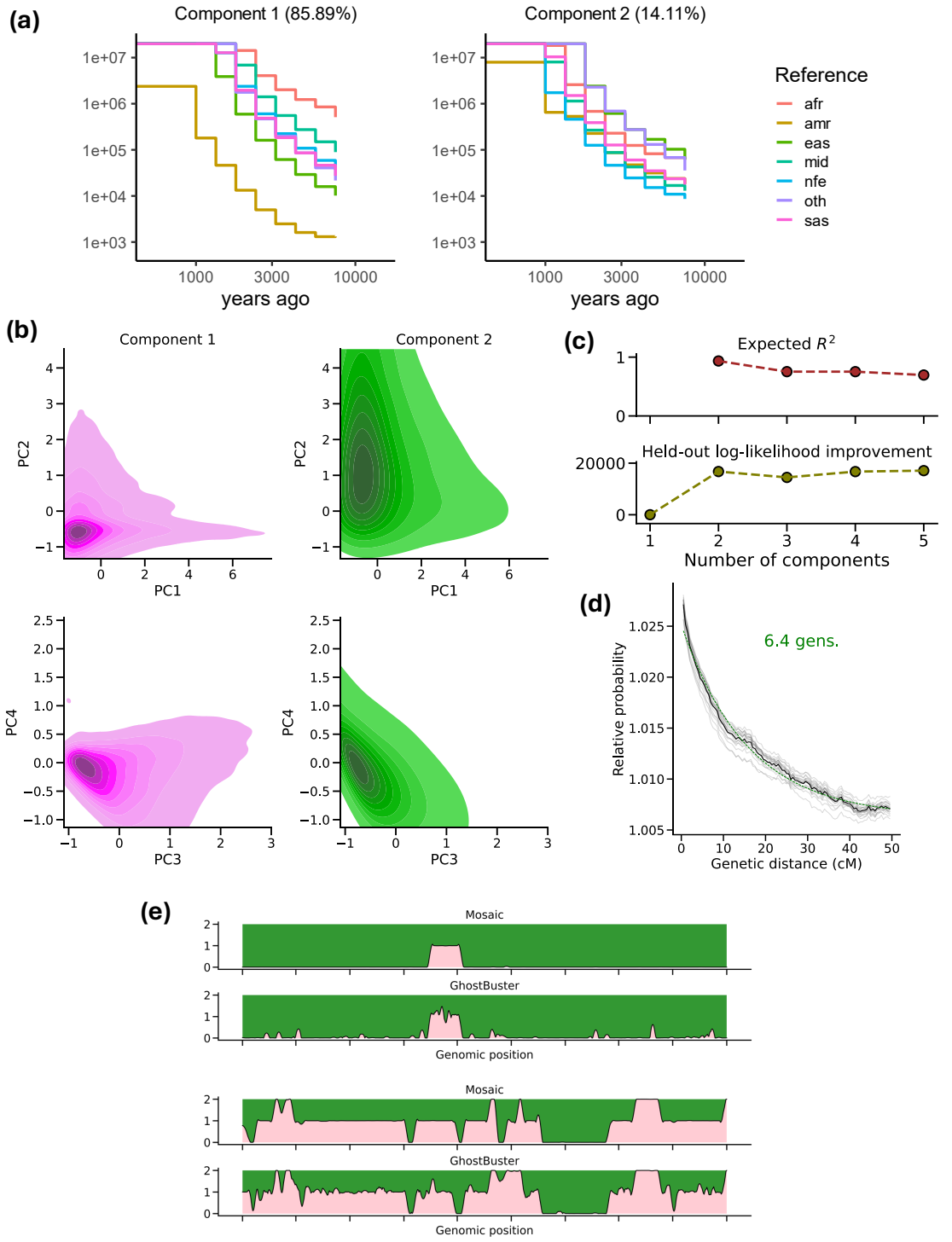


Figure 2.14: Decomposing Maya individuals in Human Genome Diversity Project. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Local ancestry inference for GhostBuster vs. Mosaic. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. The diploid local ancestry in (e) is based on chromosome 1 for HGDP samples HGDP00856 and HGDP00860.

2.7 Identifying multi-way admixture in holocene Europe

Europe's prehistory is among the most thoroughly understood, owing to the wealth of ancient DNA data, archaeological evidence, and linguistic studies conducted in the region. One of the most significant breakthroughs in ancient DNA research over the past decade has been the identification of a large-scale migration from the Steppe into Europe (and Asia), which fundamentally altered the genetic landscape of the continent over the past 5,000 years [21]. Prior to this discovery, archaeological evidence indicated the existence of only two major ancestral groups in Europe: the indigenous hunter-gatherer populations and the agricultural communities that migrated from Anatolia. However, ancient DNA studies revealed the presence of a third ancestry, which in many cases now constitutes the dominant component of modern European genomes and is also likely the origin of the Indo-European languages [152]. This third ancestry is most strongly associated with the Yamnaya people, who inhabited the Steppe region north of the Caspian Sea, in what is present-day Ukraine and Russia. The Yamnaya are thought to have lived primarily as nomadic herders and were among the earliest adopters of wheeled carts and wagons. Their influence extended to the final Neolithic cultures, such as the Corded Ware and Bell Beaker cultures, which spread throughout Europe, both culturally and genetically. See Figure 1.5B for an illustration of this 3-way admixture in Europe.

We compiled a dataset of high-coverage ancient DNA samples corresponding to 17 ancient Anatolian and Balkan Neolithic farmers, 2 ancient Western hunter-gatherers, and one Yamnaya sample. We additionally added 5 present-day British samples from 1,000 GP, which we decomposed considering a time span from 1,000 to 30,000 years ago. With the abundance of ancient DNA in Europe, it is increasingly clear that most populations, including the British samples, have undergone at least two waves of admixture. The first wave, corresponding to the origin and spread of farming from Anatolia, led to the mixing of local hunter-gatherer groups with farmers from the Near East. The second wave is believed to correspond to the later large-scale Yamnaya-like ancestry from the Steppe. When we ran GhostBuster with

three components to decompose the 5 present-day British samples with ancient farmer-like, hunter-gatherer-like and Yamnaya-like populations as references; we found that each component was closer to one of the ancient groups. Component 1, contributing 12%, is closest to Western hunter-gatherers, component 2, contributing 45.6%, is closer to farmers, and component 3, explaining the remainder, is closest to Yamnaya (see Figure 2.15a). The ancestry separates in the PC plots, most strikingly for component 3 versus the rest (see Figure 2.15b), and the coancestry curves suggest admixture dates of around 330 generations for component 1, and 262 generations for component 3 (see figures 2.15c and B.7). The proportions of hunter-gatherer, farmer, and Yamnaya ancestry in British samples roughly match previous estimates [21]. In terms of held-out log-likelihood and the expected coefficient of determination, the three-component model did not outperform a single-component model and had uncertain posterior calls (see Figure 2.15d-e). This could be attributed to the small sample size of reference populations for hunter-gatherers and Yamnaya used in this analysis, making the method unsure about the event. We discuss this shortcoming and explore future directions to enhance statistical power and robustness for this and similar scenarios in the discussion section of Chapter 3.

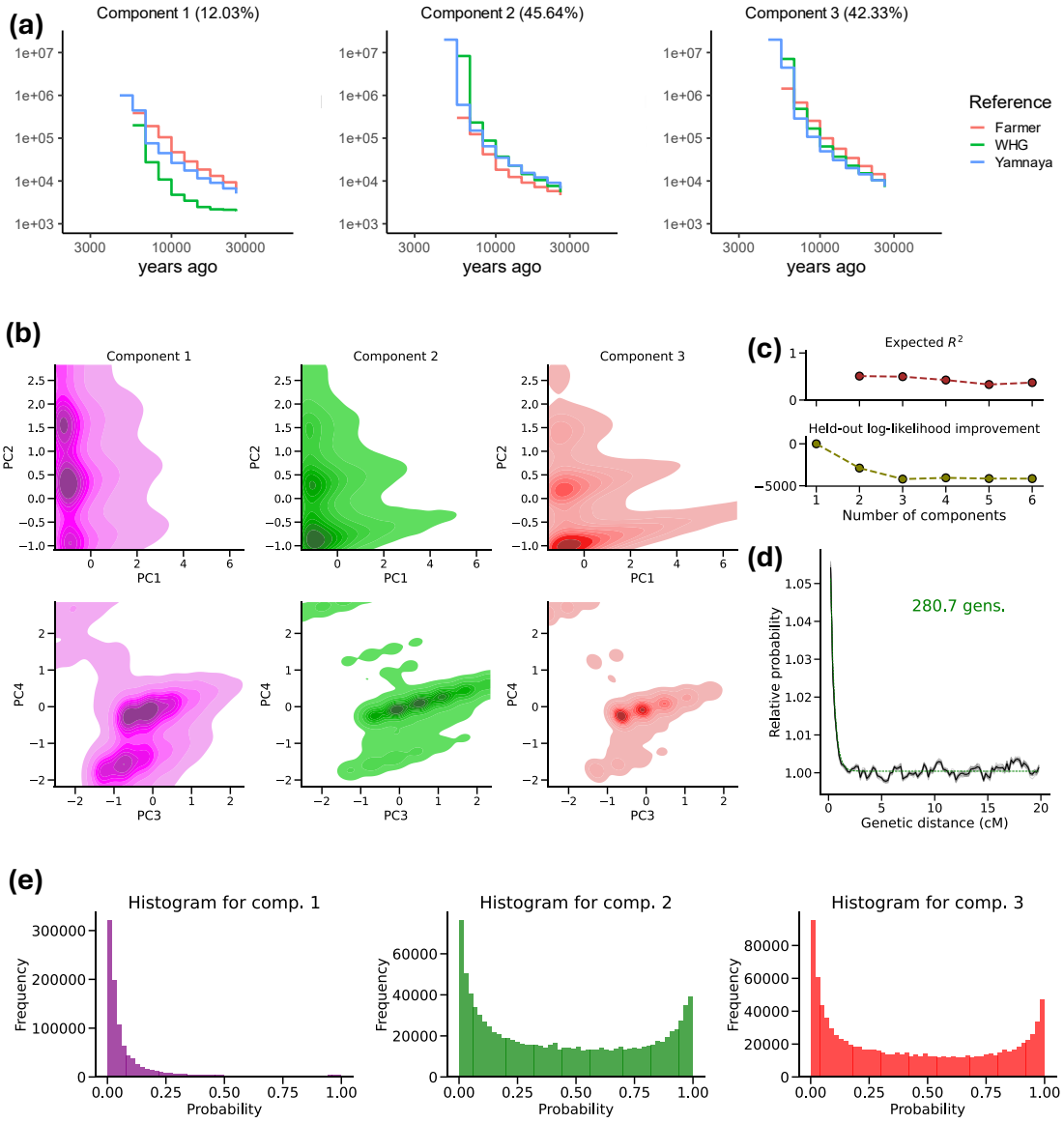


Figure 2.15: Decomposing British individuals in 1,000GP with ancient farmers, hunter-gatherers and Yamnaya. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Coancestry curve for normalized joint probability of component 1 and the inferred admixture dates (in generations). (e) Histogram of the inferred local ancestry posteriors. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5.

3

Understanding population structure in Africa using GhostBuster

Contents

3.1 Chapter overview 104
3.2 Overview of African population history 104
3.2.1 Colonization, maritime trade and Bantu expansion . . . 105
3.2.2 Back-to-Africa migration 107
3.2.3 Deep population structure within Africa 109
3.3 Evidence for widespread back-to-Africa migration . . . 111
3.3.1 Finding Eurasian-like ancestry in Africans 111
3.3.2 Dating the admixture event 116
3.3.3 Eurasian ancestry explains the Neanderthal ancestry in Africans 119
3.3.4 Mutational profiles validate Eurasian segments in Africans 126
3.3.5 Identifying source populations for Eurasian ancestry . . 131
3.4 Evidence of a deep admixture event in Africans 134
3.4.1 Ancient admixture across sub-Saharan African populations 134
3.4.2 Dating the admixture event 137
3.4.3 Mutational profiles validate the admixture event 139
3.4.4 Variation in PRDM9 activity across populations 141
3.4.5 Characterizing the admixture event using coalescence rates 146
3.4.6 Impact of the admixture event on polygenic score portability in Africa 149
3.5 Discussion 154
3.5.1 Limitations and future work 156
3.5.2 Understanding admixtures in non human organisms . . 157

3.1 Chapter overview

In this chapter, we utilize GhostBuster to analyze real genetic data from various publicly available modern and ancient human datasets to gain insights into the evolutionary history of humans in Africa.

In Section 3.2, we start by providing an overview of previous population genetics research that has contributed to our understanding of both recent and ancient history in Africa. In Section 3.3, we uncover evidence of Eurasian-like ancestry within African populations using GhostBuster. Our analysis identifies two distinct sources for this ancestry: repeated back-migration events from Eurasia over the past 10,000 years and a deeper event more than 15,000 years ago involving a population that likely remained in Africa but shares a close genetic relationship with populations outside Africa. In the remainder of the section, we characterize the back-to-Africa event and present independent strategies to validate it. These include estimating the timing of the admixture, identifying likely source populations, analyzing correlations with Neanderthal ancestry in African populations, and investigating population-specific mutational signatures. In Section 3.4, we investigate the older admixture event involving early interactions between populations that differentially relate with the out-of-Africa lineage. We propose several validation strategies for this ancient event, including distinguishing it from recent back-to-Africa migrations by the absence of Neanderthal ancestry, leveraging GC-biased mutational signatures at recombination hotspots, and demonstrating higher trait predictability in segments corresponding to the component closely related to out-of-Africa groups. Finally, in Section 3.5, we conclude the chapter by synthesizing the key findings from applying GhostBuster to both simulations and real data, and we outline potential directions for future research.

3.2 Overview of African population history

Africa is widely recognized as the cradle of modern humans, with fossil evidence of Homo sapiens dating back approximately 300,000 years [153–155]. This view is

further supported by population genetics data, which reveal Africa's extraordinary genetic diversity – more than twice that of any other continent [156] – and some of the oldest population splits in human history. Despite Africa's crucial role in human genetics, unraveling its population history poses significant challenges. The preservation of ancient DNA is hindered by the region's hot and humid climate, resulting in a scarcity of ancient genetic samples. Additionally, recent demographic events, including back-to-Africa migrations, the Bantu expansion [157], and European colonization, have obscured ancient genetic signals in modern samples, complicating efforts to reconstruct Africa's deeper genetic history.

In this section, we provide an overview of population genetics advancements in understanding the continent's population history. We begin with relatively recent events, such as European colonization, maritime trade, and the Bantu expansion in Section 3.2.1, then explore older back-to-Africa migrations in Section 3.2.2, and finally delve into the deep population structure that characterizes Africa's genetic landscape in Section 3.2.3.

3.2.1 Colonization, maritime trade and Bantu expansion

Africa has experienced significant demographic changes in the recent past. Among the most recent ones were the ones due to European colonization and the trans-Atlantic slave trade in the last five hundred years. The trans-Atlantic slave trade forcibly displaced millions of Africans, primarily from West Africa to the Americas. Many died en route, while others survived and admixed with Indigenous Americans and Europeans [158]. Within Africa, the genetic impact of European colonization was more limited due to sociopolitical barriers, segregation policies, and cultural divisions [157]. Nonetheless, some localized admixture occurred between Europeans and Africans, particularly in regions around trade ports and settler colonies, such as South Africa.

The maritime trade and Indian Ocean trading network over the past two thousand years have profoundly influenced the genetic diversity and cultural heritage of African populations. One striking example is the Malagasy people of Madagascar, who

speak an Austronesian language – remarkably different from other African languages. Several studies, including Pierron et al. [159, 160], provide evidence that the genetic composition of present-day Malagasy populations reflects a significant admixture event involving East African Bantu-speaking populations and Southeast Asians. This admixture is thought to have occurred in the first millennium CE, facilitated by the Indian Ocean trading system, which brought Austronesian-speaking seafarers from the Indonesian archipelago (approximately 7,500 kilometers away) to Madagascar. These Southeast Asian seafarers carried with them their language, agricultural techniques, and cultural practices, which later blended with those of East African populations, creating the unique genetic and cultural identity of the Malagasy people.

Apart from Madagascar, maritime trade profoundly influenced several coastal regions in eastern Africa. Notably, the Swahili coast – which includes parts of modern-day Kenya, Tanzania, and Mozambique, as well as the islands of Zanzibar, Pemba, and the Comoros, was significantly shaped by the Indian Ocean trade network. Brielle et al. [161] utilized ancient DNA from the Swahili coast to demonstrate that these extensive trade routes, connecting Africa with the Middle East, South Asia, and Europe, facilitated not only the exchange of goods but also the movement of people. This movement resulted in gene flow between African populations and traders from distant regions. Their study uncovered substantial Persian and Arab ancestry among both ancient and present-day individuals from the Swahili coast.

The Bantu expansion predates the maritime migrations and colonization of Africa and is recognized as one of the largest and most impactful migrations in African history. Bantu is one among over 500 closely related languages within the Niger-Congo language family, collectively spoken by more than 300 million people today, across an expanse of roughly 5,000,000 square kilometers. Several population genetic studies [157, 162–165] have shown the Bantu-expansion not only led to spread of the language but was also accompanied by population movement and advent of farming in Africa. The expansion is believed to be have started 4,000 to 6,000 years ago from West Africa and eventually spread to East and then South

Africa. There is still considerable debate about the exact migration paths chosen during the expansion, but nonetheless the Bantu expansion led to wide-spread mixing or replacement of local hunter gatherer and pastoralist groups across Africa.

3.2.2 Back-to-Africa migration

Migrations from Eurasia into Africa have been identified before. North African groups such as the Mozabites, Moroccans, Tunisians, and Egyptians are primarily descended from ancient inhabitants of North Africa, who have been part of the West Eurasian clade for over 10,000 years. Additionally, the Arab slave trade introduced some sub-Saharan African ancestry into these populations within the last millennium [23, 51, 58]. One of the primary barriers to migration into Africa is the vast Sahara Desert, which would have impacted migrations from Eurasia into sub-Saharan Africa. However, it is well known that the Sahara underwent periodic climate cycles due to changes in Earth's orbit, transitioning between arid and humid phases. The humid phase, often referred to as the "Green Sahara" period, could have made migrations into and out of Africa more feasible. The last Green Sahara period occurred approximately 5,000 to 11,000 years ago [166, 167].

One of the first studies to identify potential Eurasian ancestry in sub-Saharan Africans was conducted by Pickrell et al. [168, 169]. The authors utilized ALDER [53], a method designed to identify and date admixture signals through coancestry curves (as described in Section 2.3.3), and found evidence of West Eurasian ancestry in South and East African populations, but not in West or Central Africans. The estimated proportions of West Eurasian ancestry were up to 14% in South African populations and up to 50% in East African populations. The admixture events were dated to approximately 900–1,800 years ago in South Africa and 2,700–3,300 years ago in East Africa. Furthermore, the study revealed that the Eurasian ancestry in both South and East Africa was shared, leading to the theory that the most likely source of Eurasian ancestry in South Africa originated from East Africa. While the findings strongly supported a "back-to-Africa" migration event, the method's limited statistical power and reliance on reference populations may have hindered

the detection of similar events in other African populations. These populations may have experienced smaller proportions of Eurasian ancestry or older admixture events, which were beyond the resolution of the study.

In another study, Llorente et al. [170, 171] sequenced an ancient individual from the Mota caves in Ethiopia, who lived approximately 4,500 years ago. This ancient DNA sample provided direct evidence of Eurasian ancestry in African populations, as it was considered to predate the back-to-Africa migration event and thus served as a reference population without any Eurasian admixture. Using the Mota individual as a reference, the study identified 4 to 7% West Eurasian ancestry in East African populations, with lower proportions observed in some southern African populations. However, this analysis relied on a reference out-group population and summary statistics derived from allele frequency differences, which tend to be underpowered for detecting subtle admixture signals.

Several studies have identified small amounts of Neanderthal ancestry in African individuals, suggesting a back-to-Africa migration as a plausible explanation [172, 173]. Chen et al. [172] developed a novel method called IBDmix to detect introgressed Neanderthal sequences in African individuals. The authors discovered that per individual, modern-day Africans carry between 16.4 and 18 megabases (Mb) of Neanderthal sequence per individual, which is roughly one-third of the Neanderthal sequence found in modern-day non-Africans. Interestingly, 94% of the Neanderthal sequences identified in African samples were shared with non-Africans, with a significantly higher overlap observed with Neanderthal sequences found in Europeans compared to East Asians. To explain the large amount of Neanderthal ancestry that they identify, combined with very short tract lengths, the authors propose a model involving an ancient migration of humans to Neanderthal regions, followed by a more recent back migration to Africa, as the most likely explanation for the presence of Neanderthal segments in African populations. Overall, the study provides independent validation of a backflow event into Africa, though it does not have the resolution to precisely characterize the proportions and timings of this event.

Finally, ancient DNA from North Africa has revealed significant ancient migrations from the Middle East and Europe into the African continent [174–176]. The Taforalt individual, sequenced from eastern Morocco and dated to 15,000 years ago, exhibited genetic composition that was two-thirds similar to Middle Eastern hunter-gatherers and one-third similar to sub-Saharan Africans [174]. Recent studies have uncovered additional ancient DNA in North Africa, indicating not only migrants from the Middle East but also European Neolithic farmers who introduced new lifestyles to North Africa between 7,000 and 9,000 years ago [175, 176]. These migrations from the Middle East and Europe likely influenced the genetics of populations in regions north of the Sahara. However, the extent to which such migrations impacted populations south of the Sahara remains an open question, given the Sahara's historical role as a barrier to large-scale gene flow.

3.2.3 Deep population structure within Africa

The population structure in Africa prior to the Out-of-Africa event has been a topic of considerable debate, with compelling evidence supporting the existence of structured populations even before humans expanded out of Africa. Scerri et al. [177] argue that fossil data from Africa do not reflect contemporary humans emerging progressively in a single region. Instead, these fossils suggest that humans appeared at different times and in various combinations, exhibiting diverse ancestral features. This perspective is supported by findings suggesting that modern human cognition may have developed independently in multiple regions, alongside paleo-climatic models that reveal how periodic climate fluctuations likely influenced migration and adaptation [178, 179]. The structured African metapopulation model proposed by Scerri et al. challenges the simple tree-like model of human evolution. This model suggests that ancient populations could coalesce, split, experience gene flow, or go extinct over time. Scerri et al. found that the structured metapopulation model, also referred to as the "population fragmentation and coalescence model", better explains the genetic and archaeological data compared to a simple panmixia model, which assumes a single, interbreeding population.

There has also been considerable research showing evidence for unsampled or "ghost" archaic hominids in Africa [180–184]. For instance, Durvasula et al. utilized the conditional site frequency spectrum (CSFS) to identify an archaic component in West Africans that diverged from modern humans over 700,000 years ago and contributed up to 19% to the genetic makeup of contemporary West African populations. Similarly, the Relate study reported an excess of long genetic branches in Africa, akin to patterns of Neanderthal ancestry in Europe [6]. Despite these studies, archaic introgression in Africa is still debated as methods struggle to account for the possibly complex population structure in Africa.

Hollfelder et al. [185] and Ragsdale et al. [186] argue that the model of archaic admixture in Africa does not fully explain the genetic diversity of present-day African populations. Ragsdale et al. applied a maximum likelihood framework to infer demographic parameters that best explain the one- and two-locus statistics for various African groups. The authors found that the "population fragmentation and coalescence model" [177] consistently outperformed other models, including those involving archaic admixture in Africa. The model proposed by Ragsdale et al. describes a weakly structured stem, with three co-existing stem populations more than 100,000 years ago. These populations underwent migration among themselves and eventually merged in different proportions to form present-day African populations. The proposed model aligns more closely with the archaeological data from that period but remains limited by model misspecification, as it imposes constraints on the range of possible models when inferring demographic parameters.

Ancient DNA in Africa is scarce due to poor preservation in hot and humid conditions, but there are a few samples spanning the past 18,000 years. Lipson et al. [187] argue that analyzing these ancient samples helps reduce confounding factors from more recent events such as back-to-Africa migrations, the Bantu expansion, and European colonization. By studying the ancient DNA of 31 samples, the study aimed to understand the population structure of hunter-gatherers in central and southern Africa, which represent some of the deepest splits in modern human lineages. Using supervised PCA, the researchers categorized the ancient samples based on three

modern African groups: the San from southern Africa, the Mbuti from central Africa, and the Dinka from northeastern Africa. The authors found increasing regionalization among the ancient samples, with genetic similarities best predicted by geographic locations, suggesting minimal long-range migrations within these groups. The observed patterns of variation in the ancient samples were best explained by admixture involving at least three ancestries between 20,000 and 50,000 years ago.

Finally, Cousins et al. [188] provide statistical evidence for archaic admixture events that potentially affected all human populations. The study proposes a novel method based on the pairwise sequentially Markovian coalescent (PSMC [189]) to identify events of population splits and mergers. Their analysis suggests that human populations underwent an admixture event approximately 300,000 years ago with two deeply divergent human species, which diverged around 1.5 million years ago. The admixture proportions in all analyzed human populations were roughly 80:20.

3.3 Evidence for widespread back-to-Africa migration

To investigate the admixture history within Africa, we used GhostBuster to analyze genetic data from 17 modern sub-Saharan African groups from the Human Genome Diversity Project (HGDP), the 1,000 Genomes Project (1,000 GP), and Simon's Genome Diversity Project (SGDP). In addition, we examined five ancient African samples (coverage between 10.52 - 18.51), including a 7,900-year-old sample from Shum Laka in Cameroon [190]; 1,900- and 400-year-old samples from Ballito Bay, Eland Cave, and Newcastle in South Africa [191]; and a 4,500-year-old sample from Mota Cave in Ethiopia [170]. The groups we analyzed are listed in Table 3.1, and a map of the geographic locations of all modern and ancient samples is provided in Figure 3.1c.

3.3.1 Finding Eurasian-like ancestry in Africans

To detect Eurasian-like ancestry in African samples, we combined non-Finnish European (NFE), East Asian (EAS), and South Asian (SAS) individuals into a

Population	Dataset	Location	Samples
Esan	1,000GP	Nigeria	20
Gambian	1,000GP	Gambia	20
Luhya	1,000GP	Kenya	20
Mende	1,000GP	Sierra Leone	20
Biaka	HGDP	Central African Republic	22
Mandenka	HGDP	Senegal	21
Mbuti	HGDP	Democratic Republic of Congo	12
San	HGDP	Namibia and South Africa	6
Yoruba	HGDP	Nigeria	21
Bantu Kenya	HGDP	Kenya	11
Bantu SouthAfrica	HGDP	Botswana or Namibia	8
Somali	SGDP	Kenya	1
Luo	SGDP	Kenya	2
Masai	SGDP	Kenya	2
Ju Hoan North	SGDP	Namibia	4
Dinka	SGDP	Sudan	3
Khomani San	SGDP	South Africa	2
Total			**195**

(a)

ID	Publication	Location	Coverage	Age
I10871	Lipson et al. [190]	Shum Laka, Cameroon	18.51	7890
new001	Schlebusch et al. [191]	Newcastle, South Africa	11.073	418
ela001	Schlebusch et al. [191]	Eland Cave, South Africa	10.52	493
baa001	Schlebusch et al. [191]	Ballito Bay A, South Africa	12.93	1909
I5950	Llorente et al. [170]	Mota cave, Ethiopia	11.28	4472
Total				**5**

(b)

Table 3.1: Sub-Saharan African populations analyzed. (a) Modern samples and, (b) Ancient samples

single "super-population" and labeled them collectively as Eurasians. Each African population was decomposed using GhostBuster using this Eurasian and other African populations in the dataset as reference groups. The analysis was conducted with two components and pre-specified coalescence rates based on genome-wide coalescence rates for Eurasians and the specific African population being analyzed. This approach is justified by our empirical observation that Africans carry only

a small proportion of Eurasian ancestry, making genome-wide ancestry a reliable proxy for each component. To avoid potential confounding factors, we excluded recently admixed groups such as individuals from the Middle East and Americas from the reference panel. We focused on coalescence events occurring between 0 and 50,000 years ago, dividing this time span into 20 epochs on a log scale, ranging from 1,000 to 50,000 years. Coalescence events after 50,000 years were disregarded, as they are likely associated with older events before the out of Africa migration.

We identified between 1.2% and 19.3% Eurasian-like ancestry among the modern African groups analyzed, with the highest proportions observed in the Somali and Masai from East Africa (approximately 19.3% and 13.5% respectively, see Figure 3.1d). In ancient samples, we detected around 1.2% Eurasian-like ancestry in the 4,500-year-old Mota individual from Ethiopia and 1.6% in the 7,900-year-old Shum Laka individual from Cameroon. In contrast, only 0.37% Eurasian-like ancestry was found in the 1,900-year-old Ballito Bay individual from South Africa, an ancestor of modern-day San populations. It is important to note that these proportions represent ancestry where the target lineage coalesces more quickly with Eurasian populations than with other African populations, rather than indicating the specific geographic origins of the source groups. As we will demonstrate later, this Eurasian-like ancestry results from two distinct events: a repeated back-migration from Eurasia within the last 10,000 years and an ancient admixture event over 20,000 years ago between populations that are differentially related to the out-of-Africa population.

We estimated coalescence rates using the inferred local ancestry, which demonstrated strong concordance with the genome-wide coalescence rates used for model fitting, confirming the robustness of our approach (see figures 3.1a, B.2, B.3, and B.5). The Eurasian and African-like components are clearly separated in a PCA visualization of coalescence counts and opportunities, further validating the successful identification of two ancestries (see Figure 3.1b). Moreover, the high coefficient of determination underscores the reliability of the method's local ancestry assignments (see Figure 3.1e). A comparison of inferred local ancestry with background selection B-statistics [192] revealed no significant correlation, indicating

that the method is not substantially influenced by biases associated with genomic variation in background selection (see Figure B.8).

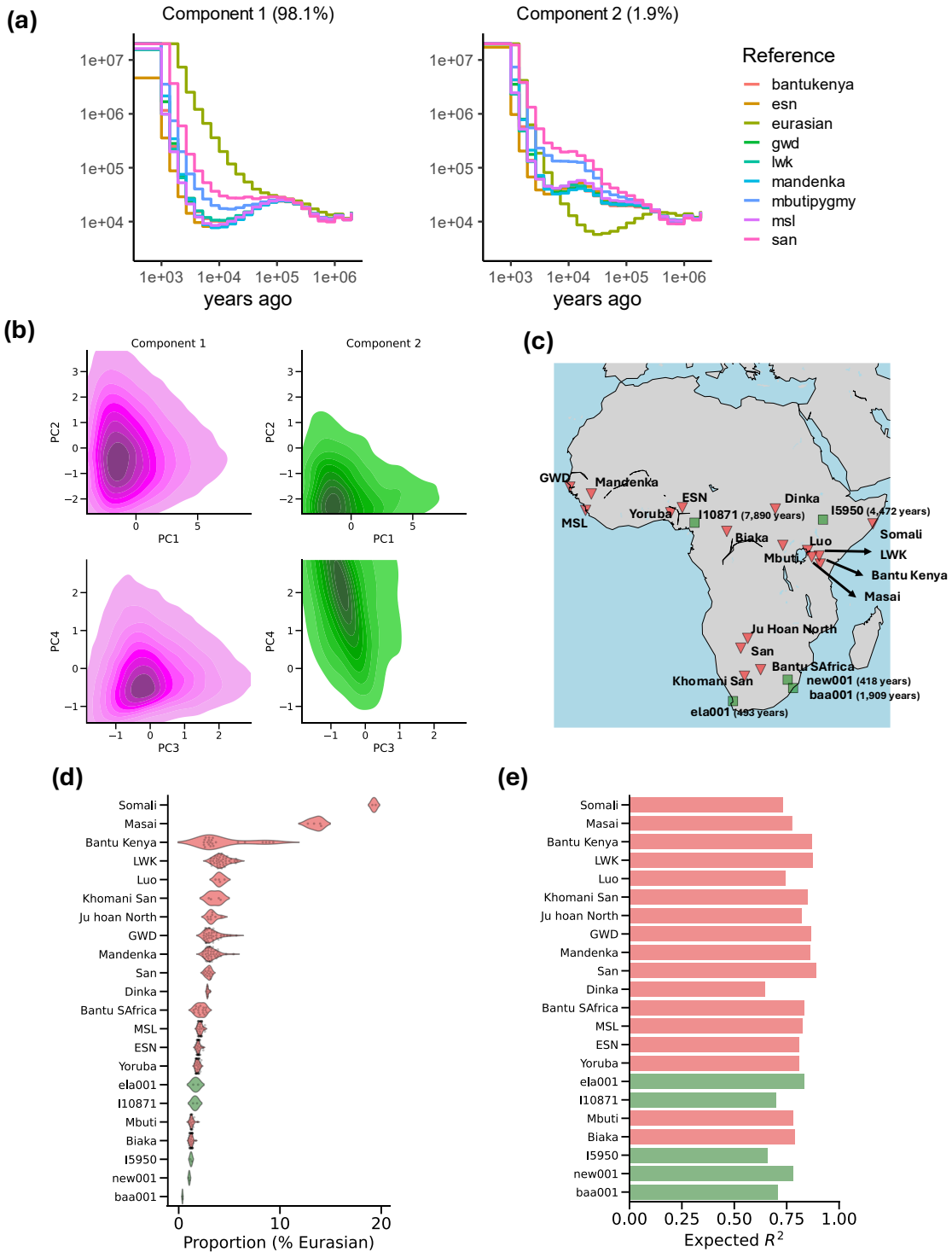


Figure 3.1: Finding Eurasian-like ancestry in sub-Saharan Africans. (a) Inferred inverse coalescence rates and proportions when decomposing Yorubans: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Map of the African samples analyzed with red inverted triangles corresponding to modern samples in HGDP, 1,000GP and SGDP, and green squares for 5 ancient African samples (with their estimated sampling ages in brackets). (d) Proportion of Eurasian-like ancestry and (e) Expected coefficient of determination across all African populations analyzed. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. Individual data-points overlaid on the violin plots.

3.3.2 Dating the admixture event

We dated these admixture events using coancestry curves derived from the local ancestry inferred by GhostBuster. To capture population-specific variation, we estimated admixture dates separately for each African population. However, for single-sample individuals, we could not apply the pseudo-individual normalization method described in Section 2.3.3, which is critical for addressing population-specific drift and ensuring unbiased admixture date inference. We therefore do not report admixture dates for ancient individuals and the Somali population.

We fit the coancestry curves using three models: a single-date admixture model, a two-date admixture model, and a continuous admixture model with a constant migration rate. Our analysis revealed that a single admixture date did not adequately fit the coancestry curves for several African populations, as indicated by the visual fit of the coancestry curves and significant in-sample log-likelihood difference between the single-date and two-date admixture models ($> 20\times$ difference in number of parameters). This finding suggests a more complex genetic history (see Figure 3.2a-c). Both the two-date and continuous admixture models provided an improved fit to the coancestry curves, as measured by the log-likelihood of the least squares fit. Notably, the two-date admixture model outperformed the continuous migration model in 14 out of 16 African populations (see Figure 3.2d).

The two admixture dates exhibited considerable variation across geographic regions. In East African populations such as the Luhya (LWK), Bantu Kenya, Masai, and Luo, the more recent admixture event occurred $\leq$ 3,000 years ago. In contrast, for West and Central African populations, including the Mende (MSL), Gambian (GWD), Mandenka, Esan (ESN), Yoruba, Mbuti, and Biaka, the recent admixture event is estimated to have occurred between 3,000 and 7,000 years ago. South African populations exhibited intermediate admixture dates ranging from 400 to 4,500 years ago, with the Khomani San showing the most recent admixture event at approximately 400 years ago. The older admixture event dates were at least 15,000 years old, ranging from 15,000 to 35,000 years across different African populations. Population-specific admixture dates and contributions are summarized

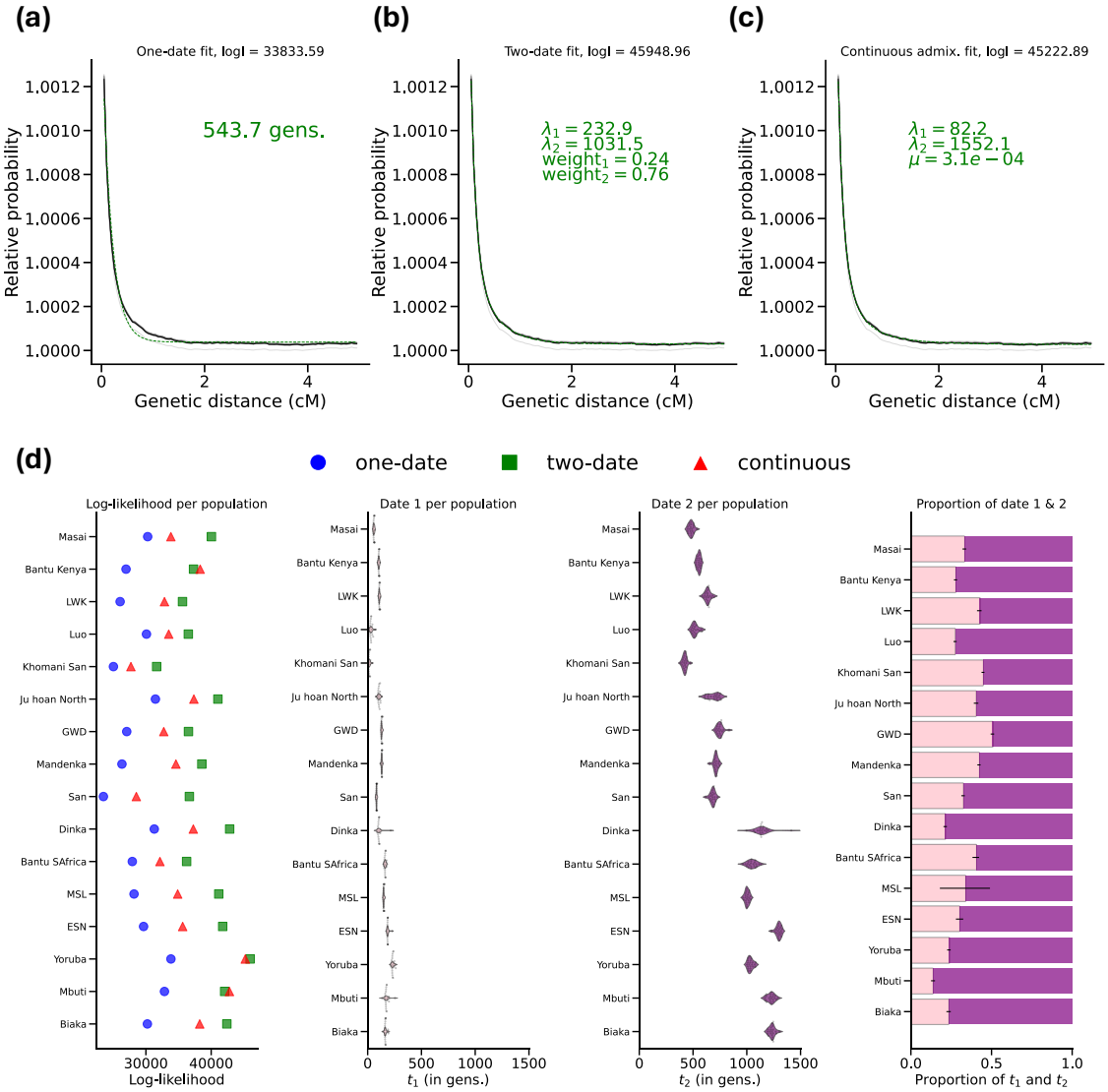


Figure 3.2: Dating the Eurasian-like ancestry in sub-Saharan Africans. Coancestry curves and (a) one-date admixture model fit, (b) two-date admixture model fit and (c) constant migration rate continuous admixture model fit for Yorubans. (d) Per-population least squared log-likelihood corresponding to various admixture models along with the admixture dates and weights inferred using two-date admixture model. The log-likelihood values are aggregated across 20 jackknife runs. Populations are sorted based on highest to lowest proportion of Eurasian-like ancestry, pink represents the younger admixture date and purple represents the older admixture date in (d). Individual data-points overlaid on the violin plots. Admixture dates in the plot are in generations.

in Figure 3.2d and Table B.2. It is important to note that these admixture dates have not been adjusted for potential inaccuracies in the recombination map, as reliable error estimates were unavailable. This limitation may introduce a downward bias in the dating of the older admixture event [137].

Given the distinct dates of these two admixture events, we perform separate validation analyses for each. To differentiate between the events, we first calculate

the segment length of Eurasian ancestry segments by applying a threshold to the local ancestry posterior at 50%, then combining adjacent matching ancestry segments. Next, we estimate the posterior probability that a segment belongs to either the younger (typically in the last 10,000 years) or older (typically between 15,000 to 35,000 years) admixture event for each population, using segment length as the determining factor. In particular, we use the admixture dates and proportions as inferred using per-population coancestry curve dating to assign this posterior. Given the assumption that segment lengths follow an exponential distribution conditioned on the admixture time, we proceed with the following model, given segment length L, admixture times t_{younger} and t_{older} and priors p_{younger} and p_{older} from coancestry curve dating:

$$\begin{aligned}
\textbf{Likelihood: } & \mathbb{P}(L|t) = te^{-t*L} \\
\textbf{Prior: } & \mathbb{P}(t = t_{\text{younger}}) = p_{\text{younger}} = 1 - p_{\text{older}} \\
\textbf{Posterior: } & \mathbb{P}(t = t_{\text{younger}}|L) = \frac{\mathbb{P}(L|t_{\text{younger}}) \cdot p_{\text{younger}}}{\mathbb{P}(L|t_{\text{younger}}) \cdot p_{\text{younger}} + \mathbb{P}(L|t_{\text{older}}) \cdot p_{\text{older}}} \\
& \mathbb{P}(t = t_{\text{older}}|L) = \frac{\mathbb{P}(L|t_{\text{older}}) \cdot p_{\text{older}}}{\mathbb{P}(L|t_{\text{younger}}) \cdot p_{\text{younger}} + \mathbb{P}(L|t_{\text{older}}) \cdot p_{\text{older}}}
\end{aligned} \tag{3.1}$$

We estimate the posterior probability of each position in the local ancestry belonging to either the younger or older admixture event. By combining these posterior estimates with GhostBuster's initial local ancestry inference, we obtain the probabilities that each segment represents either a younger or older Eurasian ancestry. Additionally, we construct histograms reflecting the segment length distribution for each population analyzed (see Figure B.9). In subsequent sections, we demonstrate that these younger and older Eurasian segments differ in their mutational signatures and levels of Neanderthal enrichment. We hypothesize that the younger Eurasian segments correspond to a back-to-Africa migration originating from Eurasia, while the older segments likely represent an ancient admixture event between African groups – those who likely remained in Africa and those who migrated out of Africa. For much of the analysis in subsequent subsections, we focus

primarily on the younger segments to highlight their distinct properties relative to the older segments. However, a detailed analysis of the older admixture segments is conducted in further Section 3.4.

3.3.3 Eurasian ancestry explains the Neanderthal ancestry in Africans

Previous research has shown that modern Africans carry small amounts of Neanderthal ancestry in their genomes, likely due to a back-migration event from Europe [139, 172, 193, 194]. We look for this Neanderthal ancestry as a way to validate the Eurasian segments identified in our analysis. Specifically, we used GhostBuster to detect Neanderthal ancestry in African individuals by jointly analyzing genealogies inferred from both modern African samples and ancient samples, including three Neanderthal genomes from the Altai, Chagyrskaya, and Vindija caves.

Our analysis focused on coalescence events that occurred between 50,000 and 2 million years ago, partitioning this time frame into 20 epochs on a logarithmic scale. To specifically detect Neanderthal segments rather than other ancient events, we restricted the analysis to coalescence events with sequenced Neanderthals only, ignoring the rest of the tree. We began by setting the cluster coalescence rates to genome-wide coalescence rates observed between sequenced Neanderthals themselves and the target African populations. Using these predefined coalescence rates, we employed the EM algorithm to estimate admixture proportions, admixture time, and local ancestry across the sample. In the final iteration of the EM algorithm, we update the coalescence rates based on the inferred local ancestry instead of using the pre-specified genome-wide coalescence rates. This adjustment fine-tunes the estimates to account for differences between the Neanderthal sources that contributed ancestry to Africans and the sequenced Neanderthals, allowing us to obtain unbiased and accurate estimate of Neanderthal local ancestry in African genomes.

To validate our approach for detecting Neanderthal ancestry, we first estimated Neanderthal ancestry in Eurasians, recovering approximately 1–2% across six modern

Eurasian groups (see Figure B.10). These proportions align with prior findings [195, 196]. East Asian populations exhibited approximately 20% higher Neanderthal ancestry, with an average of 1.5%, compared to South Asians and Europeans, which showed about 1.25% Neanderthal ancestry – consistent with previous studies [197, 198]. Furthermore, we observed little to no Neanderthal ancestry in genomic regions previously identified as "archaic deserts" [173, 199] (see Figure B.10d). It is important to note our approach circumvents the need for un-admixed reference samples or identity-by-descent (IBD) segments with Neanderthals, and utilizes the entire coalescence history with sequenced Neanderthal directly, making it more robust and powerful for finding Neanderthal ancestry in Africans.

We found that African samples from the HGDP+1,000GP dataset exhibit approximately 0.004% to 0.056% Neanderthal ancestry, with the lowest levels observed in Mbuti individuals and the highest levels in Bantu Kenyans, among the populations analyzed. Figure 3.3a shows the inverse coalescence rates for Neanderthal ancestry in Yoruba (as a representative population). Our analysis revealed a strong correlation between the genome-wide Neanderthal proportion and the Eurasian-like ancestry proportion inferred with GhostBuster, yielding a correlation $R^2 = 0.51$. To examine Neanderthal ancestry correlations associated with the younger and older admixture events, we fitted a joint regression model to predict Neanderthal ancestry proportion based on the proportions of younger and older Eurasian ancestry within individuals. The younger Eurasian ancestry proportion had a slope of 1.22%, whereas the older Eurasian ancestry showed a slope of 0.67%. These slopes represent the increase in Neanderthal ancestry per unit increase in Eurasian-like ancestry, effectively quantifying the proportion of Neanderthal ancestry within these Eurasian segments. The joint regression also inferred an intercept not significantly different from 0 at 0.009%. This suggests a probable absence of Neanderthal ancestry in non-Eurasian-like segments and highlights a disparity in Neanderthal ancestry between the two Eurasian-like segments identified in Africa (see Figure B.11).

At the local ancestry level, we defined the enrichment of Neanderthal ancestry within Eurasian segments as the mean local ancestry posterior for Neanderthal ancestry in Eurasian segments (determined by a local ancestry threshold of 0.5) divided by the mean posterior in non-Eurasian segments. Under a model with no correlation between Neanderthal and Eurasian-like ancestry in Africa, this enrichment should be close to unity. However, we observed a 21.9× enrichment of Neanderthal ancestry within Eurasian segments compared to non-Eurasian segments (see Figure 3.3b). Further stratification of Eurasian segments by length showed even higher enrichment for longer segments and lesser enrichment for shorter segments: 31.4× for segments longer than 0.5cM, 23.1× for segments between 0.2 and 0.5cM, 11.3× for segments between 0.1 and 0.2cM, and 4.3× for segments shorter than 0.1cM.

The disparity in estimated coefficients for older and younger admixture in Figure B.11, combined with reduced enrichment in shorter segments, suggests the possibility that Neanderthal ancestry is entirely absent from the older admixture event and the reduced signal is only driven by misassignment of segments originating from the more recent event to the older event. To obtain more robust estimates of Neanderthal local ancestry while minimizing false-positive rates, we recalculated Neanderthal ancestry in Africans by masking non-Eurasian-like segments and only using the Eurasian-like segments during coalescence rate inference. This masking produced genome-wide proportions of Neanderthal ancestry that were comparable to the original run without masking (paired t-test, $p = 0.18$).

We dated the admixture event using Eurasian segments that specifically contained at least one Neanderthal segment. This analysis aimed to determine whether Neanderthal ancestry was present in both the younger and older Eurasian segments. If a date exceeding 20,000 years is detected, it would indicate that the older Eurasian segments also carried Neanderthal ancestry. Conversely, detecting only more recent admixture dates would suggest that Neanderthal ancestry arrived exclusively through the more recent admixture event. We constructed coancestry curves using hard-called local ancestry, where Eurasian-like segments with Neanderthal ancestry

(posterior probability ≥ 0.1) at at least one position were assigned a value of 1, while all other regions, including Eurasian-like segments without Neanderthal ancestry, were assigned a value of 0. We then fit multiple admixture models to the generated coancestry curves. Our analysis revealed that a single admixture date provides an adequate (in-sample log-likelihood difference between single- and two-date admixture models $< 2 \times$ difference in number of parameters, see Figure B.12) fit for most African populations, with some exceptions. For instance, populations such as Bantu Kenya from East Africa fit better with a two-admixture-date model. Similarly, West African populations like Gambian (GWD) and Mende (MSL) also showed a slight improvement with a two-admixture-date model, though these estimates were more noisy (see Figure B.12). The admixture dates corresponding to both single- and two-date models for some populations ranged approximately from 4,000 to 10,000 years ago (see Figure 3.3c-d). These admixture dates fall within the range of the younger admixture dates inferred previously in Section 3.2, and are not consistent with the older admixture date. This suggests that the older admixture date, which occurred approximately 15,000 - 35,000 years ago, did not carry Neanderthal ancestry. Furthermore, admixture dates varied based on the present-day geographic location of samples. West African populations, such as Yoruba and Esan, exhibited the oldest admixture dates, approaching 10,000 years ago – aligning with the green Sahara period [166, 167] and well before the expansion of Bantu-speaking peoples. Central African hunter-gatherers, including Mbuti individuals, exhibit minimal Neanderthal ancestry, leading to noisier coancestry curve dating and were therefore excluded from the main figure. More detailed results, including the dating of Neanderthal ancestry in Mbuti individuals, can be found in Figure B.12. These findings suggest that the Neanderthal ancestry present in Africans is predominantly associated with the younger admixture event, which occurred during the Holocene, starting shortly after the end of the last Ice Age.

To further validate our hypothesis that the older Eurasian segments do not carry substantial Neanderthal ancestry, we analyzed the mean posterior probability of Neanderthal ancestry as a function of segment length. We compared the empirical

estimates of mean Neanderthal posterior probabilities to theoretical predictions from a two-date admixture model, under the assumption that only the younger admixture event contributes Neanderthal ancestry. The theoretical model is defined as follows:

$$\begin{aligned} P(L \leq l|\text{model}_{\text{younger}}) &= 1 - e^{-lt_{\text{younger}}} \\ P(L \leq l|\text{model}_{\text{older}}) &= 1 - e^{-lt_{\text{older}}} \\ P(\text{Neanderthal}|\text{model}_{\text{younger}}) &= \alpha \\ P(\text{Neanderthal}|\text{model}_{\text{older}}) &= 0 \\ P(\text{model}_{\text{younger}}) = 1 - P(\text{model}_{\text{older}}) &= p_{\text{younger}} \end{aligned} \tag{3.2}$$

where, $\text{model}_{\text{younger}}$ and $\text{model}_{\text{older}}$ correspond to Eurasian-like ancestry in Africans derived from the younger and older admixture events, respectively. Here, α represents the proportion of Neanderthal ancestry in the Eurasian source population contributing to the younger admixture event; t_{younger} and t_{older} are the admixture dates, and p_{younger} is the proportion of ancestry contributed by the younger admixture event. These parameters are estimated using the coancestry curves presented in Section 3.3.2. Based on the Equation 3.2, the probability of Neanderthal ancestry conditioned on segment length can be expressed as:

$$P(\text{Neanderthal}|L \leq l) = \frac{\alpha p_{\text{younger}}(1 - e^{-lt_{\text{younger}}})}{p_{\text{younger}}(1 - e^{-lt_{\text{younger}}}) + (1 - p_{\text{younger}})(1 - e^{-lt_{\text{older}}})} \tag{3.3}$$

We estimated confidence intervals for the theoretical models by incorporating uncertainties in coancestry dating. Our results showed that the theoretical model, assuming no Neanderthal ancestry from the older admixture event, fit the empirical data well across several populations (see Figure B.13). This finding further supports the hypothesis that the older admixture event did not contain substantial Neanderthal ancestry.

Finally, we analyzed Neanderthal content in ancient African samples, finding 0.026% and 0.017% Neanderthal ancestry in ela001 (493 years) and new001 (418 years) individuals respectively, and 0.01% in the 1,900-year-old Ballito Bay individual (baa001). In contrast, we detected no Neanderthal ancestry ($< 10^{-9}$%) in ancient samples from the Mota Cave (4,472 years) and Cameroon (7,890 years). This

again suggests that the younger admixture event which might be missing in Mota and Cameroon individuals contributed Eurasian-like ancestry with detectable Neanderthal ancestry, whereas the older admixture event might be associated with reduced or no Neanderthal ancestry.

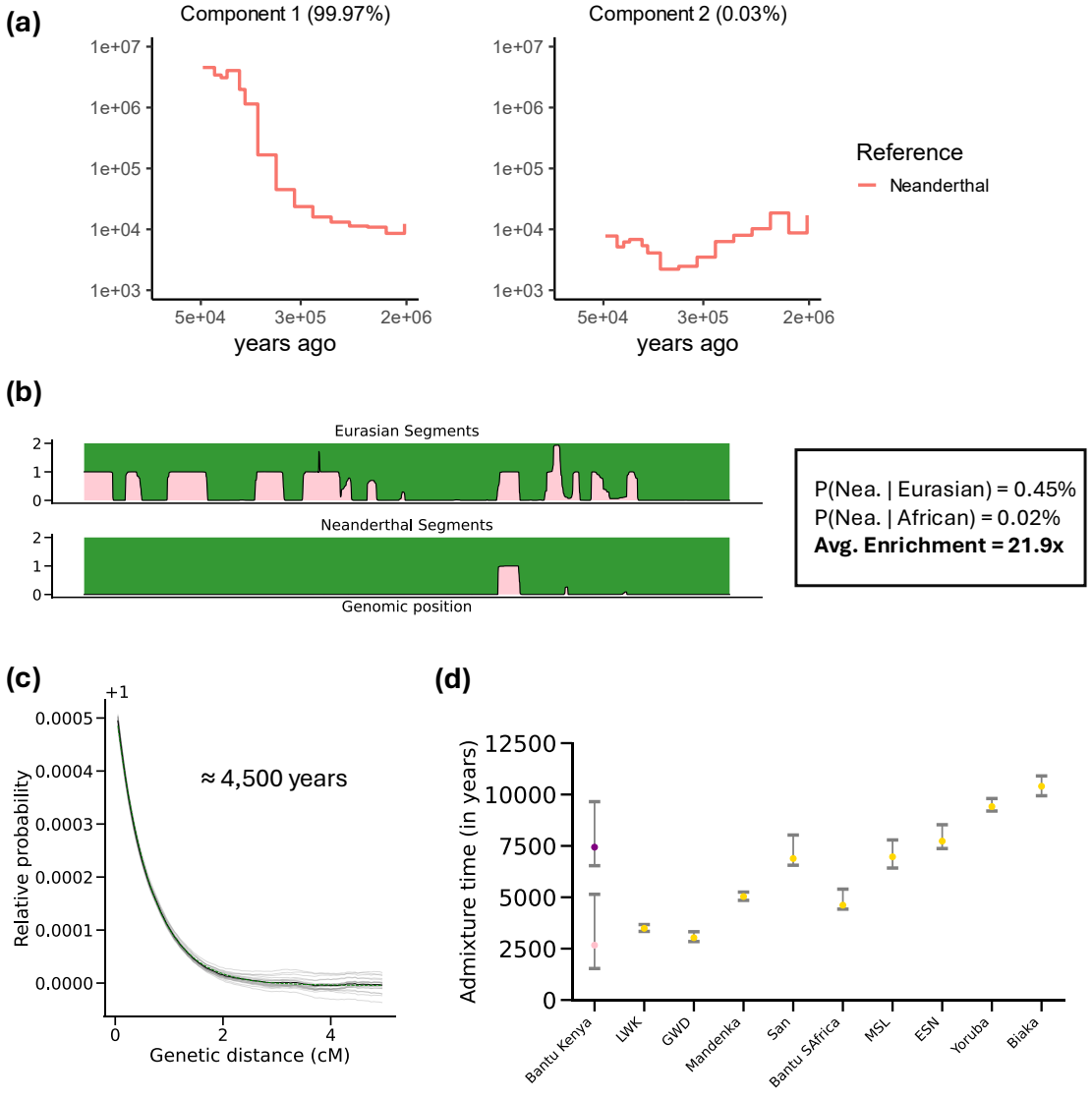


Figure 3.3: Eurasian ancestry explains the Neanderthal ancestry in Africans. (a) Inferred inverse coalescence rates and proportions for Neanderthal ancestry in Yorubans: the line represents the inverse coalescence rate profiles with sequenced Neanderthals. (b) Sample HGDP01405 showing a 20 Mb region enriched for Neanderthal ancestry, overlapping Eurasian-like segments. The average enrichment is calculated as the ratio of the mean posterior probability of Neanderthal ancestry in Eurasian segments, $P(\text{Nea} \mid \text{Eurasian})$, to the mean posterior probability in African segments, $P(\text{Nea} \mid \text{African})$. (c) Coancestry curve showing the normalized joint probability and single admixture date fit for Eurasian segments carrying Neanderthal ancestry, averaged across all African samples analyzed. (d) Admixture dates derived from coancestry curves generated by Eurasian segments carrying Neanderthal ancestry for each population. For Bantu Kenya a two-date admixture model fit is shown, as it better explains the coancestry patterns. More detailed results are provided in Figure B.12. Grey lines in (c) and error bars in (d) represent jackknife estimates and 95% confidence intervals around the mean. For panel (d) single date admixtures shown in golden, whereas two date admixtures shown in pink and purple. Results derived from HGDP+1,000GP+aDNA+archaic genealogies.

3.3.4 Mutational profiles validate Eurasian segments in Africans

Evidence suggests that human mutation rates have evolved over short time scales and independently across different populations [76, 200]. One notable example is the TCC to TTC mutation, which occurs at up to twice the rate in West Eurasian groups compared to East Asians or Africans. Another example is GC-biased gene conversion (GCbGC) [201], which preferentially selects GC alleles (termed "strong") over AT alleles (termed "weak"). GCbGC is influenced by factors such as population size and recombination hotspots: populations with larger effective population sizes or fewer bottlenecks exhibit higher rates of weak to strong mutations than populations with reduced population sizes. Figure B.14 illustrates the evolution of weak to strong mutations relative to strong to weak mutations across various continental populations.

To investigate this within the context of the back-migration event, we define the normalized mutation rate similar to the Relate [6] paper for a specific mutation type (example, TCC to TTC, or weak to strong) in a given population. This normalized mutation rate (NMR) is calculated as the ratio of the number of mutations of that specific type that have arisen in the population to the number of mutations of all other types that have arisen in the same population. Formally, this is expressed as:

$$\mathrm{NMR}_{\text{pop, type}} = \frac{\sum_{mut} \mathbb{1}(mut \in \text{type})\, N_{mut,pop}}{\sum_{mut} \mathbb{1}(mut \notin \text{type})\, N_{mut,pop}} \tag{3.4}$$

where, $\sum_{mut}$ sums all the mutations in the genealogy, $N_{mut,pop}$ counts the number of individuals belonging to population 'pop' carry the mutation 'mut'. We further stratify mutation rates by allele age, dividing them into 10 epochs ranging from 1,000 years to 1 million years. To facilitate meaningful comparisons, we further divide these normalized mutation rates (NMR) and the normalized mutation rates (NMR) in African populations:

$$\mathrm{NMRE}_{\text{pop, type}} = \frac{\mathrm{NMR}_{\text{pop, type}}}{\mathrm{NMR}_{\text{AFR, type}}} \tag{3.5}$$

This measure, referred to as the Normalized Mutation Rate Enrichment (NMRE), quantifies the relative enrichment of specific mutation types in a given population compared to their normalized mutation rates in Africans. A significant deviation from unity indicates that the specific mutation type is either enriched (greater than 1) or depleted (less than 1) in the population relative to the African baseline. We examine this enrichment across various continental populations, including all non-Finnish Europeans (NFE), East Asians (EAS), South Asian (SAS) and Middle Eastern and North African (MID) groups in HGDP. Notably, we analyze the enrichment for Eurasian and non-Eurasian segments in African samples separately. To correct for mutational biases in local ancestry fitting, we down-sample Eurasian and non-Eurasian segments at each location to equalize the number of descendants in both categories. We perform 1,000 chromosome-wise bootstraps to estimate the confidence interval around the normalized mutation rate enrichment.

Examining the TCC to TTC mutation types, we observe significant enrichment not only in European, Middle Eastern, and South Asian populations but also in the Eurasian-like segments within African individuals. In Figure 3.4a,b, this enrichment is plotted separately for Eurasian segments, conditioned on their segment lengths. Both long and short Eurasian segments in Africans exhibit significant enrichment in TCC to TTC mutations during the period from 1,000 to 40,000 years ago, consistent with the evolutionary trajectory of these mutations, which rose sharply around 15,000 years ago and declined approximately 2,000 years ago [200]. The enrichment is particularly pronounced in longer Eurasian segments, which are more likely to be associated with the younger admixture date, peaking at approximately 2× around 20,000 years ago – levels comparable to those observed in Europeans. This pattern suggests a European source as the most probable origin for the back-to-Africa migration.

We conducted this analysis across various African populations and consistently found enrichment in Eurasian segments across all groups. Populations with more recent admixture, such as the San, Bantu Kenya, and Mandenka, exhibited the highest enrichment, while those with older admixture events, such as the Mbuti,

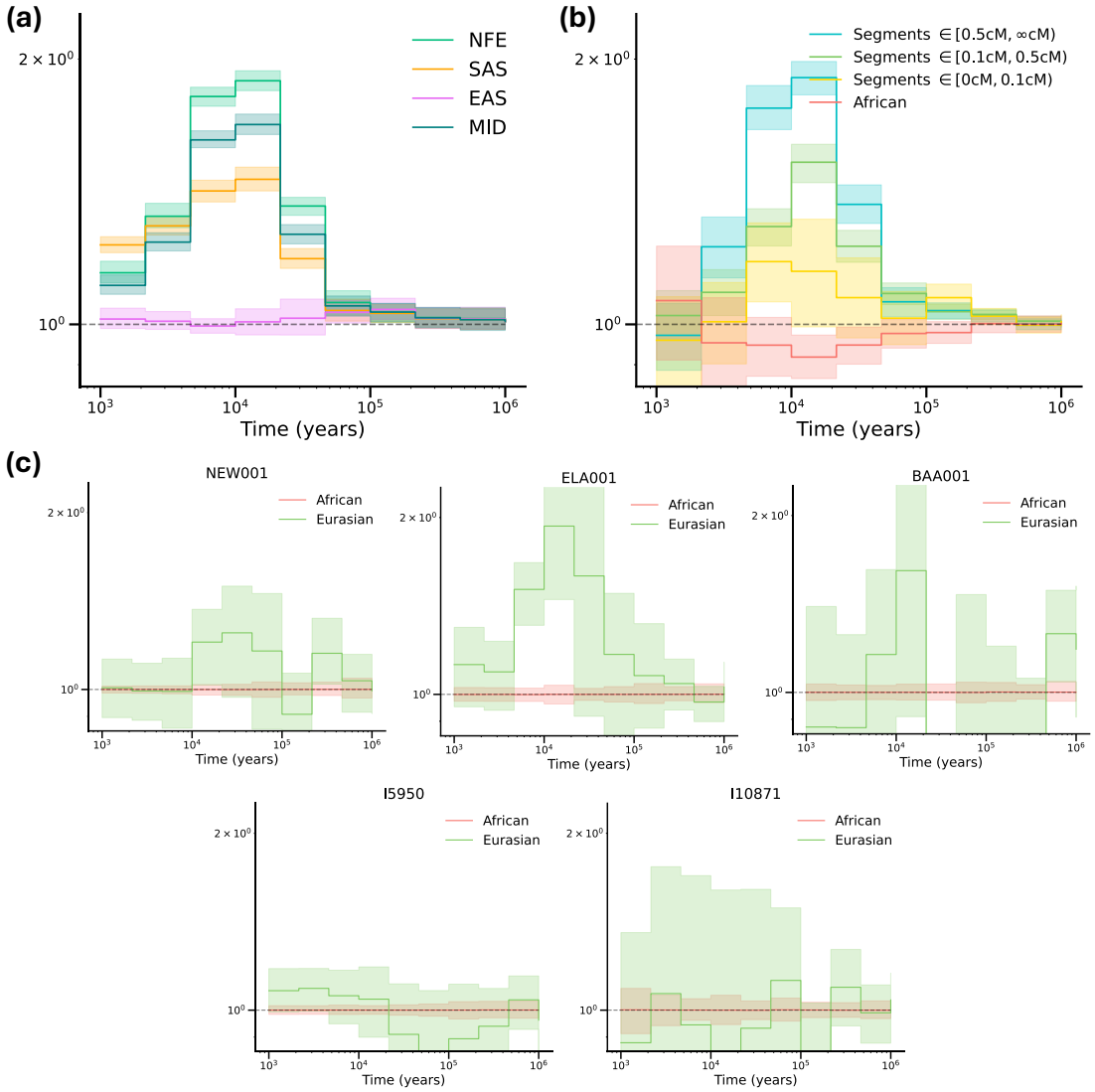


Figure 3.4: Enrichment of TCC to TTC mutation rates in African components. (a) Enrichment of TCC to TTC mutation rates across various continental groups, including NFE (non-Finnish Europeans), SAS (South Asians), EAS (East Asians), and MID (Middle Eastern and North African populations). (b) Enrichment of TCC to TTC mutation rates in Eurasian and non-Eurasian (African) segments in African populations. (c) TCC to TTC mutation rate enrichment in five ancient African samples, including new001, ela001, baa001, I5950, and I10871 (see population descriptions in Table 3.1b). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. In panel (b), the Eurasian segments in Africans are further categorized based on their lengths: segments shorter than 0.1 cM, those between 0.1 and 0.5 cM, and segments longer than 0.5 cM. Results derived from HGDP+1,000GP and HGDP+1,000GP+aDNA+archaic genealogies.

Biaka, and Yoruba, showed weaker enrichment for TCC to TTC mutations. Notably, some African populations, including the San and Bantu Kenyans, displayed empirical enrichment levels exceeding the average across all non-Finnish Europeans (see Figure B.15). In our analysis of ancient African samples, similar enrichment patterns emerged, though with increased variability due to smaller sample sizes. We normalized mutation rates in ancient samples relative to the mutation rates in their African segments, rather than using mutation rates from modern African samples, to minimize the impact of differences in sampling times and biases associated with ancient DNA genealogy construction. We observed modest enrichment in more recent ancient samples from South Africa, whereas older samples from Mota Cave and Cameroon showed no detectable enrichment (see Figure 3.4c). This absence of enrichment in older samples could be attributed to statistical noise when detecting TCC to TTC mutation enrichment in individual ancient samples or might indicate a lack of back-to-Africa admixture signals in these older populations.

In addition to examining the TCC to TTC mutations, we also investigated the enrichment of mutations related to weak to strong mutation types. For our analysis, we excluded CpG sites when defining weak or strong mutations. Similar to patterns observed in European, East Asian, and South Asian populations, the Eurasian segments within African genomes also showed reduction for weak to strong mutations relative to African genomes as a whole. The reduction in weak-to-strong mutations reached its peak approximately 50,000–100,000 years ago, aligning with patterns observed in other Eurasian populations. Interestingly, Eurasian segments of all lengths in Africans exhibit a similar pattern of weak to strong mutation reduction, and with a slightly lower amplitude compared to that observed in Europeans and East Asians. This pattern suggests that the Eurasian-like ancestry in African genomes may have experienced an even more pronounced bottleneck than the population that initially migrated out of Africa (see Figure 3.5a,b). As with TCC to TTC enrichment, we stratified the weak to strong mutation reduction by population and ancient samples. We found consistent levels of weak to strong reduction across all modern populations and ancient samples, suggesting that both the younger and

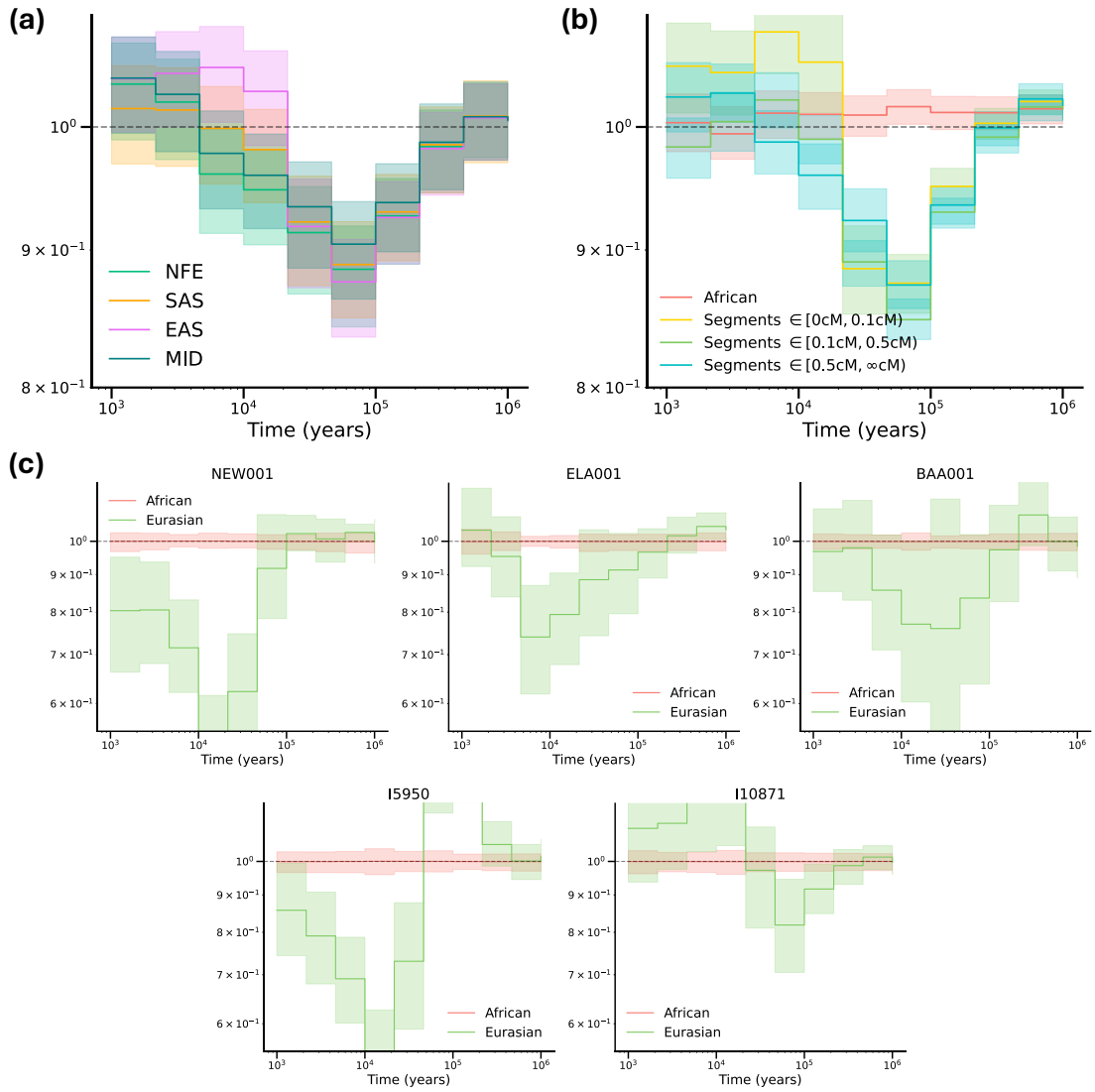


Figure 3.5: Enrichment of weak to strong mutation rates in African components. (a) Enrichment of weak to strong mutation rates across various continental groups, including NFE (non-Finnish Europeans), SAS (South Asians), EAS (East Asians), and MID (Middle Eastern and North African populations). (b) Enrichment of weak to strong mutation rates in Eurasian and non-Eurasian (African) segments in African populations. (c) weak to strong mutation rate enrichment in five ancient African samples, including new001, ela001, baa001, I5950, and I10871 (see population descriptions in Table 3.1b). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. In panel (b), the Eurasian segments in Africans are further categorized based on their lengths: segments shorter than 0.1 cM, those between 0.1 and 0.5 cM, and segments longer than 0.5 cM. Results derived from HGDP+1,000GP and HGDP+1,000GP+aDNA+archaic genealogies.

older admixture events may have undergone a population bottleneck approximately 50,000–100,000 years ago (see Figure 3.5c and B.16). Finally, we present the mutation rate enrichment for Eurasian-like versus non-Eurasian-like segments in Africans across all 96 trimer mutation types, as shown in Figure B.17.

3.3.5 Identifying source populations for Eurasian ancestry

To characterize the admixture event and identify potential source populations contributing Eurasian-like ancestry to Africa, we calculated coalescence rates using modern samples from the HGDP+1,000GP dataset and a subset of 40 high-coverage ancient samples with estimated sampling ages less than 10,000 years and sampled in Eurasia or Africa (see Table B.1 for list of all ancient samples). For clarity, we grouped ancient samples corresponding to Neolithic Anatolian and Balkan farmers into one group, and Western Hunter-gatherers into one group. We used Colate [202] with the `-mode local_ancestry` flag to infer the coalescence rates conditioned on the local ancestry in African samples. We analyzed Eurasian-like segments associated with both recent and ancient admixture events separately, with a primary focus on understanding the recent Eurasian-like segments. Furthermore, to capture population-specific variations in Eurasian ancestry, we conducted analyses for each African population separately.

To understand coalescence relationships across populations, we focused on the epoch spanning 10,000–16,000 years and constructed a pairwise coalescence rate matrix. This matrix was visualized using a heatmap, a clustering dendrogram, and PCA (see 3.6a-c, respectively). The clustering dendrogram is based on inverse coalescence rates distance matrix. We observed that both modern and ancient samples primarily clustered by sampling time and geography, forming distinct continental groups for East Asian, European, and African populations. The five ancient African populations (ela001, new001, baa001, I5950, I10871) clustered closely with modern African segments, while the European ancients coalesced with modern Europeans, South Asians, Middle Eastern groups, and the Eurasian-like segments in Africans. The Eurasian-like segments in Africa clustered more closely with modern Eurasian populations than with African segments. The younger segments exhibited stronger clustering with Europeans, whereas the older segments displayed roughly equal affinity to all out-of-Africa populations, including Europeans, South Asians, and East Asians (see Figure 3.6a,c). The Eurasian segments for both recent and ancient admixture events, exhibited high coalescence rates within their own group

than with other populations, suggesting a significant bottleneck in the Eurasian population that contributed to Eurasian-like ancestry in Africans (see Figure 3.6a). Further examining the pairwise coalescence rate patterns among Eurasian-like segments in Africans reveals geographic patterns, indicating a more pronounced bottleneck in West African populations such as Yoruba and Esan (see Figure B.18).

We performed PCA on the inferred coalescence rates matrix, initially using only modern populations and African segments in Africans. We then projected the ancient samples and the Eurasian-like segments in Africa onto the same plot. We identified signatures of prior admixture: South Asian samples aligned along the European–East Asian cline, Middle Eastern groups positioned along the European–African cline, and the Hazara clustered along the South Asian–East Asian cline. The younger Eurasian segments in Africans clustered closer to Europeans, lying on the gradient connecting European and the older Eurasian segments in Africans. When looking at ranked coalescence rates across modern samples, the younger Eurasian-like segments in Africa coalesced most rapidly with North African, Middle Eastern (specifically, Mozabite and Bedouin), and southern European groups, such as Sardinians and Tuscans. Notably, West and East African populations exhibited distinct affinities toward Mozabite and Bedouin populations. West African groups showed closer genetic ties to the Mozabite from Algeria, while East African groups displayed greater affinity with Bedouins sampled from the Middle East. This pattern, combined with the differing admixture dates observed in East and West Africa (as discussed in Section 3.3.2), implies distinct sources of Eurasian ancestry in these regions and highlights the ancient nature of back-migration into West Africa, which dates back to approximately 10,000 years.

Among ancient samples, the highest coalescence rates were observed with Anatolian and Balkan farmers (see Figure 3.6d-e, Figure B.19). This indicates that the Eurasian ancestry in Africa, particularly the younger segments, may be associated with European farming groups. It also suggests a possible trajectory of farming spreading from Europe into the Middle East and North Africa, followed by the further dissemination of this ancestry across Africa [174–176].

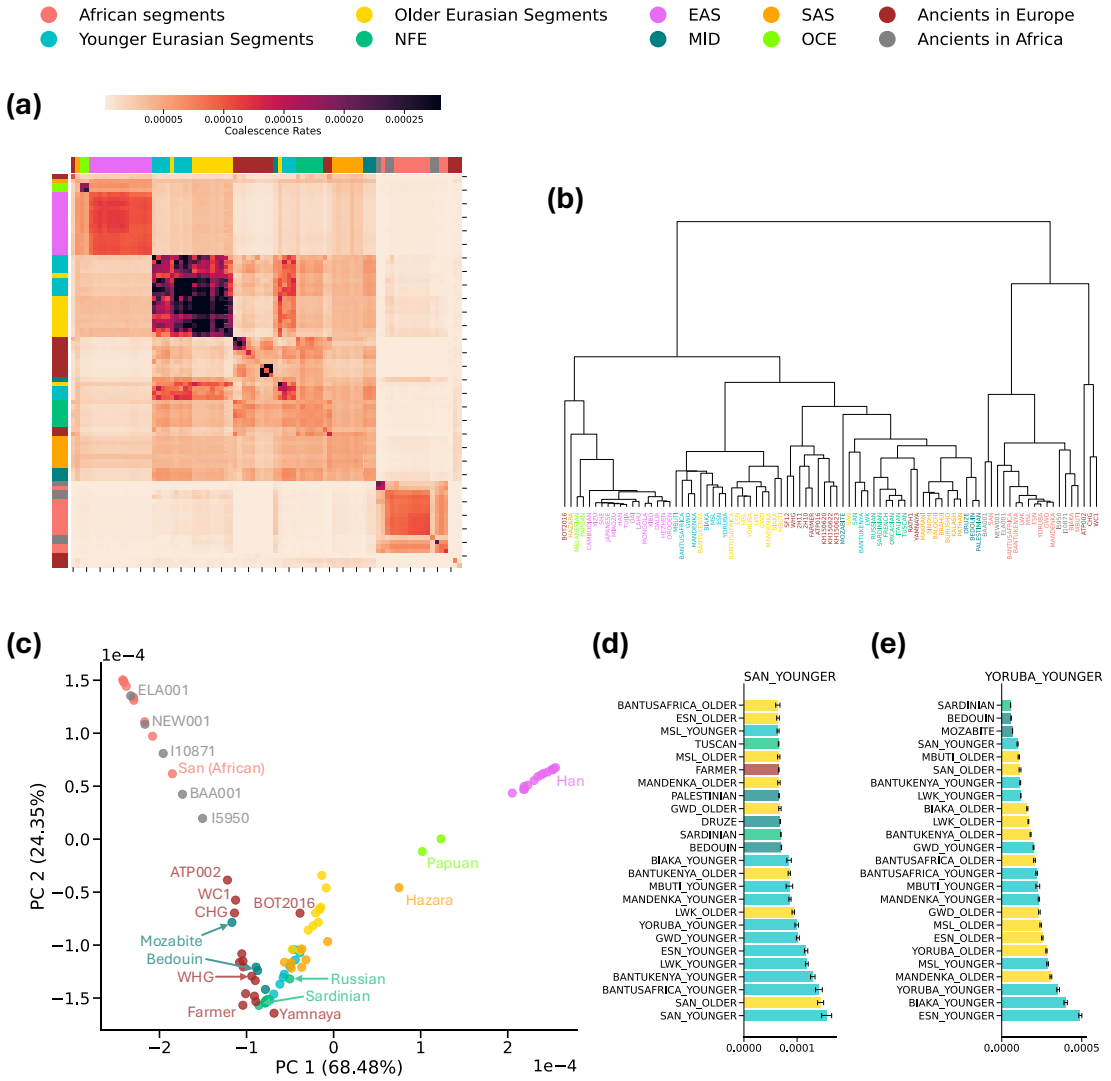


Figure 3.6: Coalescence rates with modern and ancient samples. (a) Matrix of pairwise coalescence rates for all modern and ancient groups in the epoch spanning 10,000–15,848 years. (b) Dendrogram corresponding to the UPGMA hierarchical clustering of the pairwise coalescence rate matrix in (a). (c) PCA visualization of coalescence rates for the epoch spanning 10,000–15,848 years, with PC1 and PC2 representing the eigenvectors capturing the greatest variance among modern samples. (d-e) Top 25 groups based on coalescence rates with (d) younger Eurasian-like segments in the San, and (e) younger Eurasian-like segments in Yoruba during the epoch spanning 10,000–15,848 years. Row and column colors in (a), dendrogram leaves in (b), and dots and bars in (c-e) are colored according to the common legend. NFE = non-Finnish Europeans, EAS = East Asians, MID = Middle Easterners and North Africans, SAS = South Asians, OCE = Oceanians. The numbers in brackets in (c) indicate the proportion of variance explained by each principal component. The error bars in (d-e) represent 95% confidence intervals calculated using 100 bootstraps. Results derived from HGDP+1,000GP+aDNA+archaic genealogies.

3.4 Evidence of a deep admixture event in Africans

We used GhostBuster to understand the deep evolutionary history in Africa. Our analysis incorporates the same set of modern and ancient African samples as used in the "Back-to-Africa" section, detailed in Table 3.1.

3.4.1 Ancient admixture across sub-Saharan African populations

To examine ancient events, we focused on coalescence events that occurred between 10,000 and 5,000,000 years ago, dividing this time frame into 20 epochs on a logarithmic scale. We excluded coalescence events more recent than 10,000 years, as these likely represent more recent events like previously discussed back-to-Africa migrations. Consistent with previous analyses of back-to-Africa migration, we grouped non-Finnish Europeans (NFE), East Asians (EAS), and South Asians (SAS) as Eurasians. For each African population, analyses were conducted using all Eurasian and African populations in the dataset as reference groups, while excluding Middle Eastern and American populations from the reference panel. To further concentrate on broader lineage relationships, coalescence events between individuals within the target population were excluded.

Starting from a random initialization, the coalescence rates converged to reveal two distinct components. Figure 3.7a shows the inferred coalescence rates in Yorubans, used as a representative group. These rates suggest two components that primarily differ in their coalescence rates with Eurasians. We refer to the component closer to Eurasian populations as the out-of-Africa-like (OOA-like) component and the one relatively closer to African populations as the non-out-of-Africa-like (non-OOA-like) component. The Eurasian-like ancestry previously identified in Section 3.3 strongly correlates with the OOA-like component in this deep admixture, with more than 98% of segments attributed Eurasian-like lying within this component. This correlation is largely explained by the component's faster coalescence rate with Eurasians and supports the presence of two distinct types of Eurasian ancestries in Africans, as discussed in Section 3.3. To specifically examine this deeper admixture

events, we masked regions corresponding to previously identified Eurasian-like segments (posterior probability Eurasian-like > 10%) and a 100kb flanking region around them to avoid biases from recent back-to-Africa migration.

Both the held-out likelihood and the expected coefficient of determination indicate that two components best capture the observed coalescence differences without ambiguity regarding the admixture event (see Figure 3.7b and Figure B.20). We identified varying proportions of OOA-like and non-OOA-like components across modern and ancient African groups after removing ancestry corresponding to the back-to-Africa migration. Among modern groups in the HGDP+1,000GP and SGDP datasets, Dinka individuals exhibited the highest proportion of OOA-like ancestry at approximately 36.8%, while Khomani San individuals showed the lowest at 10.6%. In ancient individuals, we detected 35.9% OOA-like ancestry in the 4,500-year-old Mota individual, 27.6% in a 7,900-year-old ancient African from Cameroon, and 30.1% in younger ancient samples from South Africa, namely ela001 and new001. In contrast, the Ballito Bay individual, baa001, exhibited 12.3% OOA-like ancestry. Overall, we observed a higher proportion of the OOA-like component in East Africa, followed by West and South Africa, with the lowest levels in Central African hunter-gatherers (see Figure 3.7d). The first four principal components of the coalescence count and opportunity matrix did not reveal substantial differences between clusters (see Figure 3.7b). This is attributable to the large number of African reference panel groups compared to a single group representing Eurasians, which exhibit less variation between the components inferred by GhostBuster.

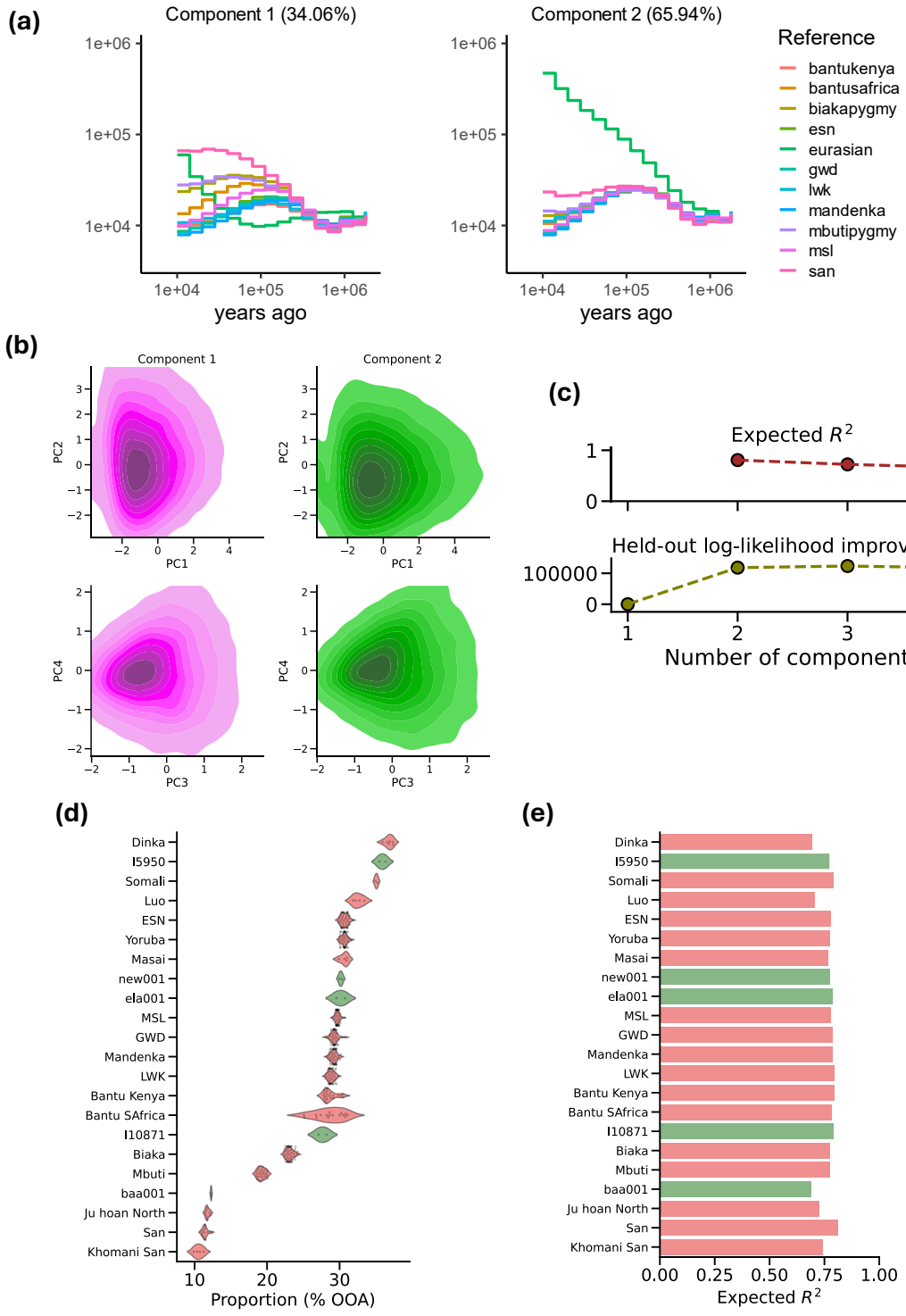


Figure 3.7: Decomposing sub-Saharan Africans to find a deep admixture event. (a) Inferred inverse coalescence rates and proportions when decomposing Yorubans: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Proportion of the out of Africa-like ancestry and (e) Expected coefficient of determination across all African populations analyzed. The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5. Individual data-points overlaid on the violin plots.

3.4.2 Dating the admixture event

We constructed coancestry curves using the inferred local ancestry data without filtering out previously identified back-to-Africa segments. These curves were then used to fit various admixture dating models, including single-date, two-date, and continuous admixture models. Similar to the back-to-Africa migration, we found that a single-date model did not provide a good fit to the coancestry curves. Notably, the two-date admixture model provided the best fit for the coancestry data across 10 out of 16 African populations (see Figure 3.8). Consistent with the back-to-Africa migration, the younger admixture dates ranged from 1,500 to 7,000 years ago, while the older admixture event occurred approximately 23,000 to 40,000 years ago (see Figure 3.8).

Next, we removed the local ancestry in and around the Eurasian-like segments previously identified in Section 3.3 and recalculated the coancestry curves. After removing the local ancestry corresponding to the more recent back-to-Africa migration, we found that the single-date admixture model fit the coancestry data as well as the two-date and continuous admixture models in most scenarios (see Figure B.21). This suggests that the younger date in the initial fit most likely captures the more recent back-to-Africa migration. Therefore, we focus on the older date as the estimated time of admixture between OOA-like and non-OOA-like populations and exclude the Eurasian-like ancestry identified in the back-to-Africa section from all subsequent analyses. This admixture is dated to 23,000-40,000 years across groups (see Table B.3). It is important to note that these admixture dates have not been adjusted for potential errors in the recombination map, as we lacked reliable error estimates for it. We believe this may introduce a downward bias in the dating estimates [137]. However, it might be reasonable to assume that the admixture occurred around or after the out-of-Africa migration dated around 50,000–100,000 years ago [18].

Given the deep nature of the admixture, we performed additional "sanity checks", such as running GhostBuster with the HMM disabled and using only African reference populations (removing Eurasians). These analyses produced

similar coalescence rates and correlated but more uncertain local ancestry estimates compared to the initial run with the HMM enabled and Eurasian samples included in the reference population (see Figure B.22 and B.23).

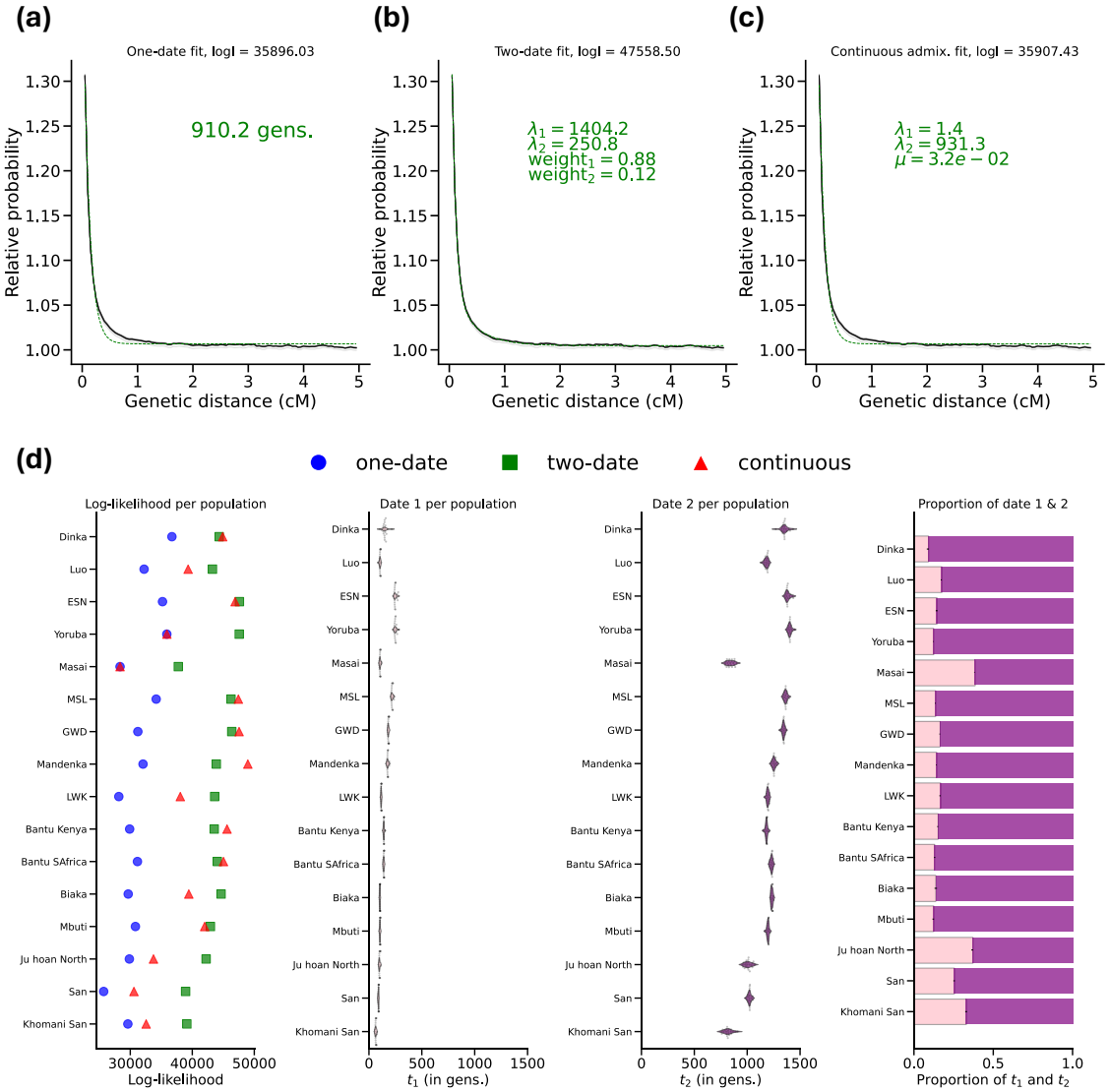


Figure 3.8: Dating the ancestral components corresponding to deep admixture event. Coancestry curves and (a) one-date admixture model fit, (b) two-date admixture model fit and (c) constant migration rate continuous admixture model fit for Yorubans. (d) Per-population least squared log-likelihood corresponding to various admixture models along with the admixture dates and weights inferred using two-date admixture model. The log-likelihood values are aggregated across 20 jackknife runs. Populations are sorted based on highest to lowest proportion of Eurasian-like ancestry, pink represents the younger admixture date and purple represents the older admixture date in (d). Individual data-points overlaid on the violin plots. Admixture dates in the plot are in generations.

3.4.3 Mutational profiles validate the admixture event

We used mutation profiles as near orthogonal evidence that GhostBuster has identified two distinct components that are unlikely to be data or method artefacts. Similar to the analysis in Section 3.3.4, we calculated the normalized mutational enrichment profiles for continental populations and inferred mixture components relative to genome-wide normalized mutation rates in the African population.

In contrast to our analysis of the more recent back-to-Africa migration, we did not observe enrichment for TCC to TTC mutations (see Figure B.24). This is likely because this admixture event predates the emergence of TCC to TTC mutations in Eurasian populations. However, we detected distinct patterns in the enrichment of weak to strong mutations between the two ancestry components. To classify mutations as weak or strong, CpG sites were excluded, following the approach used in analyzing the back-to-Africa migration. We observed a reduction of weak to strong mutations in the out-of-Africa-like (OOA-like) component compared to genome-wide weak to strong mutation rates within Africa. This reduction peaked approximately 50,000 – 100,000 years ago, aligning with patterns observed in other Eurasian populations (see Figure 3.9a-b). These findings suggest a notable divergence in the trajectories of existing mutations during this period, driven by differential GC-biased gene conversion between the two ancestral components. Among ancient samples, the 4,500-year-old individual from Mota Cave (I5950) and the 400-year-old individual from South Africa (new001) exhibited approximately double the weak-to-strong mutation reduction signal compared to modern samples (see Figure 3.9c). This increase cannot be attributed to our estimated uncertainty and may suggest that the OOA-like group ancestors for these individuals experienced a particularly small effective population size. In contrast, the 1,900- and 500-year-old individuals from South Africa (baa001 and ela001) displayed a subtle weak-to-strong signal reduction in the OOA-like group, similar to what is observed in modern-day samples (see Figure 3.9c). Meanwhile, the 7,900-year-old individual from Cameroon (I10871) showed no significant difference in weak-to-strong enrichment between the two components (see Figure 3.9c).

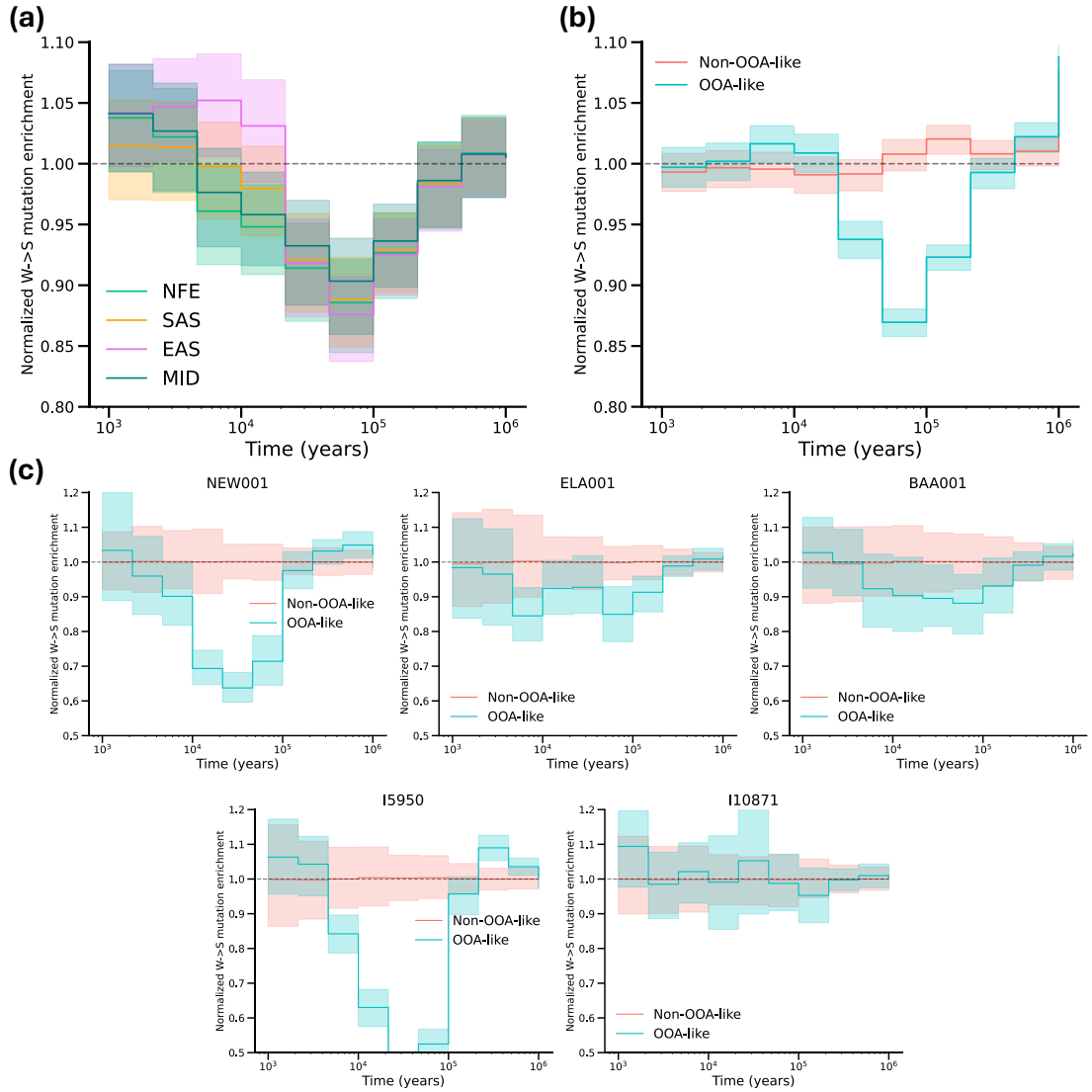


Figure 3.9: Enrichment of weak to strong mutation rates in the deep admixture African components. (a) Enrichment of weak to strong mutation rates across various continental groups, including NFE (non-Finnish Europeans), SAS (South Asians), EAS (East Asians), and MID (Middle Eastern and North African populations). (b) Enrichment of weak to strong mutation rates in OOA-like and non-OOA-like (African) segments in African populations. (c) Weak to strong mutation rate enrichment in five ancient African samples, including new001, ela001, baa001, I5950, and I10871 (see population descriptions in Table 3.1b). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. Results derived from HGDP+1,000GP and HGDP+1,000GP+aDNA+archaic genealogies.

3.4.4 Variation in PRDM9 activity across populations

In humans, recombination predominantly occurs at specific regions in the genome known as recombination hotspots. The placement of these hotspots is largely governed by the PRDM9 gene, which encodes a zinc finger protein responsible for identifying and binding to distinct DNA sequence motifs. Once bound, PRDM9 modifies histones near the binding sites, creating an environment conducive to recombination initiation.

Multiple types of PRDM9 have been identified in humans, distinguished by variations in the zinc finger array configuration. These PRDM9 types can bind to different sequence motifs across the genome, resulting in distinct recombination hotspot locations. Of all the known PRDM9 types in humans, PRDM9-A is the most common, binding to a specific sequence motif ("CCNCCNTNNCCNC") [203, 204]. It is present in all modern human populations but absent in chimpanzees [205]. PRDM9-B shares most of its binding motif with PRDM9-A and thus targets similar hotspot locations. Another notable PRDM9 type, PRDM9-C, binds to a distinct sequence motif, a unique 17-base pair sequence ("CCgCNgtNNNCgtNNCC"), which directs recombination events to different genomic locations. The small letters in the motif indicate less conserved bases. PRDM9-C is particularly common in African populations but occurs less frequently in Eurasian populations [206–208]. Importantly, the probability of crossover hotspots associated with PRDM9-C is strongly influenced by the alleles an individual carries at the PRDM9 gene (rs6889665, $p < 10^{-245}$ [206]).

The recombination machinery repairs double-strand breaks with a slight preference for incorporating GC alleles over AT alleles, leading to GC-biased gene conversion and an increased rate of weak to strong mutations around recombination hotspots. Separately, frequent DNA repair at these hotspots gradually erodes the sequence motifs that PRDM9 zinc fingers bind to, a phenomenon known as hotspot death. The continual loss of hotspots places evolutionary pressure on PRDM9 to form new hotspots through alterations in its zinc finger domains, changing its binding specificity. Consequently, PRDM9 has become one of the fastest-evolving

genes, constantly adapting to balance hotspot erosion and renewal [205, 209]. While the locations of recombination hotspots are determined by the specific PRDM9 allele an individual carries, examining GC-biased gene conversion and signatures of hotspot death at these sites provides insight into the evolutionary history of a lineage. We leverage this evolutionary information to improve our understanding and provide independent evidence of the ancient admixture event under investigation.

We begin by examining the PRDM9-C allele (rs6889665), which has previously been shown to strongly predict the probability of crossover at PRDM9-C hotspots [206]. In our dataset, we found the SNP rs6889665 to have a frequency of 33.7% in Africans, 8.4% in East Asians, 7.4% in South Asians, and 6.3% in non-Finnish Europeans. Additionally, its frequency was 44.1% in the non-out-of-Africa (non-OOA-like) component but only 2.7% in the out-of-Africa-like (OOA-like) component. Figure B.25 depicts the tree at PRDM9, identifying rs6889665 as one of the oldest mutations on the tree. Notably, the mutation shows an excess of descendants inferred to possess ancestry from the non-OOA-like component compared to the OOA-like component.

Next, we analyzed the PRDM9-A and PRDM9-C hotspots along the genome (provided by Dr. Anjali Hinch) to investigate differences in their evolutionary histories across various populations and ancestral components involved in the deep admixture event. Specifically, we examined weak to strong mutations in and around these hotspots and their impact on different populations. PRDM9-A and PRDM9-C hotspots were analyzed separately, excluding regions where both hotspots occur within a 5kb interval. Consistent with the mutational enrichment analysis, CpG sites were excluded when classifying mutations as weak or strong.

To understand population-level differences in PRDM9 activity, we began by examining the number of descendants carrying the weak to strong mutations at these hotspots per population. This quantity was normalized by the total number of population samples in the dataset and the AT content of the region. Furthermore, we only considered mutations with an age of less than one million years. Figure 3.10a,c presents the normalized descendant count per continental population as

a function of the distance from the hotspot center for weak to strong mutations. For hotspots associated with PRDM9-A, we observed similar peak intensities around the hotspot center across all continental populations analyzed. In contrast, PRDM9-C hotspots exhibited differences in activity across continental populations, with Africans showing a higher peak compared to non-African populations. This discrepancy could be attributable to population-specificity of PRDM9-C allele previously identified [206, 208].

To explore differences in weak to strong mutation rates across the ancestral components involved in the deep admixture event, we analyzed weak to strong mutations within 1,000 bp of the hotspot center and calculated the time-stratified mutation rate enrichment as described in Section 3.4.3. However, it is important to note that while Relate provides estimates of mutational ages, its accuracy within hotspots is limited. These estimates are likely correlated with true ages, but they should not be interpreted as highly precise chronological markers. Unlike the genome-wide mutational enrichment analysis, we normalized the enrichment relative to the mutation rate in East Asians (rather than Africans) and focused exclusively on regions around the hotspots.

Figure 3.10b,d illustrates the weak to strong mutation rate enrichment for the OOA-like and non-OOA-like components at PRDM9-A and PRDM-C hotspots respectively. Relative to the weak to strong mutation rate in East Asians, both the OOA-like and non-OOA-like components exhibited enrichment for weak to strong mutations around 50,000 years ago at PRDM9-A hotspots. This enrichment was of a similar magnitude for both components. However, in regions near PRDM9-C hotspots, we observed a significant difference in enrichment between the two components (see Figure 3.10d), with the non-OOA-like component exhibiting higher weak to strong mutations. This suggests that the PRDM9-C allele historically exhibited greater activity relative to the PRDM9-A allele in the non-OOA-like group compared to the OOA-like group. This finding, along with the broader differences in GC-biased gene conversion (GCbGC), provides compelling evidence that the two groups represent genuine, ancient population stratification within

African ancestral groups. It is difficult to conceive of confounding factors that could misclassify ancestry in a way that would result in distinct mutation types appearing on separate lineages. This is particularly true given that information about mutation type is invisible to GhostBuster and that our normalization procedure for identifying mutational biases explicitly matches the genomic regions used to define OOA-like and non-OOA-like mutation profiles. Furthermore, these patterns align precisely with a priori expectations of ancient stratification. The present-day frequency of PRDM9-C is lower than PRDM9-A in both non-African and inferred OOA-like African individuals, while it is comparatively higher in non-OOA-like African lineages. These observations strongly support the hypothesis of deep population structure within African ancestral groups. Finally, it is worth noting that the mutation rate enrichments shown in Figure 3.10b,d are based on mutations mapped to the genealogy. A small fraction of mutations (less than 3%) do not map to the genealogy and therefore cannot be dated. Nevertheless, we believe the observed signal is robust and unlikely to be affected by these non-mapping mutations.

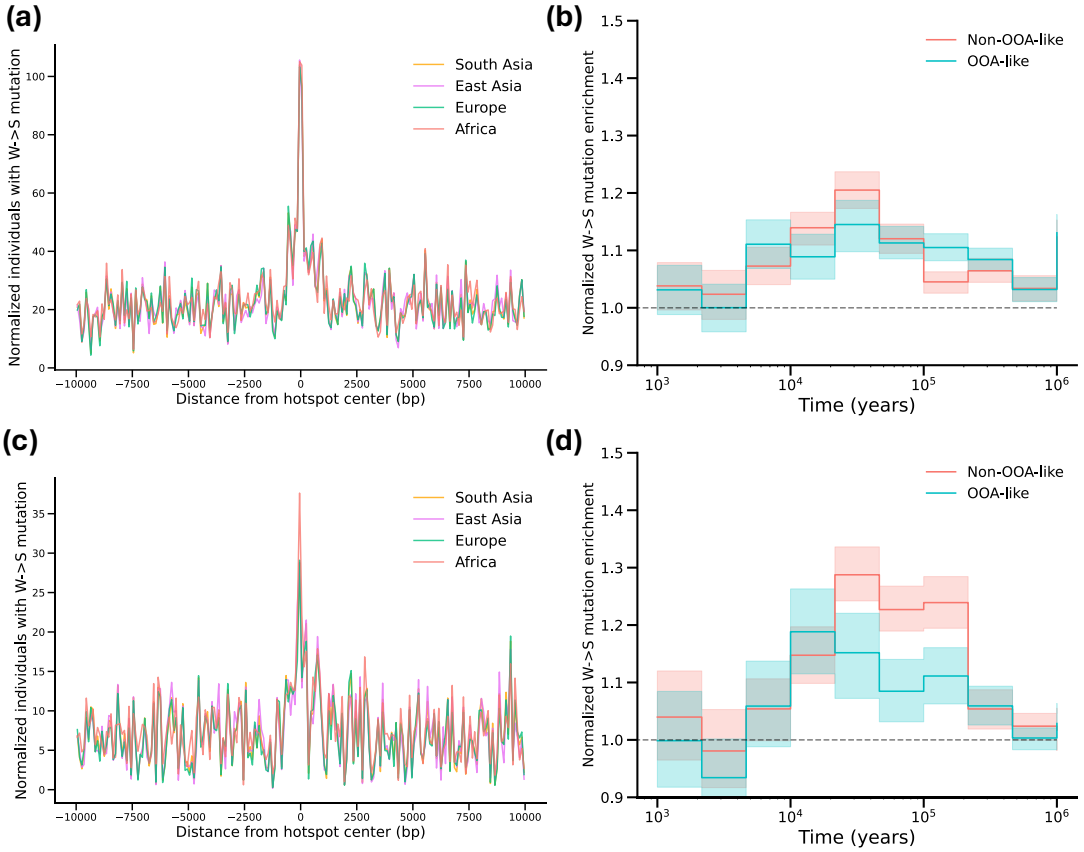


Figure 3.10: Weak to strong mutation rates at recombination hotspots across populations. Weak to strong mutation counts across populations and deep admixture component in (a-b) PRDM9-A hotspots, (c-d) PRDM9-C hotspots. Panel (a) and (c) measure the number of individuals carrying weak to strong mutations around the PRDM9 hotspots, normalized by the total number of individuals and the AT content of the region. Whereas, Panel (b) and (d) measure the time-stratified normalized mutation rate enrichment with respect to weak to strong mutation rates in East Asians. In panel (b) and (d) the shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment. Results derived from HGDP+1,000GP and HGDP+1,000GP+aDNA+archaic genealogies.

3.4.5 Characterizing the admixture event using coalescence rates

In order to characterize the deep admixture event better, we look at the relationship of two ancestral components with various modern, ancient and archaic populations. Similar to analysis in Section 3.3.5 we calculate coalescence rates with modern samples from HGDP+1,000GP dataset and 49 high-coverage ancients and 4 archaic human groups (see Table B.1) conditioned on the inferred local ancestry for each African population.

We consider coalescence events in epoch corresponding to 64,000-100,000 years in the coalescent history. Following the analysis of the back-to-Africa migration (Section 3.3.5), we constructed a pairwise coalescence rate matrix, which is visualized using a heatmap, UPGMA dendrogram, PCA, and coalescence rate bar plots for this epoch (see Figure 3.11a-e, respectively). The PCA plots in Figure 3.11c were inferred by calculating PCA on the coalescence rate matrix corresponding to modern populations and non-OOA-like segments in Africans. We then projected the ancients, archaic and OOA-like segments onto the plot. We found the populations to cluster based on geography and sampling times with distinct continental ancestries corresponding to East Asians, Europeans and non-OOA-like African segments visible in the heatmap and PCA (see Figure 3.11a,c). The OOA-like segments clustered better with Eurasian populations than non-OOA-like African population. On the PCA plot, they lie along the cline connecting non-OOA-like segments with Europe - East Asia cline. In the inverse coalescence rate plots, the Eurasian population aligns more closely with the OOA-like segment and is estimated to have split from them approximately 100,000 years ago.

We analyzed coalescence rates in relation to archaic human groups, including three Neanderthal and one Denisovan sample. These samples clustered with the non-OOA-like component (see Figure 3.11b). To confirm this observation, we examined coalescence rates across different time epochs and found that both Denisovan and Neanderthal samples consistently exhibited closer relationships to the non-OOA-like component until approximately one million years ago, when all human

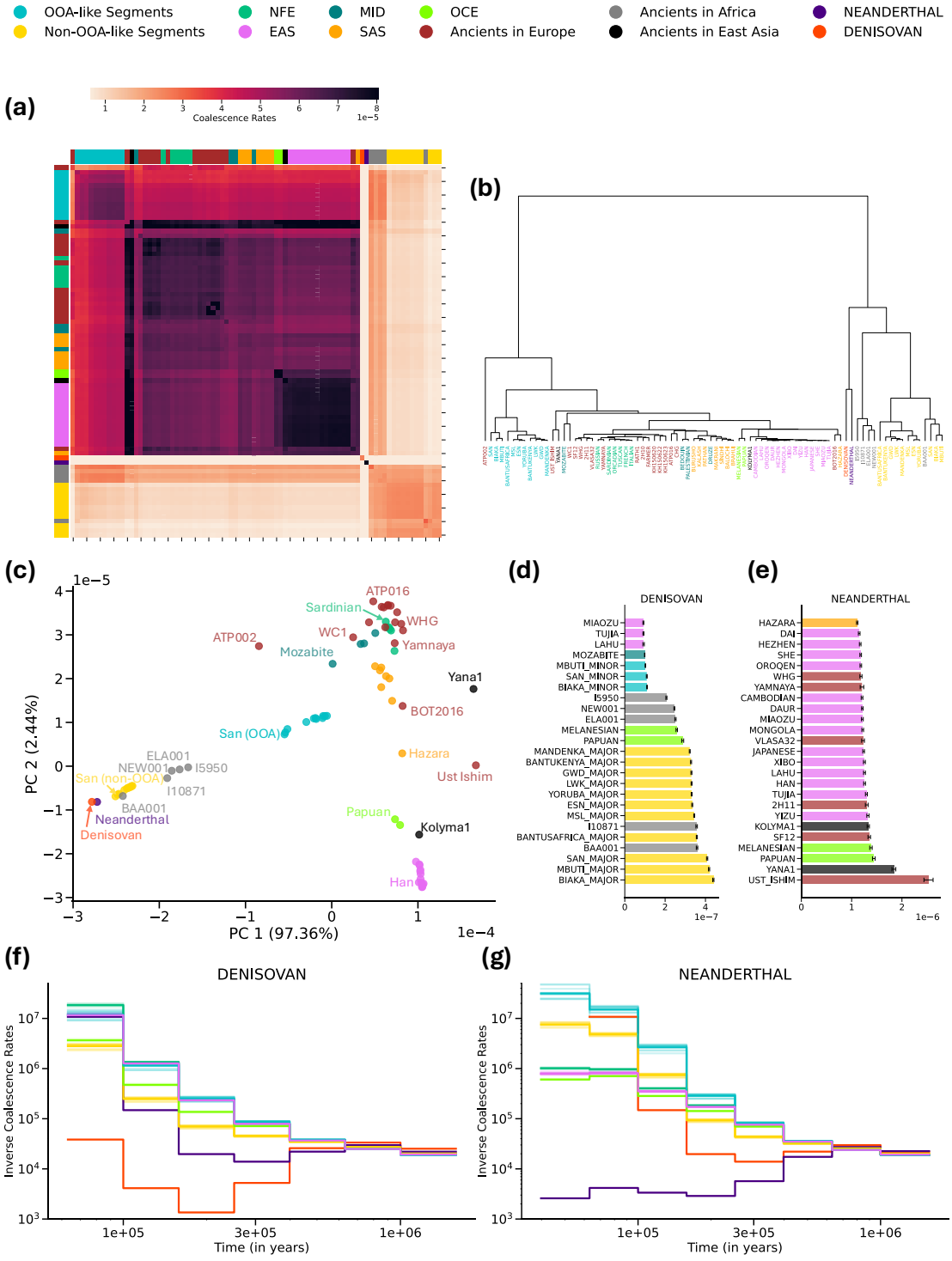


Figure 3.11: Coalescence rates with modern and ancient samples. (a) Matrix of pairwise coalescence rates for all modern and ancient groups in the epoch spanning 64,000–100,000 years. (b) Dendrogram corresponding to the UPGMA hierarchical clustering of the pairwise coalescence rate matrix in (a). (c) PCA visualization of coalescence rates for the epoch spanning 64,000–100,000 years, with PC1 and PC2 representing the eigenvectors capturing the greatest variance among modern samples. (d-e) Top 25 groups based on coalescence rates with (d) Denisovan, and (e) Neanderthals during the epoch spanning 64,000–100,000 years. (f-g) Inverse coalescence rates (ICR) of (f) Denisovan and (g) Neanderthals with other modern human groups from 64,000-1.6million years. Darker lines represent population-averaged ICRs. Row and column colors in (a), dendrogram leaves in (b), dots and bars in (c-e), and lines in (f-g) are colored according to the common legend. NFE = non-Finnish Europeans, EAS = East Asians, MID = Middle Easterners and North Africans, SAS = South Asians, OCE = Oceanians. The numbers in brackets in (c) indicate the proportion of variance explained by each principal component. The error bars in (d-e) represent 95% confidence intervals calculated using 100 bootstraps. Results derived from HGDP+1,000GP+aDNA+archaic genealogies.

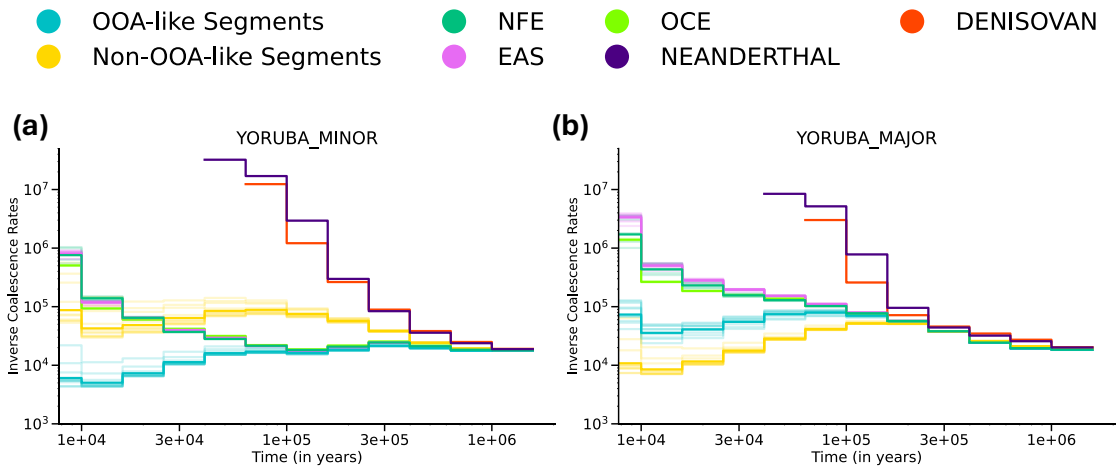


Figure 3.12: Inverse coalescence rates (ICRs) of OOA-like and Non-OOA-like component in Yoruba. ICRs of OOA-like (YOROBA_MINOR) and Non-OOA-like (YORUBA_MAJOR) components in Yoruba with other modern human and archaic groups from 64,000-1.6million years ago. Darker lines represent population-averaged ICRs and fainter lines represent within population ICRs. NFE = non-Finnish Europeans, EAS = East Asians, MID = Middle Easterners and North Africans, SAS = South Asians, OCE = Oceanians.

lineages converge. The coalescence rate patterns suggest ancient migration or admixture events involving these four human and archaic groups (see Figure 3.11d-g). Furthermore, inverse coalescence rates between Out-of-Africa-like (OOA-like) and non-OOA-like components reveal a deep ancestral split, with the two rates converging at least 300,000 years ago (see Figure 3.12). This predates the divergence of the Eurasian population from the OOA-like component and occurs after the split of Neanderthals and Denisovans from the modern human lineage, although there may be considerable uncertainty in the exact split time.

3.4.6 Impact of the admixture event on polygenic score portability in Africa

Polygenic scores (PGS) evaluate an individual's genetic predisposition to specific diseases or traits, providing valuable insights into personalized health risks. A PGS is derived from large-scale genome-wide association studies (GWAS), which identify associations between genetic variants and traits of interest. These associations are then used to calculate the risk for a disease or the predisposition to a complex trait. PGSs have been transformative in personalized medicine and is increasingly being applied in clinical settings [95, 210]. However, despite their success, a major limitation of PGS is their reduced accuracy and predictive power across diverse ancestral backgrounds [211, 212]. The majority of polygenic scores (PGS) are derived from genome-wide association studies (GWAS) conducted predominantly on populations of European ancestry, resulting in limited "portability" when applied to non-European populations. One key reason for this lack of portability is the difference in Linkage Disequilibrium (LD) patterns and allele frequencies between ancestrally diverged groups. Since PGS are typically based on common variants that serve as proxies for underlying causal mutations, differences in LD between the causal and tag variants or shifts in the frequency of these variants can introduce bias. This bias is most pronounced in individuals who are genetically distant from the GWAS reference population [212]. We aim to investigate the differences in PGS portability for African individuals in the UK Biobank. Specifically, in the context of the deep admixture event, we aim to determine whether the two ancestral components in Africa exhibit differences in PGS portability across a broad range of phenotypes. Furthermore, given our hypothesis that the OOA-like population significantly contributed to the ancestry of European individuals, we anticipate this analysis will reveal higher PGS portability for segments derived from the OOA-like population compared to those derived from the non-OOA-like population.

We used posterior mean effect estimates from the recent GWAS method Quick-draws to calculate polygenic scores (PGS) for several quantitative traits in the UK Biobank, utilizing data from up to 405,000 white British individuals for model fitting.

Detailed information on Quickdraws and the PGS calculation methodology can be found in Chapter 4 and Section 5.4.4. The traits analyzed included blood-related phenotypes, anthropometric measurements, and other quantitative traits (see Table B.5 for a full list). We focused our analysis on 32 phenotypes selected based on the predictive power of polygenic scores (PGS) in Europeans ($R^2 > 0.1$) and a high phenotyping rate in self-identified Africans (phenotyping rate > 0.9). Figure 3.13a illustrates the prediction R^2 achieved using the calculated PGS across various self-identified ancestry groups and phenotypes. For several phenotypes, we observed that the predictive power in African individuals dropped to approximately 20% of the prediction R^2 observed in European populations.

To examine differences in PGS portability across the two ancestries identified using GhostBuster – namely, OOA-like and non-OOA-like – we began by phasing the UK Biobank SNP array data using `SHAPEIT4` [213]. We filtered the data to retain only 4,557 African samples with less than 10% recent European admixture, as inferred using `hapmix` [23, 214]. The phased haplotypes were then lifted over to the hg38 genome build. To infer the ancestral components within these African samples, we utilized an optimized implementation of `Chromopainter` [38], using African samples from the HGDP and 1,000 Genomes Project as the reference panel. Chromosome painting was performed with parameters optimized through initial fitting on a subset of the data: a mutation probability per SNP of $\theta = 0.0011$ and a recombination scaling factor $\rho = 368.43$. This process assigned each genomic site in the UK Biobank African samples a posterior probability of copying from individuals in the reference panel. For each site, we selected the top 20 reference individuals with the highest posterior probabilities and normalized these probabilities so their sum equaled one. Using local ancestry inference from GhostBuster and the HGDP+1,000GP reference panel, we then estimated local ancestry in the UK Biobank African samples by calculating a weighted average of local ancestry posteriors, with weights determined by the `Chromopainter` copying probabilities. Overall, we obtained an estimate of the local ancestry corresponding to the two ancestral components in 4,557 phased African samples from the UK Biobank.

To examine differences in the predictive ability of PGS across the two ancestral components, we decomposed the PGS of an African individual into OOA-like and non-OOA-like components, conditioned on the local ancestry of the sample, following the pipeline described in [214]. The decomposition begins by estimating the expected allele counts for each ancestry background as follows:

$$\begin{aligned} H_{i,j}^{c} &= H_{i,j} * P(L_{i,j} = c) \\ G_{i,j}^{c} &= H_{i_1,j}^{c} + H_{i_2,j}^{c} \end{aligned} \tag{3.6}$$

where, $H_{i,j}$ represents the phased haplotype of sample i (with i_1 and i_2 indexing the haplotypes of sample i) at variant position j, $L_{i,j}$ is the local ancestry of sample i at position j, and $G_{i,j}^{c}$ is the expected allele count for ancestry background $c \in \{\text{OOA-like}, \text{non-OOA-like}\}$. As suggested in [214], we mean-center these genotypes before calculating ancestry-specific PGS as follows:

$$\begin{aligned} \bar{G}_{ij}^{\text{OOA-like}} &= G_{ij}^{\text{OOA-like}} - 2f_j^{\text{OOA-like}} \left(P_{ij}^{\text{OOA-like, OOA-like}} + P_{ij}^{\text{OOA-like, non-OOA-like}}\right), \\ \bar{G}_{ij}^{\text{non-OOA-like}} &= G_{ij}^{\text{non-OOA-like}} - 2f_j^{\text{non-OOA-like}} \left(P_{ij}^{\text{non-OOA-like, non-OOA-like}} + P_{ij}^{\text{OOA-like, non-OOA-like}}\right) \end{aligned} \tag{3.7}$$

where, $\bar{G}_{ij}^{c}$ is the mean-centered genotype for ancestry background c. $P_{ij}^{c_1,c_2}$ denotes the local ancestry product across the two haplotypes, calculated as $P_{ij}^{c_1,c_2} = P(L_{i_1,j} = c_1) * P(L_{i_2,j} = c_2)$, where i_1 and i_2 represent the two haplotypes of an individual. The frequency $f^{\text{OOA-like}}$ is determined using an OLS regression of the observed genotype given the local ancestry:

$$\begin{aligned} G_{ij} = I_j + S_j * \left(2P_{ij}^{\text{OOA-like, OOA-like}} + 2P_{ij}^{\text{OOA-like, non-OOA-like}}\right) + \varepsilon_i \\ f^{\text{OOA-like}} = S_j + \frac{I_j}{2} \quad \text{and} \quad f_j^{\text{non-OOA-like}} = \frac{I_j}{2} \end{aligned} \tag{3.8}$$

Finally the ancestry specific PGS can be obtained by multiplying the posterior mean effect estimates and mean-centered genotypes as follows:

$$\begin{aligned} PGS_{\text{OOA-like},i} &= \sum_j \hat{\beta}_j \bar{G}_{ij}^{\text{OOA-like}} \\ PGS_{\text{non-OOA-like},i} &= \sum_j \hat{\beta}_j \bar{G}_{ij}^{\text{non-OOA-like}} \end{aligned} \tag{3.9}$$

where, $\hat{\beta}_j$ is the posterior mean effect or PGS beta for variant j. To estimate the role of ancestry specific PGS in predicting the phenotype we fit a linear regression of the form:

$$Y_i = I + \beta_{\text{OOA-like}} PGS_{\text{OOA-like},i} + \beta_{\text{non-OOA-like}} PGS_{\text{non-OOA-like},i} + \alpha \text{Covariates}_i \quad (3.10)$$

Here, $\beta_{\text{OOA-like}}$ and $\beta_{\text{non-OOA-like}}$ represent the predictive ability of the PGS based on the OOA-like and non-OOA-like ancestry components for phenotype Y. These coefficients were estimated using ordinary least squares (OLS) regression, and their confidence intervals were derived from 1,000 bootstrap resamples of the individuals. Finally, the effect estimates were normalized by the corresponding effect estimates obtained from non-British self-identified European individuals, providing a measure of predictive ability relative to European samples.

We evaluated the predictive ability of PGS based on the two ancestry components across 32 quantitative phenotypes in 4,557 African samples. The predictive ability of the OOA-like component, $\beta_{\text{OOA-like}}$, was higher than that of the non-OOA-like component, $\beta_{\text{non-OOA-like}}$, in 28 out of 32 traits (paired t-test $p = 5.3 \times 10^{-6}$, see Figure 3.13b). An inverse-variance meta-analysis across 32 phenotypes showed a 35% higher effect size ($\beta_{\text{OOA-like, meta}} = 0.81$, $\beta_{\text{Non-OOA-like, meta}} = 0.6$) and approximately 82% greater predictive power as measured by R^2), which scales with the square of the effect size. Among the phenotypes analyzed, total bilirubin levels exhibited predictive power greater than 1 in both components, as its predictive ability in Africans was comparable to that in Europeans (see Figure 3.13a). Certain phenotypes, such as gamma-glutamyl transferase (GGT) and Apolipoprotein B, showed substantial differences in predictive ability between the two components. GGT demonstrated significantly higher predictive power in the OOA-like component ($\beta_{\text{OOA-like}} = 0.98$, $\beta_{\text{non-OOA-like}} = 0.06$) but low prediction R^2 overall, whereas Apolipoprotein B showed greater predictive power in the non-OOA-like component ($\beta_{\text{OOA-like}} = 0.34$, $\beta_{\text{non-OOA-like}} = 0.92$). This discrepancy is likely attributable to the oligogenic nature of these traits, with only a small number of genomic regions controlling the phenotype [215]. Manhattan plots for these traits are presented in Figure B.26.

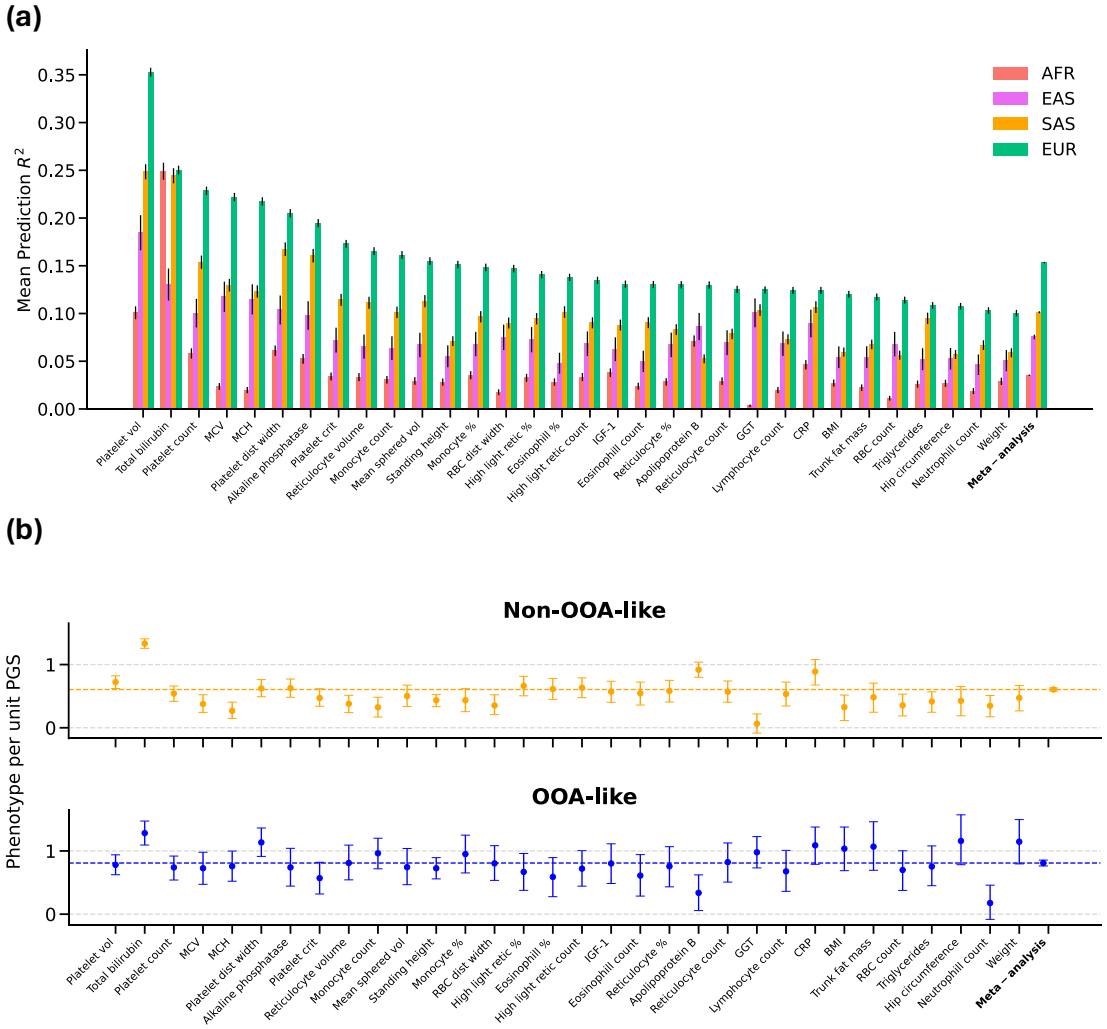


Figure 3.13: Differences in PGS predictive power across populations and deep admixture components. (a) PGS prediction R^2 across 32 quantitative phenotypes and four population groups: self-identified non-British Europeans (EUR), South Asians (SAS), East Asians (EAS), and Africans (AFR). Error bars represent the 95% confidence intervals for the prediction R^2. (b) Normalized predictive ability of the PGS based on the OOA-like and non-OOA-like ancestry components across the same 32 quantitative phenotypes. Error bars represent the 95% confidence intervals calculated using 1,000 bootstrap replicates. Traits are sorted based on their prediction R^2 in Europeans. PGS was calculated using posterior mean effect estimates obtained during Quickdraws' model fitting step. A meta-analysis using inverse variance-weighted average across traits is presented in (a) and (b).

3.5 Discussion

Genealogies provide a powerful and interpretable framework for understanding human history. In this work, we introduce GhostBuster, a mixture model that leverages inferred genealogies to detect and analyze historical admixture events. GhostBuster identifies patterns in the coalescence history of target individuals, clustering these patterns to detect ancestry differences across the genome. In simulations involving both modern and ancient gene flow, GhostBuster demonstrated high power for detecting admixture events and provided more accurate local ancestry inference than traditional HMM-based methods [151]. In real data examples, including admixture events within the past 2,000 years, GhostBuster's results aligned closely with previous haplotype-based methods [58], underscoring its robust capability to detect recent gene flow events.

We applied GhostBuster to investigate the population history of Africa by analyzing joint genealogies inferred from publicly available modern and ancient populations. Our analysis began by focusing on admixture history in the context of back-to-Africa migration [168, 170, 172]. Using GhostBuster, we uncovered widespread back-to-Africa migration from Eurasia, contributing an average of approximately 2.5% and up to 10% Eurasian ancestry across all present-day African populations analyzed. We identified previously unrecognized ancient back-migration events in several populations, particularly in West and Central Africa, while corroborating previously documented back-migration events in East and South Africa [169, 170]. We estimated the timing of these admixture events ranged from the past millennium to approximately 10,000 years ago, coinciding with the end of the last Ice Age and aligning with the Green Sahara period [166, 167], when the Sahara was likely more navigable. Characterizing and validating the admixture event, we also found that this back-migration event alone provides a plausible explanation for the presence of small amounts of Neanderthal ancestry in African genomes. Furthermore, the Eurasian segments in Africans exhibited distinct mutational profiles, including elevated TCC to TTC mutations, and showed genetic proximity to populations in modern-day North Africa, the Middle East, and

Anatolian farmer-related groups. These findings suggest a connection between this migration and the spread of farming practices from Europe into Africa.

In addition to the back to Africa migration, we applied GhostBuster to examine deeper population history in Africa, focusing on events from 10,000 years ago and earlier. Previous studies have proposed various hypotheses including presence of ghost populations [180, 184], weak structured stem [186], and ancient migrations [190] to explain the complex population structure in sub-Saharan Africa. Using GhostBuster, we identified two distinct ancestral components in the African genome, differentiated by their coalescence rates with Eurasians. We hypothesize that these components reflect an ancient admixture event occurring around or after the out-of-Africa migration, involving two ancestral populations. One of these populations likely represents a group closely related to the out-of-Africa population. We observed significant differences in population sizes and historical hotspot activity between these ancient groups, as evidenced by variations in weak to strong mutation rates across the genome and at specific recombination hotspots. Additionally, we identified differential affinities of the two components with archaic human groups, including Neanderthals and Denisovans, and found evidence of a deep population split between the two components, occurring at least 300,000 years ago. Finally, our analysis of African individuals in the UK Biobank revealed substantial differences in polygenic score (PGS) portability between these two ancestral components when European GWAS data were used for PGS construction.

Overall, GhostBuster proved effective in characterizing previously known admixtures while uncovering new insights into more ancient admixture events in Africa. We believe GhostBuster can be leveraged to investigate a wide range of admixture events, both recent and ancient, offering valuable insights with numerous downstream applications. For instance, admixture inference is crucial for understanding the evolutionary history of species, shedding light on the complex dynamics of migration, adaptation, and gene flow that have shaped modern populations. By inferring local ancestry along the genome, GhostBuster enables the identification of regions involved in adaptive admixture, where genes from one population may confer

selective advantages in another. Furthermore, GhostBuster's ability to detect and analyze fine-scale admixture has significant implications for medical genetics, helping researchers uncover the genetic basis of phenotypic variation across populations, including differences in disease susceptibility and other health-related traits.

3.5.1 Limitations and future work

We identify several current limitations and areas for future improvement. First, GhostBuster uses genome-wide genealogies, which generally requires high-coverage sequencing data to accurately infer genetic relationships between samples. However, many ancient DNA samples are sequenced at low coverage, limiting their integration into current genealogy inference methods. Addressing this challenge through scalable "threading" techniques to incorporate low-coverage ancient DNA into genealogies, or through imputation methods [216, 217], could significantly expand the number of ancient samples available for analysis. This would greatly improve GhostBuster's capacity to provide a comprehensive view of population history, particularly in cases where admixture signals are subtle and may not be apparent from limited high-coverage ancient DNA or modern samples – such as in the complex 3-way admixture observed in Europe (see Section 2.7).

Second, biases in genealogy inference can occasionally produce false admixture signals, as shown in our simulations (see Section 2.5). Typically, these biases are mitigated by evaluating the held-out log-likelihood, assessing confidence in local ancestry inference, and analyzing coancestry curve dating. Advancements in genealogy inference methods to reduce biases and increase reconstruction accuracy would further strengthen our model's reliability, making this an important direction for future research.

Third, GhostBuster is computationally slower than some other ancestry inference software that rely on low-dimensional summary statistics. While these alternatives often lack the statistical power to detect subtle admixture signals, there are promising strategies to improve GhostBuster's computational efficiency. For instance, using PCA on the coalescence count matrix generated by GhostBuster can provide a fast,

approximate inference of admixture. Additionally, implementing GhostBuster in a low-level programming language or leveraging modern computing architectures like GPUs could significantly improve its speed, making it more feasible for large-scale analyses. We propose these improvements as directions for future work.

Finally, expanding the use of GhostBuster to study the evolutionary history of other species would also be highly beneficial. In the next section, we demonstrate its effectiveness in uncovering an admixture event in American wolves. We believe that GhostBuster could similarly aid in understanding the population dynamics and admixture history of several other species, paving the way for broader applications in evolutionary biology.

3.5.2 Understanding admixtures in non human organisms

To demonstrate the ability of GhostBuster to detect admixture beyond humans, we applied it to understand the admixture history of North American wolves. We analyzed five American Great Lakes wolves alongside one Coyote and 29 European grey wolf genomes. Previous research has shown that Great Lakes wolves have undergone admixture with Coyotes over the last few thousand years [218–220]. We used the Relate trees and dataset from [221], along with a canid recombination map from [222]. GhostBuster was run with two components, considering coalescence events over a wide time interval ranging from 100 years to one million years. Our analysis revealed that the major component, contributing approximately around three quarters, was closer to grey wolves, while the minor component was closer to the coyote sample in the dataset. Using coancestry curve fitting, we identified the two-date admixture model as the best fit for the observed coancestry curves. The inferred admixture dates were 186.7 generations (560.1 years) and 1777.7 generations (5,333.1 years), with 79% of the observed admixture attributed to the older admixture date. See Figure 3.14 for further details.

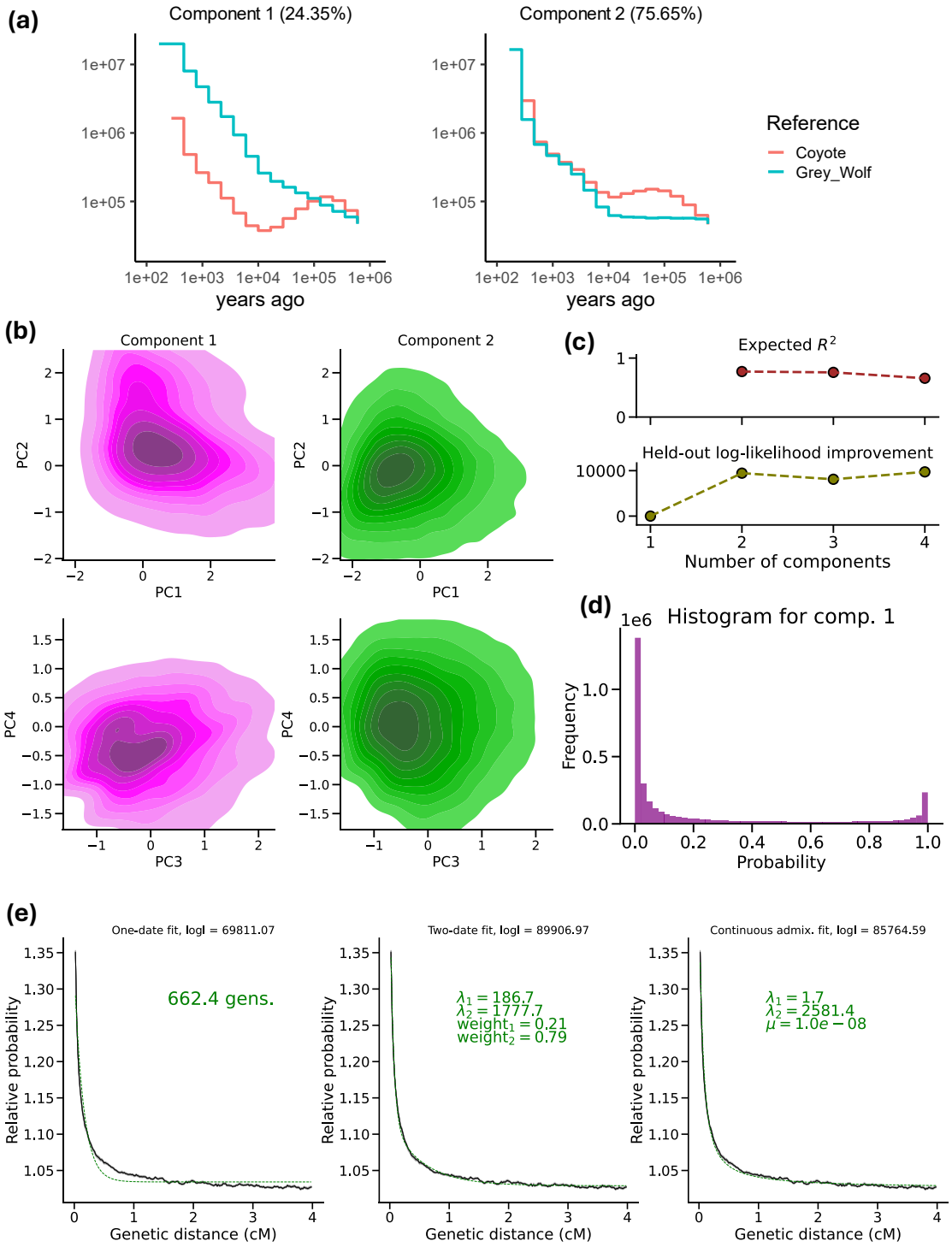


Figure 3.14: Decomposing American wolves with grey wolves and coyote as reference. (a) Inferred inverse coalescence rates and proportions: each line represents inverse coalescence rate profiles with a reference population. (b) PCA visualization of the coalescence count and opportunity matrix derived from the genealogies plotted separately for each component. (c) Expected coefficient of determination and held-out log-likelihood improvement with varying number of components. (d) Histogram of the inferred local ancestry posteriors. (e) Coancestry curves and model fit for one-date, two-date, and continuous migration scenarios (admixture dates in generations). The figure title includes 'logl', which refers to the log-likelihood of the fit (higher values indicate a better fit). The PCA visualization in (b) is based on a KDE plot with a threshold of 0.05, and binary local ancestry estimates are obtained by thresholding the inferred posteriors at 0.5.

4

A scalable variational inference approach for increased mixed-model association power

Contents

4.1 Chapter overview 160
4.2 Theoretical framework for our method 161
4.2.1 Bayesian inference 161
4.2.2 Variational inference 162
4.2.3 Evidence lower bound and mean-field variational inference 162
4.2.4 Coordinate ascent variational inference 164
4.2.5 Stochastic variational inference 166
4.3 Method description 169
4.3.1 Overview of the method 169
4.3.2 Bayesian regression with spike-and-slab prior 171
4.3.3 Heritability estimation 176
4.3.4 Test statistics calculation 177
4.3.5 Performance optimization 181
4.3.6 Hyperparameter tuning 183
4.4 Summary of the Quickdraws algorithm 185

4.1 Chapter overview

In this chapter, we introduce a novel method called Quickdraws, which we developed to perform scalable and statistically powerful genome-wide association studies. Quickdraws leverages advancements in machine learning and the processing power of modern computing architectures, such as graphical processing units (GPUs), to conduct genome-wide association studies on a large scale. We begin by presenting the theoretical framework for Quickdraws in Section 4.2, which incorporates modern variational inference techniques, including stochastic variational inference. After establishing the theoretical foundation, we will delve into the algorithm's details in Section 4.3, covering several steps such as heritability estimation, Bayesian regression, and the estimation and calibration of test statistics. Finally we summarize the key steps in the algorithm in Section 4.4.

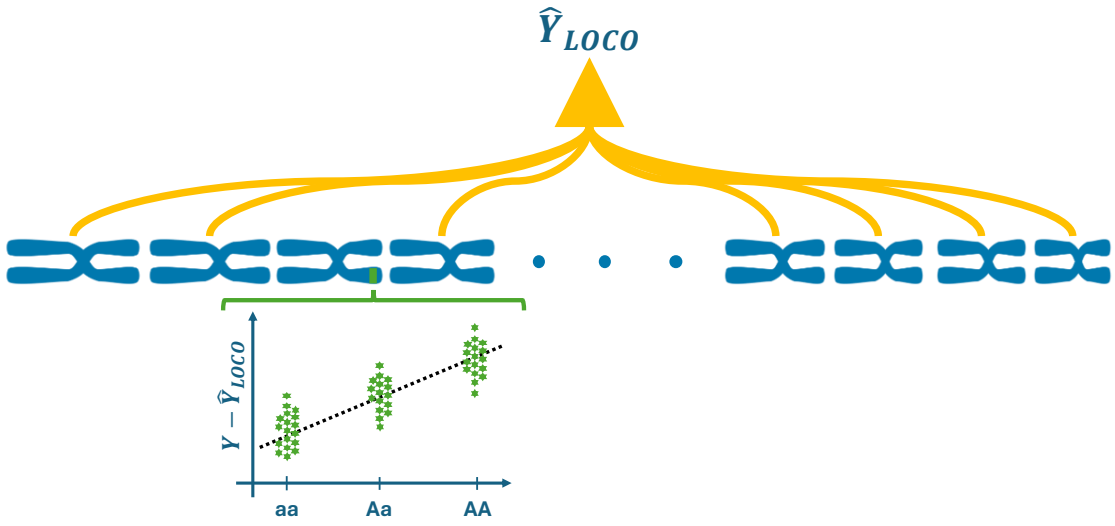


Figure 4.1: A simplified overview of Quickdraws. Quickdraws is a scalable mixed model designed for genome-wide association studies (GWAS). It improves the power of association testing through a two-step process. First, it estimates the genomic prediction for a trait by excluding the testing chromosome, employing a leave-one-chromosome-out method. The prediction models sparsity of effects and is optimized using advanced variational inference techniques ensuring scalability for large sample sizes. Second, the model leverages this genomic prediction to improve the signal-to-noise ratio for finding association between a trait and a variant.

4.2 Theoretical framework for our method

4.2.1 Bayesian inference

Bayesian inference is a popular statistical inference approach that takes into account prior beliefs about the model along with evidence obtained from the data. We introduce some basic terminology related to Bayesian inference, which we utilize in the following sections to describe our inference approach. We refer the reader to [223, 224] for a more in-depth review on Bayesian inference and variational Bayes approaches. We use the following notation and terminology:

- $X = (x_1, x_2, ..., x_n)$ is a set of n observed data-points
- θ parameterizes the probability distribution of the data-points, $x \sim P(x \mid \theta)$.
- The **prior** distribution, $P(\theta)$, is the distribution of the parameters before any data is observed.
- The **data likelihood** or likelihood, $P(X \mid \theta)$, is the distribution of the data conditioned on model parameters.
- The **evidence** or marginal likelihood is the distribution of the observed data marginalized over parameters,

$$P(X) = \int P(X \mid \theta)P(\theta)d\theta.$$

- The **posterior**, $P(\theta \mid X)$, is the distribution of the parameters after taking into account the observed data,

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)} = \frac{P(X \mid \theta)P(\theta)}{\int P(X \mid \theta)P(\theta)d\theta}.$$

In the context of Bayesian inference, we are most interested in deriving the posterior distribution given prior beliefs on the parameters. Unless the likelihood and prior are conjugate, however, we generally cannot derive a closed-form solution for the posterior. Thus, in more general and usually more complex scenarios, Bayesian Inference heavily relies on sampling approaches, such as MCMC, or approximate methods, such as variational Bayes, to infer the posterior.

4.2.2 Variational inference

Variational inference is a popular approximate inference technique often used to infer posterior estimates for models that do not allow for a closed-form solution for their posterior. In variational inference, posterior inference is formulated as an optimization problem. As the exact posterior probability is often intractable, we assume an approximate posterior family that is easy to evaluate. We then optimize the parameters of this posterior family so that the Kullback–Leibler (KL) divergence between the approximate and true posteriors is sufficiently small:

$$KL\left(q_\omega(\theta) \| P(\theta \mid X)\right) = \int q_\omega(\theta) \log\left(\frac{q_\omega(\theta)}{P(\theta \mid X)}\right) d\theta, \tag{4.1}$$

where $P(\theta \mid X)$ is the true posterior and $q_\omega(\theta)$ is the approximate posterior, parameterized by ω (also known as variational parameters). Variational inference aims to minimize the KL divergence in Equation 4.1, to obtain optimal variational parameters ω^*:

$$\omega^* = \underset{\omega}{argmin} KL(q_\omega(\theta)||P(\theta|\mathbf{X})). \tag{4.2}$$

4.2.3 Evidence lower bound and mean-field variational inference

Note that Equation 4.2 poses an optimization problem that requires the calculation of KL divergence between approximate posterior and the true posterior. However, this objective cannot be computed because it requires information about the posterior, which is not available in closed form. We thus use the Bayes rule to simplify the above equation,

$$\begin{aligned} KL(q_\omega(\theta)||P(\theta|\mathbf{X})) &= \mathbb{E}[log(q_\omega(\theta))] - \mathbb{E}[log(P(\theta|\mathbf{X})] \\ &= \mathbb{E}[log(q_\omega(\theta))] - \mathbb{E}[log(P(\theta, \mathbf{X})] + log(P(\mathbf{X})) \\ &= KL(q_\omega(\theta)||P(\theta)) - \mathbb{E}[log(P(\mathbf{X}|\theta)] + log(P(\mathbf{X})) \geq 0 \end{aligned} \tag{4.3}$$

$$\implies log(P(\mathbf{X})) \geq \mathbb{E}[log(P(\mathbf{X}|\theta)] - KL(q_\omega(\theta)||P(\theta)). \tag{4.4}$$

In the above, the term on the left is the logarithm of the evidence, so the term on the right is called the evidence lower bound, or ELBO. Variational inference algorithms try to minimize the KL divergence between the true posterior and approximate posterior by maximizing the ELBO in Equation 4.4. The ELBO can also be seen as the difference between the expected log-likelihood of the data and the KL divergence between the approximate posterior and prior,

$$\mathcal{L}_{VI}(\omega) = \mathbb{E}[log(P(\mathbf{X}|\theta)] - KL(q_\omega(\theta)||P(\theta)). \tag{4.5}$$

In the context of Bayesian regression, our sampling distribution is $P(y \mid X)$ instead of $P(X)$, where y is the outcome variable and X is the input variable. In that case, the ELBO is written as

$$\mathcal{L}_{VI}(\omega) = \underbrace{\mathbb{E}[\log(P(\mathbf{y}|\theta, \mathbf{X}))]}_{\text{Expected LL}} - \underbrace{KL(q_\omega(\theta)||P(\theta))}_{\text{KL-divergence}} \tag{4.6}$$

Mean-field variational inference

The fully-factorized or mean-field posterior is a commonly used approximate posterior in variational inference. In this approach, the posterior can be represented as a product of factors with only one parameter per factor. This enables the use of coordinate-ascent variational inference algorithms and makes operations such as sampling and computation of the ELBO, used in other variational inference strategies, more efficient.

Using this approach, the model we adopt when estimating genetic effects in quantitative traits, that is, a Bayesian linear regression with spike-and-slab prior and a fully-factorized spike-and-slab approximate posterior, takes the form

$$\begin{aligned}
&\textbf{Likelihood:} && P(y \mid X, \beta) \sim \mathcal{N}(\beta^T X, \sigma_e^2) \\
&\textbf{Prior:} && P(\beta) = \prod_j P(\beta_j),\ P(\beta_j) \sim (1 - p_0)\mathcal{N}(0, \sigma^2) + p_0\delta(0) \\
&\textbf{Approximate Posterior:} && q(\beta) = \prod_j q(\beta_j),\ q(\beta_j) \sim (1 - \psi_j)\mathcal{N}(\mu_j, \sigma_j^2) + \psi_j\delta(0)
\end{aligned} \tag{4.7}$$

In the above, y and X are the output and input variables, respectively, β are the unknown effects of the linear regression, which have a spike-and-slab prior $P(\beta)$, σ_e^2 and σ^2 are the variance component corresponding to noise and per predictor effect, p_0 is the sparsity and j indexes all the predictors in the regression. The approximate posterior in this model assumes a fully-factorized spike-and-slab structure with variational parameters μ, ψ, and σ. As shown in [225], a spike-and-slab prior is conjugate to the normal likelihood, making a fully-factorized spike-and-slab posterior a good choice for the approximate posterior family.

4.2.4 Coordinate ascent variational inference

Coordinate ascent variational inference (CAVI), introduced in [224], is one of the most popular mean-field variational inference algorithms. CAVI iteratively optimizes each factor in the mean-field posterior while fixing the other factors, thereby improving the ELBO in each iteration and finding a local optimum. CAVI updates each factor in the approximate posterior as follows:

$$\log q(\beta^*{}_j) = E_{\beta_{(-j)}}[\log(P(y, \beta))] + C \tag{4.8}$$

The optimal update for a variational parameter is proportional to the exponential of the expected logarithm of the complete conditional, where the expectation is taken with respect to all the variational parameters except the one which is being updated. The hyperparameters, which include σ^2, σ_e^2 and p_0 are assumed fixed. It can be shown that the above updates increase the ELBO each iteration (see [223] for a concise proof). We next derive the CAVI update equations for the Bayesian linear regression model described in Equation 4.7.

Derivation of CAVI updates for spike-and-slab Bayesian linear regression:

$$\begin{aligned}
\log q(\beta^*{}_j) &= E_{\beta_{(-j)}}[\log(P(y, \beta))] + C \\
&= E_{\beta_{(-j)}} \log P(y \mid \beta) + E_{\beta_{(-j)}} \log P(\beta_j) + C \\
&= \sum_{n=1}^{N} \Big(E_{\beta_{(-j)}} \log P(y_n \mid \beta) \Big) + E_{\beta_{(-j)}} \log P(\beta_j) + C.
\end{aligned} \tag{4.9}$$

We will remove terms which do not depend on β_j and absorb them in the constant C; the constant is thus not necessarily the same from one step to the other, but we simplify the notation by always expressing it as C. We start by simplifying the equation by substituting the normal log-likelihood and spike-and-slab prior,

$$\log q(\beta^*{}_j) = \sum_{n=1}^{N} E_{\beta_{(-j)}}\Big(-\frac{(y_n - \beta^T x_n)^2}{2\sigma_e^2} - log(\sigma_e)\Big) + E_{\beta_{(-j)}} \log P(\beta_j) + C. \quad (4.10)$$

Because the spike-and-slab distribution can be seen as mixture of non-overlapping Gaussians, [225] shows that it forms an exponential family (see theorems 1 and 2 in [225]). The natural parameters for the spike-and-slab distribution $(1-p_0)\mathcal{N}(\mu,\sigma^2)+p_0\delta(0)$ are $-\frac{1}{2\sigma^2}$, $\frac{\mu}{\sigma^2}$, and $\log p_0 - \log(1-p_0) + \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2$, with corresponding sufficient statistics $\mathbb{I}\{\beta_j \neq 0\}\,\beta_j^2$, $\mathbb{I}\{\beta_j \neq 0\}\,\beta_j$, and $\mathbb{I}\{\beta_j = 0\}$. We can substitute the natural parameters for the prior to simplify further:

$$\begin{aligned}
&\log q(\beta_j^*) = \\
&= \sum_{n=1}^{N} E_{\beta_{(-j)}}\left(-\frac{(y_n - \beta^T x_n)^2}{2\sigma_e^2}\right) + E_{\beta_{(-j)}}\Bigg(- \mathbb{I}\{\beta_j \neq 0\}\,\frac{\beta_j^2}{2\sigma^2} \\
&\quad - \mathbb{I}\{\beta_j = 0\}\,(\log p_0 - \log(1-p_0) + \log(\sigma))\Bigg) + C \\
&= -E_{\beta_{(-j)}}\left(\frac{\sum_{n=1}^{N}(y_n - \beta_j x_{n,j} - \sum_{i\neq j}\beta_i x_{n,i})^2}{2\sigma_e^2}\right) - \mathbb{I}\{\beta_j \neq 0\}\,\frac{\beta_j^2}{2\sigma^2} - \mathbb{I}\{\beta_j = 0\}\left(\log\frac{p_0\sigma}{1-p_0}\right) + C \\
&= -E_{\beta_{(-j)}}\left(\frac{\sum_{n=1}^{N}(\beta_j x_{n,j} - (y_n - \sum_{i\neq j}\beta_i x_{n,i}))^2}{2\sigma_e^2}\right) - \mathbb{I}\{\beta_j \neq 0\}\,\frac{\beta_j^2}{2\sigma^2} - \mathbb{I}\{\beta_j = 0\}\left(\log\frac{p_0\sigma}{1-p_0}\right) + C \\
&= -E_{\beta_{(-j)}}\left(\frac{\sum_{n=1}^{N}(\beta_j^2 x_{n,j}^2 - 2(y_n - \sum_{i\neq j}\beta_i x_{n,i})\beta_j x_{n,j})}{2\sigma_e^2} + \frac{\beta_j^2}{2\sigma^2}\right)\mathbb{I}\{\beta_j \neq 0\} - \left(\log\frac{p_0\sigma}{1-p_0}\right)\mathbb{I}\{\beta_j = 0\} + C \\
&= -\left(\frac{\sum_{n=1}^{N} x_{n,j}^2}{2\sigma_e^2} + \frac{1}{2\sigma^2}\right)\mathbb{I}\{\beta_j \neq 0\}\,\beta_j^2 - \left(\log\frac{p_0\sigma}{1-p_0}\right)\mathbb{I}\{\beta_j = 0\} \\
&\quad + E_{\beta_{(-j)}}\left(\frac{\sum_{n=1}^{N}(y_n - \sum_{i\neq j}\beta_i x_{n,i})x_{n,j}}{\sigma_e^2}\right)\mathbb{I}\{\beta_j \neq 0\}\,\beta_j + C
\end{aligned} \quad (4.11)$$

Substituting the expected value of the approximate spike-and-slab posterior $(1-\psi_i)\mathcal{N}(\mu_i,\sigma_i) + \psi_i\delta(0)$ with $(1-\psi_i)\mu_i$,

$$\begin{aligned}
\log q(\beta^*{}_j) = &-\left(\frac{\sum_{n=1}^{N} x_{n,j}^2}{2\sigma_e^2} + \frac{1}{2\sigma^2}\right)\mathbb{I}\{\beta_j \neq 0\}\,\beta_j^2 + \\
&+\left(\frac{\sum_{n=1}^{N}(y_n - \sum_{i\neq j}\mu_i(1-\psi_i)x_{n,i})x_{n,j})}{\sigma_e^2}\right)\mathbb{I}\{\beta_j \neq 0\}\,\beta_j + \\
&-\left(\log\frac{p_0\sigma}{1-p_0}\right)\mathbb{I}\{\beta_j = 0\} + C.
\end{aligned} \quad (4.12)$$

Now comparing the coefficients of the sufficient statistics $\mathbb{I}\{\beta_j \neq 0\}\,\beta_j^2$, $\mathbb{I}\{\beta_j \neq 0\}\,\beta_j$ and $\mathbb{I}\{\beta_j = 0\}$ to the natural parameters of the spike-and-slab $-\frac{1}{2\sigma_j^2}$, $\frac{\mu_j}{\sigma_j^2}$ and $\log\psi_j - \log(1-\psi_j) + \frac{\mu_j^2}{2\sigma_j^2} + \frac{1}{2}\log\sigma_j^2$, respectively [225], we get the following update rules:

$$
\begin{aligned}
\sigma_j^2 &= \frac{1}{\frac{\sum_{n=1}^{N} x_{n,j}^2}{\sigma_e^2} + \frac{1}{\sigma^2}} \\
\mu_j &= \frac{\sum_{n=1}^{N}(y_n - \sum_{i\neq j}\mu_i(1-\psi_i)x_{n,i})x_{n,j})}{\sum_{n=1}^{N} x_{n,j}^2 + \frac{\sigma_e^2}{\sigma^2}} \\
\psi_j &= 1 - \frac{1}{1 + \frac{p_0}{1-p_0}\sqrt{1+\sigma^2/\sigma_e^2\sum_{n=1}^{N} x_{n,j}^2}\exp\left\{-\frac{\left(\sum_{n=1}^{N}\left(y_n - \sum_{i\neq j}\mu_i(1-\psi_i)x_{n,i}\right)x_{n,j}\right)^2}{2\sigma_e^4/\sigma^2 + 2\sigma_e^2\sum_{n=1}^{N} x_{n,j}^2}\right\}}
\end{aligned}
\tag{4.13}
$$

4.2.5 Stochastic variational inference

CAVI provides a principled way to perform variational inference but requires a full pass through the data for each iteration, which is computationally expensive for large datasets. BOLT-LMM [106] [107] observed that CAVI scales approximately as $\mathcal{O}(MN^{1.5})$, because the number of CAVI steps increase by $\sqrt{N}$, where M is the number of predictors and N is the number of samples. Stochastic variational inference [226] uses ideas from stochastic optimization [227] to update the variational parameters through a noisy but unbiased estimate of ELBO's gradient. Like other modern machine learning optimizers, stochastic variational inference is also amenable to working with batches of data, rather than with a full pass through the data, making it more efficient when applied to large datasets. Overall, stochastic variational inference provides a more efficient and scalable approach for posterior inference and has been widely applied to many machine learning problems, ranging from topic modeling [226] to generative models such as variational autoencoders (VAE) [228].

We now derive the stochastic variational inference objective for the Bayesian linear regression model described in Equation 4.7 and a similar Bayesian logistic regression model. In later Section 4.3.2, we highlight several practical considerations linked to implementing the algorithm.

Derivation of Stochastic VI objective for spike-and-slab Bayesian linear regression:

Stochastic VI tries to find the optimal variational parameters by performing gradient descent directly on the ELBO objective in Equation 4.6,

$$L_{VI}(\psi, \mu, \sigma) = E[log(P(y \mid \beta, X)] - KL(q_{\psi,\mu,\sigma}(\beta) \mid P(\beta)) \tag{4.14}$$

Assuming the same model as in Equation 4.7 with a fully factorized spike and slab prior, the ELBO can be simplified as follows:

$$\begin{aligned} L^Q_{VI}(\psi, \mu, \sigma) &= \sum_{n=1}^{N} E[log(p(y_n \mid \beta))] - KL(q_{\psi,\mu,\sigma}(\beta) \mid P(\beta)) \\ &= \sum_{n=1}^{N} E[log(p(y_n \mid \beta))] - \sum_{j=1}^{M} KL(q_{\psi_j,\mu_j,\sigma_j}(\beta_j) \mid P(\beta_j)) \\ &= -\sum_{n=1}^{N} \int \frac{(y_n - \beta^T x_n)^2}{2\sigma_e^2} q_{\psi,\mu,\sigma}(\beta) d\beta - \sum_{j=1}^{M} KL(q_{\psi_j,\mu_j,\sigma_j}(\beta_j) \mid P(\beta_j)) \end{aligned} \tag{4.15}$$

where N is the number of samples, M is the number of predictors, and β are the effect estimates. We derive the KL divergence between two spike-and-slab distributions as a special case of the KL divergence between non-overlapping mixture distributions (see Appendix A for details). We do not have a closed form for the log-likelihood term. The usual approach in stochastic variational inference is to replace the integral with an approximate Monte Carlo estimate. This results in a noisy yet unbiased estimate of the ELBO, which can be shown to converge to a local minimum under the Robbins-Munro condition [227] on step-size:

$$\begin{aligned} L^Q_{VI}(\psi, \mu, \sigma) \approx &- \sum_{n=1}^{N} \sum_{s=1}^{S} \frac{(y_n - \beta(s) x_n)^2}{2\sigma_e^2} + \\ &- \sum_{j=1}^{M} \left(\frac{\psi_j}{2} \left(-1 + \frac{\mu_j^2 + \sigma_j^2}{\sigma^2} - \log \frac{\sigma_j^2}{\sigma^2} \right) + (1 - \psi_j) \log \frac{1 - \psi_j}{1 - p_0} + \psi_j \log \frac{\psi_j}{p_0} \right) \end{aligned} \tag{4.16}$$

We take S Monte Carlo samples from the approximate posterior $\beta(s)$ to calculate the approximate log-likelihood. Similar to the use of stochastic optimization in machine learning, we can also have a batched version of the above ELBO, where we update the variational parameters after each mini-batch of data. We

divide the input data into B equally sized batches and maximize the batched-version of ELBO given below:

$$L_{VI}^{Q}(\psi,\mu,\sigma) \approx -\sum_{b=1}^{B}\left(\sum_{s=1}^{S}\frac{(y_b-\beta(s)X_b)^2}{2\sigma_e^2}+\right.$$
$$\left.-\frac{1}{B}\sum_{j=1}^{M}\left(\frac{\psi_j}{2}\left(-1+\frac{\mu_j^2+\sigma_j^2}{\sigma^2}-\log\frac{\sigma_j^2}{\sigma^2}\right)+(1-\psi_j)\log\frac{1-\psi_j}{1-p_0}+\psi_j\log\frac{\psi_j}{p_0}\right)\right) \tag{4.17}$$

Derivation of Stochastic VI objective for spike-and-slab Bayesian logistic regression:

The Bayesian logistic regression model is similar to the linear regression case, only differing in the likelihood term. The model for Bayesian logistic regression with spike-and-slab prior can be written as

$$\begin{aligned}
&\textbf{Likelihood:} && P(y \mid X,\beta)=\prod_{n=1}^{N} c_n^{y_n}\{1-c_n\}^{1-y_n},\ c_n=\sigma(\beta^T X_n)\\
&\textbf{Prior:} && P(\beta)=\prod_j P(\beta_j),\ P(\beta_j)\sim(1-p_0)\mathcal{N}(0,\sigma^2)+p_0\delta(0)\\
&\textbf{Approximate Posterior:} && q(\beta)=\prod_j q(\beta_j),\ q(\beta_j)\sim(1-\psi_j)\mathcal{N}(\mu_j,\sigma_j^2)+\psi_j\delta(0)
\end{aligned} \tag{4.18}$$

Similar to the linear model in Equation 4.7, X_n and y_n represent the input vector and output values for the n-th datapoint. Additionally, σ is the sigmoid function and c_n is the output of the logistic regression after applying the sigmoid. We still assume a spike-and-slab approximate posterior, although it is not conjugate to the logistic regression model, because we empirically find it to still perform well in this case. The ELBO for Bayesian logistic regression can be simplified as in the linear case:

$$\begin{aligned}
L_{VI}^{Bi}(\psi,\mu,\sigma) &= \sum_{n=1}^{N} E[log(p(y_n \mid \beta))] - KL(q_{\psi,\mu,\sigma}(\beta) \mid P(\beta))\\
&= -\sum_{n=1}^{N}\int\{y_n\log c_n+(1-y_n)\log(1-c_n)\}q_{\psi,\mu,\sigma}(\beta)d\beta-\sum_{j=1}^{M}KL(q_{\psi_j,\mu_j,\sigma_j}(\beta_j)\mid P(\beta_j))
\end{aligned} \tag{4.19}$$

where $c_n = \sigma(\beta^T X_n)$, that is, the output after sigmoid activation. We estimate the expected log-likelihood with a Monte Carlo estimate, whereas we write a closed-form solution for the KL divergence term similar to Equation 4.17. The batched version of the ELBO for Bayesian logistic regression can be written as

$$L_{VI}^{Bi}(\psi, \mu, \sigma) \approx -\sum_{b=1}^{B}\left(\sum_{s=1}^{S} y_b \log(\sigma(\beta^T X_b)) + (1-y_b)\log(1-\sigma(\beta^T X_b)) + \right.$$
$$\left. - \frac{1}{B}\sum_{j=1}^{M}\left(\frac{\psi_j}{2}\left(-1 + \frac{\mu_j^2 + \sigma_j^2}{\sigma^2} - \log\frac{\sigma_j^2}{\sigma^2}\right) + (1-\psi_j)\log\frac{1-\psi_j}{1-p_0} + \psi_j \log\frac{\psi_j}{p_0}\right)\right) \tag{4.20}$$

where we take S Monte Carlo samples to approximate the log-likelihood term.

4.3 Method description

4.3.1 Overview of the method

Quickdraws is a mixed-model association algorithm that models phenotypes as a combination of fixed effects, random genetic effects, and random environmental effects (see Section 1.3.3 for introduction to mixed models). Like all modern GWAS algorithms [106–108, 119, 122, 123], Quickdraws detects association using a two-step approach. The first step, sometimes called model fitting, focuses on estimating the random genetic effects, while the second, which we refer to as the testing step, computes an association statistic for all candidate variants.

During the first model fitting step, Quickdraws estimates the variance due to the genetic and environmental random effects. For quantitative traits, we estimate variance components for multiple traits in parallel using randomized Haseman-Elston regression (RHE), a fast method-of-moments approach [109–111]. For binary traits, we instead perform a grid search over a predefined set of heritability values. Next, Quickdraws estimates genetic effects by using a leave-one-chromosome-out scheme [104–106, 117], where phenotypic values for all samples are predicted from raw genotype data, using predominantly common genomic variants. The accuracy of this prediction step is a key determinant of association power [105], and is performed using Bayesian regression, where the previously obtained variance

component estimates are used to set the prior. Similarly to BOLT-LMM, which uses a mixture-of-Gaussians prior on variant effects [106, 107], Quickdraws increases association power by modeling trait architectures that are not fully polygenic. It does so by adopting a spike-and-slab prior on variant effects. Other scalable methods, such as REGENIE and FastGWA, on the other hand, rely on Gaussian priors that are more computationally tractable but assume fully polygenic traits.

Quickdraws incorporates several technical advancements to achieve scalability while maintaining high power. First, Quickdraws uses stochastic variational inference [226, 229], optimized with first-order optimizers [227, 230], to enable linear scaling with sample size, whereas BOLT-LMM requires $\mathcal{O}(N^{1.5})$ computation for N individuals. Second, gradient-based stochastic variational inference allows Quickdraws to leverage modern computing architectures, such as GPUs. By offloading matrix multiplications and gradient evaluations to GPUs, Quickdraws achieves significant speedups over CPU-based computation. Finally, Quickdraws accelerates convergence by initializing leave-one-chromosome-out (LOCO) runs with effect estimates from the whole-genome model, reducing computation without compromising accuracy [231]. Further details on Bayesian regression and its computational optimizations are provided in the later Section 4.3.2.

During the testing step, Quickdraws uses these estimated genetic effects to compute a score-based test statistic for a linear or logistic mixed model, which is approximated up to a constant of proportionality [106, 232, 233]. This constant is later estimated by matching an estimate of the scaled effective sample size [234] from linear or logistic regression on an unrelated and homogeneous subset of the data. We additionally correct for potential instability in the score-based test statistic that may be due to case/control imbalance in binary traits [119], using approximate Firth logistic regression [123]. We further optimize the calculation of test statistics using Numba [235] and parallelize matrix operations across multiple cores.

A graphic overview of the method can be found in Figure 4.2.

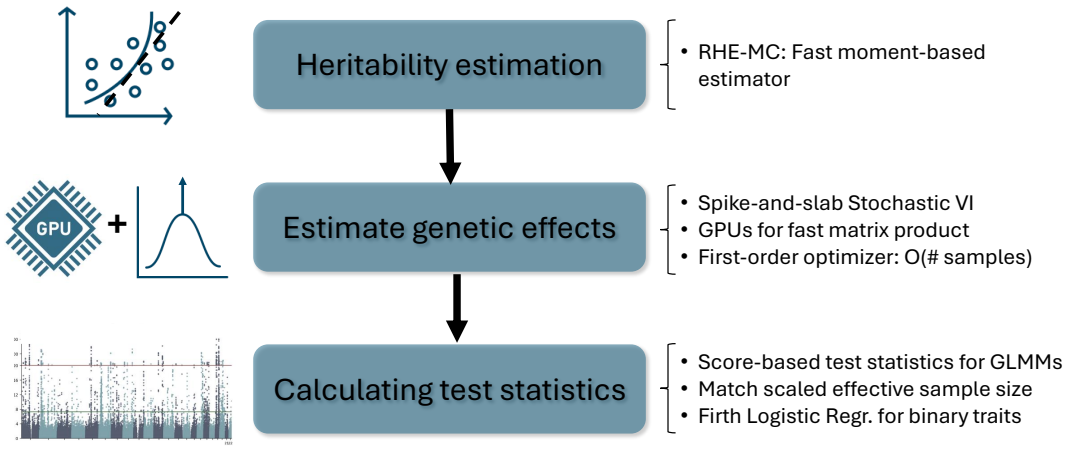


Figure 4.2: Overview of the Quickdraws algorithm. The algorithm is divided into two main steps, first step deals with estimating genetic effects and the second step with calculating and calibrating per-variant test statistics. The first step is further divided into heritability estimation using moment-based methods [111] and Bayesian regression for phenotype prediction which uses stochastic variational inference and GPU-based matrix multiplication for increased scalability. In the test statistics calculation step we use a novel calibration strategy which estimates and matches effective sample size with linear regression model run on unrelated and homogeneous subset of the data, along with approximate Firth logistic regression to correct for biases in binary traits. The illustrations were downloaded from Flaticon.

4.3.2 Bayesian regression with spike-and-slab prior

Quickdraws uses a Bayesian regression to estimate the genetic effects. This Bayesian regression uses a spike-and-slab distribution on the variant effects, which results in increased association power compared to approaches that assume an infinitesimal model and rely on a Gaussian prior. We optimize the posterior using stochastic variational inference [226], which significantly improves the scalability of this approach.

Quantitative traits

Quickdraws adopts a Bayesian linear regression model as discussed in Section 4.2.3:

$$\begin{aligned}
&\textbf{Likelihood:} && P(y \mid X, \beta) \sim \mathcal{N}(\beta^T X, \sigma_e^2) \\
&\textbf{Prior:} && P(\beta) = \prod_j P(\beta_j),\ P(\beta_j) \sim (1-p_0)\mathcal{N}(0, \sigma^2) + p_0\delta(0) \\
&\textbf{Approximate Posterior:} && q(\beta) = \prod_j q(\beta_j),\ q(\beta_j) \sim (1-\psi_j)\mathcal{N}(\mu_j, \sigma_j^2) + \psi_j\delta(0)
\end{aligned} \tag{4.21}$$

Similar to Equation 4.7, given N individuals and M variants, X is an $N \times M$ covariate-adjusted genotype matrix, y is a $N \times 1$ covariate-adjusted phenotype vector, and β is a $M \times 1$ vector of variant effects. $\delta(0)$ is the Dirac delta function at 0, representing the spike in the spike-and-slab prior used to model the sparsity in the genetic architecture. σ_e^2 and σ^2 are the variance component corresponding to environment and per predictor effect, p_0 is the sparsity and j indexes over all model SNPs. As we discussed in Section 4.2.3, we also assume a fully-factorized spike-and-slab approximate posterior, which leads to efficient and low-variance sampling from the posterior. And, we use stochastic variational inference to directly optimize the evidence lower bound (ELBO). The ELBO for Bayesian linear regression with spike-and-slab prior, normal likelihood, and spike-and-slab approximate posterior (derived in Section 4.2.5) can be approximated as follows:

$$\begin{aligned}
L_{VI}^{Q}(\psi, \mu, \sigma) \approx -\sum_{b=1}^{B} \Bigg(\sum_{s=1}^{S} \frac{(y_b - \beta(s) X_b)^2}{2\sigma_e^2} + \\
- \frac{1}{B} \sum_{j=1}^{M} \left(\frac{\psi_j}{2} \left(-1 + \frac{\mu_j^2 + \sigma_j^2}{\sigma^2} - \log \frac{\sigma_j^2}{\sigma^2} \right) + (1-\psi_j) \log \frac{1-\psi_j}{1-p_0} + \psi_j \log \frac{\psi_j}{p_0} \right) \Bigg)
\end{aligned} \tag{4.22}$$

In this expression, $L_{VI}^{Q}(\psi, \mu, \sigma)$ is the variational inference objective to be maximized, ψ, μ, σ are the variational parameters to be optimized, X_b and y_b are a genotype matrix and a phenotype vector containing the b-th (out of B) batch of the data, batched along the number of samples, while $\beta(s)$ are the effect estimates sampled from the approximate posterior $q(\beta)$. Note that, contrary to other variational inference schemes, we aim to optimize a stochastic objective, since $\beta(s)$ is randomly sampled.

Binary traits

For binary traits, we adopt a Bayesian logistic regression model, so that the likelihood of Equation 4.21 becomes:

$$P(y \mid X, \beta) = \prod_{n=1}^{N} c_n^{y_n} \{1 - c_n\}^{1-y_n},\ c_n = \sigma(\beta^T X_n), \tag{4.23}$$

while the prior and approximate posterior are unchanged. In the above, X_n and y_n represent the covariate-adjusted genotype vector and phenotype values for the n-th individual, σ is the sigmoid function, which maps the output of the regression to a value between 0 and 1, and all other quantities are as previously defined. We also assume that the approximate posterior has a spike-and-slab distribution; although it is not conjugate to the logistic distribution, we observed it to provide a good approximation to the true posterior. The ELBO, which we again optimize using stochastic variational inference, now takes the form:

$$\begin{aligned} L_{VI}^{Bi}(\psi, \mu, \sigma) \approx -\sum_{b=1}^{B} \Bigg(\sum_{s=1}^{S} y_b \log(\sigma(\beta^T X_b)) + (1 - y_b) \log(1 - \sigma(\beta^T X_b)) + \\ - \frac{1}{B} \sum_{j=1}^{M} \left(\frac{\psi_j}{2} \left(-1 + \frac{\mu_j^2 + \sigma_j^2}{\sigma^2} - \log \frac{\sigma_j^2}{\sigma^2} \right) + (1 - \psi_j) \log \frac{1 - \psi_j}{1 - p_0} + \psi_j \log \frac{\psi_j}{p_0} \right) \Bigg) \end{aligned} \tag{4.24}$$

where all terms are defined as in Equation 4.22 and derivation is provided in Section 4.2.5.

Practical considerations

Stochastic VI methods provide a principled and efficient way to perform variational inference. This approach is simple but requires particular care to make sure that the variance of the ELBO estimate is not too large, as that may have a negative impact on convergence. We make use of multiple recently developed strategies to reduce variance in the stochastic estimation of the ELBO, including continuous relaxation [236, 237], local reparameterization trick [238], and antithetic variates [239].

Reparameterization: When dealing with stochastic objectives, such as the ELBO, it is often difficult to differentiate the objective due to intrinsic randomness in the

parameters. One solution to this problem is provided by the reparameterization trick [228, 240], which models the stochasticity in the parameters as an input to the model rather than an intrinsic property of the model. In more detail, the reparameterization trick assumes that distributions can be reparameterized in the form $g(\theta, \epsilon)$ where θ is the variational parameter, ϵ is the independently sampled random variable, and g is a differentiable function. As an example, consider the case of an isotropic normal distribution. To sample from a mean-field normal distribution, we may reparameterize the sample as follows:

$$\begin{aligned} X \sim \mathcal{N}(\mu, \sigma^2 I) \implies X = \mu + \epsilon \cdot \sigma, \\ \epsilon \sim \mathcal{N}(0, I), \end{aligned} \tag{4.25}$$

where ϵ is i.i.d. normal noise with no dependence on μ and σ. Relying on the reparameterization and treating ϵ as an independent input, we can differentiate with respect to μ and σ.

Local reparamaterization: The local reparamaterization trick [238] was introduced as a way to provide computationally fast and low-variance gradient estimators in parameter-heavy stochastic VI methods. It was originally applied to perform variational inference in Bayesian neural networks but is easily extended to Bayesian regression. Using the original reparamaterization trick [228, 240] leads to sampling i.i.d. noise at least once for each parameter, making this approach computationally intensive in high-dimensional settings, which may involve millions of parameters. The local reparamaterization trick translates the uncertainty on individual effect estimates, β, into local noise in the output, $\beta^T X$, thus reducing the computational overhead. In the simple mean-field Gaussian VI, the local reparameterization trick can be used to sample the outputs directly,

$$w_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2_{i,j}), \quad b_{m,j} = \sum_i a_{m,i} w_{i,j}, \tag{4.26}$$

where $a_{m,i}$ is the m-th input data point and i-th input feature and $b_{m,j}$ is the output corresponding to the m-th data point and j-th output variable in the Bayesian linear

regression model. $b_{m,j}$ is a weighted-sum of normally distributed random variables, implying that $b_{m,j}$ is also normally distributed, with the following parameters:

$$\begin{aligned} &w_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2_{i,j}) \implies b_{m,j} \sim \mathcal{N}(\gamma_{m,j}, \delta^2_{m,j}), \\ &\gamma_{m,j} = \sum_i a_{m,i}\mu_{i,j}, \quad \delta^2_{m,j} = \sum_i a^2_{m,i}\sigma^2_{i,j}. \end{aligned} \tag{4.27}$$

In order to sample this output directly, one can apply the original reparameterization trick for the normal distribution:

$$b_{m,j} = \gamma_{m,j} + \epsilon_{m,j}\delta_{m,j}, \quad \epsilon_{m,j} \sim \mathcal{N}(0,1). \tag{4.28}$$

Note that in Bayesian linear/logistic regression the local reparameterization removes the need to sample all the weights in a layer and only requires sampling proportional to the number of output variables (= number of phenotypes in our case) for each mini-batch of data, substantially improving the computational complexity of the algorithm.

The local reparamaterization trick was originally only derived for mean-field Gaussian variational inference, since the weighted sum of arbitrary distributions cannot always be analytically derived. In the case of mean-field spike-and-slab distributions, we use the central limit theorem to approximate the weighted sum of independent spike-and-slab distributions as approximately Gaussian. Note that the input dimension for Bayesian regression in our setup is quite high (usually $> 100,000$ markers), justifying this approximation. This idea was first explored in [241] to sample discrete weights for binary and ternary neural networks. We thus reparametrize the output of our model (as described in equations 4.21 and 4.23) as follows:

$$\begin{aligned} &w_{i,j} \sim p_{i,j}\mathcal{N}(\mu_{i,j}, \sigma^2_{i,j}) + (1-p_{i,j})\delta(0) \implies b_{m,j} \sim \mathcal{N}(\gamma_{m,j}, \delta_{m,j}), \\ &\gamma_{m,j} = \sum_i a_{m,i}\mu_{i,j}p_{i,j}, \quad \delta_{m,j} = \sum_i a^2_{m,i}(p_{i,j}\sigma^2_{i,j} + p_{i,j}\mu^2_{i,j} - p^2_{i,j}\mu^2_{i,j}). \end{aligned} \tag{4.29}$$

Antithetic variates for variance reduction: The use of antithetic variates is a popular approach to speed up Monte Carlo computations of the expectation of a random function [239]. This strategy relies on taking the antithetic path of

the sampled path to reduce the variance in the overall estimator. For example, in order to sample from a normal random variable (with mean μ and variance σ^2), one might use the reparameterization trick to sample $\mu + \epsilon\sigma$. The antithetic path for a normally distributed random variable is $\mu - \epsilon\sigma$. In addition to providing a valid sample from the underlying distribution (in this case, normal with mean μ and variance σ^2), the antithetic path also reduces the variance of the Monte Carlo estimate, as the sampled path and antithetic path are often negatively correlated.

We use both antithetic variates and the local reparameterization trick to reduce the variance of the ELBO. As described above, the antithetic path for a normally distributed random variable is simply obtained by replacing ϵ with $-\epsilon$. For each forward pass, we thus obtain two Monte Carlo samples for the output of the model as follows:

$$\begin{aligned}
\beta(s_1) &= (\sum_i a_{m,i}\mu_{i,j} + \sqrt{\sum_i a_{m,i}^2(p_{i,j}\sigma_{i,j}^2 + p_{i,j}\mu_{i,j}^2 - p_{i,j}^2\mu_{i,j}^2}\epsilon), \\
\beta(s_2) &= (\sum_i a_{m,i}\mu_{i,j} - \sqrt{\sum_i a_{m,i}^2(p_{i,j}\sigma_{i,j}^2 + p_{i,j}\mu_{i,j}^2 - p_{i,j}^2\mu_{i,j}^2}\epsilon), \\
\beta(s) &= [\beta(s_1), \beta_2(s_2)].
\end{aligned} \tag{4.30}$$

4.3.3 Heritability estimation

To set the prior variance and likelihood contribution in equations 4.21 and 4.23, we compute an estimate of the narrow-sense heritability for the trait of interest. For quantitative traits, we estimate heritability using a parallelized implementation of the randomized Haseman–Elston regression algorithm[109, 110] (RHE, PyPI version 1.0) that allows the simultaneous estimation of variance components for multiple traits [111]. This approach requires a single pass through the data and allows estimating heritability in time linear in N and M. We set the number of random RHE-mc vectors to 50 while assigning markers to 8 components based on their LD scores and MAF. We also exclude the HLA region, as recommended in [110]. For Binary traits, we perform a grid search over a set of heritability values, $h^2 \in \{0.01, 0.25, 0.5, 0.75\}$, running the Bayesian regression for each value and selecting the heritability corresponding to the highest likelihood.

4.3.4 Test statistics calculation

After estimating the random genetic effects as described in sections 4.3.2 and 4.3.3, we compute association statistics for quantitative or binary traits as follows.

Quantitative traits

We model the phenotype of interest as a combination of fixed and random effects, which include genetic effects and environmental effects:

$$y = \alpha C + x_{test}\beta_{test} + g + \epsilon, \tag{4.31}$$

where y is a $N \times 1$ vector of phenotype values, x_{test} is a $N \times 1$ vector of variant values, C is a $N \times C$ matrix of C covariates, g and ϵ are random genetic and environmental effects. Note that using a covariate-adjusted genotype and phenotype by first regressing out any covariates is equivalent to including covariates as fixed effects in the above equation. Based on this model, we aim to compute the following χ^2 association statistic (a derivation is provided in Section 1.3.3):

$$\frac{(x_{test}^T \hat{V}^{-1} y)^2}{x_{test}^T \hat{V}^{-1} x_{test}} \sim \chi_1^2, \tag{4.32}$$

where $\hat{V}$ is the estimated variance matrix, defined as $\hat{V} = \frac{\hat{\sigma_g}^2}{M} X_{GRM} X_{GRM}^T + \hat{\sigma_e}^2 I_N$, $\hat{\sigma_g}^2$ and $\hat{\sigma_e}^2$ are estimated genetic and environmental variance components, and X_{GRM} is the genotype matrix used for model fitting. Obtaining the test statistic of Equation 4.32 requires calculating the inverse of a $N \times N$ matrix, which creates a computational bottleneck for large sample sizes. This test statistic can be linked to the best linear unbiased predictor (BLUP) of a trait, by re-writing Equation 4.32 as described in [106]:

$$\frac{(x_{test}^T \tilde{y}/\sigma_e^2)^2}{x_{test}^T \hat{V}^{-1} x_{test}} \sim \chi_1^2. \tag{4.33}$$

In the above, $\tilde{y}$ is a residual phenotype estimated using BLUP. We adopt a similar strategy and replace the $\tilde{y}$ with $\tilde{y}_{LOCO}$, leave-one-chromosome-out residual phenotype from Bayesian regression. In addition, the denominator $x_{test}^T \hat{V}^{-1} x_{test}$ of Equation 4.33 can be shown to be well approximated using a constant multiple of

$x_{test}^T x_{test}$ [232], further reducing computational costs. We can therefore write the test statistics for a quantitative trait up to a constant of proportionality as:

$$\chi_Q^2 \propto \frac{(x_{test}^T \tilde{y}_{LOCO})^2}{x_{test}^T x_{test}}. \tag{4.34}$$

Binary traits

For binary traits, we model the association between genotype and phenotype using a logistic mixed model:

$$logit(p_i) = \alpha C + x_{test}\beta_{test} + g + \epsilon, \tag{4.35}$$

where $p_i = P(y_i = 1 \mid x_{test}, g, C)$ is the probability of i-th individual being a case given x_{test}, covariates C, and random genetic effects g. A score test statistic for the null hypothesis $\beta_{test} = 0$ is then computed as $T = x_{test}^T(y - \hat{p})$, where $\hat{p}$ is the estimated mean under the null model. The normalized test statistic for logistic mixed model can then be written as:

$$T = \frac{x_{test}^T(y - \hat{p})}{\sqrt{x_{test}^T \hat{P} x_{test}}}, \tag{4.36}$$

where $\hat{P} = \hat{V}^{-1} - \hat{V}^{-1}C(C^T\hat{V}^{-1}C)^{-1}C^T\hat{V}^{-1}$ is a dense $N \times N$ matrix with $\hat{V} = \frac{\hat{\sigma_g}^2}{M} X_{GRM} X_{GRM}^T + \hat{W}^{-1}$ and $\hat{W} = diag\{\hat{p}(1-\hat{p})\}$. Similar to quantitative traits, the test statistic in Equation 4.35 can be written up to a constant of proportionality [119] as:

$$T_B \propto \frac{x_{test}^T(y - \hat{p})}{\sqrt{x_{test}^T \hat{W} x_{test}}}, \tag{4.37}$$

where $\hat{W}$ is a diagonal matrix (with diagonal entries $\hat{p}_i(1-\hat{p}_i)$) from the predictions of the null model in Step 1. The test statistic from 4.37 is assumed to be normally distributed, which is often not the case with binary traits that have low prevalence or for rare variants. We therefore instead perform Firth logistic regression, which penalizes the likelihood by using a Jeffreys' prior, removing most of the asymptotic bias in maximum-likelihood estimation. In order to save computation, we use the likelihood ratio test with an approximate version of Firth logistic regression

introduced in [123] on variants below a p-value threshold of 0.05. We further narrow down the list of variants by only applying Firth logistic regression to rare variants (default MAF $< 5\%$) or rare traits (default prevalence $< 5\%$) or both.

Calibration of test statistics

Calibrating the test statistics generated up to a constant of proportionality in Equation 4.34 and 4.37 is a crucial part of a GWAS algorithm, which ensures the false positives are controlled. We calibrate the summary statistics by estimating the effective sample size (ESS) increase compared to running linear regression on a homogeneous subset of unrelated individuals. To this end, we estimate the increase in ESS linked to the use of non-infinitesimal (i.e., sparse) Bayesian linear regression (γ_{blr}), as well as the reduction in ESS due to the presence of close relatives (γ_{rel}). We then adjust the summary association statistics by matching the mean χ^2 statistic minus one [234] with that observed for linear or logistic regression run on a homogeneous subset of unrelated individuals. In more detail, we multiply the uncalibrated Quickdraws summary statistics from equations 4.34 and 4.37 by a correction term c, computed as:

$$c = \frac{\gamma_{rel}\gamma_{blr}\frac{N_{qd}}{N_{lr}}(< \chi^2_{lr} > -1) + 1}{< \chi^2_{qd} >}. \tag{4.38}$$

In the above, $< \chi^2_{qd} >$ and $< \chi^2_{lr} >$ are the mean χ^2 test statistics from uncorrected Quickdraws and from linear/logistic regression run on a homogeneous unrelated subset of individuals, respectively. $\frac{N_{qd}}{N_{lr}}$ is the ratio of the total number of samples to the number of homogeneous unrelated samples used for linear/logistic regression. γ_{rel} and γ_{blr} are correction terms to account for relatedness and the use of Bayesian linear regression, which we describe in detail below. Note that the γ_{blr} term is estimated for each LOCO run, so the correction in Equation 4.38 is performed separately for each chromosome. For binary traits, we use logistic regression and apply the correction separately for Firth-corrected and Firth-uncorrected variants.

Correcting the ESS for the presence of relatedness (γ_{rel} term). To estimate the effective number of samples in the presence of relatedness, we follow an approach

similar to that of [242]. To this end, we first estimate genetic relatedness between samples using KING [243] as implemented in PLINK [113]. We only retain relationships up to the 3rd degree (KING score $> 2^{-\frac{9}{2}}$), as we found the ESS multiplier, described below, to not be significantly impacted by the inclusion of higher degree relatives. For each pair of relatives, we estimate the degree of the relationship using KING's output values as follows:

Range for KING ψ value	**Relatedness**
$\psi \geq 2^{-\frac{3}{2}}$	1 (Monozygotic Twins)
$2^{-\frac{3}{2}} > \psi \geq 2^{-\frac{5}{2}}$	0.5 (1^{st} degree)
$2^{-\frac{5}{2}} > \psi \geq 2^{-\frac{7}{2}}$	0.25 (2^{nd} degree)
$2^{-\frac{7}{2}} > \psi \geq 2^{-\frac{9}{2}}$	0.125 (3^{rd} degree)

Table 4.1: Genetic relationship estimates based on KING values.

Based on this, we compute the reduction factor for the effective sample size, γ_{rel}, using the following estimator, for which a derivation can be found in [242]:

$$\gamma_{rel} = \frac{tr((K\sigma_g^2 + I(1-\sigma_g^2))^{-1}K)}{N}. \tag{4.39}$$

In this expression, N is the total number of samples, K is the estimated kinship matrix, and σ_g^2 is the heritability of the trait. To efficiently compute γ_{rel}, we eigendecompose the matrix K, obtained using KING utilizing the block-like structure of the kinship matrix. In particular, the effective sample size multiplier in the presence of relatedness can be approximated as:

$$\gamma_{\text{rel}} \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\lambda_i}{(\lambda_i - 1)\sigma_g^2 + 1}, \tag{4.40}$$

where λ_i is the i-th eigenvalue of the kinship matrix K. And to facilitate this computation, we separately sum across each block B_i of related individuals:

$$\gamma_{\text{rel}} \approx \frac{1}{N}\sum_{i=1}^{B}\sum_{j=1}^{N_{B_i}} \frac{\lambda_{i,j}}{(\lambda_{i,j} - 1)\sigma_g^2 + 1}. \tag{4.41}$$

For binary traits, we convert the heritability from the liability scale to the observed scale.

Correcting the ESS for the use of Bayesian linear regression (γ_{blr} term): The effective sample size multiplier, γ_{blr}, has been shown to be proportional to the inverse of the residual trait variance in a linear regression model [106, 242]. To account for the increase in effective sample size in the model fitting step, we therefore calculate the inverse variance of the residual trait

$$\gamma_{blr} = \frac{Var(y)}{Var(y - \hat{y})} \tag{4.42}$$

where y and $\hat{y}$ are true and predicted trait values from the whole-genome regression regression on a held-out set. In more detail, we use the effect estimates from the best sparsity parameter to calculate the variance explained on held-out data. As the γ_{blr} term needs to be computed during the LOCO step, we compute these variance terms separately for each chromosome and sum across all chromosomes, excluding the one currently being tested.

4.3.5 Performance optimization

Memory optimization

1. **Raw genotypes:** for a diploid individual, the raw genotypes take values in $0, 1, 2$. We store the $N \times M$ genotype matrix in $NM/4$ bytes, that is using 2 bits per genotype entry. We do so by encoding 8 consecutive entries of the genotype matrix using two 8-bit integers using the built-in `numpy` features in Python, `np.packbits()` and `np.unpackbits()`. We also store the column-wise mean and variance of the genotype matrix, used to mean-center and standardize, on the fly.

2. **Genotype streaming during the calculation of test statistics:** the calculation of test statistics is usually performed on many more variants than those used for model fitting. We therefore optimize this step by streaming genotype blocks rather than loading all the variants into memory. The block size is determined based on the maximum memory allowed at runtime (default: 32GB) and the data IO is parallelized using multiple cores.

Speed optimization

1. **GPUs for fast Bayesian regression:** we use GPUs along with the `pytorch` machine learning package [244] to speed-up matrix multiplication and sampling operations during stochastic variational inference.

2. **Transfer learning:** rather than training the 22 LOCO models independently, we initialize each model training using the effect estimates previously computed for the whole-genome regression. This transfer learning approach requires fewer iterations to converge, making model fitting 2× to 2.5× faster, and leads to similar association power when compared to training each LOCO model separately. We compare the advantage of transfer learning over independent random initialization in simulations, by looking at the power improvement vs number of iterations in Figure 4.3.

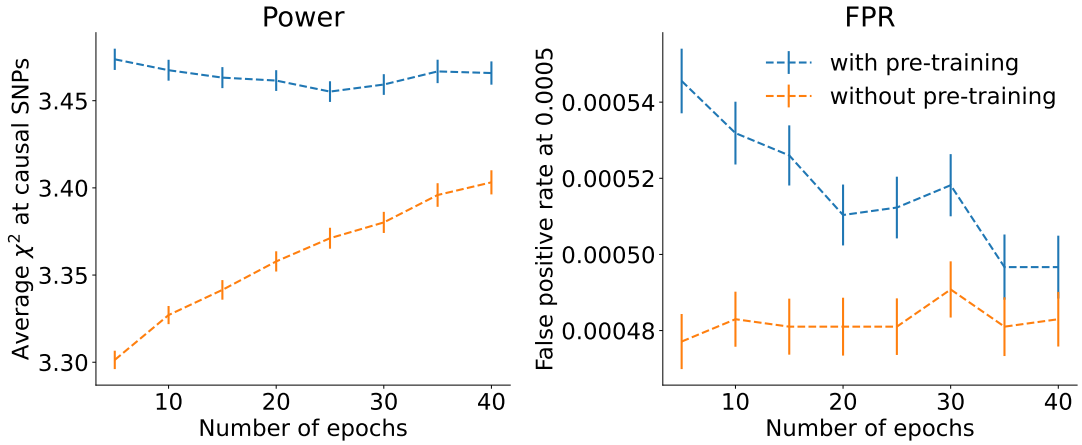


Figure 4.3: Transfer-learning in Bayesian regression. We measure the causal χ^2 and false positive rate at 5×10^{-4}, with and without pretraining (for 90 epochs), and varying the number of LOCO training epochs. Error bars represent the standard error of the mean χ^2 or false positive rate across 50 independent quantitative traits with GB-unrel structure and 2% polygenicity.

3. **Numba optimization for the calculation of test statistics:** we implement a fully-vectorized version of linear regression for multiple traits using the `parallel` and `njit` functionality from the Numba just-in-time (JIT) compiler [235].

4. **Calibrating test statistics using genotyped SNP data:** instead of running linear regression or logistic regression on all imputed variants, we calibrate our test statistics only using the genotype data provided in step 1 ($\sim 458k$ variants). We find that matching scaled effective sample size estimates using this subset of variants provides consistent calibration in the larger set of imputed variants, while greatly reducing running time, as we only need to evaluate the logistic or linear regression model on a subset of variants.

5. **HDF5 and PySnpTools for data loading:** we store the raw genotype matrix in a compressed HDF5 file [245], which provides fast sample-wise access (that is, it allows obtaining all variants for an individual) for Bayesian regression. We instead use PySnpTools (see `https://github.com/fastlmm/PySnpTools`), which provides parallelizable variant-wise access, to access `bgen` or `bed` files for the calculation of test statistics.

6. **Approximate Firth logistic regression:** Approximate Firth regression is only applied to variants with logistic regression p-value below 0.05, as done in REGENIE and FastGWA-GLMM. We further narrow down the list of variants by only applying Firth logistic regression to rare variants (MAF $< 5\%$) or rare traits (prevalence $< 5\%$).

7. **Multiple traits:** Most of our operations are vectorized using Numpy [246] and Numba [235], which leads to computational gains in the parallel analysis of multiple traits.

4.3.6 Hyperparameter tuning

We perform cross-validation by randomly splitting the data into training (80%) and testing (20%) sets to estimate the optimal sparsity hyperparameter $1 - p_0$ (see Equation 4.7) using the values $\{0.5, 0.2, 0.1, 0.05, 0.02, 0.01\}$. For heritability in binary traits, we evaluate $\{0.01, 0.25, 0.5, 0.75\}$. Compared to coordinate-ascent variational inference, stochastic variational inference relies on additional hyperparameters that affect the final testing performance. These include the learning

rate of the optimizer, the batch-size, and the number of training epochs. Setting a high learning rate results in unstable convergence of the optimizer, while a low learning rate leads to the need for more training epochs. Larger batch sizes lead to better performance but increased GPU memory usage. Considering these trade-offs, we set the batch-size to 128; the learning rate corresponding to each value of the sparsity hyperparameter was set to 4×10^{-4}, 2×10^{-4}, 2×10^{-4}, 1×10^{-4},2×10^{-5}, 5×10^{-6} for sparsity $(1 - p_0) \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$. We also set the number of training epochs to 80 for the initial cross-validation step, and to 40 for the final LOCO step. We observed these hyperparameters to be robust across various sample sizes, providing a good balance between statistical power and computational costs.

4.4 Summary of the Quickdraws algorithm

We provide a summary of the Quickdraws algorithm, which performs two main steps:

1. **Estimating genetic effects (model fitting):** This is further divided into two parts:

   (a) **Variance component estimation:** For quantitative traits, we use a multi-trait version of the RHE-mc algorithm [109–111] to obtain estimates of narrow-sense heritability. For binary traits, we directly perform the Bayesian regression (see below) while testing several heritability values in a grid search.

   (b) **Bayesian regression:** We perform Bayesian linear/logistic regression to estimate the genetic effects, using a spike-and-slab prior on the effect estimates:

   $$P(\beta_j) \sim (1 - p_0)\mathcal{N}(0, \sigma^2) + p_0\delta(0). \tag{4.43}$$

   There is one free hyperparameter in the prior (the sparsity term $1 - p_0$) for quantitative traits and two free hyperparameters for binary traits (the sparsity term and h^2). We perform cross-validation to estimate these hyperparameters, splitting the input genotype/phenotype data into 90% for training and 10% for validation and parallelly running the regression using $1 - p_0 \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ and, for binary traits, $h^2 \in \{0.01, 0.25, 0.5, 0.75, 0.9\}$. We choose the sparsity (and h^2 for binary traits) parameter that maximizes (in the validation set) the test log-likelihood for the traits.

   After obtaining the optimal hyperparameters for each trait, we retrain the regression using a leave-one-chromosome-out approach. We obtain a $2.5 - 3\times$ speed-up for the training for this LOCO step by using transfer learning, initializing the effect estimates from the cross-validation run performed using all chromosomes.

2. **Calculation and calibration of association statistics (testing):** Given the estimated genetic effects, we calculate linear mixed-model or logistic mixed-model test statistics up to a constant of proportionality. We rely on optimizations such as just-in-time compilation and code vectorization in Numba. We estimate the constant of proportionality by matching the estimated scaled effective sample size of linear regression on a subset of unrelated homogeneous samples for quantitative traits, and logistic regression on a subset of unrelated homogeneous samples for binary traits.

5

Application to simulated and real phenotypes in UK Biobank

Contents

5.1 Chapter overview 188
5.2 Data and quality control 188
5.2.1 UK Biobank data 188
5.2.2 Quality control and preprocessing for GWAS 189
5.3 Performance in simulated data 190
5.3.1 Simulation design 190
5.3.2 False positive rate evaluation 193
5.3.3 Statistical power evaluation 199
5.4 UK Biobank analysis 204
5.4.1 Analysis of false positive rates 204
5.4.2 Number of associations 207
5.4.3 Replication analysis with Biobank Japan, Finngen and other large-scale GWAS 209
5.4.4 Predictive power in out-of-sample data 212
5.4.5 Correction of participation bias 216
5.5 Cost analysis on UK Biobank RAP 216
5.6 Discussion 218
5.6.1 Sharing and combining posterior effect estimates 220
5.6.2 Limitations and future work 220

5.1 Chapter overview

In this chapter we apply Quickdraws to simulated and real phenotypes in the UK Biobank [3]. We start by discussing data and quality control steps we used to perform GWAS in the UK Biobank in Section 5.2. In order to assess the calibration and demonstrate statistical power we then perform simulation studies in Section 5.3 across a wide variety of designs varying relatedness, population structure, prevalence and the genetic architecture of the simulated trait. Subsequently in Section 5.4, we apply Quickdraws on 79 quantitative and 50 self-reported disease phenotypes from UK Biobank where we show Quickdraws leads to state-of-the-art statistical power while controlling for false positives. We perform a replication analysis with summary statistics from external biobanks like Biobank Japan and Finngen, followed by predictive power analysis and comparing with PGS methods. Finally in Section 5.5, we perform a thorough cost analysis on UK Biobank's cloud platform comparing Quickdraws against recent scalable methods like FastGWA and REGENIE. We conclude the chapter with discussing the key findings and directions for future work.

5.2 Data and quality control

5.2.1 UK Biobank data

We used the UK Biobank SNP array data for 405,088 white British samples [3] (in PLINK bed format [112, 113]) for model fitting and HRC+UK10k-imputed data [3, 247, 248] (in bgen v1.2 format [249]) for association testing. We filtered the set of available autosomal variants to have minor allele frequency (MAF) $\geq 1\%$, Hardy-Weinberg equilibrium $p \geq 1 \times 10^{-15}$, and a genotyping rate above 99%, obtaining a set of 458,620 markers. We used these markers as input for the model-fitting step of BOLT-LMM, REGENIE, and Quickdraws, while for SAIGE we performed LD pruning using PLINK (v1.9 [112, 113] setting `-window_size` to 50kb, `-step_size` to 5, and the R^2 threshold to 0.05) as recommended, which resulted in $89,177$ markers. We computed association statistics on a much larger set of $\sim$13.3 imputed variants with MAF $\geq 0.1\%$ and INFO score ≥ 0.8.

For our main analyses we considered 79 quantitative traits (comprising blood-related, anthropometric, and other traits) and 50 binary self-reported disease traits. The traits we selected had a phenotyping rate higher than 80% and, for quantitative traits, a statistically significant estimated narrow-sense heritability ($p < 5 \times 10^{-4}$, using LD-score regression estimates available at `https://nealelab.github.io/UKBB_ldsc/h2_browser.html`). Quantitative traits were also standardized, mean-centered, and quantile normalized to have an approximately Gaussian distribution. All methods we considered were provided with the top 20 principal components of the genotype matrix [3], age, sex, age^2, $age * sex$, $age^2 * sex$, and smoking status as covariates during model fitting and testing.

We additionally considered 2,923 plasma protein traits from the UK Biobank, which we pre-processed as done in [250, 251]. We downloaded normalized protein expression values from the UK Biobank RAP (field 30900), with measurements for up to 53,074 participants. For GWAS, we considered up to 49,441 European participants (based on self-reported field 21000). We ran Quickdraws with covariates including the top 20 principal components [3], age, sex, age^2, $age * sex$, $age^2 * sex$, smoking status, collection site, batch, and time difference between blood sampling and protein measurement. We also ran REGENIE on 250 randomly sampled plasma proteins, focusing on a subset of $N = 43{,}293$ individuals overlapping the British ancestry subgroup described in [3]. Association was performed in independent batches of 250 traits in parallel.

5.2.2 Quality control and preprocessing for GWAS

Our analyses made use of the following filtering criteria:

1. **Genotype QC:** for model fitting, we use markers with minor allele frequency $\geq 1\%$, Hardy-Weinberg equilibrium test $P > 1 \times 10^{-15}$, and genotyping rate $> 99\%$.

2. **Missing data:** we remove samples with phenotype missingness above 50% and mean-impute the remaining missing values. During the Bayesian regression

step, we also median-impute the genotype matrix to enable 2-bit genotype encoding. We allow missing data during the computation of test statistics.

3. **Mean-centering and transforming:** all analyses are performed on mean-centered and standardized genotypes. Quantitative phenotypes are mean centered, and quantile normalized to a normal distribution.

4. **Covariates:** we regress covariates (top 20 PCs, age, sex, age^2, age×sex, age^2×sex, smoking status) from both genotype and phenotype on the fly, which is equivalent to including covariates as fixed effects in the model. We also remove individuals with any missing covariates ("complete case analysis").

5.3 Performance in simulated data

5.3.1 Simulation design

To assess the robustness and power of Quickdraws and other methods, we performed simulations with varying levels of relatedness and population stratification. We used SNP array data from the UK Biobank, using all autosomal variants with minor allele frequency $\geq 0.01\%$, Hardy-Weinberg equilibrium $p \geq 1 \times 10^{-15}$, and a genotyping rate above 99%, obtaining $512,828$ variants. Of these, 54,568 had a minor allele frequency (MAF) between 10^{-4} and 10^{-2}. We either randomly sampled 50,000 individuals to assemble groups of samples matching specific relatedness and population structure criteria, or considered the entire subset of white British (N=$405k$) and self-identified Europeans (N=$460k$). In each setting, we simulated 50 quantitative and 50 binary traits with narrow-sense heritability $h_g^2 = 0.4$, a realistic MAF dependence ($\alpha = -0.3$ [252, 253]), and varying polygenicity and prevalence levels.

We also simulated phenotypes generated using different distributions of genetic effects, including spike-and-slab, mixture of Gaussian, Laplace, and Gaussian. To simulate effects following a spike-and-slab prior, given a polygenicity value p and M variants, we randomly selected Mp causal variants (out of $M = 512,828$ variants) and randomly sampled their effect sizes from a Gaussian distribution.

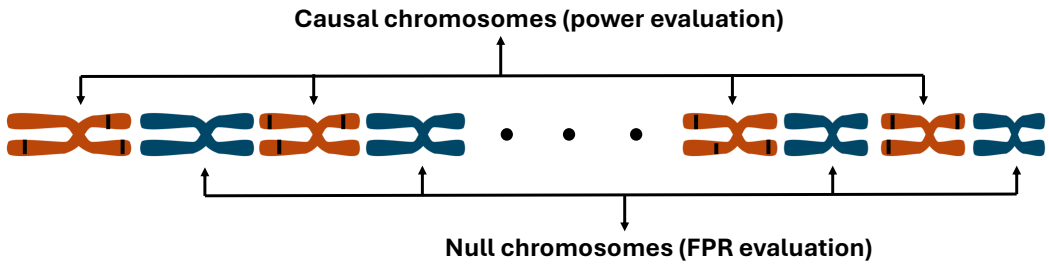


Figure 5.1: Phenotype simulation design. We simulate causal effects using a spike-and-slab, mixture of Gaussian, Laplace or Gaussian distribution only on odd chromosomes, whereas we treat even chromosomes as null controls to test for false positive rate and calibration of test statistics.

To facilitate the calculation of false positive rates and assess calibration, we only chose causal variants from the odd chromosomes, setting the effect sizes of variants on even chromosomes to zero. We simulated population stratification and shared environment between close relatives using

$$y_i = \sum_{j=1}^{Mp} G_{i,j}\beta_j + A\gamma + \tau + \epsilon,$$
$$\beta_j \sim \mathcal{N}(0, \frac{h_g^2 c_f}{Mp}) \quad \epsilon \sim \mathcal{N}(0, h_e^2), \tag{5.1}$$

where $\sum_{j=1}^{Mp} G_{i,j}\beta_j$ is a genetic effect explaining 40% of trait variance, $A\gamma$ is a population stratification term determined using self-reported ancestry, τ is the shared environmental effect, simulated by assigning a Gaussian draw to closely related individuals, and ϵ is the residual environmental effect, with variance equal to h_e^2. To simulate from a mixture-of-Gaussian prior, we added an extra parameter $f \in 0.05, 0.5$, which controls the fraction of variance explained by the Gaussian with smaller variance. For Laplace and Gaussian priors, we assumed polygenicity $p = 1$, and sampled the effects β_j in Equation 5.1 from either distribution.

In our simulations, we set the variance explained due to population stratification to be 5% of trait variance, whereas sample relatedness explained 10% of the trait variance in 2^{nd} degree relatives and 20% in 1^{st} degree relatives [108]. To simulate a realistic MAF-dependent trait architecture, we set $c_f \propto (2f(1-f))^{\alpha}$, where f is

the minor allele frequency and $\alpha = -0.3$ [252, 253]. For binary traits, we adopted a liability threshold model by synthesizing a quantitative phenotype as in Equation 5.1 and setting individuals above/below a given prevalence value as cases/controls. We varied the prevalence in binary traits from 0.3 to 0.001 and the amount of relatedness and population structure based on the following four simulation conditions:

- **Unrelated white British (GB-unrel):** We randomly sampled 50,000 individuals from the set of unrelated white British as defined in [3]. This scenario does not include population structure or relatedness.

- **Related white British (GB-rel):** We included $25,000$ individuals from the set of individuals in the related white British but not in the unrelated white British subset, along with $25,000$ unrelated white British individuals. This resulted in 552 first degree, 906 second degree, and 1396 third degree relative pairs (3.4× more relative pairs than randomly sampling from related white British subset). We also considered less and more extreme levels of relatedness, assembling sets of samples by randomly sampling from the related white British subset (referred to as GB-rel-ukb) and including high proportions of first and second degree relatives (referred as GB-rel+ having 7.3× and 4.8× more first and second degree pairs than randomly sampling from related white British subset).

- **European structure (EUR):** To simulate higher levels of population structure in the dataset, we included 25,000 non-British European samples [3] along with 25,000 unrelated British samples. This set also comprised 451 first-degree and 107 second-degree relative pairs, corresponding to 1.2× and 0.8× the amount of first- and second-degree pairs found in the subset of related white British individuals.

- **Pan-ancestry structure (PAS):** To simulate population structure comprising multiple ancestry descriptors found in the UK Biobank, we included 7,500 self-reported African samples, 7,500 self-reported South Asian samples, along

with 35,000 unrelated British samples, as defined in [3]. This set comprised 425 first-degree and 173 second-degree relative pairs, corresponding to 1.1× and 1.2× more first- and second-degree pairs than randomly sampling from the subset of related white British individuals. We simulated correlated variant effects across ancestry groups using a multivariate Gaussian, setting the covariance of effects between Africans and other subgroups to 0.4, and between the European and south-Asian subgroups to 0.7 [254].

5.3.2 False positive rate evaluation

We verified the calibration of Quickdraws and the other methods by considering varying levels of polygenicity, relatedness, population structure, and the prevalence of binary traits. We measured false positive rates (FPRs), calculated as the proportion of null variants (that is, variants on even chromosomes) below a chosen p-value threshold.

We first tested calibration in analyses involving quantitative traits, considering several simulation settings with population structure and relatedness and varying polygenicity of traits from 1% to 10%. In these experiments, linear regression was not robust to the presence of relatedness and population structure, as previously observed [105, 106, 108], while all other methods were reasonably calibrated, with a few exceptions (see Figure 5.2). Like other mixed-model methods, Quickdraws remained calibrated, not showing significant inflation in any of the simulation conditions we considered. REGENIE produced significantly inflated test statistics in data sets including high levels of relatedness (t-test $p < 1.5 \times 10^{-4}$, see GB-rel+ in Figure 5.2), which comprise 1,250 first-degree and 1,250 second-degree relative pairs (corresponding to ∼7.3× and ∼4.8× more relatives compared to the white British subset of UK Biobank samples). Finally, we observed higher variance for Quickdraws' FPR estimates in simulations involving non-homogeneous ancestry, due to higher noise in the estimated effective sample size, which relies on a smaller number of unrelated homogeneous individuals.

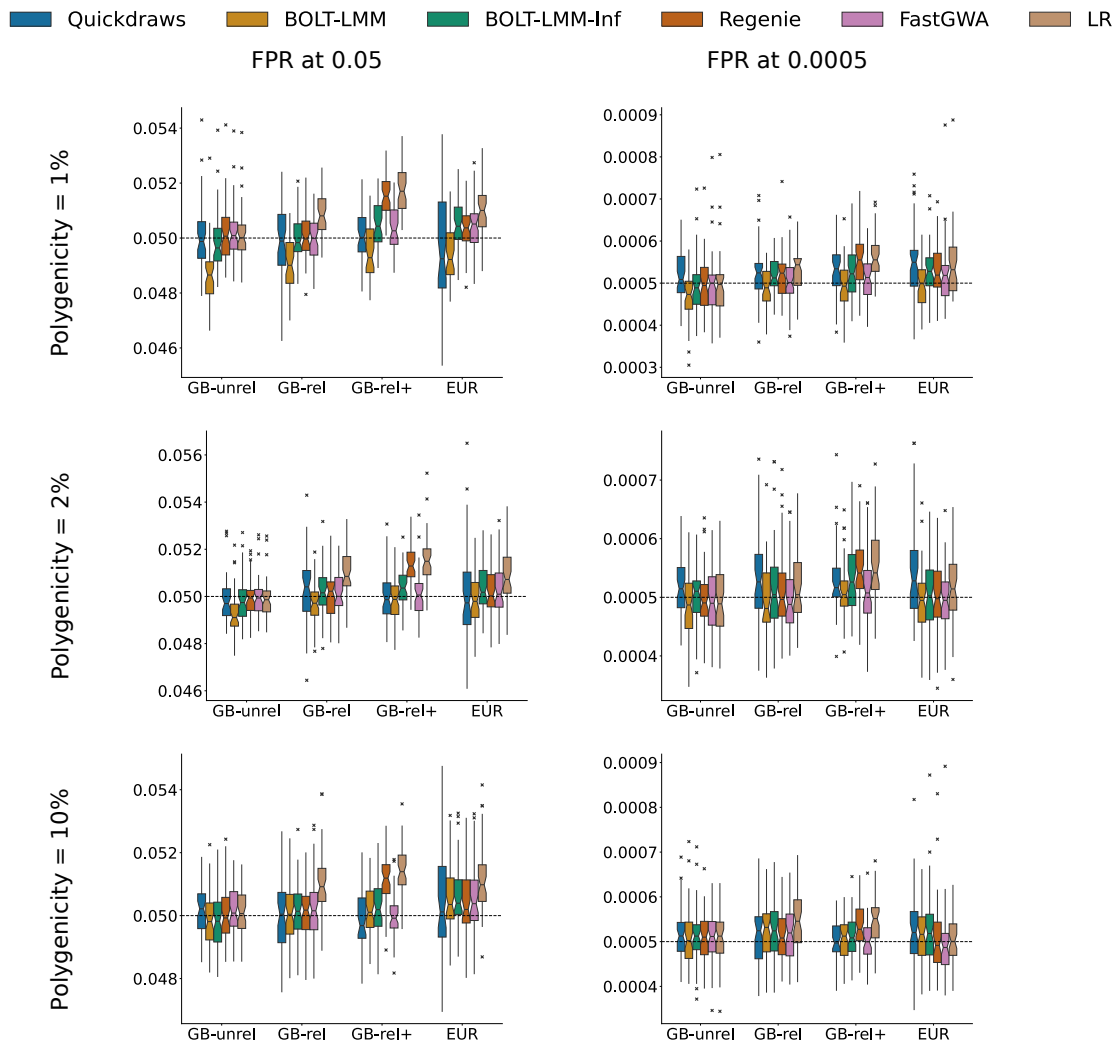


Figure 5.2: Summary of calibration in simulations for quantitative traits. False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the number of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, GB-rel+, EUR) is provided in Section 5.3.1.

We then assessed statistical robustness in the analysis of binary traits, where we varied the prevalence of the trait from 0.3 to 0.001 under varying levels of population structure and relatedness (see figures 5.3 and 5.4 for FPR of common and rare variants separately) while fixing the polygenicity of the traits to 2%. Binary trait association analysis often violates the normality assumptions made when defining the null distribution of association statistics in linear mixed models,

thereby resulting in high FPR, particularly for lower frequency variants or for traits with lower prevalence [119]. Consistent with this, BOLT-LMM and BOLT-LMM-Inf, which are not designed for the analysis of low-prevalence binary traits, were inflated across all simulated conditions (t-test $p < 3 \times 10^{-5}$) when the prevalence dropped to or below 0.1. Quickdraws produced controlled false positive rates for both common (MAF $> 1\%$) and rare (MAF $< 1\%$) variants in all the simulation settings we considered, which included population structure, relatedness, and low-prevalence binary traits (see figures 5.3 and 5.4 respectively). REGENIE with Firth logistic regression fallback was inflated for common variants in traits with high prevalence (prevalence $= 0.1$ and 0.3) for several population structure and relatedness settings we considered. When the prevalence of the trait was set to 0.01 (that is, 500 cases out of 50,000 samples) or below, both Quickdraws and REGENIE yielded deflated test statistics (t-test $p < 3 \times 10^{-5}$); in scenarios with such a low prevalence, however, all methods we considered lacked sufficient statistical power to detect association (χ^2 at causal variants $\approx$1). SAIGE did not converge for up to 28% of the traits when the prevalence dropped to 0.001, leading to significant deflation (t-test $p < 10^{-15}$).

Finally, we verified the calibration of false positive rates in various other simulated conditions, including stronger population structure with up to 30% non-European ancestry (see Figure B.27), varying levels of relatedness (see Figure B.28), and genetic architectures involving diverse causal effect size distributions (see Figure B.29). We also evaluated false positive rates in simulations involving all white British individuals (N $\approx 405k$) and all self-identified European individuals from the UK Biobank (N $\approx 460k$), varying levels of polygenicity in quantitative traits (1% to 10%), and varying levels of prevalence (0.3 to 0.001) in binary traits. We found Quickdraws to yield controlled false positive rates across varying significance thresholds, while fastGWA, REGENIE, and BOLT-LMM-Inf led to significant inflation in some cases (see figures 5.6, 5.7 and tables 5.1, B.4). This is likely the result of residual population stratification, which is not fully corrected using principal components in these simulated scenarios and, if present, may cause subtle inflation in real-data analyses [255–257].

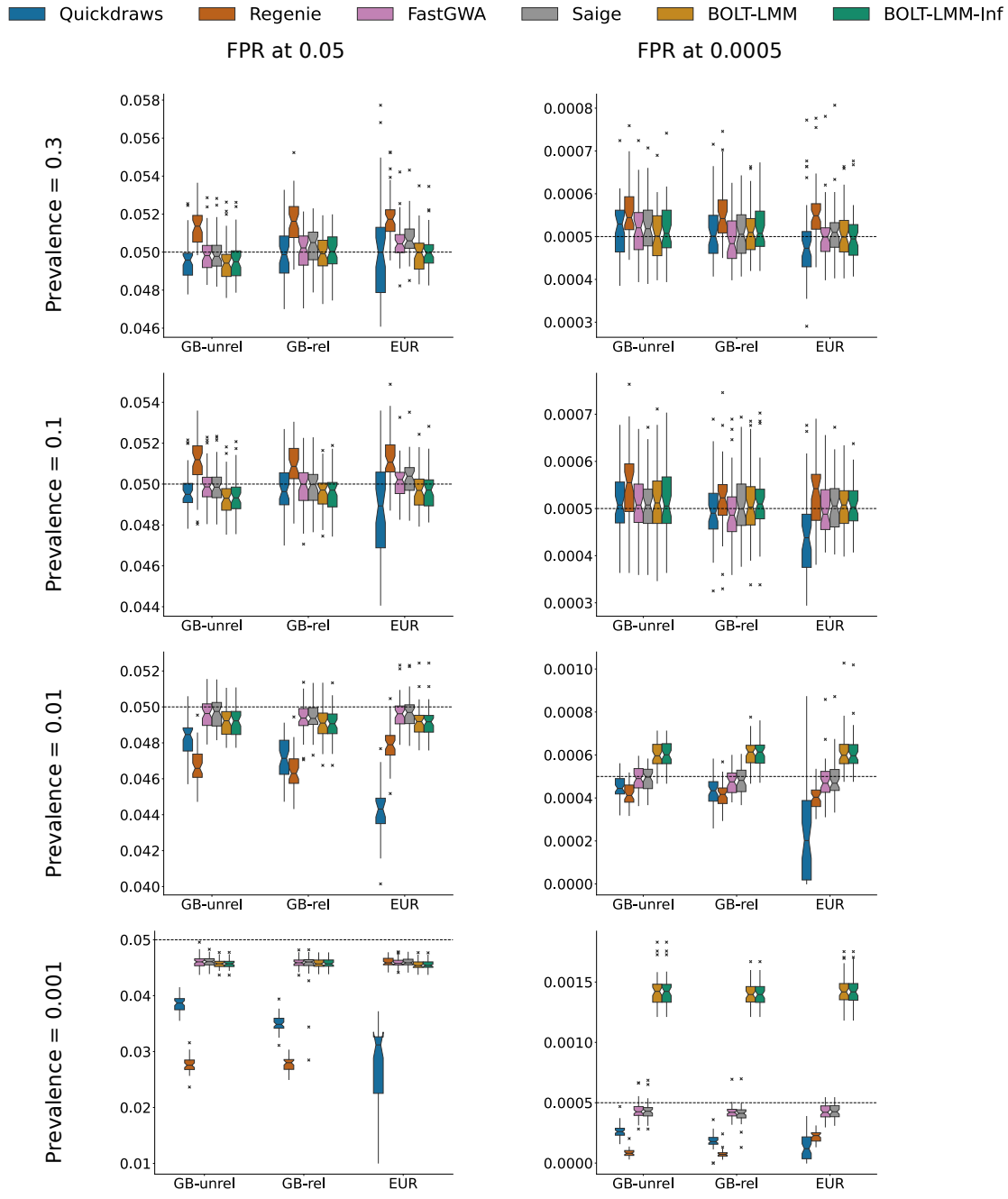


Figure 5.3: Summary of calibration in simulations for binary traits and common variants (MAF $\geq 1\%$). False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the number of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, EUR) is provided in Section 5.3.1.

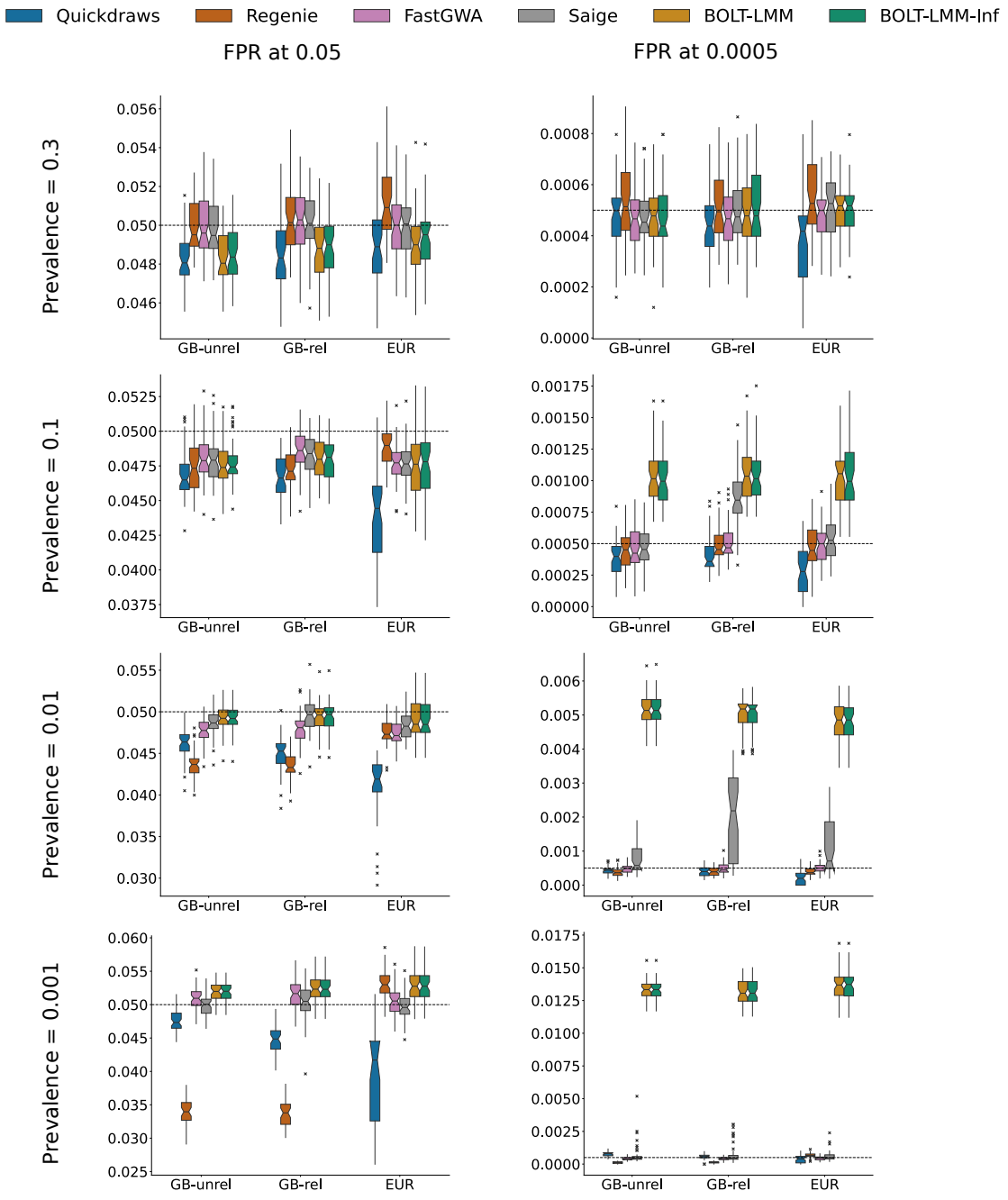


Figure 5.4: Summary of calibration in simulations for binary traits and rare variants (MAF $< 1\%$). False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the number of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, EUR) is provided in Section 5.3.1.

Table 5.1: Summary of calibration in simulations involving (a) $N = 405$k white British individuals and (b) $N = 460$k European individuals [3]. We report the false positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the fraction of variants on even chromosomes with p-value lower than α. We varied the polygenicity from 1% to 10% in quantitative traits. For binary traits, we changed the prevalence from 0.3 to 0.001, fixing the polygenicity to 2%. Mean FPR refers to the mean across 50 independent simulated traits, SE refers to standard errors, and p-values are computed for the significance of a deviation from the FPR threshold.

Simulation Type	FPR Threshold	Polygenicity (QT) / Prevalence (BT)	Mean FPR	SE FPR	T-test P
Quantitative Traits	0.05	1%	0.05034	0.00022	0.13382
		2%	0.05040	0.00029	0.17865
		10%	0.05036	0.00017	0.03658
	0.0005	1%	0.00052	0.00001	0.06363
		2%	0.00053	0.00002	0.04845
		10%	0.00052	0.00001	0.04787
Binary Traits	0.05	0.3	0.04923	0.00022	0.00103
		0.1	0.04986	0.00019	0.46758
		0.01	0.04940	0.00016	0.00044
		0.001	0.04648	0.00016	7.39×10^{-39}
	0.0005	0.3	0.00048	0.00001	0.05193
		0.1	0.00051	0.00001	0.57445
		0.01	0.00046	0.00001	0.00018
		0.001	0.00042	0.00001	5.28×10^{-13}

(a) $N = 405$k white British individuals.

Simulation Type	FPR Threshold	Polygenicity (QT) / Prevalence (BT)	Mean FPR	SE FPR	T-test P
Quantitative Traits	0.05	1%	0.05061	0.000231	0.0098
		2%	0.05019	0.00025	0.4444
		10%	0.05037	0.00026	0.1709
	0.0005	1%	0.00053	0.00001	0.0075
		2%	0.00053	0.00001	0.01728
		10%	0.00053	0.00001	0.0429
Binary Traits	0.05	0.3	0.04983	0.00015	0.3108
		0.1	0.04997	0.00017	0.9015
		0.01	0.0437	0.00058	5.19×10^{-18}
		0.001	0.03121	0.0022	3.72×10^{-13}
	0.0005	0.3	0.00051	0.00001	0.1391
		0.1	0.00053	0.00002	0.0436
		0.01	0.00033	0.00003	4.93×10^{-7}
		0.001	0.00022	0.00003	2.47×10^{-15}

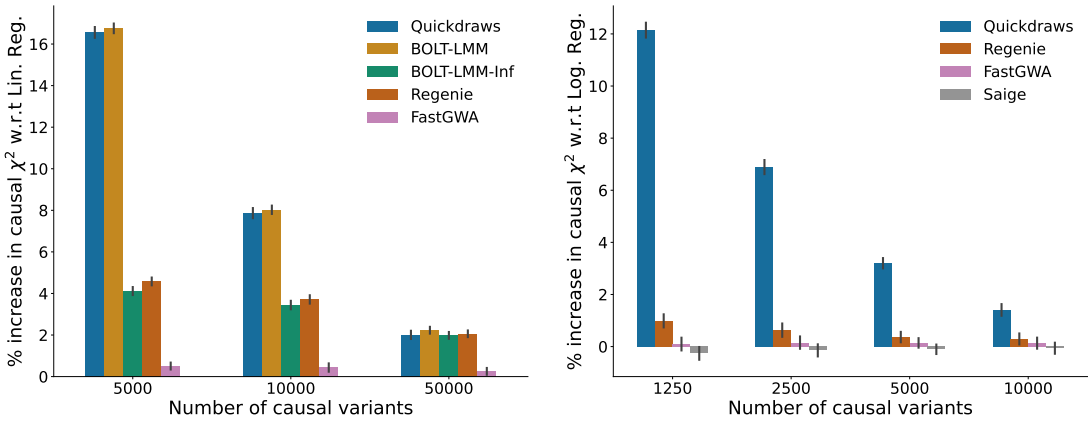
(b) $N = 460$k European individuals.

5.3.3 Statistical power evaluation

We now focus on evaluating the statistical power of Quickdraws in these simulations. We first tested the performance of Quickdraws in quantitative trait association and compared it with previous association testing methods including BOLT-LMM, REGENIE, FastGWA, and an implementation of linear regression provided in Plink [112, 113]. For BOLT-LMM, we also considered a faster implementation that uses infinitesimal priors (BOLT-LMM-Inf), which is more scalable but does not lead to higher association power in less polygenic genetic architectures [106, 107]. We measured statistical power by considering the average χ^2 test statistics at simulated causal variants, focusing on the set of unrelated white British samples and varying the number of causal variants in the simulation from 5,000 to 50,000 [252]. We normalized the average χ^2 test statistics at causal variants by the average χ^2 test statistics at null variants to remove any residual stratification. In these simulations, the modeling of trait polygenicity adopted by Quickdraws led to higher average χ^2 statistics (paired t-test $p < 10^{-20}$ for polygenicity up to 2%) compared to REGENIE, FastGWA, BOLT-LMM-Inf, and linear regression (see Figure 5.5a and Figure B.30 for more simulation conditions). The difference in association power compared to other infinitesimal methods was larger for traits with lower polygenicity (paired t-test $p < 10^{-53}$ for 5,000 causal variants). For instance, when the number of causal variants was reduced to 5,000, Quickdraws achieved $\geq 11.46\%$ higher average χ^2 compared to REGENIE and BOLT-LMM-Inf (paired t-test $p < 10^{-53}$), resulting in a similarly large gain in effective sample size [234].

Next, we applied Quickdraws to the analysis of binary traits, comparing its association power with previous binary trait association methods including SAIGE, REGENIE, and FastGWA-GLMM. We simulated phenotypes under a liability threshold model, where cases and controls are defined as individuals who are above or below a chosen liability threshold as described in Section 5.3.1. We fixed the prevalence at 0.3 and varied the number of causal variants in the simulation from 1,250 to 10,000 [252, 258, 259]. In these experiments, the use of a non-infinitesimal prior in Quickdraws led to higher statistical power compared to SAIGE, REGENIE,

and FastGWA (paired t-test $p < 10^{-12}$, Figure 5.5b and Figure B.31 for more simulation conditions). The difference in association power compared to other methods was again larger for traits with lower polygenicity (paired t-test $p < 10^{-43}$ for 1,250 and 2,500 causal variants). For instance, Quickdraws obtained 11.5% and 12.04% higher average χ^2 compared to REGENIE and FastGWA on the simulated causal variants for traits with a low polygenicity of 0.25%.



(a) **(b)**

Figure 5.5: Statistical power in simulated quantitative and binary traits for unrelated British samples. (a) % increase in average χ^2 test statistics at causal variants with respect to linear regression (Lin. Reg.) for quantitative traits, varying the number of simulated causal variants from 5,000 to 50,000 variants. (b) % increase in average χ^2 test-statistics at causal variants with respect to logistic regression (Log. Reg.) for binary traits, varying the number of simulated causal variants from 1,250 to 10,000 variants. Traits in panel (a) and (b) are simulated for 50,000 samples with $h^2 = 0.4$; the prevalence is fixed to 30% for binary traits in panel (b). Error bars represent the standard error in the percentage improvement, measured using 50 independent traits. The causal χ^2 is normalized by the mean χ^2 at null variants for each trait.

Finally, we tested Quickdraws in simulations where causal effects were not assumed to follow a spike-and-slab distribution, where a subset of variants contributes to the phenotype. Instead, we sampled effects from a Gaussian, a mixture of Gaussian, or a Laplace distribution. In the mixture of Gaussian setting, we again observed that Quickdraws and BOLT-LMM yielded higher power compared to other models. In simulations involving fully infinitesimal traits with Laplace and Gaussian effects, Quickdraws and BOLT-LMM produced results similar to other approaches that assume a fully infinitesimal trait architecture (see Figure B.32). We

also evaluated power in simulations involving all white British individuals from the UK Biobank ($N \approx 405$k) and all self-identified European individuals from the UK Biobank ($N \approx 460$k). In these settings, we found Quickdraws to provide the highest association power, with higher power than BOLT-LMM when the polygenicity of the trait was low (see figures 5.6 and 5.7).

Overall, these simulations demonstrate that the use of a non-infinitesimal spike-and-slab prior results in higher statistical power to detect association in both quantitative and binary trait simulations. Quickdraws matched or outperformed the power of BOLT-LMM in the analysis of quantitative traits and obtained higher association power than existing GWAS algorithms in binary traits. Quickdraws also yielded controlled FPR in all simulation settings we considered, which included population structure, relatedness, and low-prevalence binary traits.

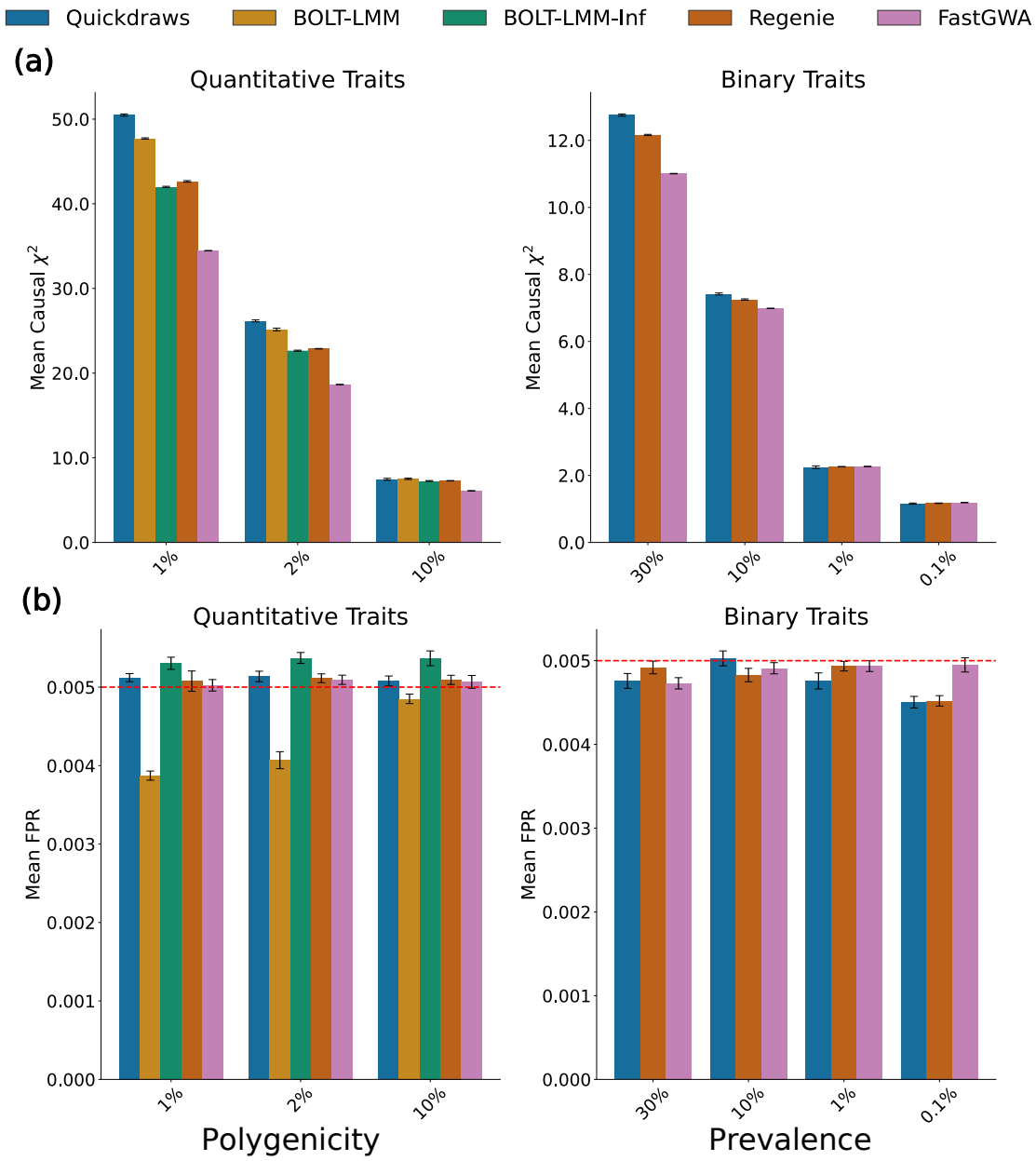


Figure 5.6: Power and FPR in N= 405k simulation. (a) The mean χ^2 at causal variants across different methods and polygenicities for Quantitative traits, and prevalence for binary traits. The χ^2 at causal variants is not normalized by the χ^2 at null variants in this plot. (b) The mean false positive rates (FPR) at 0.005 calculated at null variants across different methods and polygenicities for Quantitative traits, and prevalence for binary traits. The simulation was performed using the N= 405k set of white British individuals. The error bars represent 95% confidence intervals. The red dashed line corresponds to false positive rate = 0.005.

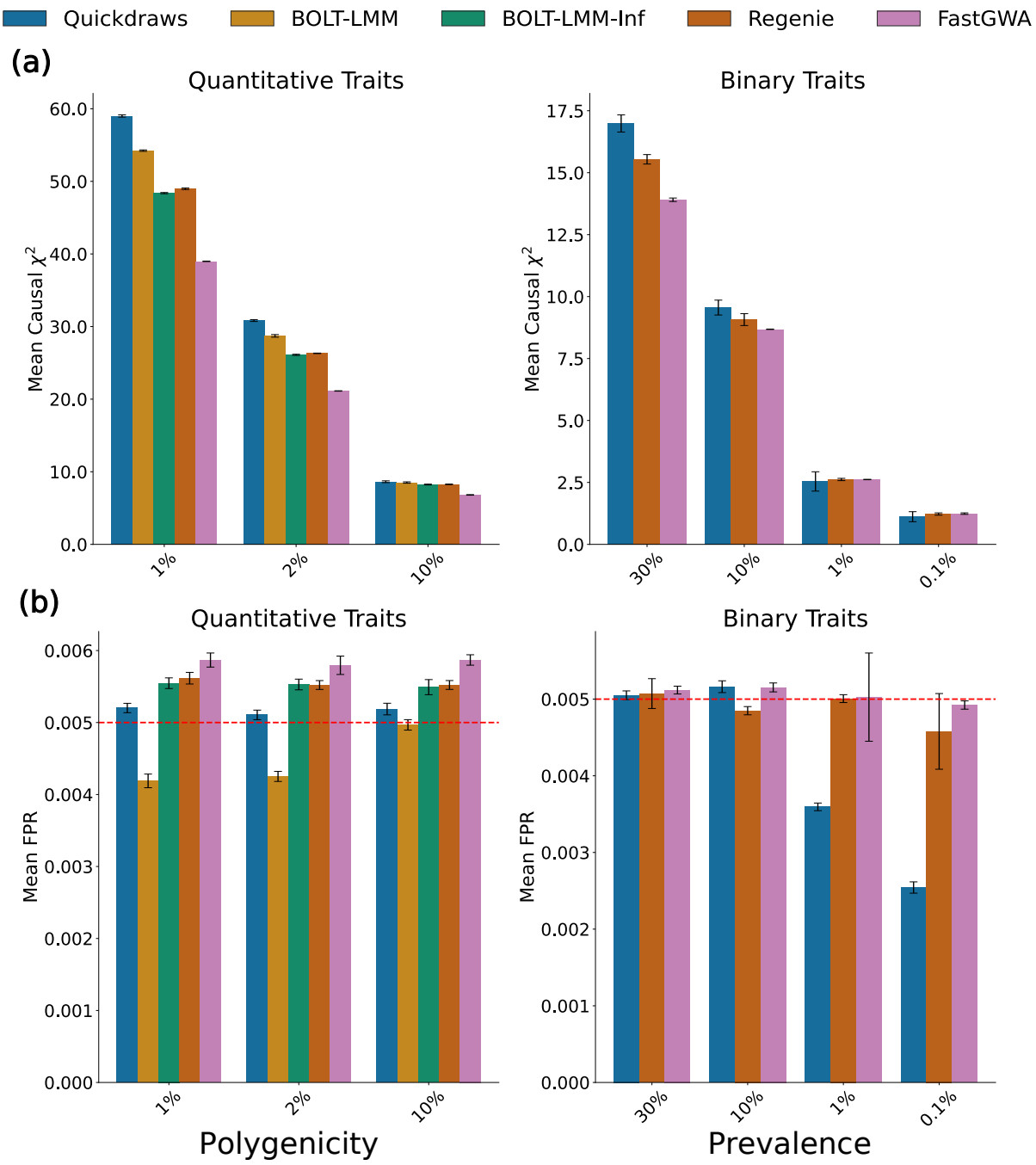


Figure 5.7: Power and FPR in N= 460k simulation. (a) The mean χ^2 at causal variants across different methods and polygenicities for quantitative traits, and prevalence for binary traits. The χ^2 at causal variants is not normalized by the χ^2 at null variants in this plot. (b) The mean false positive rates (FPR) at 0.005 calculated at null variants across different methods and polygenicities for Quantitative traits, and prevalence for binary traits. The simulation was performed using the N= 460k set of UK Biobank self-identified Europeans [3]. The error bars represent 95% confidence intervals. The red dashed line corresponds to false positive rate = 0.005.

5.4 UK Biobank analysis

We applied Quickdraws to the analysis of 79 quantitative traits (comprising of blood-related, anthropometric, and other traits) and 50 self-reported disease traits for ∼405,000 white British individuals from the UK Biobank dataset (see tables B.5 and B.6 for the list of traits), which we selected based on their high phenotyping rate and the significance of their estimated heritability as described in Section 5.2. We applied standard quality-control filtering steps, retaining 458,420 genotyped markers to be used for model-fitting and ∼13.3 million imputed variants for association testing. We compared the test statistics we obtained using Quickdraws to those obtained by other association methods, including SAIGE, FastGWA, REGENIE, and BOLT-LMM.

5.4.1 Analysis of false positive rates

We assessed the calibration of the test-statistics in real data by comparing the LD-score regression attenuation ratio estimates [107] obtained using Quickdraws to those obtained using linear regression while restricting to the set of unrelated individuals with homogeneous British ancestry [3]. Because linear regression is expected to be calibrated in this sample set, similar attenuation ratio estimates for Quickdraws provide evidence for its calibration [107, 108]. We performed LD-score regression using the baseline-LD model [260] and estimated attenuation ratios using test statistics generated using Quickdraws on all ∼13.3 million testing variants across the set of 79 quantitative traits. Quickdraws produced attenuation ratios close to those of linear regression on unrelated samples (mean attenuation ratio for Quickdraws = 0.0832, s.e. = 0.008; mean attenuation ratio for linear regression on unrelated samples = 0.0892, s.e. = 0.008; Figure 5.8). Quickdraws also remained calibrated in the analysis of low-prevalence binary traits, where it did not produce signatures of false positive association signals that are present in methods that do not perform adjustments of test statistics such as saddle-point approximation (SPA) [119] or approximate Firth logistic regression [123] (see Manhattan plots in figures B.33 and B.34, additional examples of Manhattan plots for quantitative trait are shown in figures B.35 and B.36).

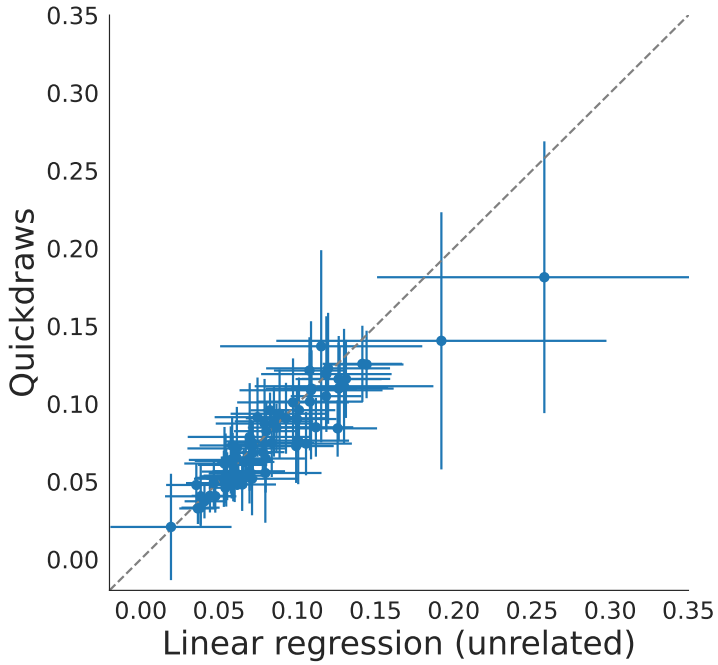


Figure 5.8: Calibration in real data analysis. Attenuation ratio of Quickdraws (N=405k) vs. linear regression in unrelated samples (N=337k); vertical and horizontal lines represent mean ± standard errors in the attenuation ratio estimate for each method.

To further validate signals found using Quickdraws but not found using REGENIE, we evaluated the functional annotation of regions where these variants are found. We used the Baseline-LD model annotations [260, 261] to assess the distribution of functional annotations for sets of variants associated using different methods. We define the functional enrichment for a set of variants as the proportion of variants being in a particular functional category, divided by the genome-wide fraction of variants assigned to that category. We considered binary and quantitative traits separately and estimated the mean functional enrichment across phenotypes by computing the ratio between the number of variants from a given set that belong to a functional category (numerator) and the total number of variants in the set (denominator). We estimated this ratio by separately summing the numerator and denominator term across all analyzed phenotypes. We estimated standard errors around the enrichment by applying the jackknife method to 50 equally sized genome blocks. We looked at functional enrichment for various sets of variants, including (1) variants found in Quickdraws but not in REGENIE, (2) variants found

in both Quickdraws and REGENIE. For (2), we matched the χ^2 distribution of the two considered sets, sampling variants detected using REGENIE so that they approximately match the empirical χ^2 distribution of the variants detected using Quickdraws. We found similar enrichments compared to variants found using both REGENIE and Quickdraws (see figures 5.9 and 5.10), suggesting no substantial differences in the functional profile of variants exclusively detected using Quickdraws.

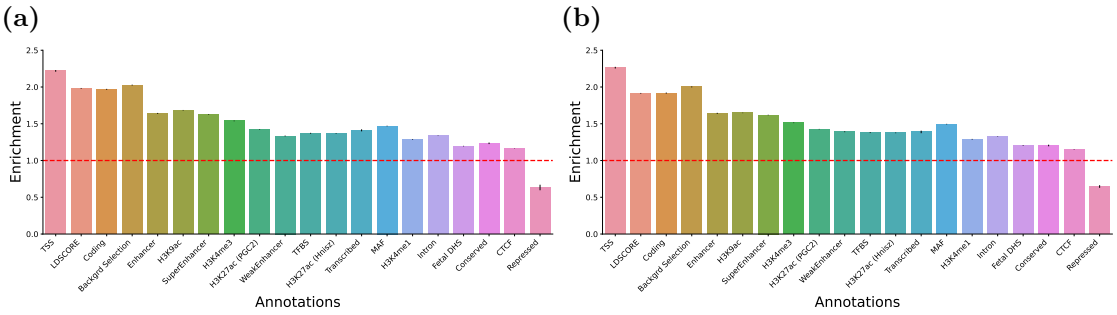


Figure 5.9: Functional enrichment at associated variants for quantitative traits. Functional enrichment profile of variants associated across 79 quantitative traits. We considered either (a) the set of variants significantly associated ($p \leq 5 \times 10^{-8}$) using both Quickdraws and REGENIE, matching the p-value distribution (see Methods), or (b) the set of variants associated using Quickdraws but not REGENIE. Error bars represent jackknife standard errors; the red dashed line corresponds to no enrichment.

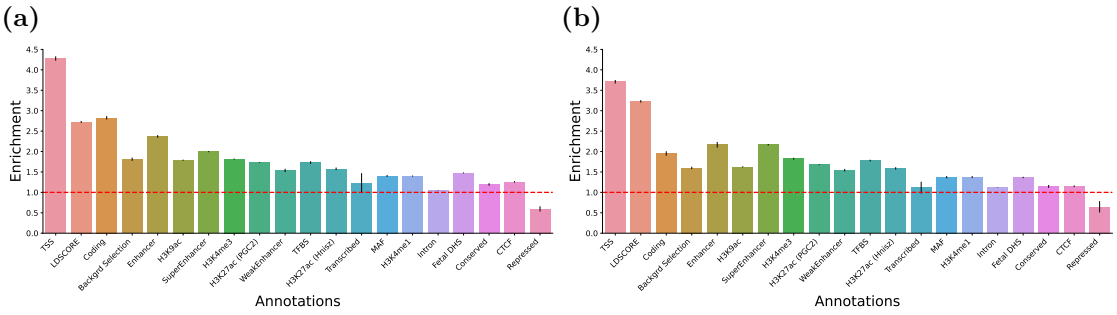


Figure 5.10: Functional enrichment at associated variants for binary traits. Functional enrichment profile of variants associated across 50 binary traits. We considered either (a) the set of variants significantly associated ($p \leq 5 \times 10^{-8}$) using both Quickdraws and REGENIE, matching the p-value distribution (see Methods), or (b) the set of variants associated using Quickdraws but not REGENIE. Error bars represent jackknife standard errors; the red dashed line corresponds to no enrichment.

5.4.2 Number of associations

We assessed the number of approximately independent associations detected using each method, performing variant clumping with strict filtering criteria (p-value threshold = 5×10^{-9}, R^2 threshold = 0.01). A summary of these results for Quickdraws, REGENIE, and FastGWA, the most scalable methods we tested, is shown in Figure 5.11a; additional data is reported in tables B.7 and B.8. The gains we observed in the number of detected associations were consistent with the increase in power observed in simulations: Quickdraws produced significantly more independent associations compared to REGENIE and FastGWA for quantitative and disease traits (binomial test $p < 1.9 \times 10^{-3}$) and a similar number of independent associations compared to BOLT-LMM for quantitative traits (total number of associations, Quickdraws = 26,236, BOLT-LMM = 26,368). In more detail, for quantitative traits, Quickdraws found 4.97% more independent associations compared to REGENIE and 22.71% more compared to FastGWA; for disease traits, Quickdraws found on average 3.25% more independent associations compared to REGENIE and 7.07% compared to FastGWA. The gain was higher in traits with high heritability or low polygenicity such as mean platelet volume (8.04% increase over REGENIE) and standing height (26.1% increase over REGENIE) [262].

We also analyzed 250 randomly sampled plasma-protein traits ($N \approx 43$k, see Methods), which have a less polygenic architecture than the other quantitative traits we considered [250]. In these analyses, Quickdraws found significantly more independent association signals compared to REGENIE (5.54% more loci, $p = 6.6 \times 10^{-3}$, see Figure 5.11b). Finally, we estimated the effective sample size, calculated as the average $\chi^2 - 1$ association statistic at variants that were found to be significantly associated ($p < 5 \times 10^{-8}$) using linear regression model applied to unrelated homogeneous subset of the data [234]. We found the effective sample size estimated for Quickdraws to be 14.7% higher ($p = 7.6 \times 10^{-4}$) than FastGWA and 3.4% ($p = 0.197$) higher than REGENIE, while being similar to that of BOLT-LMM ($p = 0.46$) when applied to quantitative traits (see Figure 5.11c-d).

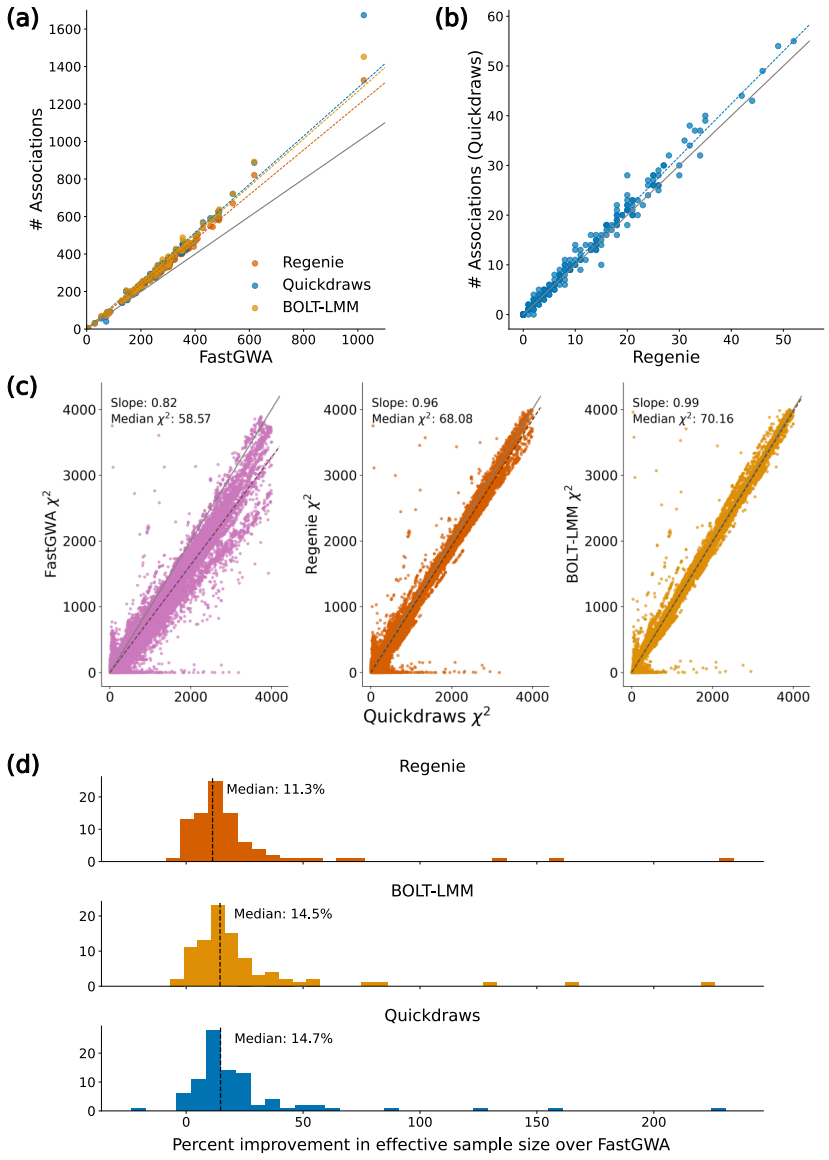


Figure 5.11: Approximately independent loci and effective sample size in UK Biobank analysis. (a) Number of approximately independent loci after plink clumping in FastGWA (x-axis) vs. REGENIE, BOLT-LMM, and Quickdraws for 79 quantitative traits. (b) Number of approximately independent loci after Plink clumping in REGENIE (x-axis) vs. Quickdraws for 250 randomly sampled plasma protein traits. (c) The χ^2 for 79 quantitative traits and N= 405k UK Biobank set conditioned on genome-wide significance ($p < 5 \times 10^{-8}$) in linear regression run on unrelated subset of the data (N= 337k). Median χ^2: FastGWA = 58.57, REGENIE = 68.08, BOLT-LMM = 70.16, Quickdraws = 70.64. (d) Histogram of the effective-sample size increase compared to FastGWA for 79 quantitative traits, measured as the mean χ^2 minus 1 at genome-wide significant variants inferred though linear regression run on unrelated subset of the data (N= 337k). For (a-c) The gray line represents the y=x line and the dashed lines represent a linear regression fit for each method, for (d) the dashed line represents the median improvement in effective sample-size across traits. In (c), the slope refers to the linear regression slope of the χ^2 values between each pair of methods.

5.4.3 Replication analysis with Biobank Japan, Finngen and other large-scale GWAS

To validate the additional associations detected using Quickdraws, we performed a replication analysis using GWAS summary statistics from the Biobank Japan data set [124], which used BOLT-LMM-Inf and SAIGE to detect associations, FinnGen, [128] which used REGENIE, and other trait-specific large-scale studies [263–267] We assessed the number of variants detected using different GWAS methods that can be replicated in the Biobank Japan, FinnGen and other large-scale studies, using downloaded summary association statistics. We considered 30 quantitative traits and 13 self-reported disease traits from Biobank Japan, as well as 21 self-reported disease traits from FinnGen for which equivalent definitions were available in both our analyses and the replication data set, and for which at least one significant association was detected in both studies. We also considered 5 disease traits (Crohn's disease [263], Type 2 Diabetes [264], Celiac disease [265], Depression [266] and Ulcerative Colitis [267]) that have publicly available summary statistics calculated on higher number of cases than in the UK Biobank. We considered variants with MAF $\geq$ 0.1%, INFO score $\geq$ 0.8, and appearing in both UK Biobank and the replicating dataset, leading to $\sim$5.89 million variants available for Biobank Japan replication, and $\sim$10.89 million variants available for FinnGen replication.

We separately considered the replication of associated variants and loci. For the replication of individual variants, we assessed the number (and proportion) of variants detected ($P_{dis} \leq 5 \times 10^{-9}$) in the UK Biobank that are also significant in Biobank Japan, using multiple replication thresholds, $P_{rep} \leq 5 \times 10^{-2}$, $P_{rep} \leq 5 \times 10^{-4}$, and $P_{rep} \leq 5 \times 10^{-6}$. Where, P_{dis} and P_{rep} are the significance thresholds for discovery and replication. For the replication of loci, we followed a procedure similar to that adopted in [268], defining a credible set for a locus to contain a lead (or sentinel) associated variant together with additional proxy variants found within a 50 kb window from the lead variant and with association significance $P \leq 100 \times P_{sentinel}$. We defined a locus as replicated if any variant in the credible set was also found to be associated (at $P \leq 5 \times 10^{-2}$) with the same direction of effect in the replication

cohort. First, we calculated the average replication rate for each trait using all GWAS methods (FastGWA, REGENIE, and Quickdraws) with a discovery threshold of 5×10^{-9}. Then, we adjusted the discovery threshold for each method, varying it from 2×10^{-9} to 8×10^{-9}, to achieve the same replication rate for each method.

Across the 53 quantitative and disease traits available in Biobank Japan and Finngen (which comprise 40 independent traits based on squared phenotypic correlation ≤ 0.1), Quickdraws yielded a significantly higher number of replicated loci compared to REGENIE (binomial test $p = 0.014$) and FastGWA (binomial test $p = 7 \times 10^{-4}$). For the set of 30 quantitative traits present in both studies, we observed a 2.5% and 15.72% increase over REGENIE and FastGWA, respectively, and a similar number of replicated loci compared to BOLT-LMM (total number of replicated loci, Quickdraws = 37,210, BOLT-LMM = 37,072). For the smaller set of 23 overlapping disease traits we analyzed, Quickdraws obtained 1.07% and 3.38% more replicated loci compared to REGENIE and FastGWA, respectively. We observed a similar increase in the number of replicated loci for 5 disease traits where we had access to meta-analyzed summary statistics from large-scale studies (see Table B.9). We verified that this increase in the number of replicated loci did not correspond to a decrease in replication rates across these methods (see Table B.10 and Figure 5.12d-e). A Venn diagram depicting the relationships between sets of variants replicated using different methods is shown in Figure 5.12b-c. Note that these results may be affected by the choice of GWAS method used in the replication cohort.

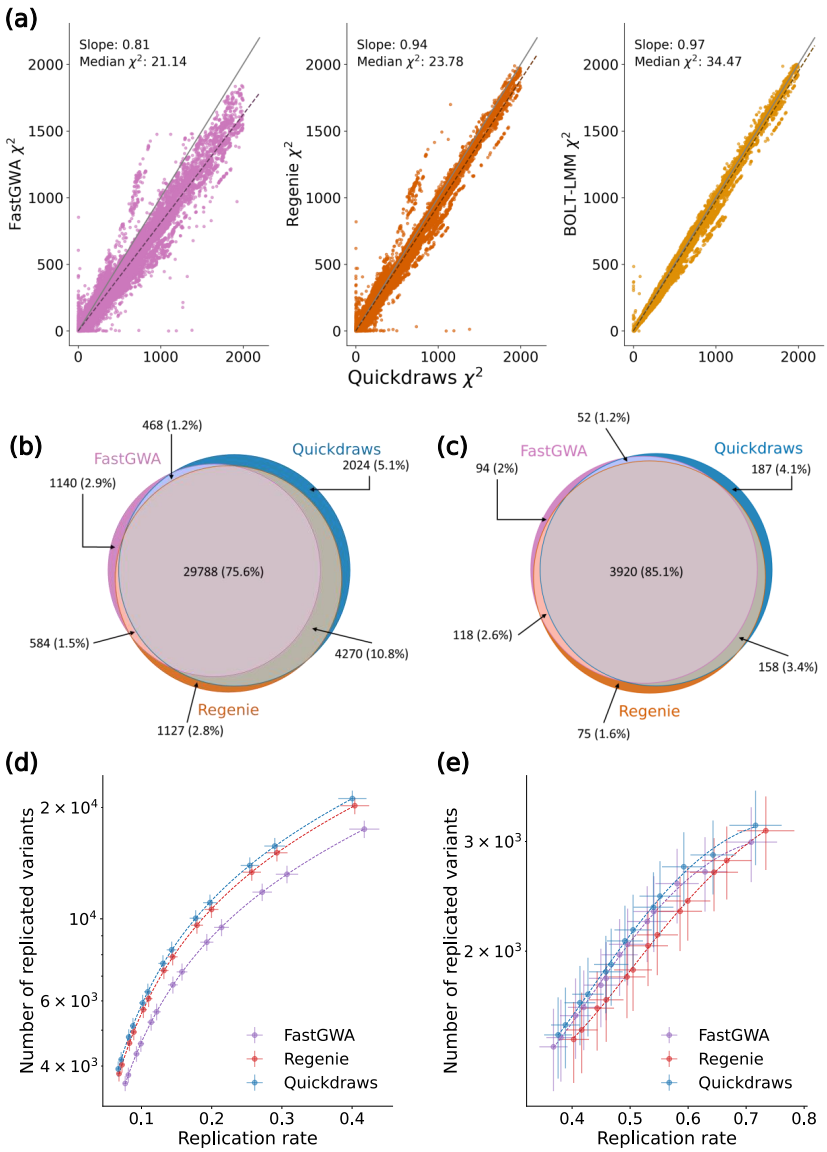


Figure 5.12: Replication analysis in Biobank Japan and Finngen. (a) The χ^2 in UK Biobank conditioned on genome-wide significance ($p < 5 \times 10^{-8}$) in replicating biobank across GWAS methods. The variants are aggregated across traits, quantitative and binary traits for REGENIE and FastGWA, and only quantitative traits for BOLT-LMM. The gray line represents the y=x line, and the dashed line the linear regression fit without intercept. Median χ^2 for quantitative traits: FastGWA = 28.36, REGENIE = 33.44, BOLT-LMM = 34.47, Quickdraws = 36.32; binary traits: FastGWA = 14.08, REGENIE = 14.34, Quickdraws = 14.51. (b-c) Replication venn diagram for (b) 30 overlapping quantitative traits in Biobank Japan and (b) 23 overlapping binary traits in Biobank Japan and Finngen. (d-e) Number of replicated variants vs replication rate for (d) quantitative and (e) binary traits. The discovery threshold was fixed to 5×10^{-9} and the replication threshold was varied from 5×10^{-2} to 5×10^{-8}. The error bars represent standard errors calculated using block jack-knife across chromosomes. The dashed line represents a cubic spline fit to the datapoints. In (a), the slope represents the linear regression slope of the χ^2 values between methods, and in (b-c), the percentages indicate the proportion of associations found with respect to the union of all methods.

5.4.4 Predictive power in out-of-sample data

To verify that these power gains are due to the modeling of non-infinitesimal trait architectures, we measured the accuracy of the polygenic predictions computed by Quickdraws, BOLT-LMM, and BOLT-LMM-Inf during the model fitting step [106, 107]. To this end, we used predictors from step 1 trained on the ∼405,000 white British subset to perform trait prediction for the remaining individuals. Across 79 quantitative traits, Quickdraws and BOLT-LMM obtained a mean correlation between true and predicted trait in the held-out European samples of 0.307 (s.e = 0.0061) and 0.313 (s.e = 0.0061), respectively, whereas BOLT-LMM-Inf yielded a lower correlation of 0.271 (s.e = 0.0061) (see Figure 5.13a). Similar improvements due to the modeling of non-infinitesimal trait architectures have also been observed in the context of polygenic prediction [106, 107, 269–271].

Comparison with PGS methods

We also compared the predictions from step 1 of Quickdraws with polygenic scores (PGS) built using recent methods [272, 273]. To this end, we constructed PGS for all the 79 quantitative traits analyzed, using either pruning and thresholding (P+T) as implemented in PRSice [272] or the PRS-CS method [273]. We used summary statistics generated using various methods on the set of ∼13.3 million imputed variants to generate PGS for held-out individuals from self-identified European, South Asian, East Asian, and African subgroups of the UK Biobank. We report the mean predictive R^2 and the 95% confidence interval across all analyzed traits, obtained using a meta-analysis of Fisher-transformed estimated correlation coefficients.

We found Quickdraws' step 1 predictors to be significantly more accurate (paired t-test $p < 5.6 \times 10^{-6}$ compared to PRS-CS and P+T) than summary-statistics derived PGS methods in the European held-out set, while achieving similar accuracy in other groups (see Figure 5.13b). In addition, we compared the accuracy of PGS estimates built based on summary association statistics obtained using different association algorithms. Consistent with a recent analysis that did not observe higher

accuracy for PGS built from summary statistics derived from non-infinitesimal modeling [274], we did not observe significant differences between using Quickdraws, BOLT-LMM, or REGENIE, but found PGS built using summary statistics from FastGWA to be significantly less predictive (paired t-test $p < 2.2\times10^{-4}$ for both PRS-CS and P+T in European and South-Asian held-out samples, see Figure 5.13c-d).

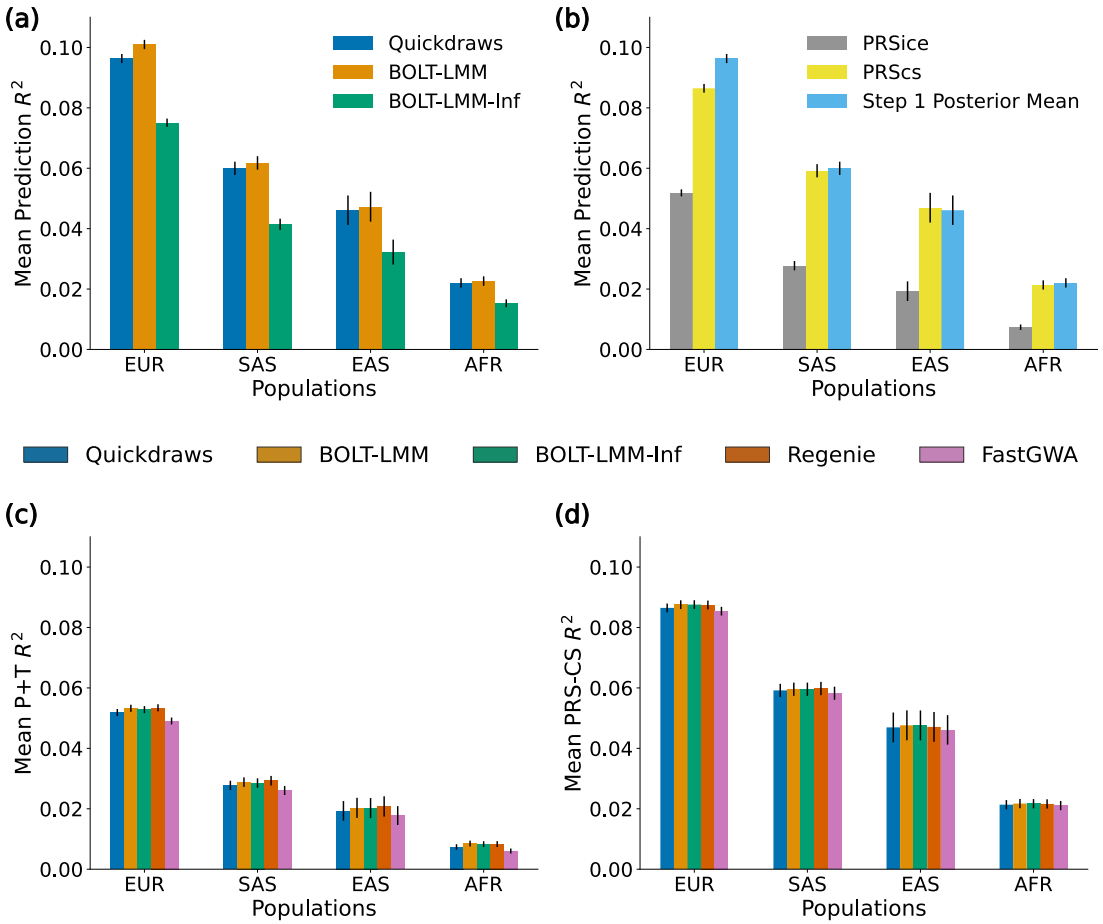


Figure 5.13: Phenotype prediction analyses in the UK Biobank. (a) Held-out mean phenotype prediction R^2 comparing step 1 posterior estimates from Quickdraws, BOLT-LMM and BOLT-LMM-Inf. (b) Comparing Quickdraws' step 1 posterior estimates with PGS calculated using Quickdraws' association statistics and P+T (pruning and thresholding as implemented in PRSice) or PRS-CS. (c-d) Comparing predictive power for PGS calculated using association statistics from different GWAS methods and different PGS methods, (c) P+T as implemented in PRSice and (d) PRS-CS. All analyses was performed on 27,683 held-out non-British Europeans, 9,044 self-identified south Asians, 1,457 self identified east Asians and 7,204 self identified African of African American samples. Results are aggregated across the 79 quantitative traits we analyzed, and the error bars represent 95% confidence interval of the mean prediction R^2 for each method in each population subgroup.

Comparison with PGS adjusted GWAS

This analysis was performed by Dr. Georgios Kalantzis, who is one of the co-authors on the paper. We briefly experimented with using polygenic scores as covariates when performing association, a strategy recently shown to effectively increase association power [275–277]. For this analysis, we focus on chromosomes 1 to 5, using up to 405,088 samples from the white British subgroup, and analyzing all 79 quantitative traits. We first generate leave-one-chromosome-out (LOCO) polygenic scores, as described in [275], using PRS-CS and summary statistics computed with FastGWA or REGENIE. We then used FastGWA and REGENIE to test for association again, this time including the LOCO PGS corresponding to the chromosome of the variant being tested as a covariate, during both the model-fitting and association steps.

Consistent with recent results [275, 277], we found that this approach increased the effective sample size for FastGWA, which relies on approximate sparse GRMs. While showing no improvement in power for REGENIE, which already uses genome-wide regression to account for polygenic effects. The effective sample size and total number of genome-wide significant hits, however, remained lower than that achieved using Quickdraws or BOLT-LMM (see Figure 5.14).

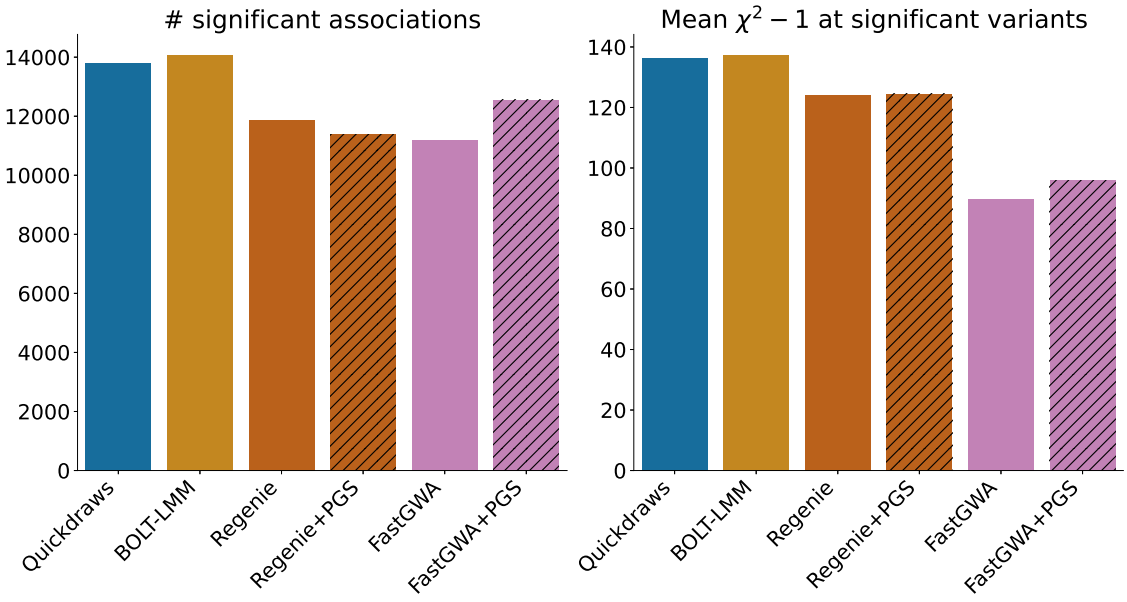


Figure 5.14: Effects of PGS-adjustment on association power. (a) Total number of independent loci found on chromosomes 1-5, using different GWAS methods with and without PGS adjustment in 79 UK Biobank quantitative traits. (b) Mean χ^2 - 1 at significant variants with and without PGS adjustment for the same traits and regions. Significant variants were defined as those with LR-unrel p-value $\leq 5 \times 10^{-8}$. FastGWA with PGS adjustment yielded an increase in the number of independent loci (with PGS adjustment, 159.1; without, 141.6; paired t-test $p = 1.9 \times 10^{-9}$) and higher mean χ^2-1 at significant variants (with PGS adjustment, 96.0; without, 89.8; $p = 3.3 \times 10^{-11}$). REGENIE with PGS adjustment yielded fewer independent loci (with PGS adjustment, 144.1; without, 150.1; $p = 7.4 \times 10^{-6}$), and similar mean χ^2 - 1 values (with PGS adjustment, 124.6; without, 124.1; $p > 0.05$).

5.4.5 Correction of participation bias

Quickdraws implements functionality to optionally correct for participation bias when computing association statistics in step 2, using an approach closely related to the one proposed in [278]. This approach relies on the use of sampling weights, which are computed by considering a range of covariates and are used to up-weight or down-weight individuals estimated to be under-sampled or over-sampled in the study [278]. Quickdraws takes these weights as input and performs a weighted linear regression, using a Huber-White estimator to compute the variance of the fixed effects in the regression. The test statistics are recalibrated using a weighted linear regression, by matching the effective sample size estimated from a weighted linear regression on an unrelated and homogeneous sample, similar to the unweighted case.

To apply this approach in the analysis of Table B.11, we analyzed 79 quantitative traits and 338,738 individuals from the white British subgroup for which all the covariates used in [278] to estimate sampling weights were available. These covariates included sex, age, education age, alcohol consumption frequency, smoking status, income, household size, employment status, body mass index, overall health, height, urbanization, weight, assessment center, self-reported ethnicity, and years of education. We used the LASSO regression model provided in [278] to estimate these sampling weights and compared the number of approximately independent loci identified using LDAK with a similar participation bias adjustment but applied on unrelated homogeneous subset of the data.

5.5 Cost analysis on UK Biobank RAP

We compared the computational efficiency of Quickdraws in large-scale UK Biobank analyses to that of REGENIE, FastGWA, SAIGE, and BOLT-LMM, testing the speed and cost of running each method on the UK Biobank's RAP cloud platform. We tested ∼13.3 million imputed genotypes for association with the 50 quantitative and 50 binary traits varying the number of samples between $N = 50{,}000$, $N = 405{,}088$ and artificially simulated $N = 1{,}000k$. We used 458,464 markers for model

fitting for Quickdraws, REGENIE, and BOLT-LMM and an LD-pruned set of 89,177 markers for SAIGE, and used precomputed sparse GRM matrices for FastGWA and SAIGE. Unlike current approaches, Quickdraws allows leveraging GPU hardware to speed up the Bayesian regression in the model fitting step. To provide a more detailed overview of computational costs and to optimize the hardware configuration used for each approach, we tested all methods on up to four different RAP machines.

For $N = 50k$, we consider `mem3_ssd1_v2_x4`, `mem2_ssd1_v2_x8`, and `mem1_ssd1_v2_x16`; for $N = 405k$, we consider `mem2_ssd1_v2_x8`, `mem2_ssd1_v2_x16`, `mem1_ssd1_v2_x36`, and `mem1_ssd1_v2_x72` ; for $N = 1{,}000k$, we only consider `mem1_ssd1_v2_x36`. Additionally, we use `mem2_ssd2_gpu1_x8` to run step 1 of Quickdraws on an instance providing a 24GB Nvidia A10G GPU. All the jobs on RAP were run using "on-demand" priority. Each method was run using configurations that would optimize performance, where possible, example, by allowing multi-threading and supplying files using the most efficient file formats, and all methods were given the same file containing testing variants as input.

BOLT-LMM, FastGWA and SAIGE do not leverage multi-trait parallelization, so we run them for 5 phenotypes and extrapolate the results for 50 phenotypes. We also only run BOLT-LMM for one of the three RAP instances, as the large difference in running costs would not to be significantly reduced using other types of hardware. For binary traits and $N = 1{,}000k$, which are more computationally intensive, we run step 2 for Quickdraws, REGENIE, and SAIGE using a subset of $\sim$13.3 million testing variants, and linearly extrapolate the results.

Additionally, Quickdraws allows running the model-fitting step in either high-memory or low-memory mode. The high-memory mode allows for loading the genotype matrix in memory for faster access. The low-memory mode uses a multi-threaded approach to stream the genotype data from disk, decreasing the total memory footprint, while being marginally slower than the high-memory option (20% slower on the RAP instance we tested). For the analysis in table 5.2, we report the minimal cost across the RAP machines we considered for each method, along with running time and total memory usage computed on a common machine,

using the high-memory option for Quickdraws. A more detailed cost analysis can be found in table B.12, B.13, B.14 and B.15.

Samples	Method	Step 1 (h)	Step 2 (h)	Total time (h)	Total memory* (GB)	Cost on RAP** (£)
50k	Quickdraws	9.6	9	18.6	48 (16)	7.9
	BOLT-LMM	127	671.5	798.5	<15	158.4
	REGENIE	1.1	19.4	20.5	<15	4.0
	FastGWA	-	20.2	20.2	<15	3.6
405k	Quickdraws	97.7	51.5	149.3	63 (16)	93.0
	BOLT-LMM	1,150	7,250	8,400	46	7500.0
	REGENIE	4.7	45.3	50.0	<15	24.2
	FastGWA	-	128.4	128.4	<15	37.3

(a) Quantitative traits

Samples	Method	Step 1 (h)	Step 2 (h)	Total time (h)	Total memory* (GB)	Cost on RAP** (£)
50k	Quickdraws	18.3	77.8	96.1	49 (16)	25.2
	SAIGE	2.1	270.9	273	<15	47.5
	REGENIE	13.1	168.9	182.0	<15	41.0
	FastGWA	-	40.6	40.6	<15	9.2
405k	Quickdraws	162.7	519.7	682.3	67 (16)	395.9
	SAIGE	9.0	12015.5	12024.4	<15	2834.9
	REGENIE	56	813.9	869.9	<15	506.9
	FastGWA	-	254.1	254.1	<15	98.1

(b) Binary traits

Table 5.2: Computational efficiency of Quickdraws for (a) quantitative and (b) binary trait association. We compare the computational requirements for recent GWAS algorithms with Quickdraws to generate summary statistics for 13.3 million variants and 50 phenotypes with either $N = 50,000$ or $N = 405,088$ and $458,464$ genotyped markers for model fitting ($89,177$ genotyped markers for SAIGE, see Methods); *Total memory includes CPU RAM memory and GPU memory, only Quickdraws requires GPU memory which is reported separately in brackets; **A more detailed cost analysis can be found in table B.12 and B.13 Running times are computed using the same hardware for all methods (mem2_ssd1_v2_x8, with 32 GB of RAM, 8-core processor for $N = 50k$ and mem1_ssd1_v2_x36, 72 GB of RAM, 36-core processor for $N = 405k$ datasets).

5.6 Discussion

To fully harness the growing volumes of genomic, phenotypic, and environmental information contained in modern biobanks, existing GWAS algorithms need to optimize a complex trade-off between cost efficiency, statistical power, and robustness. We developed an algorithm for genome-wide association testing,

called Quickdraws, that navigates these trade-offs by combining established GWAS paradigms, such as using mixed-effect models [104, 106–108, 115–118, 130–133] to handle relatedness and population structure and the use of penalized regression [123, 279] for binary traits, with recent ideas from Bayesian machine learning, such as stochastic variational inference [226, 229], first-order gradient optimizers [227, 230], and transfer learning [231]. In addition, Quickdraws allows leveraging GPUs, which are available in modern computing platforms and have been widely adopted in machine learning applications to parallelize computationally intensive matrix operations [244]. Overall, this strategy allows Quickdraws to achieve higher average association power than existing methods in the analysis of binary traits, and to match the association power obtained in quantitative traits by BOLT-LMM, the only other approach that enables modeling of non-infinitesimal architectures. In computational benchmarks using the UK Biobank RAP cloud platform, Quickdraws required a fraction of the computational resources used by BOLT-LMM, with costs comparable to those of scalable methods with lower average association power. Summary statistics for the traits we analyzed, which comprise 79 quantitative traits, 50 self-reported diseases, and 2923 plasma protein traits, are freely available for download from [280].

The gains in statistical power achieved by Quickdraws are equivalent to analyzing data sets that contain larger sample sizes [106, 107, 234], with our simulations showing increases in effective sample size of over 19.2% for traits linked to a few thousands of causal variants. Similarly, our analyses of real traits, in which we focused on quantitative traits and diseases with high heritability and phenotyping rate, yielded more independent association signals than current scalable approaches (example, 26.1% and 8.04% increase over REGENIE for Height and Mean platelet volume, see tables B.7 and B.8). Although complex traits are known to be highly polygenic [134], their genetic architectures, in particular for molecular traits such as metabolites [281], are estimated to be far from a purely infinitesimal model, with large fractions of heritability concentrating in a relatively small set of variants [252, 258, 282]. As the range of phenotypic measurements available in biobank data sets

continues to grow, we therefore expect the increase in computational power that we observed in our analyses to also be observed in broader sets of traits and diseases.

5.6.1 Sharing and combining posterior effect estimates

Our analyses demonstrate that, in addition to increasing association power by allowing the computation of a residualized phenotype, the posterior mean estimates of the effect sizes computed during the model fitting step can be used to construct polygenic scores that perform well in held-out individuals compared to those computed from summary association statistics. Therefore, sharing these effect estimates, which we make available for the traits we analyzed (see [280]), may facilitate novel downstream analyses. However, compared to the use of summary association statistics, one disadvantage of using posterior effect estimates to construct polygenic scores is the lack of established methodology to combine them across independent cohorts. Developing strategies that allow to meta-analyze these effects across cohorts could open new avenues for the development of federated association and polygenic prediction strategies that preserve high statistical power without the need for sharing individual-level data.

5.6.2 Limitations and future work

We highlight several current limitations and areas of future work to improve the Quickdraws algorithm. First, Quickdraws leverages modern GPU hardware to speed up the Bayesian regression in the model fitting step and is therefore slower when run on traditional CPU hardware (see Figure 5.15 for a comparison). GPUs may not be as easily accessible as CPUs, and have higher costs, as evidenced in Table B.12 and B.13. Our analyses on the RAP cloud platform, however, demonstrates that the overall increase in running costs is not substantial. Furthermore, the extensive use of GPUs in machine learning applications is likely to continue to increase their availability, paving the way for additional GPU-based applications in human genetics. Future work may allow developing a faster CPU-based implementation of Quickdraws. Second, although Quickdraws is well parallelized for the analysis

of multiple traits, it currently does not capture the correlations that often exist across these traits. Extending Quickdraws to leverage genetic and environmental correlations is likely to lead to improved performance [283, 284]. Third, although we have shown that a strategy based on adjusting for polygenic scores computed within the same sample does not lead to power gains comparable to those achieved by Quickdraws and BOLT-LMM, it is possible that using polygenic scores constructed from larger external cohorts will lead to more substantial power increases [276, 277]. Fourth, association analyses have recently been shown to be affected by participation biases [285, 286], to which Quickdraws is also susceptible. To mitigate these biases, we implemented the adjustment strategy proposed in [278] in step 2 of the Quickdraws algorithm (as described in Section 5.4.5). However, further work will be needed to incorporate this adjustment in step 1 of the algorithm. Despite these current limitations and avenues for future development, we believe that Quickdraws will provide a useful tool for large-scale GWAS, demonstrating the promise of leveraging modern machine learning methodology to improve statistical power and efficiency in these analyses.

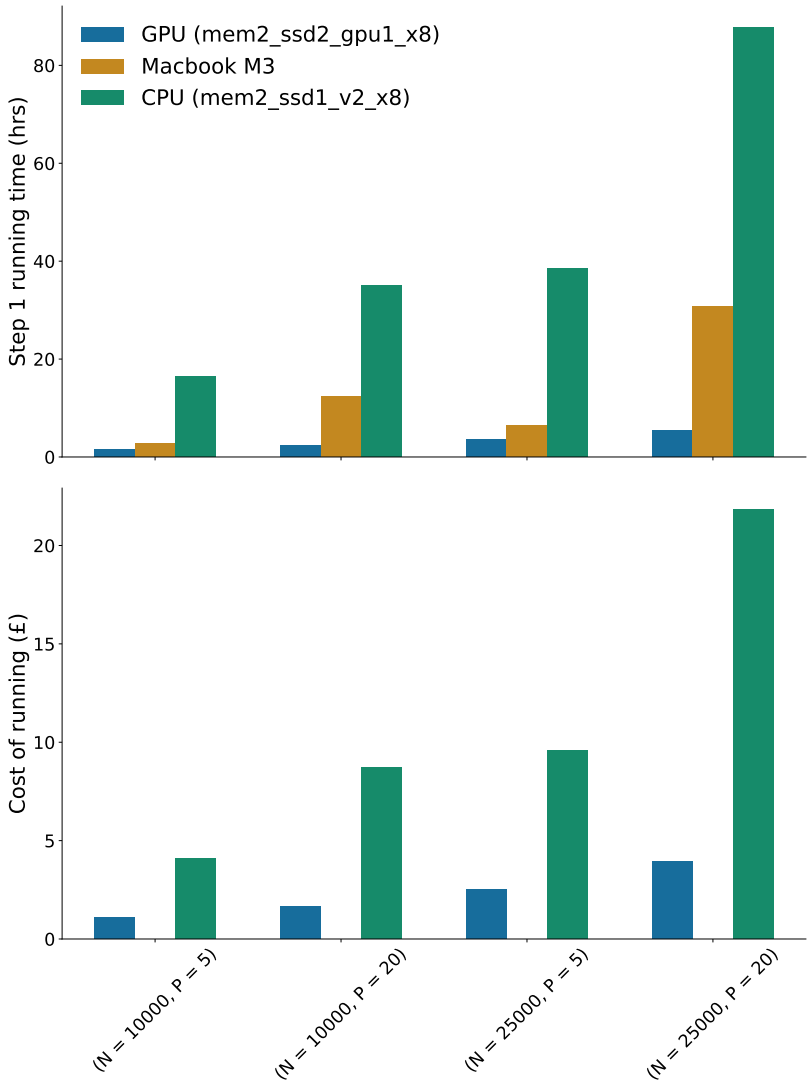


Figure 5.15: Computational performance of Quickdraws on different computing architectures. We assessed the running time and cost associated with Quickdraws' model fitting step (Step 1) across varying numbers of samples (N) and phenotypes (P). Performance was compared for three architectures: an Nvidia A10G GPU node on the UK Biobank Research Analysis Platform (RAP) (`mem2_ssd2_gpu1_x8`), a 14-core MacBook Pro equipped with the Apple M3 chip, and an 8-core CPU node on the UK Biobank RAP (`mem2_ssd1_v2_x8`). Running costs were calculated only for the GPU and CPU nodes on the RAP.

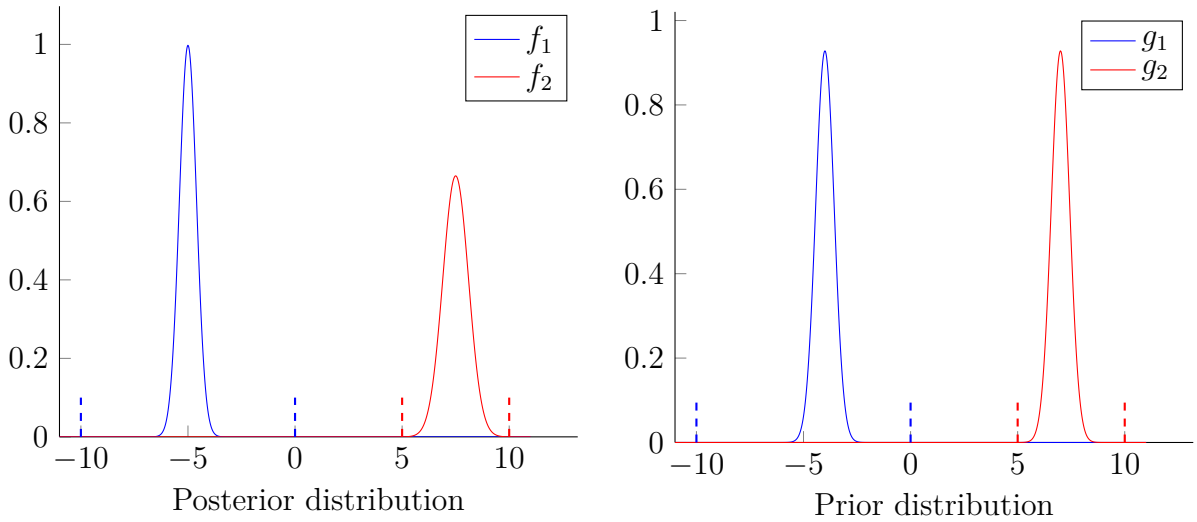
Appendices

A
Additional proofs

Contents

A.1 KL-divergence for non-overlapping mixture distributions 225

A.1 KL-divergence for non-overlapping mixture distributions



Theorem 1: Let $q = \sum_i^N \pi_i f_i$ be a mixture of non-overlapping distributions, i.e., $S_i \cap S_j = \emptyset, \forall i \neq j$, where S_i is the support of distribution f_i. If p is a probability

distribution and can be represented as a mixture of non-overlapping distributions with the same support as f_i, i.e., $p = \sum_i^N \pi'_i g_i$ and support of g_i = support of f_i, the KL divergence between q and p can be simplified as follows:

$$KL(q||p) = \sum_i^N (\pi_i KL(f_i||g_i) + \pi_i \log(\pi_i) - \pi_i \log(\pi'_i)) \tag{A.1}$$

Proof: Assuming q is 0 outside $\cup_i^N S_i$, we can write the integral as

$$\begin{aligned} KL(q||p) &= \int q\log(q) - q\log(p) = \sum_i^N \int_{S_i} q\log(q) - q\log(p) \\ &= \sum_i^N \int_{S_i} \left(\sum_i^N \pi_i f_i\right) \log\left(\sum_i^N \pi_i f_i\right) - \left(\sum_i^N \pi_i f_i\right) \log\left(\sum_i^N \pi'_i g_i\right). \end{aligned} \tag{A.2}$$

Because $S_i \cap S_j = \emptyset, \forall i \neq j$, we can simplify as

$$\begin{aligned} KL(q||p) &= \sum_i^N \int_{S_i} \pi_i f_i \log(\pi_i f_i) - \pi_i f_i \log(\pi'_i g_i) \\ &= \sum_i^N \pi_i \left(\int_{S_i} f_i \log(f_i) - f_i \log(g_i)\right) + \sum_i^N \left(\pi_i log(\pi_i) \int_{S_i} f_i - \pi_i log(\pi'_i) \int_{S_i} f_i\right) \\ &= \sum_i^N \pi_i \left(\int_{S_i} f_i \log(f_i) - f_i \log(g_i)\right) + \sum_i^N (\pi_i log(\pi_i) - -\pi_i log(\pi'_i)) \\ &= \sum_i^N (\pi_i KL(f_i||g_i) + \pi_i \log(\pi_i) - \pi_i \log(\pi'_i)) \end{aligned} \tag{A.3}$$

B

Additional figures and tables

Contents

B.1 Additional figures 227
B.2 Additional tables 261

B.1 Additional figures

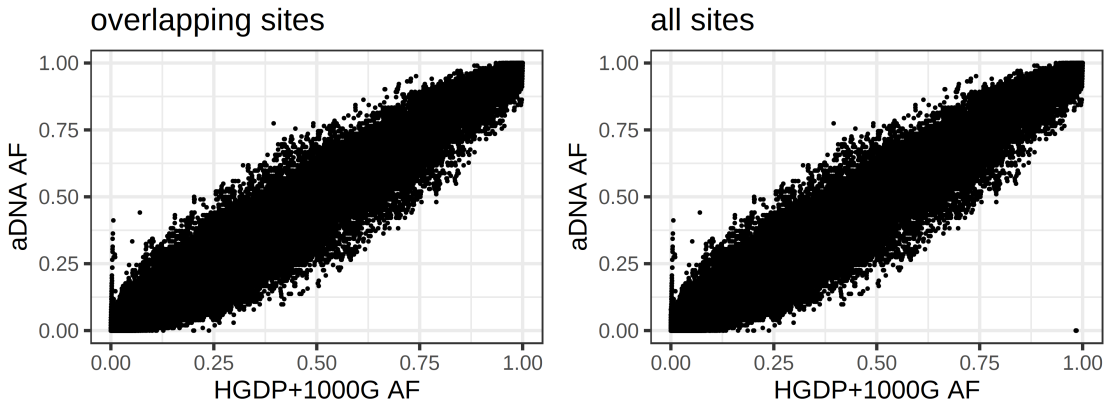


Figure B.1: Comparison of allele frequencies between HGDP+1,000GP and ancient samples. We analyze biallelic SNPs on chromosome 20 within the masked regions. The comparison includes union ("all sites") and intersection ("overlapping sites") of SNPs in the two datasets seperately.

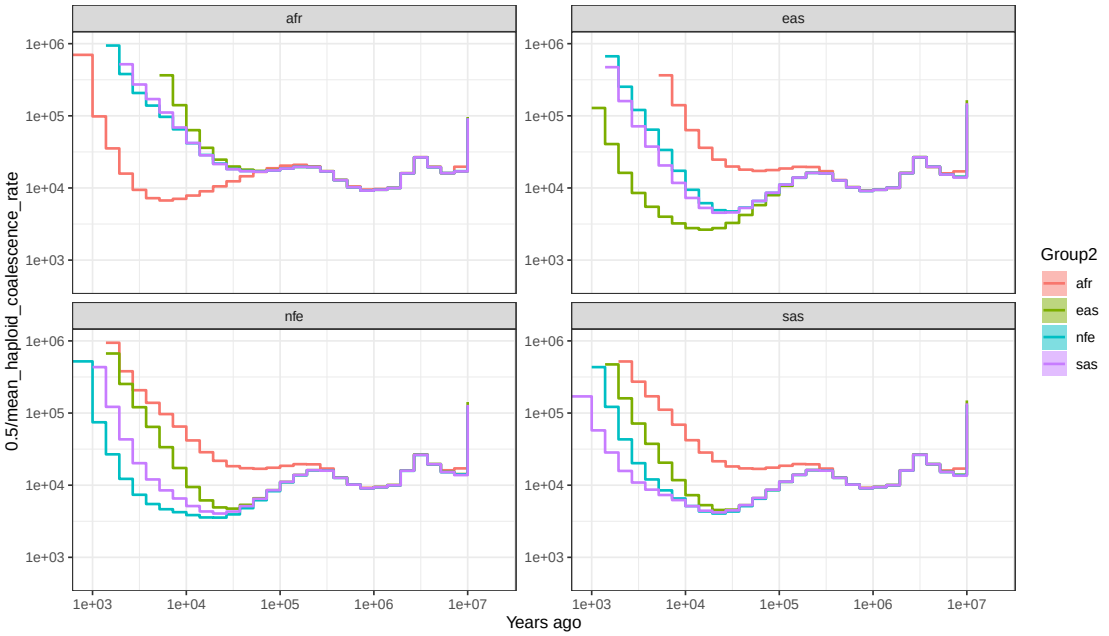


Figure B.2: Pairwise inverse coalescence rates estimated from Relate trees for the HGDP+1,000GP dataset. Genome-wide average inverse cross-coalescence rates between afr (Sub-Saharan Africans), eas (East Asians), nfe (Non-Finnish Europeans), and sas (South Asians).

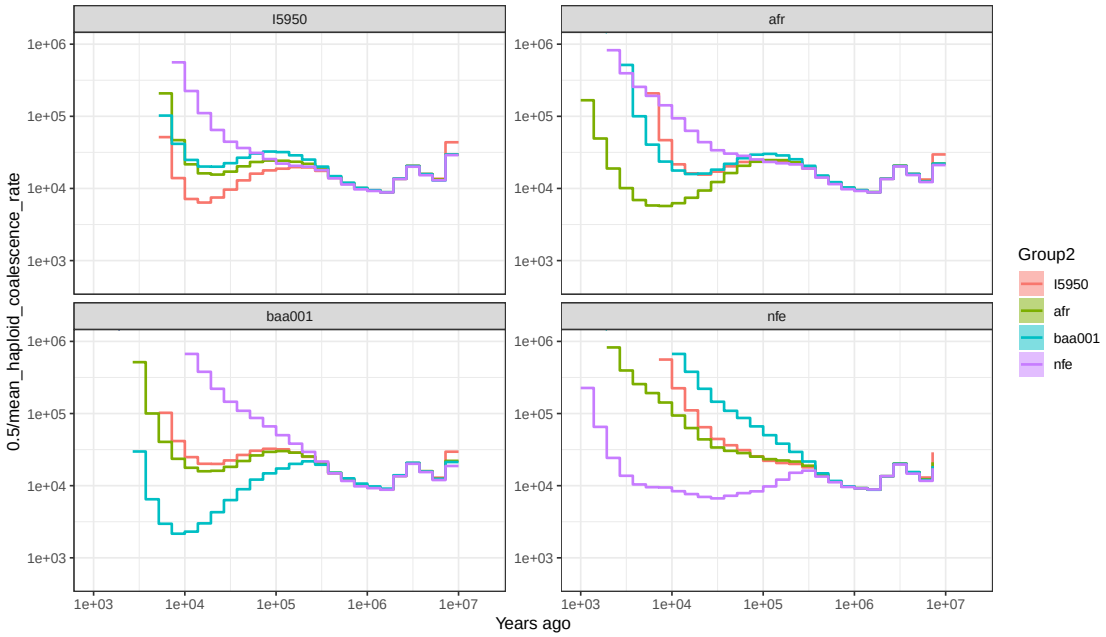


Figure B.3: Pairwise inverse coalescence rates estimated from Relate trees for the HGDP+1,000GP+aDNA dataset. Genome-wide average inverse cross-coalescence rates between I9590 (Mota individual, Ethiopia), afr (modern Sub-Saharan Africans), baa001 (Ballito Bay individual, South Africa), and nfe (modern Non-Finnish Europeans).

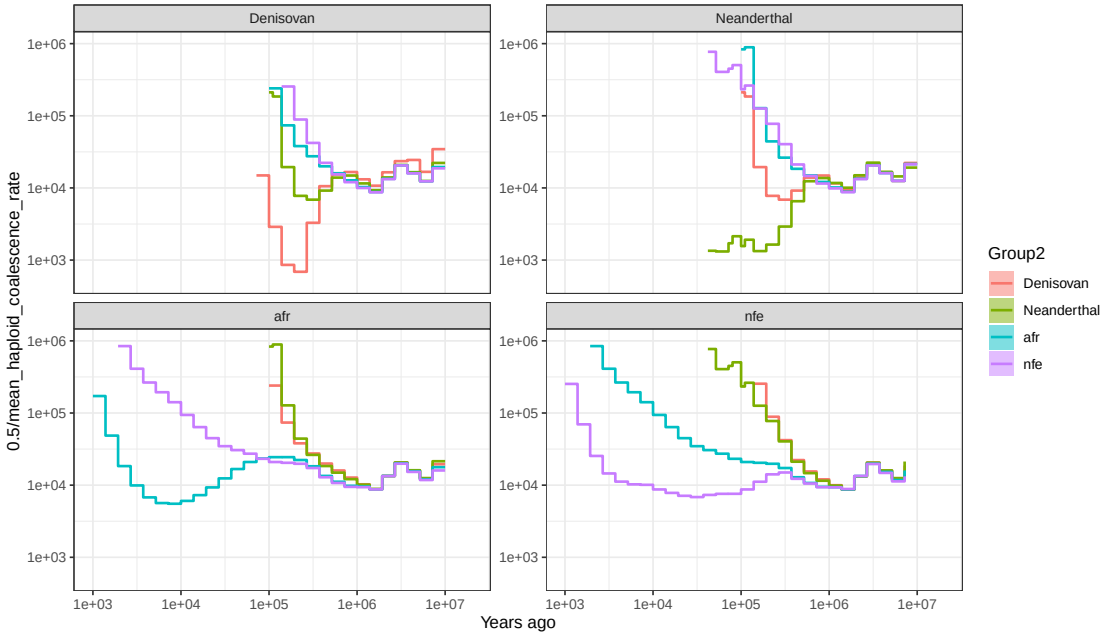


Figure B.4: Pairwise inverse coalescence rates estimated from Relate trees for the HGDP+1,000GP+archaics dataset. Genome-wide average inverse cross-coalescence rates between Denisovan, Neanderthals, afr (modern Sub-Saharan Africans), and nfe (modern Non-Finnish Europeans).

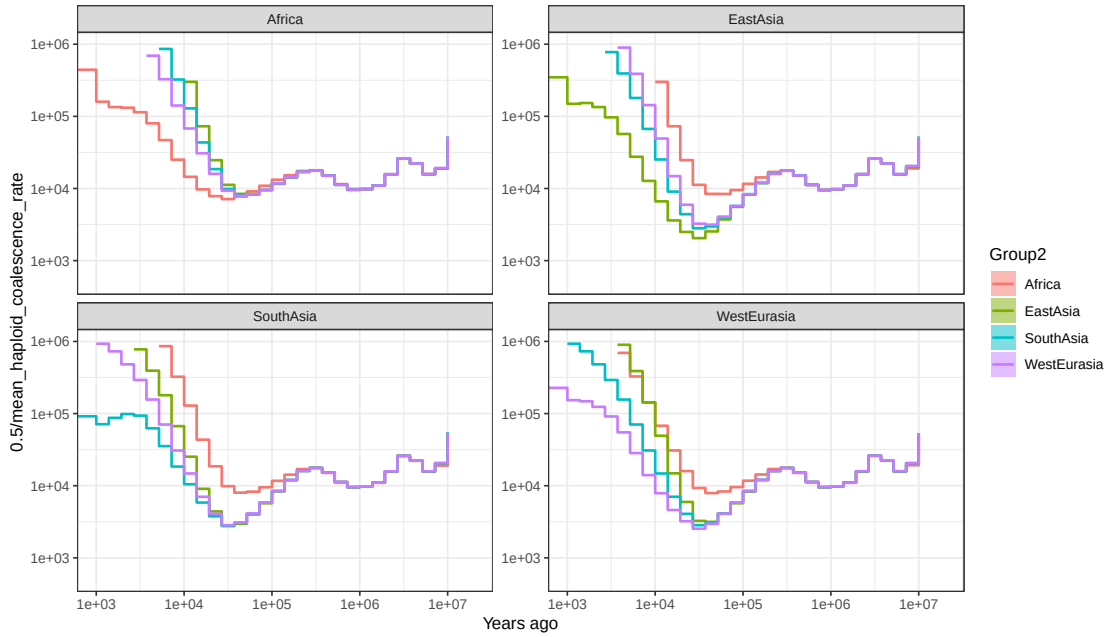


Figure B.5: Pairwise inverse coalescence rates estimated from Relate trees for the SGDP dataset. Genome-wide average inverse cross-coalescence rates between Africans, East Asians, South Asians, and West Eurasians.

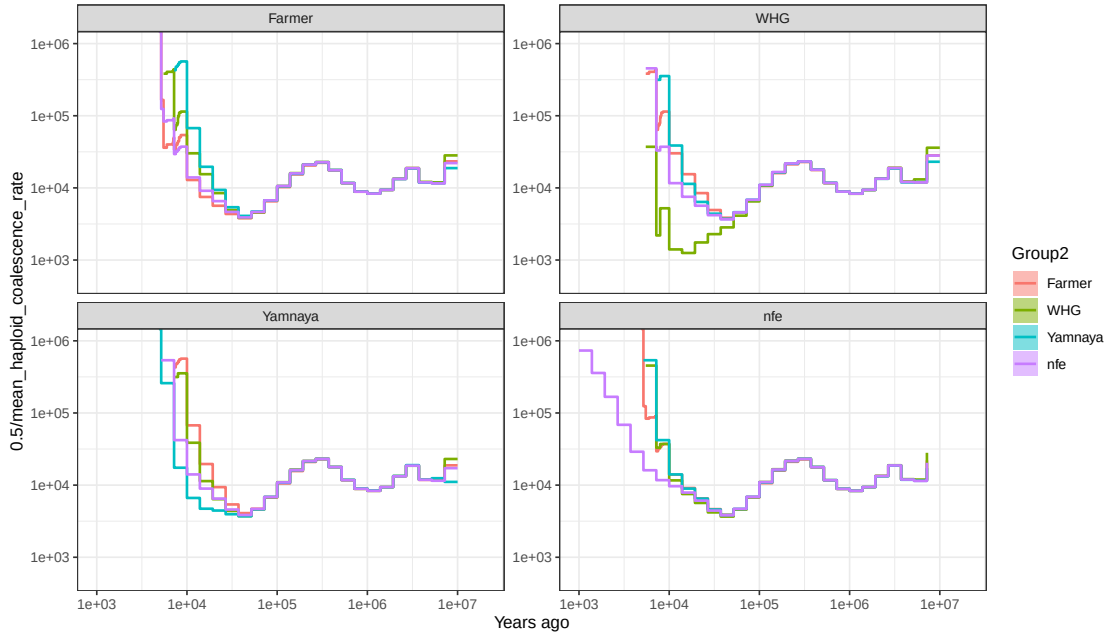


Figure B.6: Pairwise inverse coalescence rates estimated from Relate trees for the European subset of the HGDP+1,000GP+aDNA dataset. Genome-wide average inverse cross-coalescence rates between Farmers, WHG (Western Hunter-Gatherers), Yamnaya, and nfe (modern Non-Finnish Europeans).

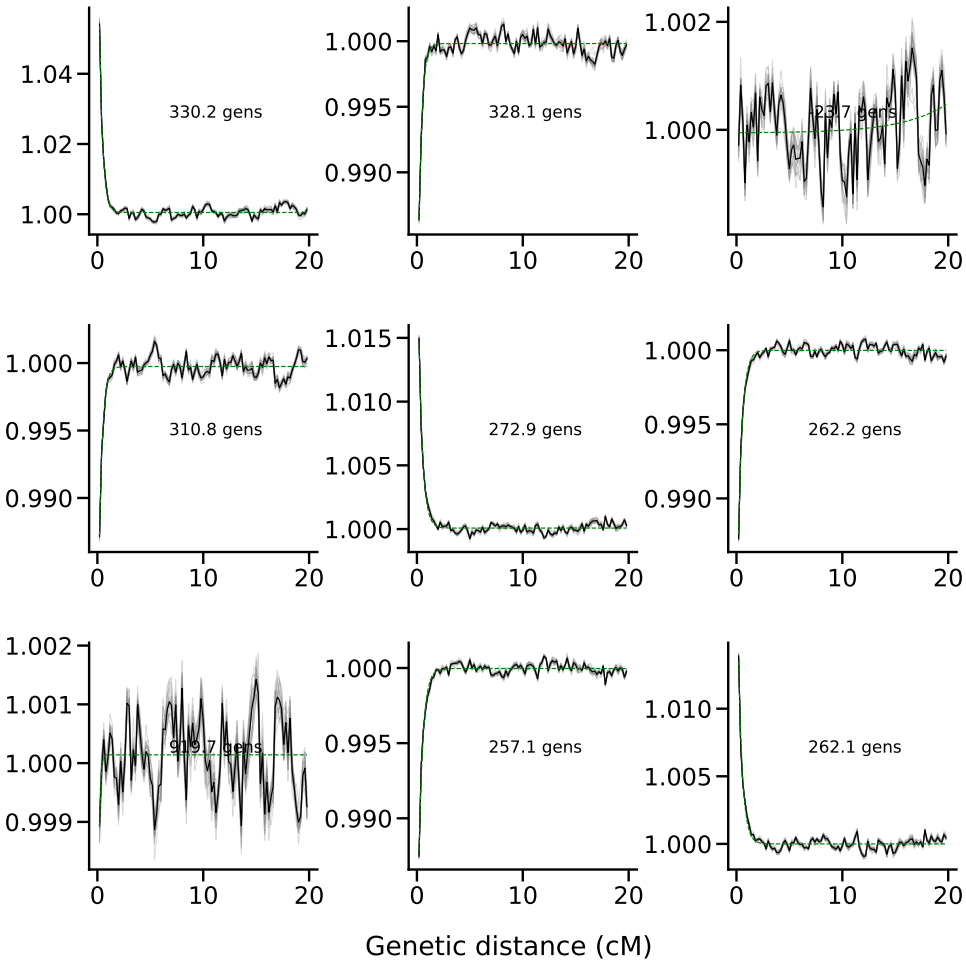


Figure B.7: Co-ancestry curves for 3-way admixture in Europe. We refit each co-ancestry curve with an exponential fit to estimate the date of a multi-way admixture event. In this case component 1 represent WHG ancestry, component 2 farmer and component 3 Yamnaya-like ancestry. Inferred admixture dates in the plot are in generations.

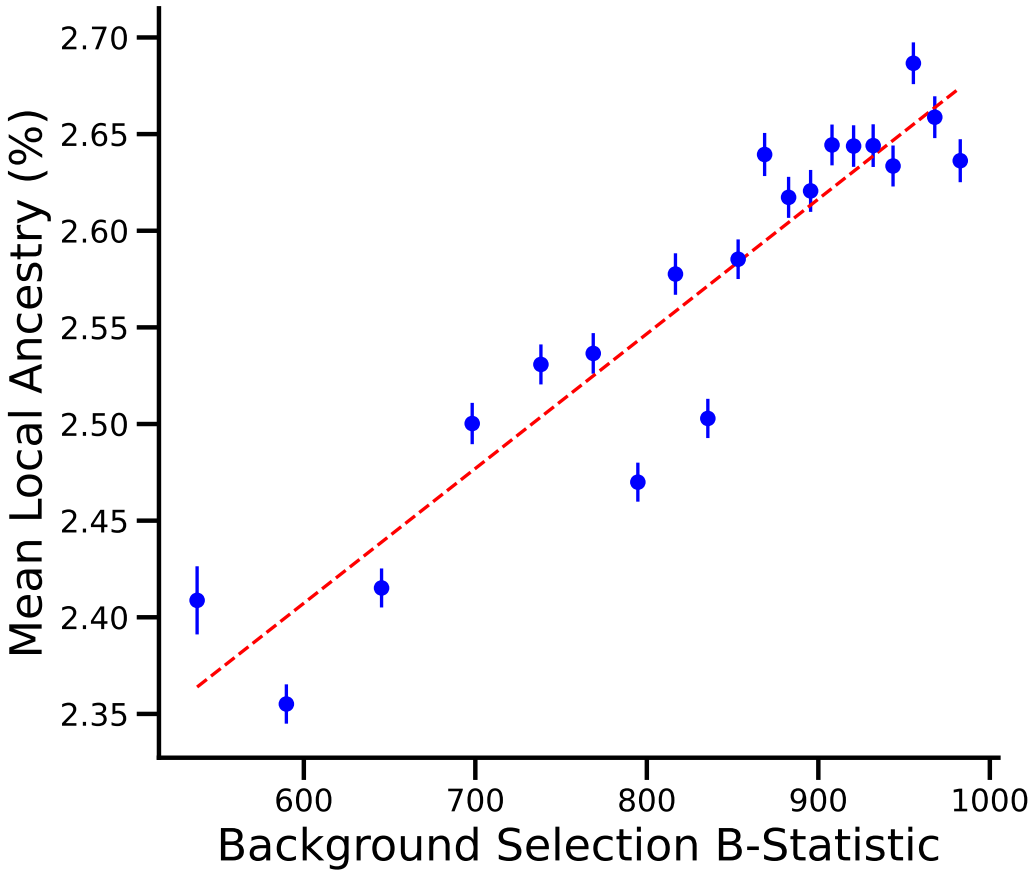
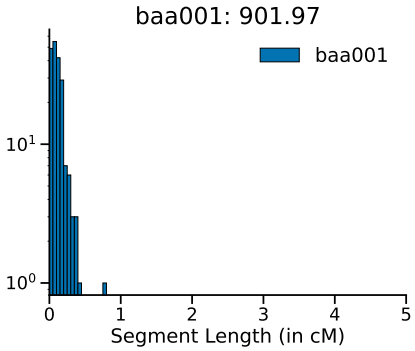
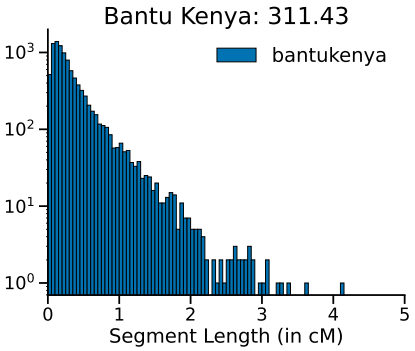


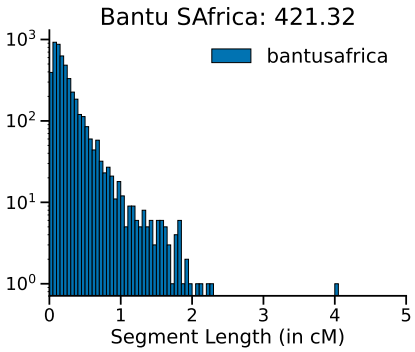
Figure B.8: Mean Eurasian-like ancestry proportion stratified by background selection intensity. The mean Eurasian-like ancestry proportion is calculated across African samples from the HGDP and 1,000GP. The background selection B-map was obtained from [192]. Error bars represent standard errors around the mean.


baa001: 901.97
baa001
Segment Length (in cM)

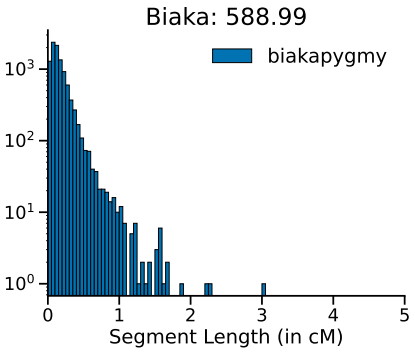
(a)


Bantu Kenya: 311.43
bantukenya
Segment Length (in cM)

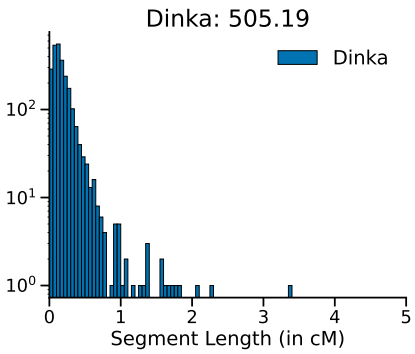
(b)


Bantu SAfrica: 421.32
bantusafrica
Segment Length (in cM)

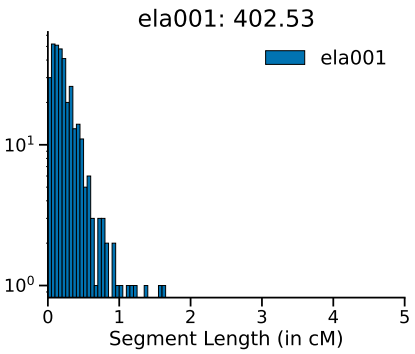
(c)


Biaka: 588.99
biakapygmy
Segment Length (in cM)

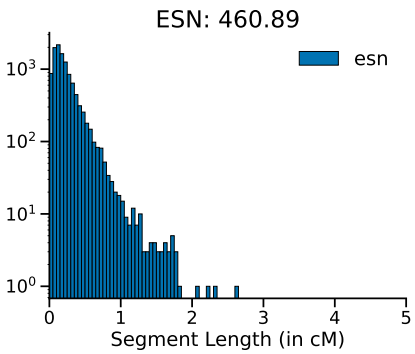
(d)


Dinka: 505.19
Dinka
Segment Length (in cM)

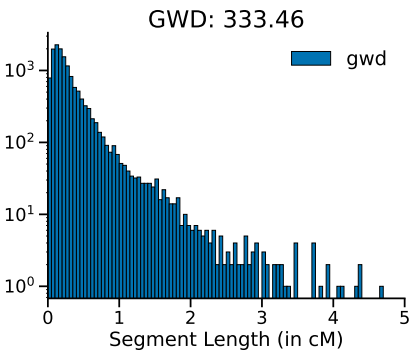
(e)


ela001: 402.53
ela001
Segment Length (in cM)

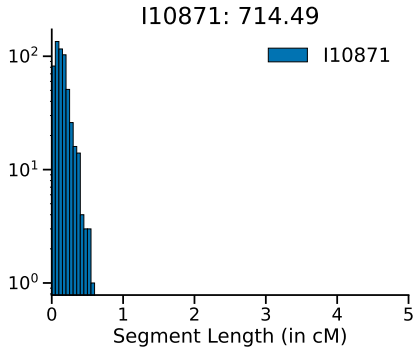
(f)


ESN: 460.89
esn
Segment Length (in cM)

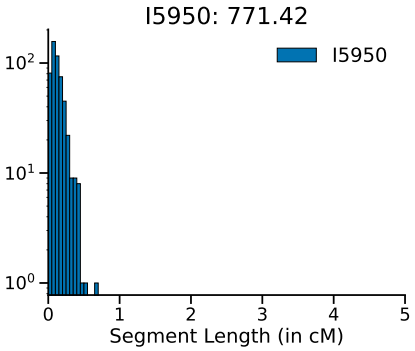
(g)


GWD: 333.46
gwd
Segment Length (in cM)

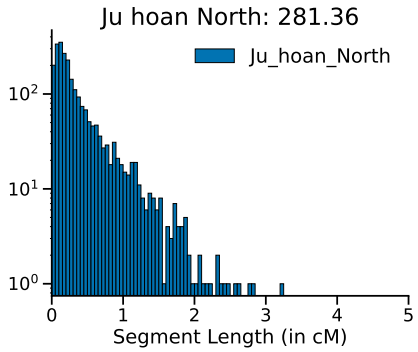
(h)


I10871: 714.49
I10871
Segment Length (in cM)

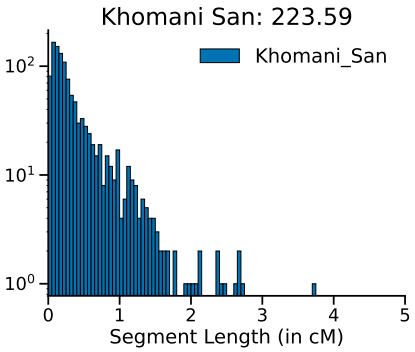
(i)


I5950: 771.42
I5950
Segment Length (in cM)

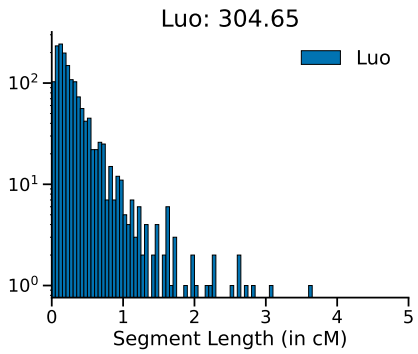
(j)


Ju hoan North: 281.36
Ju_hoan_North
Segment Length (in cM)

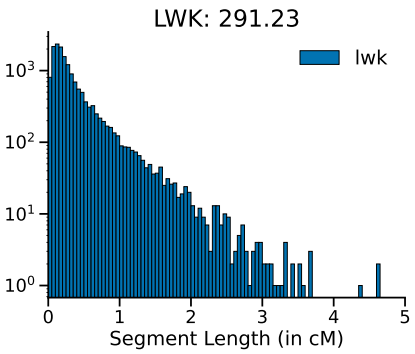
(k)


Khomani San: 223.59
Khomani_San
Segment Length (in cM)

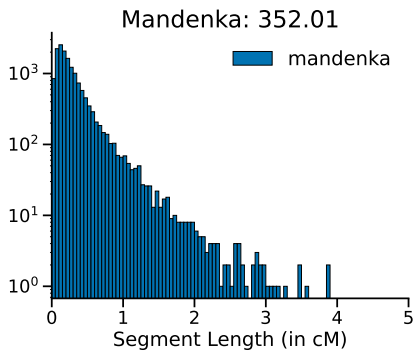
(l)


Luo: 304.65
Luo
Segment Length (in cM)

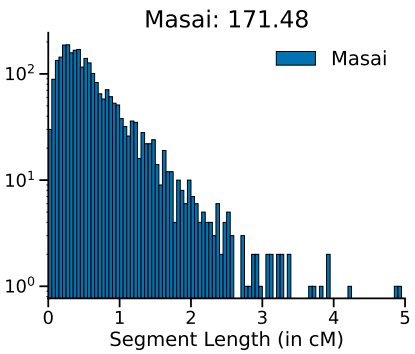
(m)


LWK: 291.23
lwk
Segment Length (in cM)

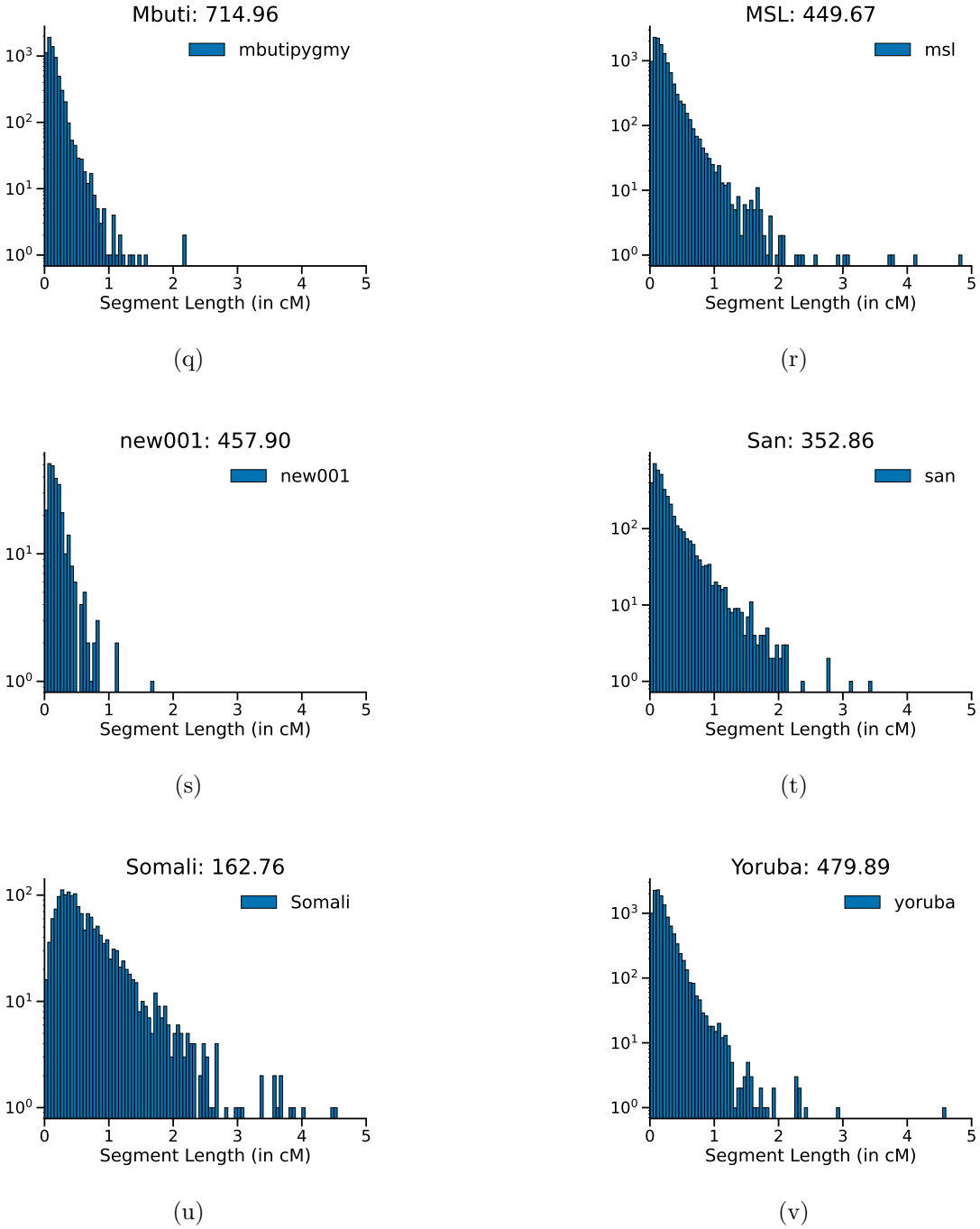
(n)


Mandenka: 352.01
mandenka
Segment Length (in cM)

(o)


Masai: 171.48
Masai
Segment Length (in cM)

(p)



(q) (r) (s) (t) (u) (v)

Figure B.9: Eurasian-like ancestry segment length distribution across African populations. Panels (a-v) show histograms of Eurasian-like ancestry segment lengths for each African population analyzed in the HGDP, 1,000GP, and SGDP datasets. Each panel title includes the population label, the reciprocal of the African proportion, and the mean segment length (an estimate of single-event admixture date in generations). The x-axis is truncated at 5 cM.

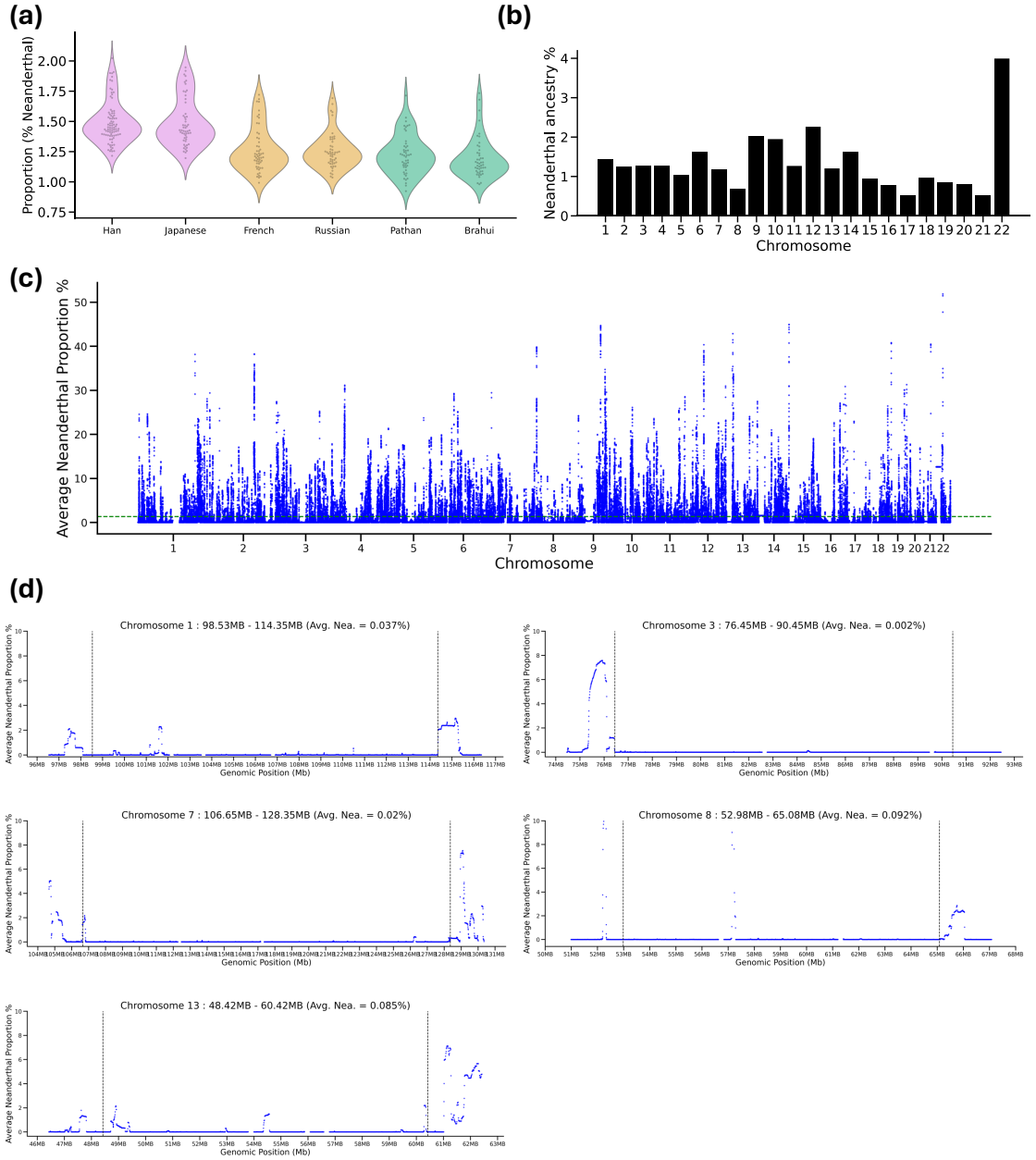


Figure B.10: Detecting Neanderthal ancestry in Eurasian populations with GhostBuster. (a) Proportion of Neanderthal ancestry estimated for two East Asian (Han, Japanese), two European (French, Russian), and two South Asian (Pathan, Brahui) populations from the HGDP+1,000GP dataset. (b) Chromosome-wide average proportion of Neanderthal ancestry. (c) Genome-wide average proportion of Neanderthal ancestry along the genome. (d) Average Neanderthal ancestry across five previously identified archaic deserts [173, 199]. In panel (c), the green dashed line represents the genome-wide average Neanderthal ancestry. In panel (d), the dashed lines mark the start and end positions of the deserts. Regions within and 100 kb around the ENCODE blacklist [287] have been excluded from the analysis. Panel (a) overlays individual genome-wide Neanderthal ancestry data points for each sample on the violin plots. The positions shown in panel (d) are based on the hg38 genome build.

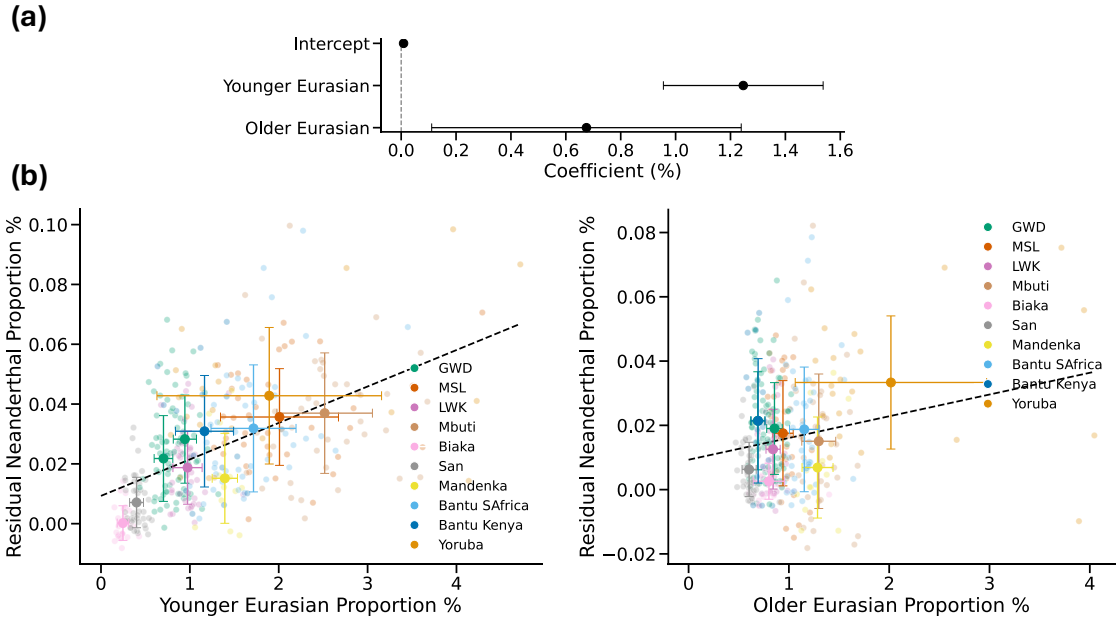


Figure B.11: Joint-regression of Neanderthal and Eurasian-like ancestry in Africans. (a) Mean coefficient estimates and confidence intervals from the joint regression analysis of genome-wide Neanderthal ancestry proportions against Eurasian-like ancestry proportions, divided into younger and older Eurasian-like segments. (b) Genome-wide residual Neanderthal ancestry vs. Eurasian-like ancestry proportions across all African samples. The residual Neanderthal proportion is defined as the genome-wide Neanderthal ancestry proportion minus the estimated proportion attributed to other Eurasian-like ancestry, as inferred from the joint regression. Error bars in (a) represent 95% confidence intervals around the mean estimated coefficients using least squares. Error bars in (b) represent the population-specific standard deviation of Eurasian and Neanderthal proportions, while the dashed line in (b) indicating the least squares fit.

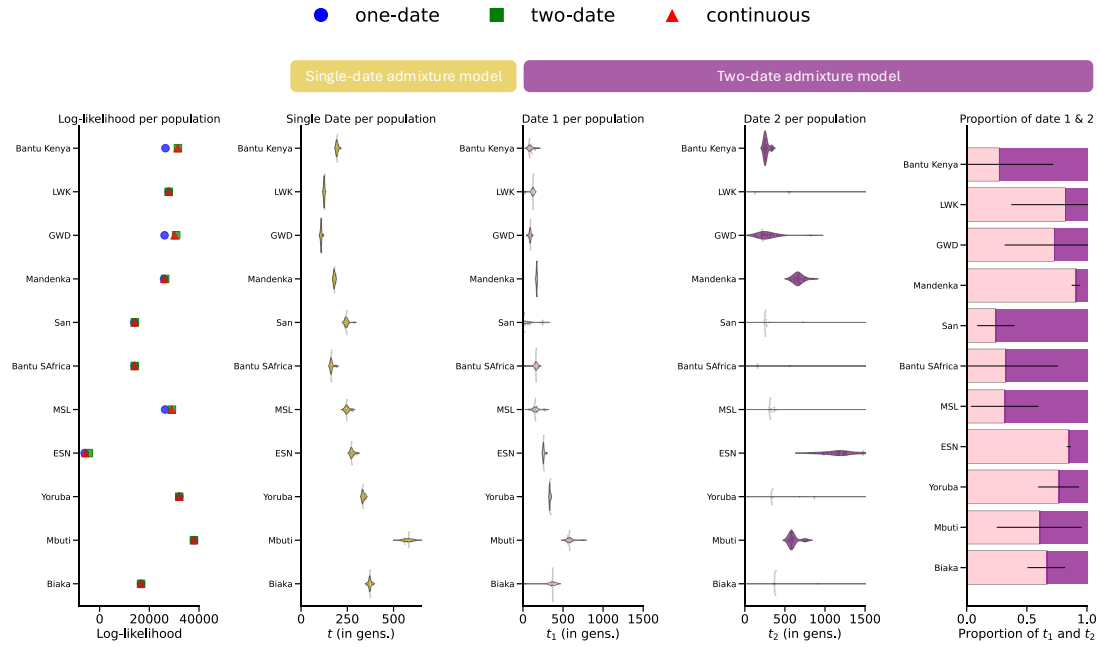


Figure B.12: Dating Eurasian-like segments carrying Neanderthal ancestry in Africans. Coancestry curves were generated by considering only Eurasian-like segments that carry Neanderthal ancestry (with posterior probability ≥ 0.1) at at least one position within the segment. Eurasian-like segments that did not carry any Neanderthal ancestry were treated as African (non-Eurasian-like) ancestry. The figure displays the log-likelihoods corresponding to various admixture models, including one-date, two-date, and constant continuous migration models. Single admixture model fits and two-date admixture model fits are shown, with populations sorted from highest to lowest proportion of Eurasian-like ancestry. The log-likelihood values are aggregated across 20 jackknife runs. For the two-date admixture model, pink represents the younger admixture date, and purple represents the older admixture date. Admixture dates in the plot are in generations.

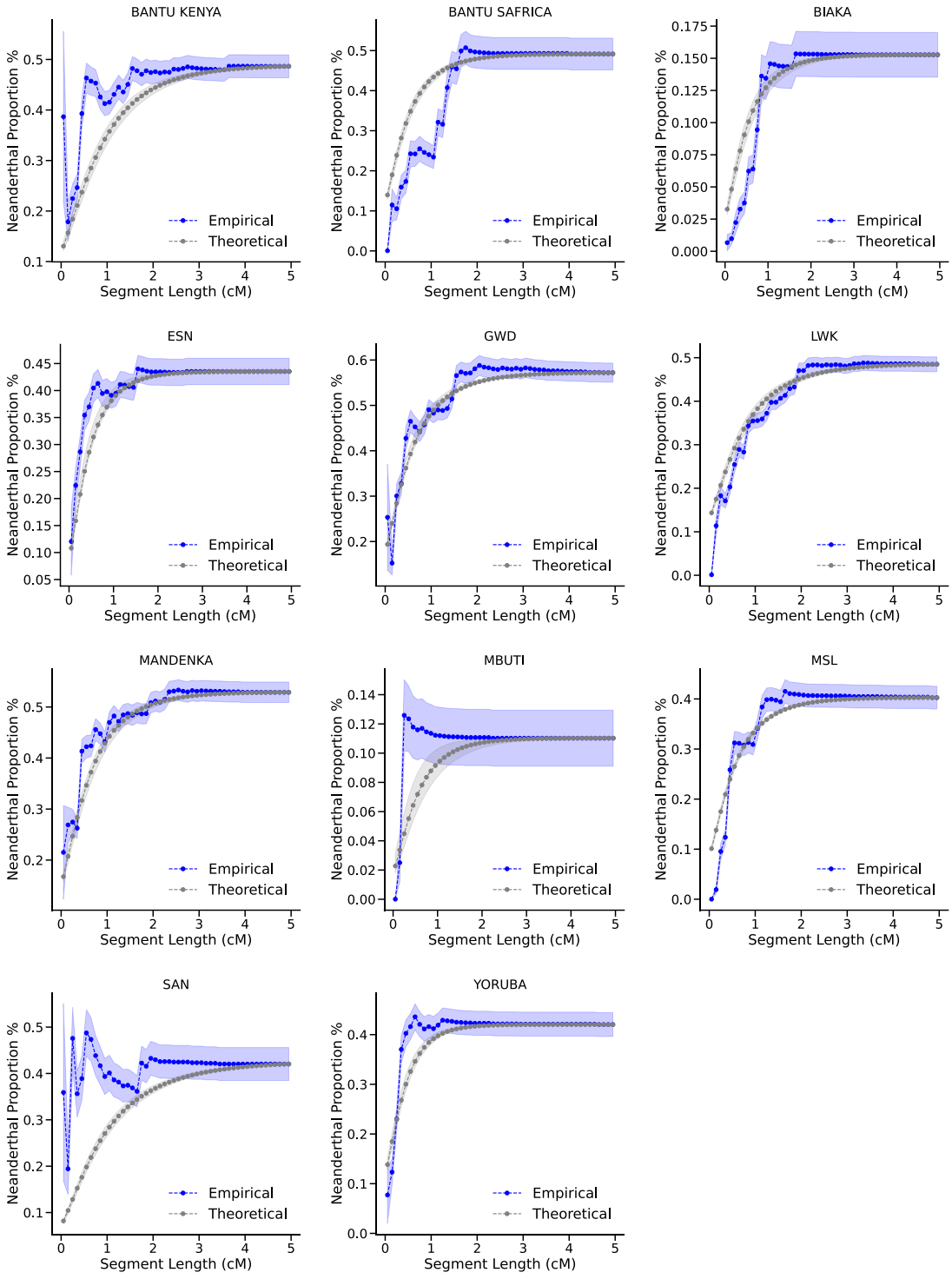


Figure B.13: Comparison of observed Neanderthal proportion with theoretical expectations under the assumption of no Neanderthal ancestry in older Eurasian admixture. The posterior probability of being Neanderthal is measured for Eurasian-like segments in Africa with segment sizes less than or equal to the specified value (x-axis). Observed values are compared to theoretical Neanderthal proportions assuming two admixture events, where only the younger event contributes to Neanderthal ancestry. Shaded regions represent 95% confidence intervals for the observed Neanderthal proportion and theoretical estimates, accounting for variability in admixture dating. The x-axis is limited to 5 cM.

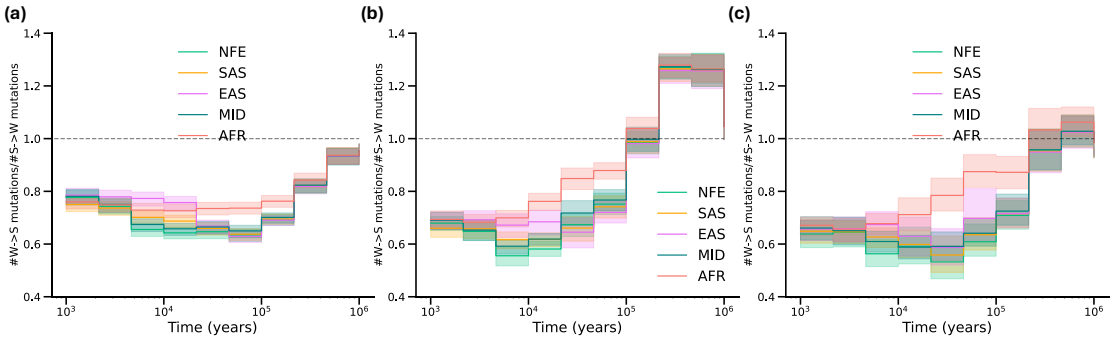


Figure B.14: Evolution of number of weak to strong mutations relative to strong to weak mutations across populations. Ratio of number of weak to strong mutations divided by strong to weak mutations across various continental groups, including NFE (non-Finnish Europeans), SAS (South Asians), EAS (East Asians), and MID (Middle Eastern and North African populations), AFR (sub-Saharan Africans). CpG sites are excluded from calling both weak and strong mutations. Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment.

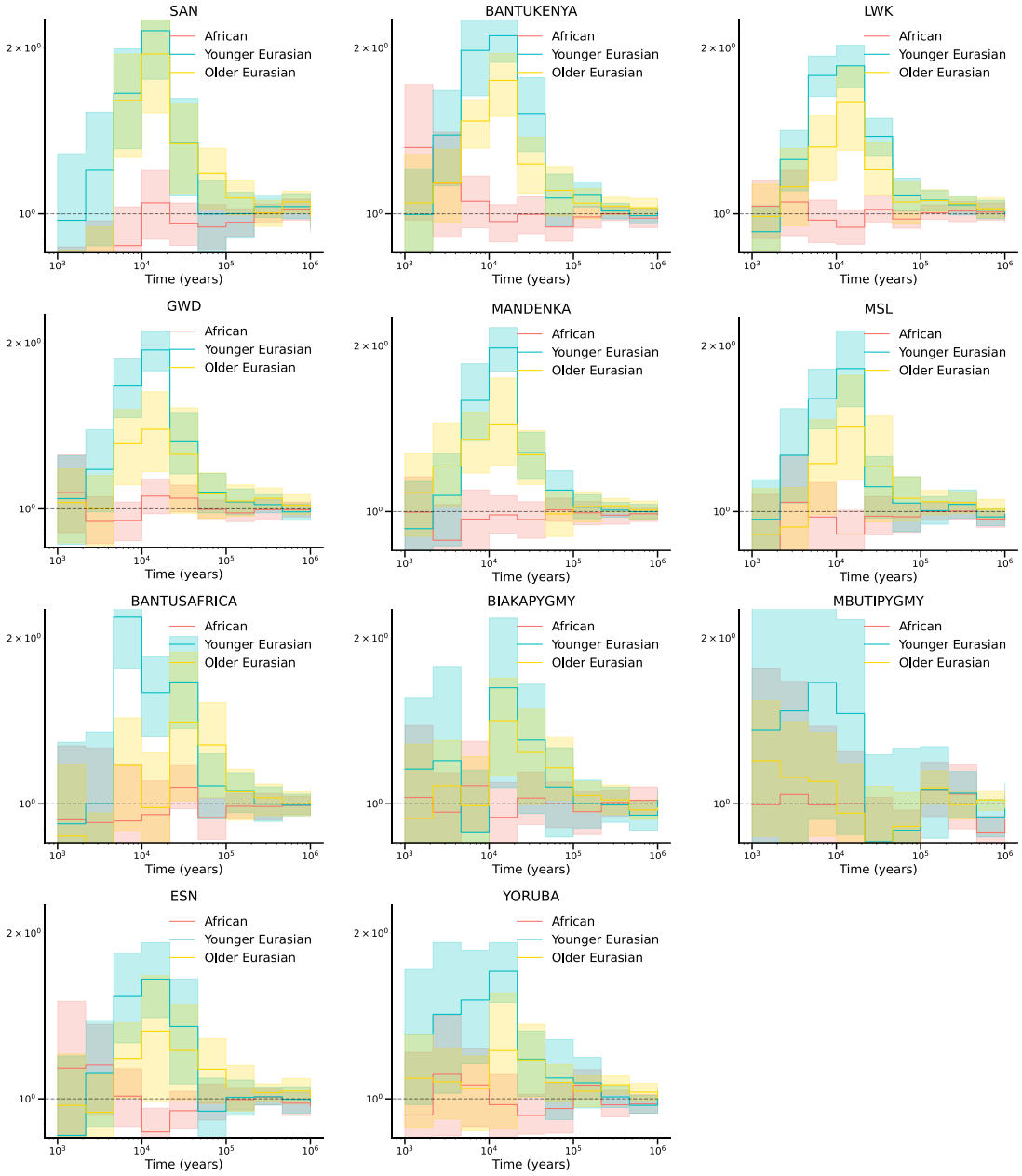


Figure B.15: Enrichment of TCC to TTC mutation rates across African ancestral components in diverse African populations. Enrichment of TCC to TTC mutation rates in younger and older Eurasian-like segments compared to non-Eurasian-like (African) segments across 11 modern African populations (see population descriptions in 3.1a). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment.

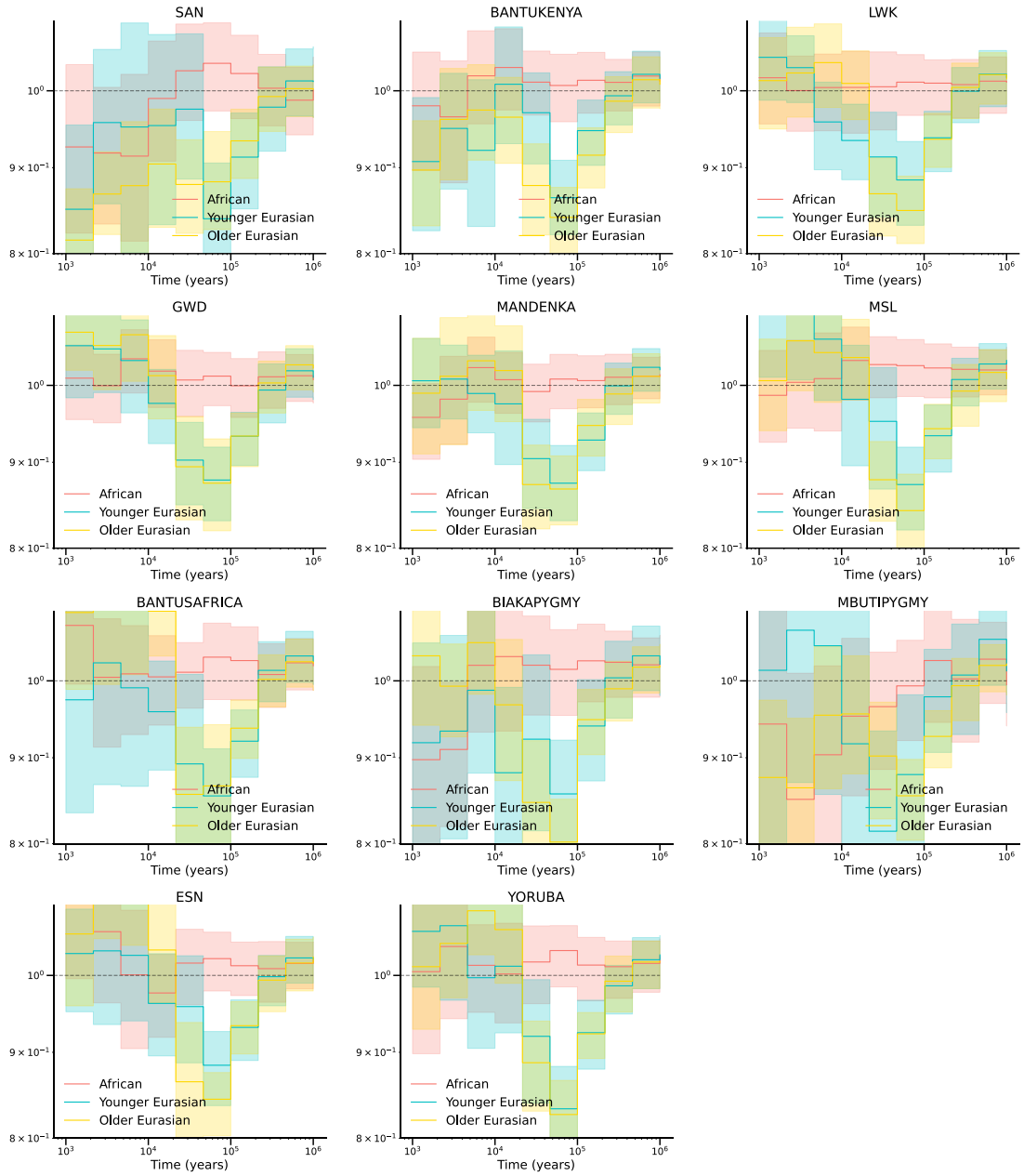
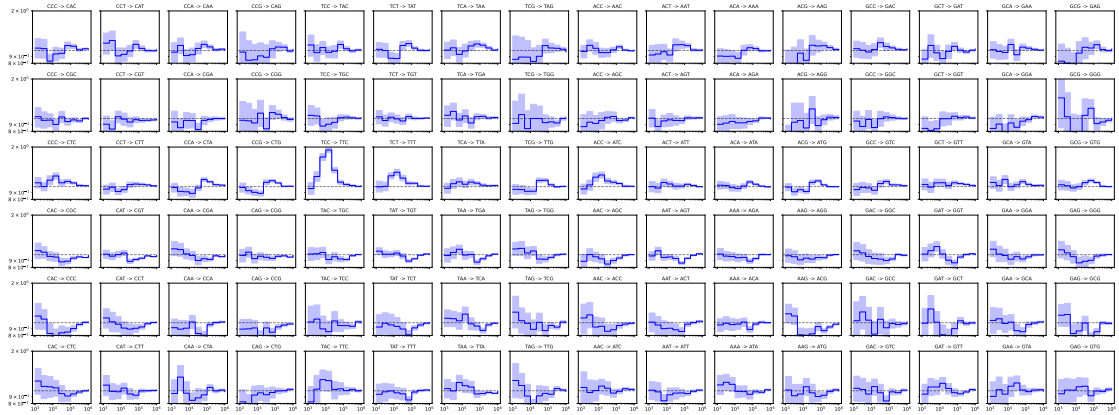
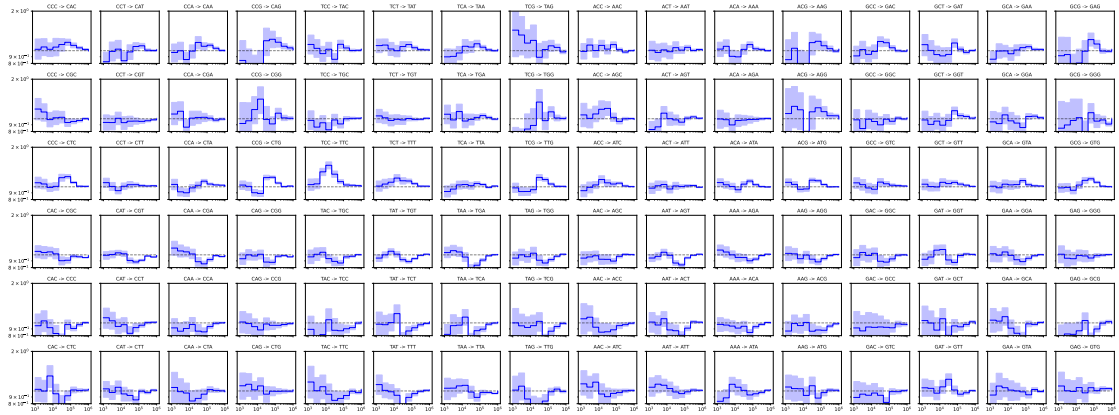


Figure B.16: Enrichment of weak to strong mutation rates across African ancestral components in diverse African populations. Enrichment of weak to strong mutation rates in younger and older Eurasian-like segments compared to non-Eurasian-like (African) segments across 11 modern African populations (see population descriptions in 3.1a). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment.



(a)



(b)

Figure B.17: Mutation count enrichment across all 96 trimer types in Eurasian-like segments within African individuals from the HGDP+1,000GP dataset. (a) Younger Eurasian-like segments and (b) older Eurasian-like segments are shown. The y-axis represents the enrichment, calculated relative to the mutation count in non-Eurasian-like segments of African individuals. Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment.

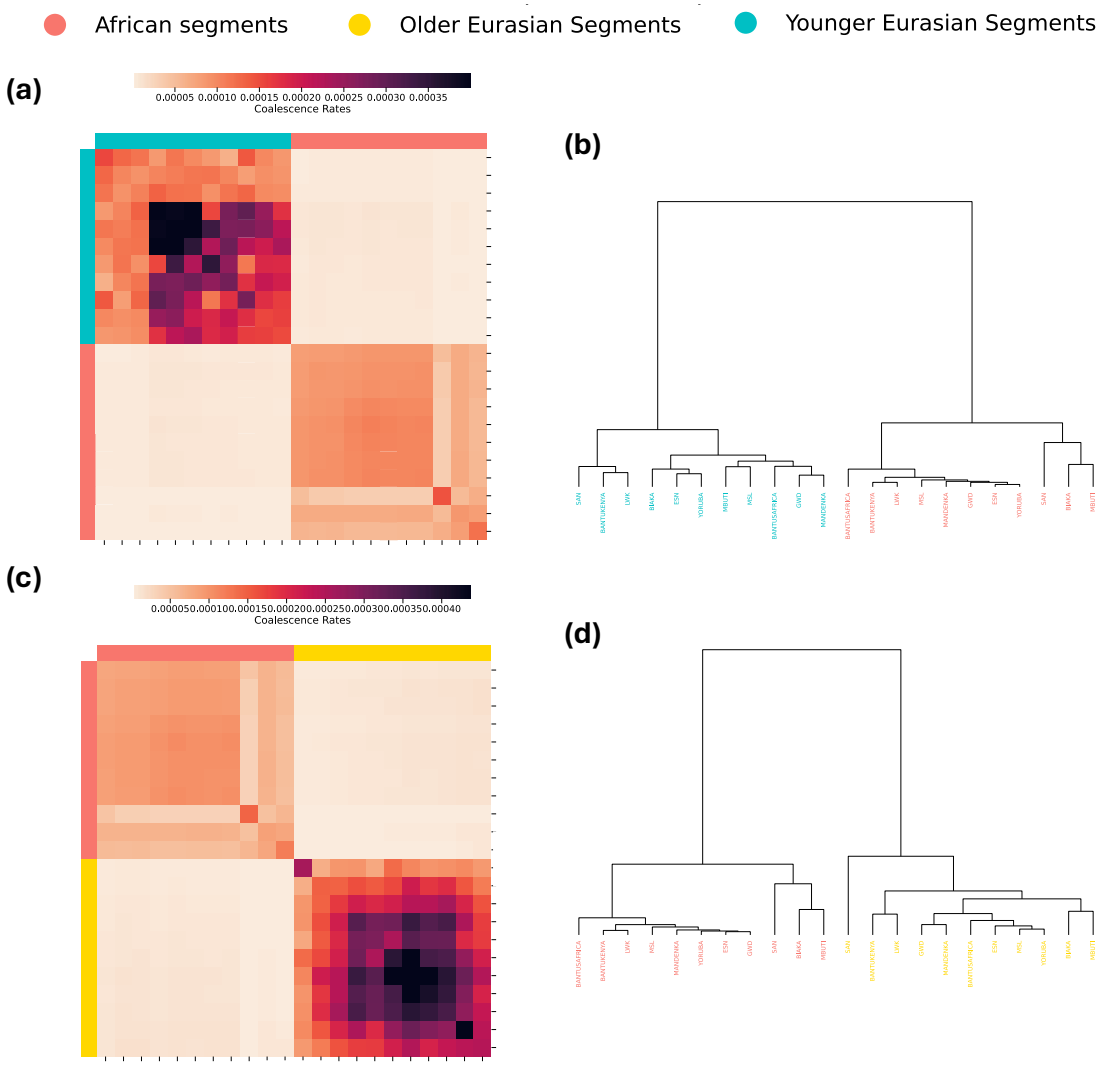


Figure B.18: Coalescence rates between Eurasian-like and non-Eurasian-like segments in Africans. Pairwise coalescence rate matrices and corresponding UPGMA dendrograms for (a-b) Younger Eurasian-like segments and (c-d) Older Eurasian-like segments. The coalescence rates correspond to the epoch spanning 10,000 to 15,848 years.



Figure B.19: Coalescence rates with modern and ancient samples. Each bar plot shows the coalescence rate for the epoch spanning 10,000–15,848 years between Eurasian-like segments from the younger admixture date and various modern and ancient samples. Only the top 25 groups are displayed. Error bars represent 95% confidence intervals, calculated from 100 bootstrap samples.

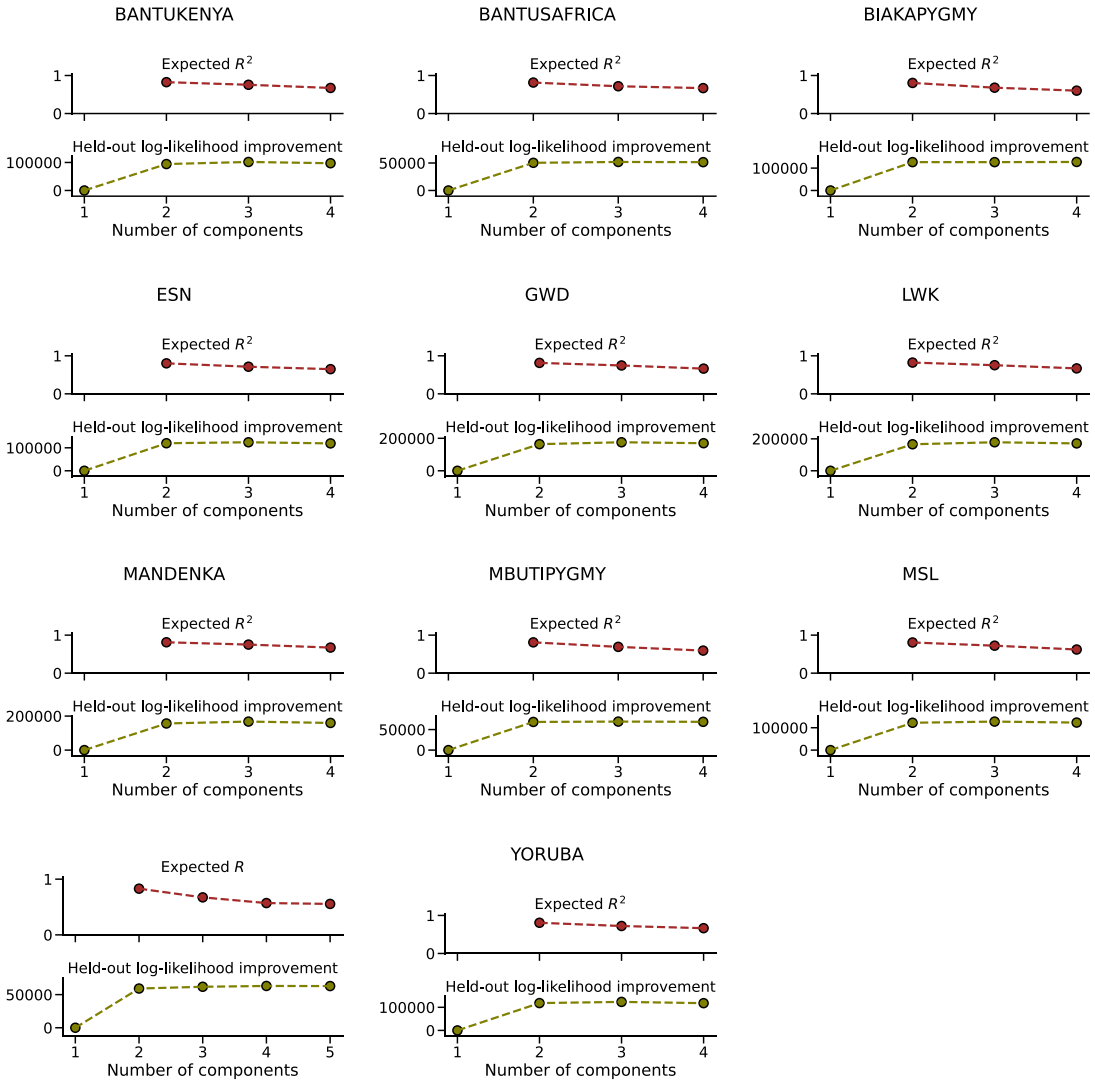


Figure B.20: Held-out likelihood and expected R^2 for the deep admixture event across African populations. The held-out likelihood was calculated by fitting the EM algorithm on chromosome 1 and testing on chromosome 2. The number of components varied between 1 and 4 or 5.

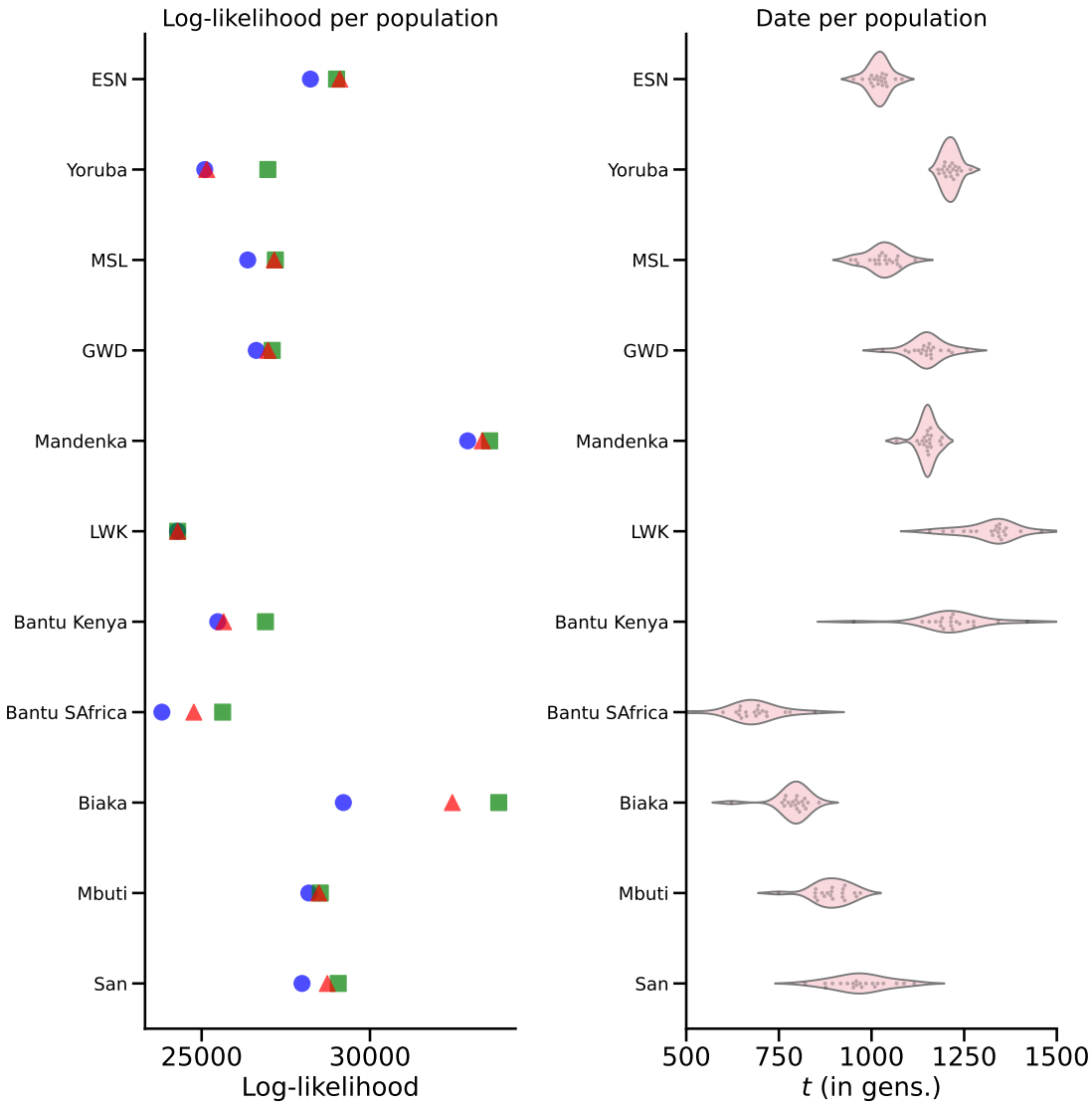


Figure B.21: Dating the ancestral components of the deep admixture event. This analysis is similar to Figure 3.8, but with local ancestry segments identified as Eurasian-like in the back-to-Africa analysis masked. Additionally, regions within 5 cM of these segments were also excluded. The plot displays the log-likelihoods for one-date, two-date, and continuous admixture models across populations, as well as the inferred admixture date for the single-date admixture model. The log-likelihood values are aggregated across 20 jackknife runs. The legend remains identical to that in Figure 3.8. Admixture dates in the plot are in generations.

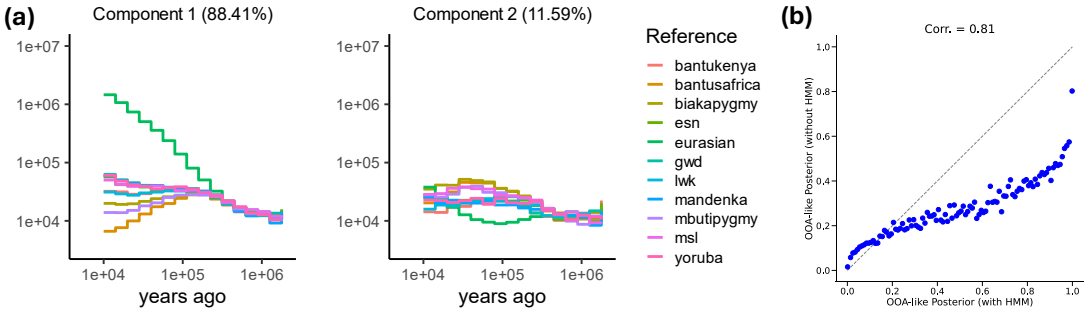


Figure B.22: Comparison of GhostBuster runs with and without the HMM. Chromosome 1 of San individuals from the HGDP+1,000GP dataset was decomposed, fixing the number of components to 2. The HMM was turned off, treating each site independently for local ancestry calculation. (a) shows the inverse coalescence rates when the HMM is turned off, and (b) compares the correlation in local ancestry inferred with and without the HMM. Plot (b) is based on binning the posterior probabilities into 500 equally spaced bins and calculating the average posterior within each bin.

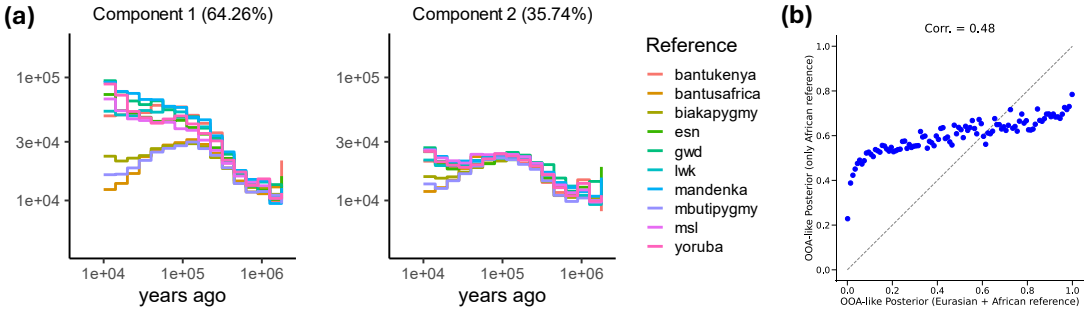


Figure B.23: Comparison of GhostBuster runs with and without Eurasian reference samples. Chromosome 1 of San individuals from the HGDP+1,000GP dataset was analyzed, fixing the number of components to 2. The Eurasian samples were excluded from the reference panel prior to ancestry decomposition. (a) shows the inverse coalescence rates when only African populations were used as the reference, and (b) compares the correlation in local ancestry inferred with and without Eurasian samples in the reference panel. Plot (b) is based on binning the posterior probabilities into 500 equally spaced bins and calculating the average posterior within each bin.

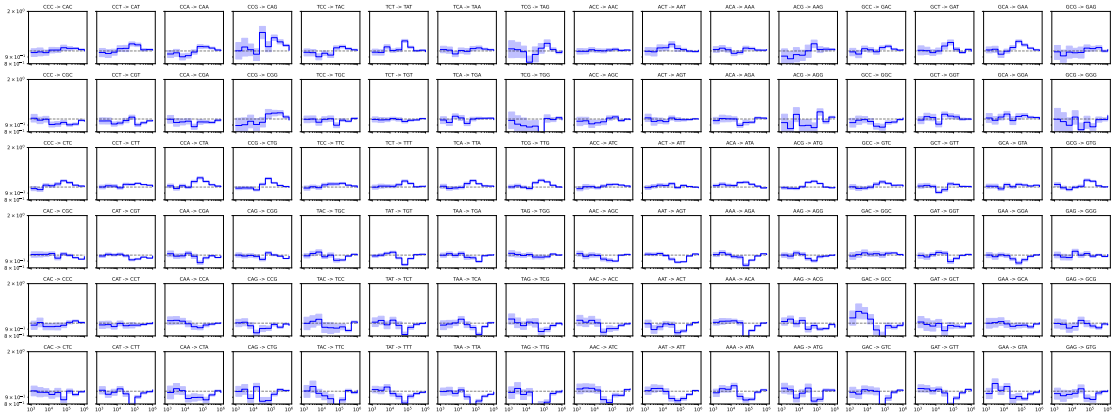


Figure B.24: Mutation count enrichment across all 96 trimer types in OOA-like segments within African individuals from the HGDP+1,000GP dataset. The y-axis represents the enrichment, calculated relative to the mutation count in non-OOA-like segments of African individuals. Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap replicates, while the dashed line represents the baseline of no enrichment.

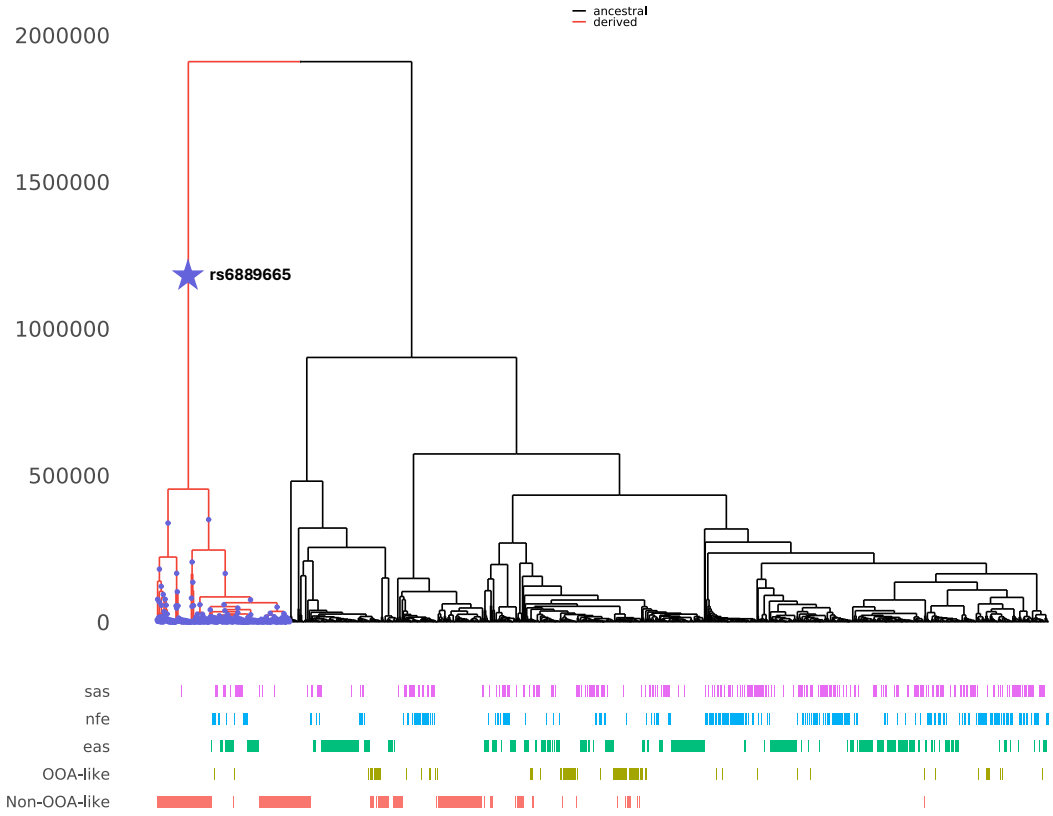


Figure B.25: Relate-inferred tree at the PRDM9 gene. Relate-inferred tree at chromosome 5:23,532,534 (hg38 build), highlighting the PRDM9-C allele (rs6889665) as a star on the lineage along with the descendant population composition. The genealogy was subset to include Africans (OOA-like and Non-OOA-like), East Asians (eas), South Asians (sas), and non-Finnish Europeans (nfe).

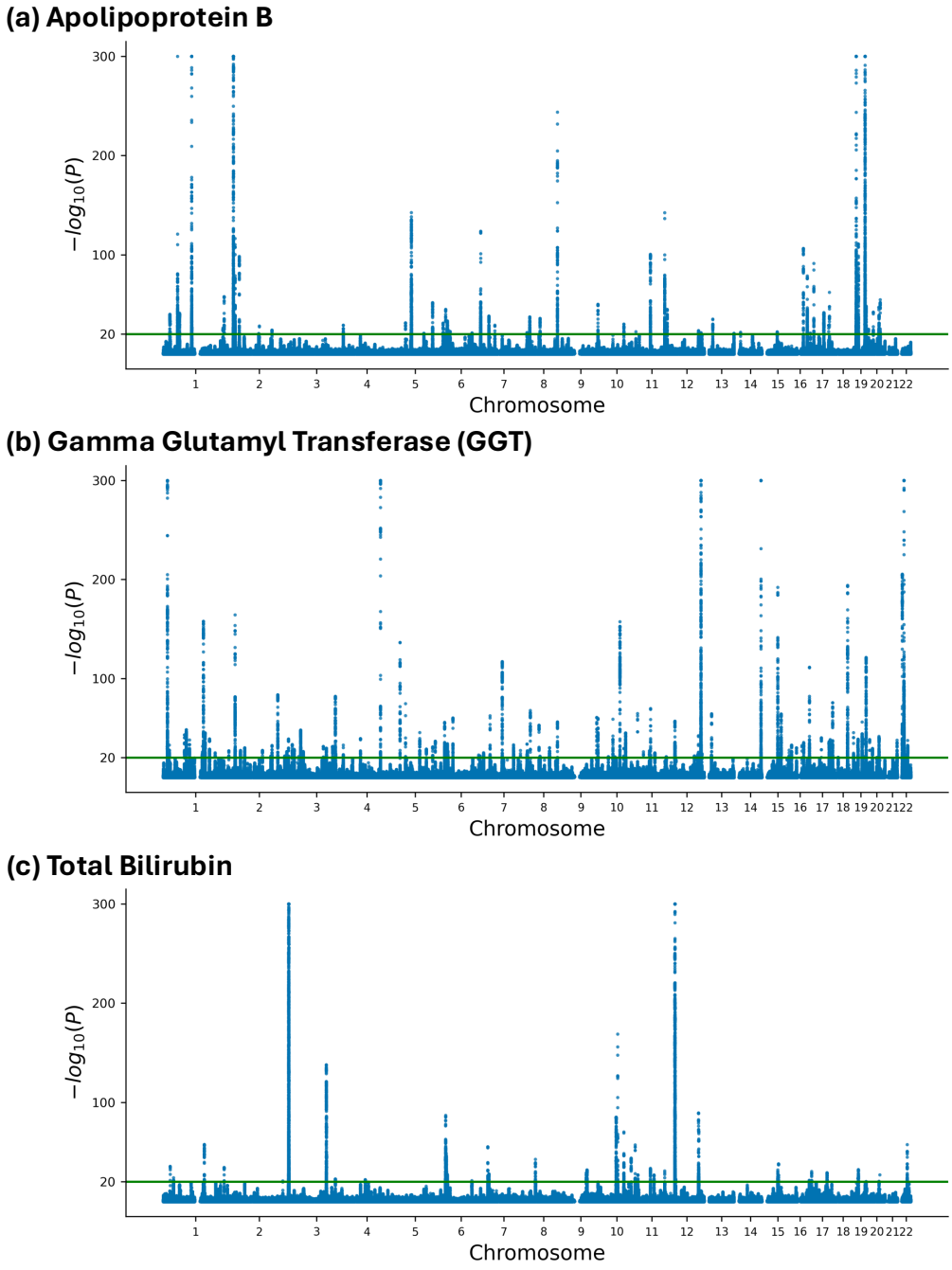


Figure B.26: Manhattan plots for some blood traits. (a) Apolipoprotein B, (b) Gamma Glutamyl Transferase (GGT), (c) Total Bilirubin. The GWAS summary statistics was calculated using Quickdraws (more details in chapter 5). The green line indicates $P = 1 \times 10^{-20}$.

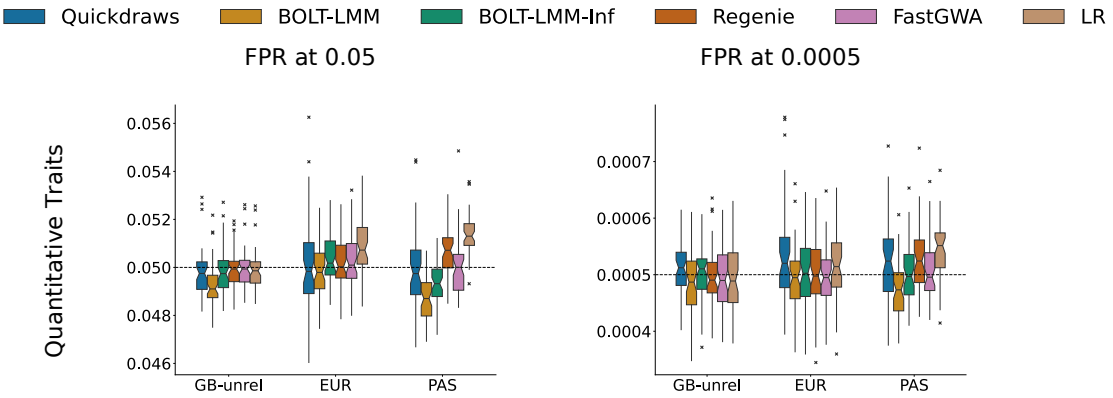


Figure B.27: Summary of calibration in simulations with varying levels of population structure. False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the fraction of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, EUR, PAS) is provided in section 5.3.1.

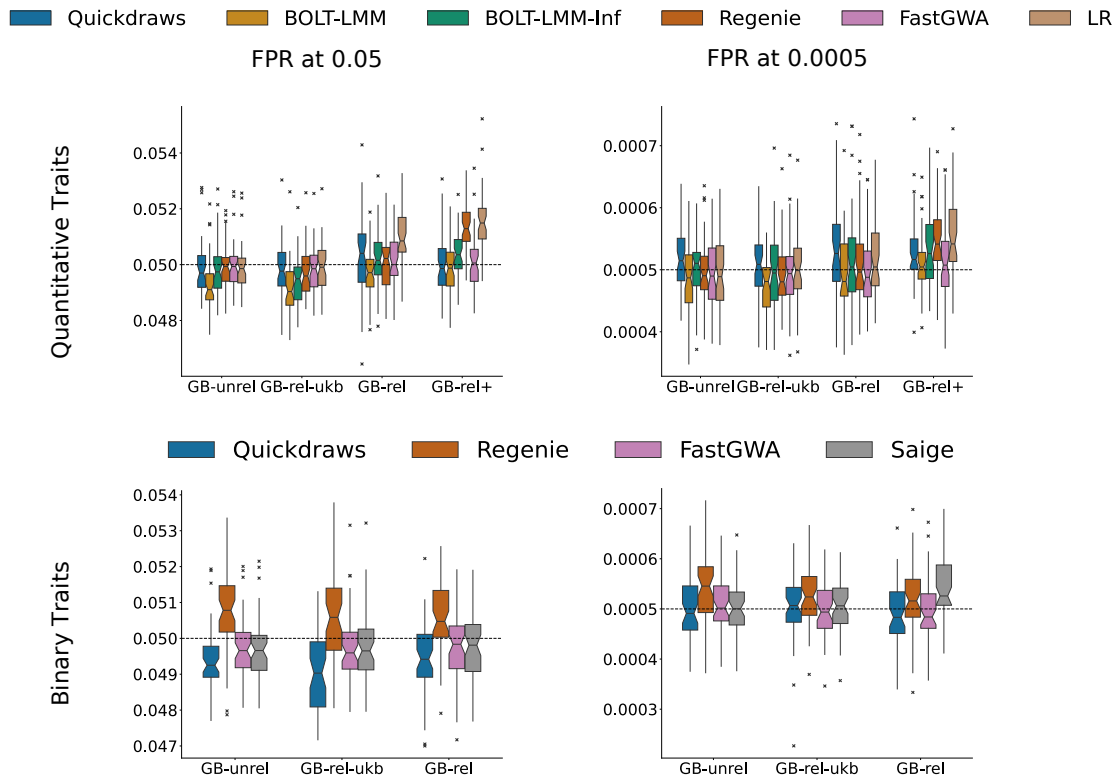


Figure B.28: Summary of calibration in simulations with varying levels of relatedness. False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the number of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. GB-unrel refers to simulations including only unrelated British individuals, GB-rel-ukb refers to randomly sampling from the related white British subset, GB-rel refers to the default relatedness setting with 3.4× more relative pairs compared to the UK Biobank, and GB-rel+ refers to the extreme relatedness case of 7.3× and 4.8× more first and second degree relative pairs compared to the UK Biobank. The prevalence of binary traits was fixed at 10%.

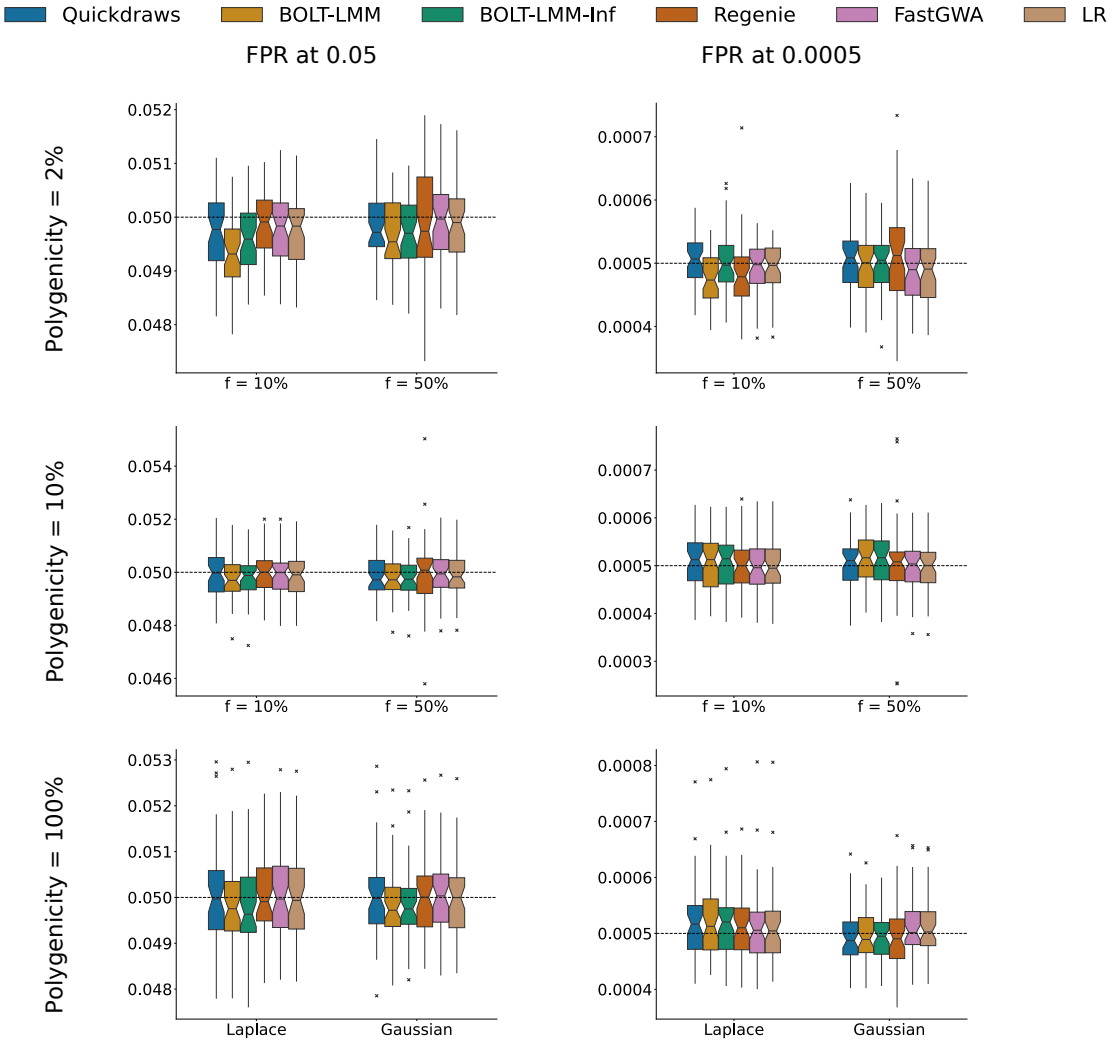


Figure B.29: Summary of calibration in simulations with varying causal effect distributions. The first two rows correspond to causal effects simulated from a mixture of two Gaussian. Polygenicity refers to the proportion corresponding to the Gaussian with higher variance, and f refers to the fraction of total variance explained by the Gaussian with lower variance. False positive rate (FPR) at a significance threshold of $\alpha \in \{0.05, 0.0005\}$, calculated as the fraction of variants on even chromosomes with p-value lower than α. The line inside each box indicates the median value, the central box indicates the interquartile (IQR) range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points.

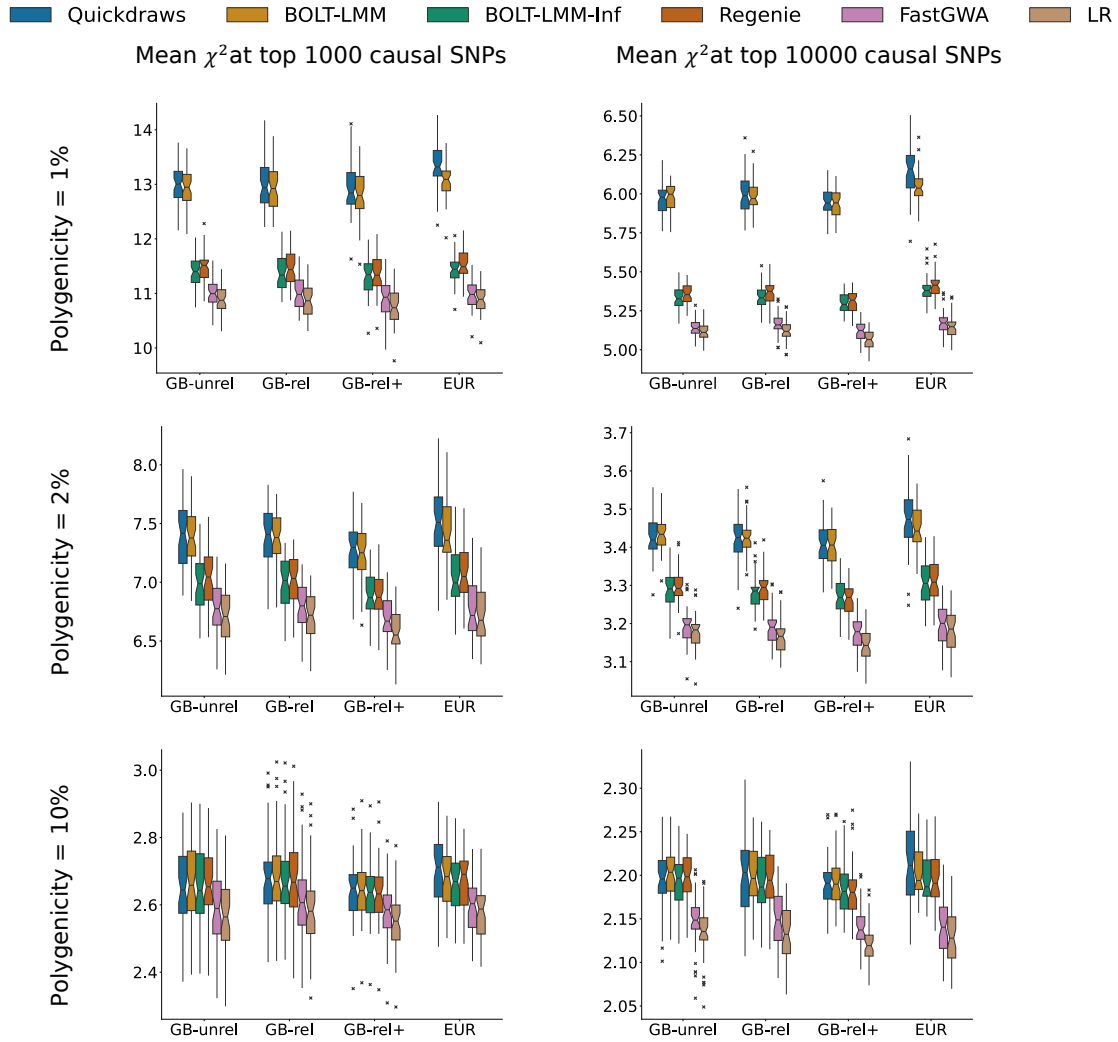


Figure B.30: Summary of statistical power in simulations for quantitative traits. We measure the normalized causal χ^2 at top 1,000 and top 10,000 causal variants. To correct for confounding, the causal χ^2 is normalized by the average χ^2 at null variants on even chromosomes. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, GB-rel+, EUR) is provided in section 5.3.1.

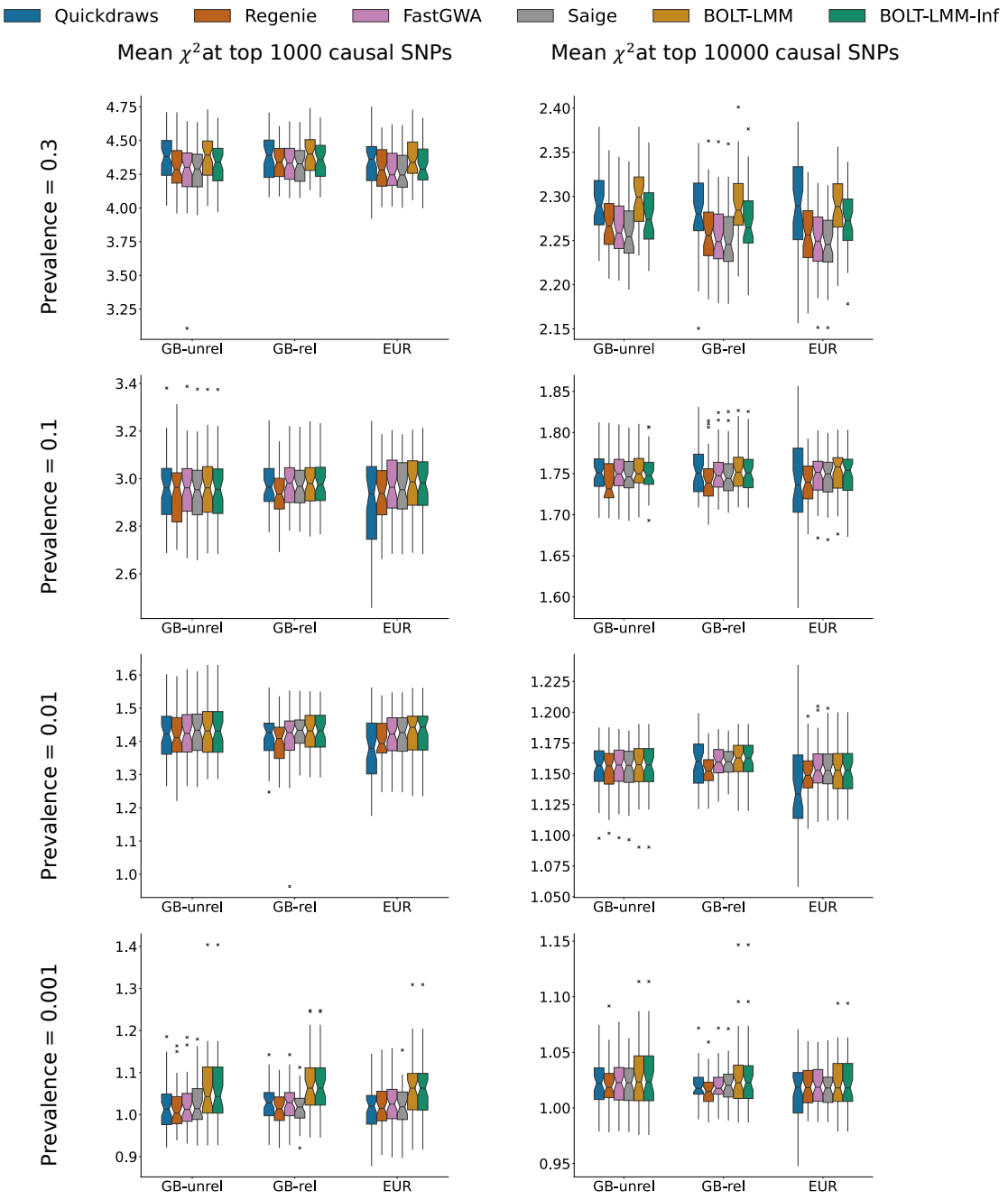


Figure B.31: Summary of statistical power in simulations for binary traits. We measure the normalized causal χ^2 at top 1,000 and top 10,000 causal variants. To correct for confounding, the causal χ^2 is normalized by the average χ^2 at null variants on even chromosomes. The line inside each box indicates the median value, the central box indicates the interquartile range, whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points. A description of the group labels (GB-unrel, GB-rel, EUR) is provided in the section 5.3.1.

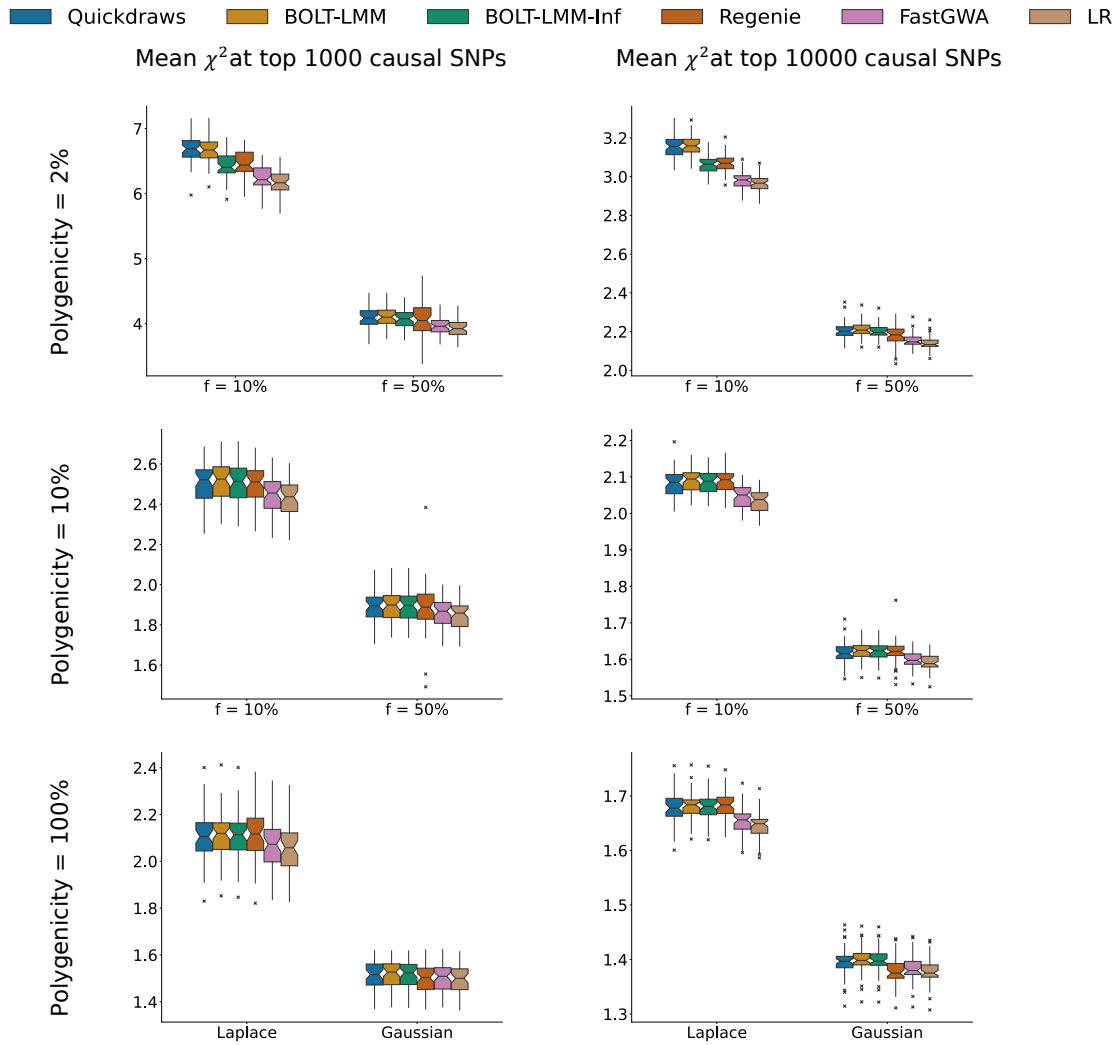


Figure B.32: Summary of statistical power in simulations with varying causal effect distributions. The first two rows correspond to causal effects simulated from a mixture of two Gaussian. Polygenicity refers the proportion corresponding to the Gaussian with higher variance, and f refers to the fraction of total variance explained by the Gaussian with lower variance. We measure the normalized causal χ^2 at top 1,000 and top 10,000 causal variants. To correct for confounding, the causal χ^2 is normalized by the average χ^2 at null variants on even chromosomes. The line inside each box indicates the median value, the central box indicates the interquartile range (IQR), whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate points.

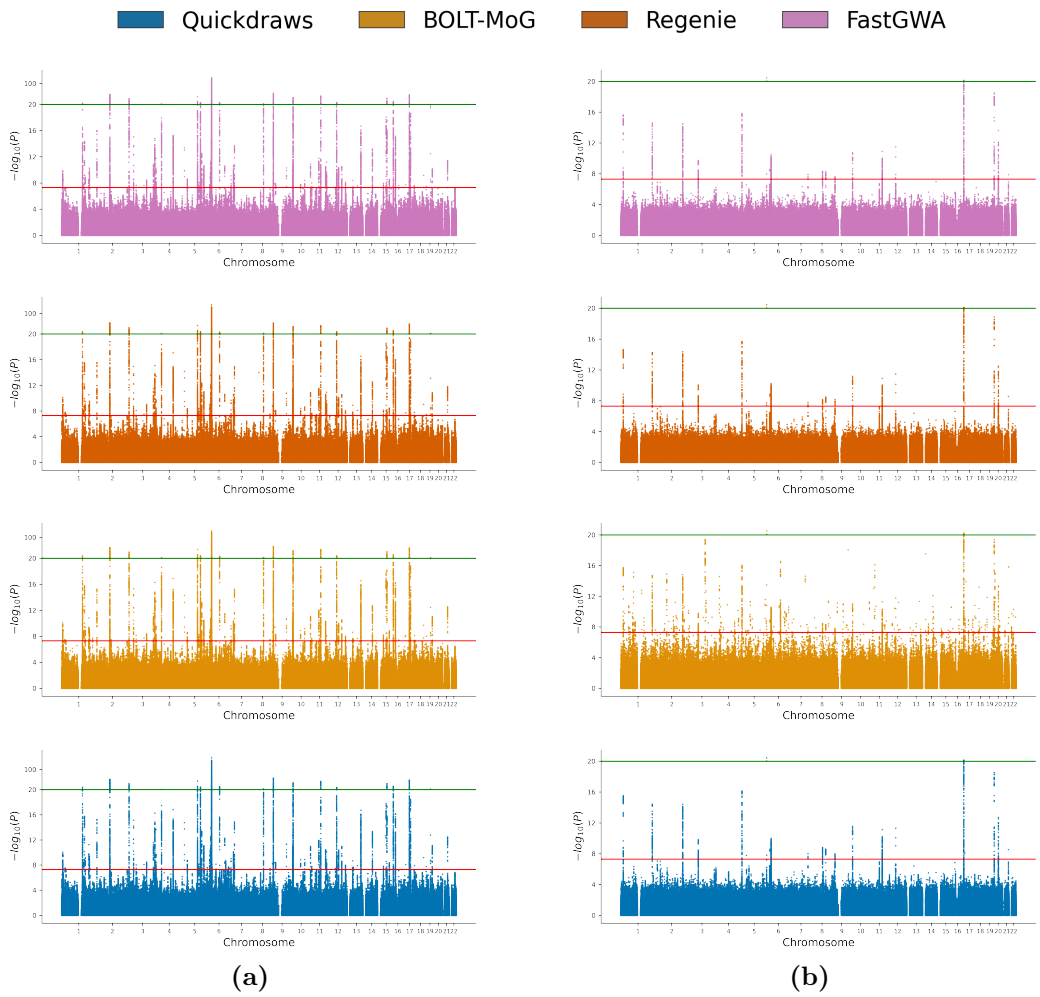


Figure B.33: Manhattan plots for two binary traits. (a) Asthma (prevalence = 0.119) and (b) Basal cell carcinoma (prevalence = 0.012). ∼13.3 million variants were tested. Red lines indicate a genome-wide significance threshold of $P = 5 \times 10^{-8}$ and the green line indicates $P = 1 \times 10^{-20}$.

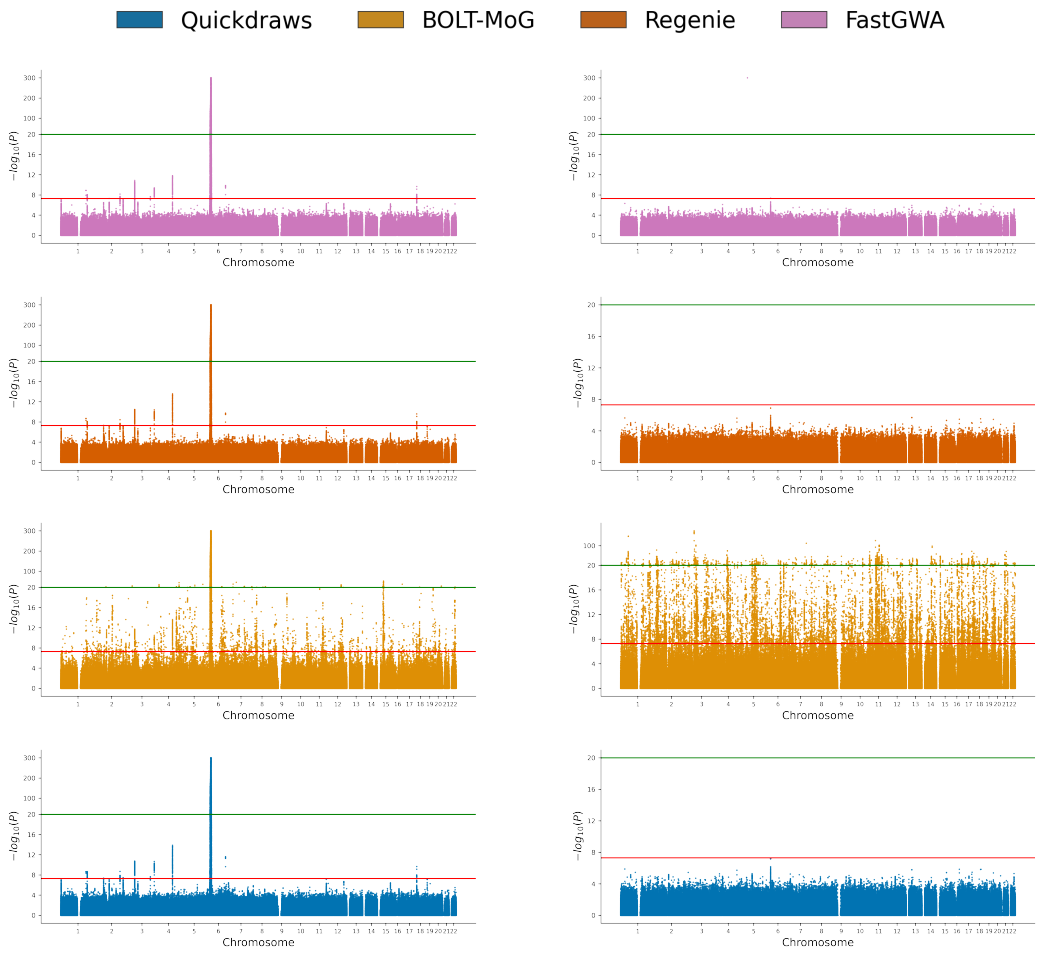


Figure B.34: Manhattan plots for two binary traits. (a) Celiac disease (prevalence = 0.0047) and (b) Vitiligo (prevalence = 0.0005). ∼13.3 million variants were tested. Red lines indicate a genome-wide significance threshold of $P = 5 \times 10^{-8}$ and the green line indicates $P = 1 \times 10^{-20}$.

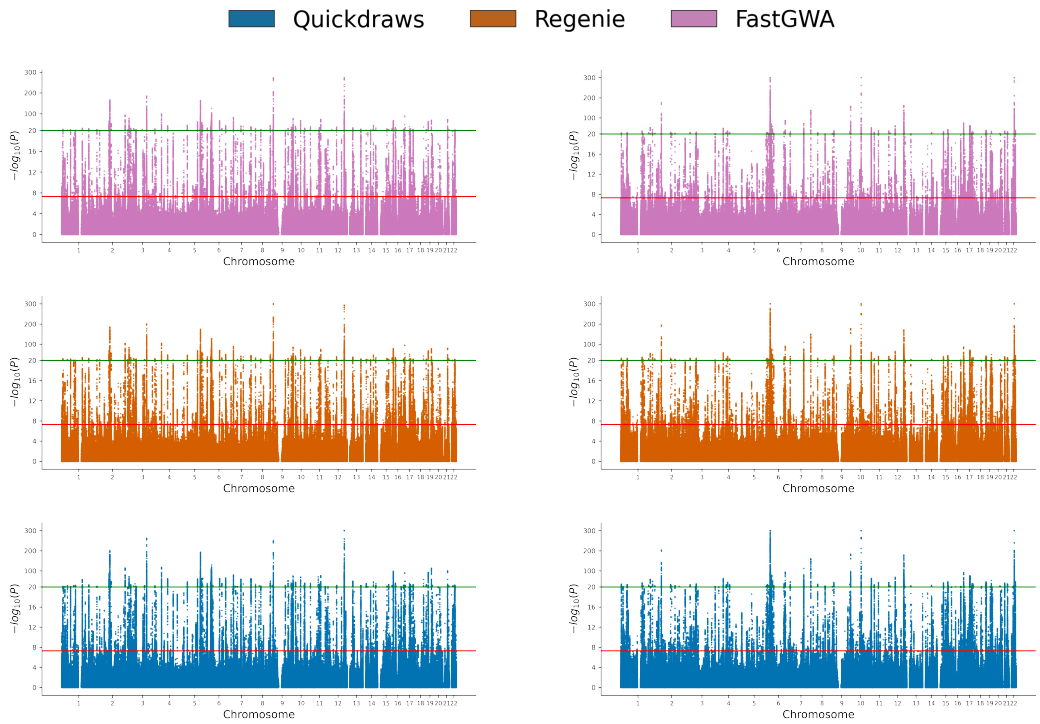
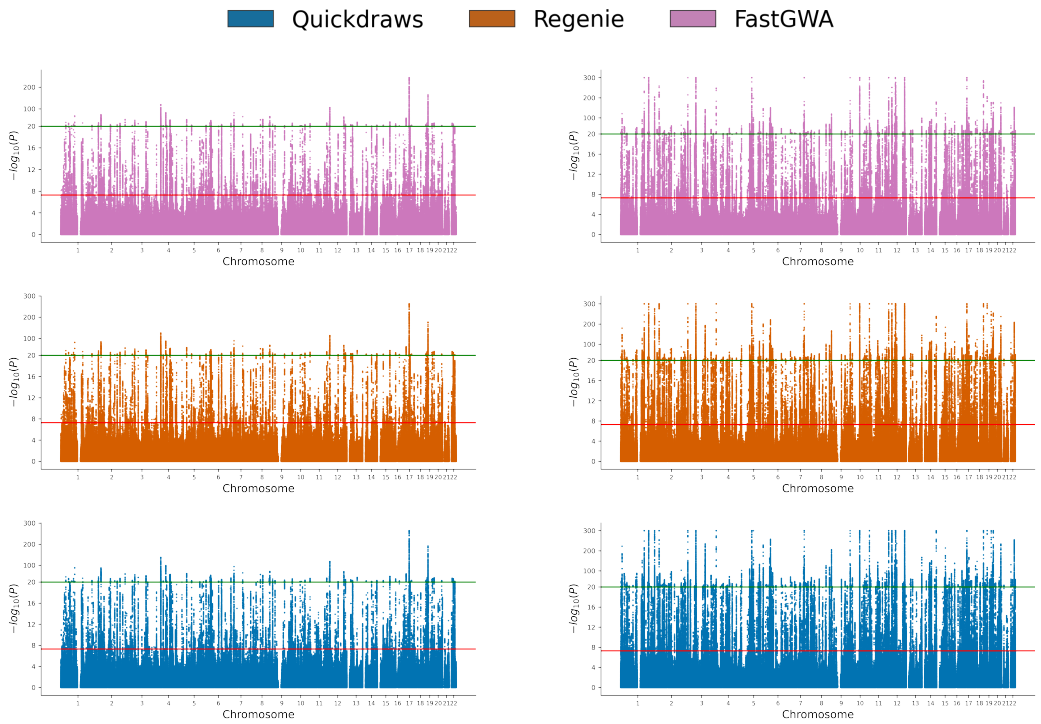


Figure B.35: Manhattan plots for two quantitative traits. (a) Eosinophil count (RHE-me $h^2 = 0.213$) and (b) Haemoglobin concentration (RHE-mc $h^2 = 0.158$). ∼13.3 million variants were tested. Red lines indicate a genome-wide significance threshold of $P = 5 \times 10^{-8}$ and the green line indicates $P = 1 \times 10^{-20}$.



(a) **(b)**

Figure B.36: Manhattan plots for two quantitative traits. (a) Lymphocyte percentage (RHE-mc $h^2 = 0.196$) and (b) Mean platelet volume (RHE-mc $h^2 = 0.448$). ∼13.3 million variants were tested. Red lines indicate a genome-wide significance threshold of $P = 5 \times 10^{-8}$ and the green line indicates $P = 1 \times 10^{-20}$.

B.2 Additional tables

Table B.1: Ancient samples built in the genealogy. Ancient sample IDs, sampling age (in years), country of origin and genomic coverage

ID	Date	Country	Coverage
AltaiNeandertal	110,450	Russia	52.0
Vindija	41,950	Croatia	30.0
Chagyrskaya	80,000	Russia	27.0
Denisova	63,900	Russia	31.0
WC1	9,219	Iran	10.4
PB675	5,416	Ireland	15.9
JP14	5,497	Ireland	16.6
SRA62	5,979	Ireland	13.2
rath1	3,906	Ireland	11.6
bally	5,132	Ireland	11.3
Yamnaya	4,890	Kazakhstan	25.2
BOT2016	5,450	Kazakhstan	13.6
Ust_Ishim	45,020	Russia	42.0
SF12	8,895	Sweden	65.2
KH150622	3,350	Germany	18.2
KH150623	3,350	Germany	17.1
KH150620	3,350	Germany	16.3
KK1	9,678	Georgia	11.3
I0018	7,140	Germany	19.0
I0001	8,025	Luxembourg	20.1
I10871	7,890	Cameroon	18.5
I5950	4,472	Ethiopia	11.3
Klein7	7,122	Austria	11.3
Asp6	7,524.5	Austria	12.1
Ess7	6,975	Germany	12.3
Herx	7,078.5	Germany	11.5
Dil16	7,116.5	Germany	10.6
Nea2	8,098	Greece	12.5
Nea3	8,183.5	Greece	11.6
VC3-2	7,495.5	Serbia	11.2
STAR1	7,532.5	Serbia	10.6
LEPE52	7,812	Serbia	12.4
LEPE48	7,939.5	Serbia	10.9
VLASA7	8,552	Serbia	15.2
VLASA32	9,604.5	Serbia	12.7
Bar25	8,294.5	Turkey	12.7
AKT16	8,547.5	Turkey	12.3

Continued on next page

Table B.1: *(Continued)*

ID	Date	Country	Coverage
USR1	11,435	USA	17.0
Sumidouro5	10,405	Brazil	13.6
Lovelock3	627	USA	18.1
Lovelock2	1,880	USA	14.4
AHUR_2064	10,970	USA	17.9
Anzick	12,649	USA	14.0
new001	418	South Africa	11.1
ela001	493	South Africa	10.5
baa001	1,909	South Africa	12.9
2H11	5,163	France	23.6
2H10	5,166	France	13.8
Kolyma1	9,786	Russia	15.0
Yana1	31,950	Russia	24.3
Sunghir3	34,093	Russia	12.0
atp016	4,971	Spain	15.3
atp002	4,746	Spain	11.3

Table B.2: Admixture date estimates for the back to Africa admixture event in Africans. Estimated dates for the younger admixture event based on two-date co-ancestry curve fitting, measured in generations. The lower and upper bounds represent the 95% confidence intervals derived from 20 jackknife replicates during the co-ancestry curve fitting process.

Population	Mean	Lower bound	Upper bound
Masai	57.7	49.5	66.7
Bantu Kenya	103.4	93.2	113.6
LWK	108.9	100.3	117.8
Luo	32.7	16.3	68.1
Khomani San	14.9	1	41.4
Ju Hoan North	105.2	82.1	126.8
GWD	130.1	125.3	141.9
Mandenka	130.6	119.2	138.9
San	81.8	75.2	88.9
Dinka	108.0	87.9	213.5
Bantu SAfrica	162.7	149.4	176.7
MSL	148.9	143.1	160.1
ESN	187.5	178.2	227.7
Yoruba	233.6	214.3	261.3
Mbuti	175.8	134.2	254.3
Biaka	165.7	143.4	192.2

Table B.3: Admixture date estimates for the deep admixture event in Africans. Estimated dates for the deep admixture event based on two-date co-ancestry curve fitting, measured in generations. The lower and upper bounds represent the 95% confidence intervals derived from 20 jackknife replicates during the co-ancestry curve fitting process.

Population	Mean	Lower bound	Upper bound
Dinka	1,353.7	1,270.1	1,433.8
Luo	1,181.2	1,139.5	1,205.8
ESN	1,384.2	1,354.5	1,436.4
Yoruba	1,402.8	1,377.4	1,446.0
Masai	841.3	792.0	899.1
MSL	1,362.7	1,338.6	1,396.8
GWD	1,343.1	1,312.2	1,367.3
Mandenka	1,253.0	1,225.6	1,283.9
LWK	1,194.2	1,168.5	1,211.1
Bantu Kenya	1,181.7	1,153.4	1,205.6
Bantu SAfrica	1,229.7	1,210.5	1,249.2
Biaka	1,233.7	1,221.7	1,251.3
Mbuti	1,196.0	1,169.3	1,217.5
Ju Hoan North	1,008.1	957.9	1,067.8
San	1,021.2	989.6	1,051.7
Khomani San	819.5	759.5	909.4

Table B.4: False positive rates (FPR) for FastGWA in quantitative trait simulations with up to 460,000 self-identified Europeans. The table presents the FPR for FastGWA under various sample sizes, controlling for residual population stratification by adjusting the number of included principal components (PCs) or by omitting them entirely. The polygenicity of the simulated trait is fixed at 1%. S.E. refers to the standard error, and * indicates a Bonferroni-significant t-test p-value. The significance threshold for the FPR was set at 0.005. Each simulation condition varied one of the following baseline parameters: sample size (N = 460k), ancestry differences explaining 5% of phenotypic variance, and the use of 10 PCs to correct for stratification.

Condition	Mean FPR	S.E. FPR	t-test p-value
Varying sample sizes			
N = 50k	0.00503	2.99×10^{-5}	0.23
N = 100k	0.00514	4.35×10^{-5}	0.0009
N = 460k	0.00586	4.11×10^{-5}	8×10^{-38} *
Varying pop. stratification			
5%	0.00586	4.11×10^{-5}	8×10^{-38} *
0%	0.00509	2.94×10^{-5}	0.0019
Varying #PCs			
3 PCs	0.01424	3.23×10^{-4}	2×10^{-48} *
10 PCs	0.00586	4.11×10^{-5}	8×10^{-38} *
20 PCs	0.00531	3.61×10^{-5}	1.8×10^{-13} *

Table B.5: Quantitative traits analyzed. Quantitative traits analyzed, including RHE-MC h^2 estimates and sample sizes (N).

UKBB ID	Phenotype	h^2	N
30850	Testosterone	0.03	355373
1438	Bread intake	0.053	396396
874	Duration of walks	0.054	346657
30740	Glucose	0.088	359138
2217	Age started wearing glasses	0.092	349337
30060	Mean corp. haemoglobin conc.	0.101	398336
30810	Phosphate	0.149	358861
30780	LDL direct	0.152	390888
3064	Peak expiratory flow (PEF)	0.152	370024
30690	Cholesterol	0.157	391567
30020	Haemoglobin conc	0.158	398342
30030	Haematocrit %	0.161	398343
46	Hand grip strength (left)	0.162	403800
47	Hand grip strength (right)	0.164	403850
30600	Albumin	0.168	359528
30670	Urea	0.172	391307
30680	Calcium	0.177	359391
30650	Asp-aminotransferase	0.178	390183
30860	Total protein	0.188	359122
4079	Diastolic blood pressure	0.188	379181
30640	Apolipoprotein B	0.189	389718
30200	Neutrophill %	0.19	397665
4080	Systolic blood pressure	0.192	379174
102	Pulse rate	0.194	379181
30180	Lymphocyte %	0.196	397665
30700	Creatinine	0.198	391377
30830	SHBG	0.2	356165
30880	Urate	0.206	391102
30630	Apolipoprotein A	0.213	357411
30150	Eosinophill count	0.213	397661
30790	Lipoprotein A	0.22	312270
30140	Neutrophill count	0.22	397661
30130	Monocyte %	0.221	397665
30870	Triglycerides	0.234	391260
30730	Gamma glutamyltransferase	0.24	391372
30710	C-reactive protein	0.241	390752
20257	Forced vital capacity (FVC) Z-score	0.242	328403
30750	Glycated haemoglobin	0.245	391597
30760	HDL cholesterol	0.248	359370
30210	Eosinophill %	0.251	397665
30130	Monocyte count	0.253	397661
30280	Immature reticulocyte fraction	0.26	391990

Continued on next page

Table B.5: *(Continued)*

UKBB ID	Phenotype	h^2	N
30250	Reticulocyte count	0.26	392210
20256	Forced expiratory volume (FEV1) Z-score	0.263	328403
30070	Red blood cell distribution width	0.264	398342
30260	Mean reticulocyte vol	0.266	392211
30240	Reticulocyte %	0.267	392210
20258	FEV1/FVC ratio Z-score	0.269	328403
30120	Lymphocyte count	0.269	397661
48	Waist circumference	0.277	404969
30010	Red blood cell count	0.279	398343
30610	Alkaline phosphatase	0.279	391586
30300	High light scatter reticulocyte count	0.289	391989
30770	IGF-1	0.29	389525
30270	Mean sphered cell vol	0.293	391990
23115	Leg fat % (left)	0.294	399130
23111	Leg fat % (right)	0.294	399151
30110	Platelet distribution width	0.296	398153
23127	Trunk fat %	0.302	398940
23119	Arm fat % (right)	0.302	399098
23123	Arm fat % (left)	0.305	399037
30290	High light scatter reticulocyte %	0.306	391990
23099	Body fat %	0.310	398955
49	Hip circumference	0.318	404972
30090	Platelet crit	0.326	398154
30840	Total bilirubin	0.327	389995
23110	Impedance of arm (left)	0.331	399149
23109	Impedance of arm (right)	0.331	399133
23128	Trunk fat mass	0.331	398918
21001	Body mass index (BMI)	0.348	405010
23106	Impedance of whole body	0.367	399144
30080	Platelet count	0.373	398339
21002	Weight	0.381	405018
30040	Mean corp. vol	0.414	398342
23105	Basal metabolic rate	0.419	399169
23102	Whole body water mass	0.428	399179
30050	Mean corp. haemoglobin	0.432	398339
30100	Mean platelet vol	0.448	398334
50	Standing height	0.745	405035

Table B.6: Self-reported disease traits analyzed. Self-reported disease traits analyzed, including their disease code and prevalence.

UKBB ID	Disease ID	Phenotype	Prevalence
20002	1065	Hypertension	0.2740
20002	1473	High cholestrol	0.1400
20002	1111	Asthma	0.1190
20002	1465	Osteoarthritis	0.0946
20002	1387	Hayfever	0.0667
20002	1286	Depression	0.0649
20002	1138	Gastric reflux	0.0529
20002	1226	Hypothyroidism	0.0518
20002	1220	Diabetes	0.0421
20002	1265	Migraine	0.0343
20002	1074	Angina	0.0335
20002	1452	Eczema	0.0320
20002	1474	Hiatus hernia	0.0266
20002	1075	Heart attack	0.0247
20001	1002	Breast cancer	0.0241
20002	1094	Deep venous thrombosis	0.0210
20002	1162	Cholelithiasis	0.0188
20002	1294	Back problem	0.0183
20002	1396	Enlarged prostate	0.0183
20002	1309	Osteoporosis	0.0182
20002	1287	Anxiety	0.0182
20002	1351	Uterine fibroids	0.0170
20002	1538	Arthritis nos	0.0148
20002	1113	Emphysema	0.0142
20002	1458	Diverticulitis	0.0137
20002	1453	Psoriasis	0.0131
20002	1277	Glaucoma	0.0125
20001	1061	Basal cell carcinoma	0.0120
20001	1044	Prostate cancer	0.0097
20002	1223	Type 2 diabetes	0.0093
20002	1197	Bladder stone	0.0092
20002	1093	Pulmonary embolism	0.0088
20001	1059	Malignant melanoma	0.0087
20002	1225	Hyperthyroidism	0.0084
20002	1202	Urinary frequency	0.0069
20002	1353	Vaginal prolapse	0.0069
20002	1330	Iron deficiency anaemia	0.0065
20002	1463	Ulcerative colitis	0.0056
20002	1417	Nasal polyps	0.0050
20002	1295	Joint disorder	0.0049
20002	1456	Coeliac disease	0.0047

Continued on next page

Table B.6: *(Continued)*

UKBB ID	Disease ID	Phenotype	Prevalence
20002	1112	Chronic obstructive airways disease	0.0045
20002	1281	Retinal detachment	0.0041
20002	1123	Sleep apnoea	0.0038
20002	1331	Pernicious anaemia	0.0033
20002	1462	Crohns disease	0.0032
20002	1291	Bipolar disorder	0.0028
20002	1446	Gout	0.0026
20002	1661	Vitiligo	0.0005
20002	1430	Hypopituitarism	0.0004

Table B.7: Number of independent associated loci for quantitative traits. Number of independent associated loci for quantitative traits after Plink clumping using summary statistics from FastGWA, Regenie, BOLT-LMM, and Quickdraws

Phenotype	FastGWA	Regenie	BOLT-MoG	Quickdraws
Testosterone	72	65	83	40
Bread intake	5	5	6	5
Duration of walks	0	0	1	0
Glucose	87	92	94	93
Age started wearing glasses	30	30	29	32
Mean corp. haemoglobin conc.	82	85	86	86
Phosphate	131	140	149	140
LDL direct	151	165	170	164
Peak expiratory flow (PEF)	82	85	83	85
Cholesterol	181	192	197	191
Haemoglobin conc	306	341	363	356
Haematocrit %	276	305	317	316
Hand grip strength (left)	75	78	79	84
Hand grip strength (right)	76	89	91	91
Albumin	198	216	222	220
Urea	145	158	161	156
Calcium	178	196	205	198
Asp-aminotransferase	254	271	282	276
Total protein	250	277	296	290
Diastolic blood pressure	129	143	151	146
Apolipoprotein B	178	205	214	206
Neutrophill %	266	290	304	306
Systolic blood pressure	150	157	162	154
Pulse rate	171	182	186	187
Lymphocyte %	305	332	344	345
Creatinine	351	400	418	403
SHBG	275	334	348	328
Urate	228	262	272	265
Apolipoprotein A	252	306	319	294
Eosinophill count	299	334	345	356
Lipoprotein A	54	65	70	56
Neutrophill count	299	330	341	356
Monocyte %	362	416	437	449

Continued on next page

Table B.7: *(Continued)*

Phenotype	FastGWA	Regenie	BOLT-MoG	Quickdraws
Triglycerides	234	287	288	294
Gamma glutamyltransferase	287	323	330	327
C-reactive protein	196	227	233	227
Forced vital capacity (FVC)	163	185	203	193
Glycated haemoglobin	345	400	418	408
HDL cholesterol	280	353	371	352
Eosinophill %	375	430	463	461
Monocyte count	402	470	490	494
Immature reticulocyte fraction	234	258	262	264
Reticulocyte count	311	361	372	367
Forced expiratory volume (FEV1)	189	218	228	228
Red blood cell distribution width	368	423	443	425
Mean reticulocyte vol	395	439	463	460
Reticulocyte %	307	353	363	361
FEV1/FVC ratio	272	301	323	307
Lymphocyte count	353	406	422	420
Waist circumference	186	206	214	212
Alkaline phosphatase	355	446	473	457
Red blood cell count	409	482	507	508
High light scatter reticulocyte count	325	373	393	390
IGF-1	364	435	459	445
Mean sphered cell vol	378	430	456	456
Leg fat % (left)	197	222	233	233
Leg fat % (right)	204	220	229	235
Platelet distribution width	456	549	576	591
Trunk fat %	197	224	238	237
Arm fat % (right)	215	240	258	251
Arm fat % (left)	206	242	256	246
High light scatter reticulocyte %	324	369	398	399
Body fat %	215	242	259	255
Hip circumference	255	292	308	306
Platelet crit	485	587	623	626
Total bilirubin	146	203	206	199
Impedance of arm (left)	257	292	309	307
Impedance of arm (right)	301	368	387	366
Trunk fat mass	302	367	384	373
Body mass index (BMI)	262	309	334	340
Impedance of whole body	354	451	488	452
Platelet count	539	670	720	721
Weight	329	398	428	423
Mean corp. vol	487	580	599	605
Basal metabolic rate	465	545	592	584
Whole body water mass	489	587	638	627
Mean corp. haemoglobin	429	538	562	569
Mean platelet vol	618	821	892	887
Standing height	1022	1327	1452	1674
Total	21,380	24,995	26,368	26,236

Table B.8: Number of independent associated loci for binary traits. Number of independent associated loci for binary traits after Plink clumping using summary statistics from FastGWA, SAIGE, Regenie, and Quickdraws. We excluded traits (enlarged prostate, prostate cancer, vaginal prolapse, and uterine fibroids) which led to matrix inversion errors or non-convergence of Firth logistic regression, or for which no associations were found using any method.

Phenotype	**FastGWA**	**SAIGE**	**Regenie**	**Quickdraws**
Hypertension	155	156	170	172
High cholesterol	52	52	56	57
Asthma	65	65	63	63
Osteoarthritis	2	2	2	2
Hayfever	20	21	20	20
Depression	1	1	1	1
Gastric reflux	1	1	1	1
Hypothyroidism	80	81	81	84
Diabetes	36	36	36	38
Migraine	9	9	8	9
Angina	8	9	9	8
Eczema	13	13	13	13
Hiatus hernia	0	0	0	0
Heart attack	9	8	9	9
Breast cancer	11	11	11	11
Deep venous thrombosis	9	9	9	9
Cholelithiasis	11	11	12	11
Osteoporosis	8	8	9	8
Diverticulitis	1	1	1	1
Psoriasis	21	21	18	25
Glaucoma	6	6	6	6
Basal cell carcinoma	12	12	12	12
Type 2 diabetes	2	2	2	2
Bladder stone	1	1	1	1
Pulmonary embolism	5	5	5	6
Malignant melanoma	3	3	2	3
Hyperthyroidism	12	12	11	12
Ulcerative colitis	8	8	8	8
Nasal polyps	5	5	6	6
Celiac disease	24	26	30	32
Chronic obstructive airways disease	1	1	0	1
Pernicious anaemia	1	1	2	2
Crohns disease	2	2	2	3
Total	594	599	616	636

Table B.9: Number of individual variants replicated using trait-specific summary statistics. Number of individual variants replicated using summary statistics for Crohn's disease [263], Type 2 Diabetes [264], Celiac disease [265], Depression [266], and Ulcerative Colitis [267]. We use a discovery threshold of 5×10^{-8} and vary the replication threshold from 5×10^{-2} to 5×10^{-6}, and compare summary statistics from FastGWA, Regenie, and Quickdraws.

Phenotype	Method	Repl. thresh. $= 5 \times 10^{-2}$		Repl. thresh. $= 5 \times 10^{-4}$		Repl. thresh. $= 5 \times 10^{-6}$	
		# repl.	repl. ratio.	# repl.	repl. ratio.	# repl.	repl. ratio.
Crohns disease	fastGWA	29	1	29	1	29	1
Crohns disease	Regenie	29	1	29	1	29	1
Crohns disease	Quickdraws	29	1	29	1	29	1
Type 2 diabetes	fastGWA	2	1	2	1	2	1
Type 2 diabetes	Regenie	2	1	2	1	2	1
Type 2 diabetes	Quickdraws	2	1	2	1	2	1
Coeliac disease	fastGWA	18	1	18	1	17	0.94
Coeliac disease	Regenie	20	1	20	1	19	0.95
Coeliac disease	Quickdraws	23	1	23	1	22	0.96
Depression	fastGWA	1	1	1	1	1	1
Depression	Regenie	0	0	0	0	0	0
Depression	Quickdraws	2	1	2	1	2	1
Ulcerative colitis	fastGWA	24	1	24	1	24	1
Ulcerative colitis	Regenie	25	1	25	1	25	1
Ulcerative colitis	Quickdraws	23	1	23	1	23	1
Total	fastGWA	74	1	74	1	73	0.99
Total	Regenie	76	1	76	1	75	0.99
Total	Quickdraws	79	1	79	1	78	0.99

Table B.10: Number of loci replicated in Biobank Japan. Number of loci replicated in Biobank Japan using summary statistics from FastGWA, Regenie, and Quickdraws. We exclude traits that did not yield any associations using any method and which do not have a matching phenotype in Biobank Japan.

Phenotype	**FastGWA**		**Regenie**		**Quickdraws**	
	# repl.	repl. ratio.	# repl.	repl. ratio.	# repl.	repl. ratio.
Eosinophill count	356	0.224	389	0.215	403	0.217
Lymphocyte count	428	0.207	471	0.200	475	0.190
Mean corp. haemoglobin	756	0.267	943	0.245	1002	0.235
Mean corp. haemoglobin conc.	138	0.280	133	0.271	137	0.273
Mean corp. volume	824	0.283	943	0.257	1010	0.253
Monocyte count	440	0.205	526	0.209	569	0.203
Neutrophill count	462	0.227	490	0.220	513	0.220
Platelet count	865	0.278	993	0.265	1111	0.250
Red blood cell erythrocyte count	714	0.278	786	0.263	850	0.260
Albumin	255	0.236	286	0.232	287	0.233
Alkaline phosphatase	497	0.187	632	0.183	649	0.179
Aspartate aminotransferase	349	0.223	380	0.227	372	0.215
C-reactive protein	262	0.144	286	0.138	339	0.152
Calcium	255	0.205	263	0.197	260	0.194
Cholesterol	240	0.229	267	0.224	264	0.208
Creatinine	626	0.281	659	0.267	652	0.267
Gamma glutamyltransferase	442	0.245	495	0.239	515	0.234
Glucose	101	0.210	101	0.208	112	0.192
Glycated haemoglobin HbA1c	331	0.165	376	0.155	403	0.157
HDL cholesterol	397	0.189	458	0.179	462	0.177
LDL direct	187	0.194	210	0.205	215	0.204
Phosphate	115	0.153	133	0.160	128	0.153
Total bilirubin	464	0.261	546	0.244	526	0.240
Total protein	349	0.235	371	0.231	391	0.233
Triglycerides	409	0.196	494	0.202	504	0.200
Asthma	97	0.263	107	0.257	108	0.267
Hypothyroidism	67	0.128	73	0.132	74	0.134
Angina	10	0.192	10	0.229	11	0.220
Depression	0	0.000	0	0.000	0	0.000
Breast cancer	18	0.419	15	0.432	19	0.422
Osteoporosis	11	0.458	11	0.355	10	0.393
Ulcerative colitis	6	0.194	2	0.088	2	0.088
Glaucoma	15	0.417	14	0.389	13	0.361
Cholelithiasis	12	0.382	13	0.361	15	0.375
Chronic obstructive airways disease	1	0.500	1	0.500	1	0.500
Hyperthyroidism	31	0.201	26	0.194	30	0.210
Type 2 diabetes	3	0.300	3	0.300	4	0.333
Nasal polyps	2	0.111	2	0.167	2	0.111
Total	10555	0.228	11931	0.220	12457	0.217

Table B.11: Number of independent associated loci for quantitative traits after applying the correction for participation bias. Number of independent associated loci for quantitative traits after PLINK clumping using summary statistics from weighted GWAS in Quickdraws ($N \approx 338,738$) compared to weighted GWAS on unrelated samples in LDAK ($N \approx 286,432$). Note that the adjustment for participation bias was only applied to step 2 of the Quickdraws algorithm.

Phenotype	LDAK	Quickdraws
Eosinophill count	66	106
Eosinophill %	70	127
Haematocrit %	39	60
Haemoglobin concentration	39	68
High light scatter reticulocyte count	69	104
High light scatter reticulocyte %	69	107
Immature reticulocyte fraction	46	66
Lymphocyte count	56	98
Lymphocyte %	44	78
Mean corp. haemoglobin	97	161
Mean corp. haemoglobin concentration	22	27
Mean corp. volume	93	173
Mean platelet volume	160	315
Mean reticulocyte volume	100	128
Mean sphered cell volume	74	126
Monocyte count	72	109
Monocyte %	74	114
Neutrophill count	42	60
Neutrophill %	34	63
Platelet count	124	203
Platelet crit	104	161
Platelet distribution width	101	150
Red blood cell count	66	111
Red blood cell distribution width	77	116
Reticulocyte count	64	95
Reticulocyte %	60	103
Albumin	29	41
Alkaline phosphatase	76	116
Apolipoprotein A	59	80
Apolipoprotein B	56	71
Asp. aminotransferase	41	54
C-reactive protein	39	58
Calcium	31	45
Cholesterol	53	71
Creatinine	57	84
Gamma glutamyltransferase	65	91
Glucose	15	21
Glycated haemoglobin	71	99
HDL cholesterol	74	101
IGF-1	60	100

Continued on next page

Table B.11: *(Continued)*

Phenotype	LDAK	Quickdraws
LDL direct	51	65
Lipoprotein A	26	29
Phosphate	22	33
SHBG	58	86
Testosterone	10	12
Total bilirubin	38	61
Total protein	35	49
Triglycerides	51	69
Urate	42	59
Urea	22	35
Standing height	201	445
Forced vital capacity (FVC) Z-score	8	15
FEV1/ FVC ratio Z-score	35	53
Basal metabolic rate	44	70
Forced expiratory volume (FEV1) Z-score	14	24
Body mass index (BMI)	23	33
Hip circumference	19	33
Waist circumference	16	22
Peak expiratory flow (PEF)	5	8
Weight	30	44
Hand grip strength (left)	1	3
Systolic blood pressure	18	22
Pulse rate	30	43
Body fat %	17	31
Diastolic blood pressure	5	12
Hand grip strength (right)	2	4
Trunk fat %	19	35
Trunk fat mass	16	30
Impedance of whole body	29	53
Bread intake	1	1
Leg fat % (left)	16	25
Leg fat % (right)	13	24
Arm fat % (left)	16	27
Arm fat % (right)	16	28
Impedance of arm (left)	25	43
Impedance of arm (right)	22	40
Whole body water mass	44	78
Age started wearing glasses	1	4
Duration of walks	0	0
Total	2629	5809

Table B.12: Computational efficiency of Quickdraws for quantitative trait association. We compare the computational requirements of several GWAS algorithms and Quickdraws across 3 to 4 RAP cloud instances for ∼13.3 million tested variants and 50 phenotypes with either $N = 50{,}000$ or $N = 405{,}088$ samples and 458,464 genotyped markers for model fitting. * The step 1 for Quickdraws is run using `mem3_ssd1_gpu_x8`; ** extrapolated value; "-" indicates that BOLT-LMM was not run in this setting.

Sample	Instance	Quickdraws* (£)	Regenie (£)	FastGWA (£)	BOLT-LMM (£)
50k	mem3_ssd1_v2_x4	**7.9**	**3.95**	**3.63**	-
	mem2_ssd1_v2_x8	8.397	4.65	4.58	**158.38**
	mem1_ssd1_v2_x16	9.018	6.18	7.01	-
405k	mem2_ssd1_v2_x8	Disk space exceeded	Disk space exceeded	**37.31**	Out of Memory
	mem2_ssd1_v2_x16	**92.993**	**24.18**	59.28	-
	mem1_ssd1_v2_x36	110.91	44.66	114.61	**7500.0**
	mem1_ssd1_v2_x72	156.908	68.72	161.88	-

(a) Total cost (GBP) on UK Biobank RAP.

Sample	Instance	Quickdraws* (h)	Regenie (h)	FastGWA (h)	BOLT-MoG (h)
50k	mem3_ssd1_v2_x4	20.18	27.07	24.84	-
	mem2_ssd1_v2_x8	18.57	20.55	20.24	798.5**
	mem1_ssd1_v2_x16	16.27	15.61	17.66	-
405k	mem2_ssd1_v2_x8	Disk space exceeded	Disk space exceeded	164.77	Out of Memory
	mem2_ssd1_v2_x16	159.78	53.4	130.91	-
	mem1_ssd1_v2_x36	149.29	50.02	128.37	8400**
	mem1_ssd1_v2_x72	149.29	38.5	90.66	-

(b) Total running time (hours).

Table B.13: Computational efficiency of Quickdraws for binary trait association. We compare the computational requirements of several GWAS algorithms and Quickdraws across 3 to 4 RAP cloud instances for ∼13.3 million tested variants and 50 phenotypes with either $N = 50{,}000$ or $N = 405{,}088$ samples and 458,464 genotyped markers for model fitting. * The step 1 for Quickdraws is run using `mem3_ssd1_gpu_x8`. All methods except FastGWA were run on 1 million test variants and extrapolated.

Sample	Instance	Quickdraws* (£)	Regenie (£)	FastGWA (£)	SAIGE (£)
50k	mem3_ssd1_v2_x4	30.5	44.20	9.28	**47.54**
	mem2_ssd1_v2_x8	29.79	41.62	**9.20**	54.48
	mem1_ssd1_v2_x16	**25.221**	**40.99**	11.64	101.94
405k	mem2_ssd1_v2_x8	Disk space exceeded	Disk space exceeded	**98.1**	**2834.91**
	mem2_ssd1_v2_x16	**395.937**	**506.85**	126.04	5703.49
	mem1_ssd1_v2_x36	571.964	769.17	226.82	10736.36
	mem1_ssd1_v2_x72	628.003	1192.22	319.66	21887.59

(a) Total cost (GBP) on UK Biobank RAP.

Sample	Instance	Quickdraws* (h)	Regenie (h)	FastGWA (h)	SAIGE (h)
50k	mem3_ssd1_v2_x4	143.87	300.78	63.58	322.24
	mem2_ssd1_v2_x8	96.14	182.03	40.63	272.97
	mem1_ssd1_v2_x16	51.21	100.49	29.34	253.28
405k	mem2_ssd1_v2_x8	Disk space exceeded	Disk space exceeded	433.29	12524.16
	mem2_ssd1_v2_x16	798.55	930.32	278.35	12584.25
	mem1_ssd1_v2_x36	682.33	869.96	254.06	12024.41
	mem1_ssd1_v2_x72	453.89	663.53	179.02	12264.12

(b) Total running time (hours).

Table B.14: Computational efficiency of Quickdraws for a single quantitative or binary trait. We compare the computational requirements of Regenie and Quickdraws, which support the parallal analysis of multiple traits, with those of FastGWA, BOLT-LMM, and SAIGE. All methods are used to compute summary statistics for ∼13.3 million variants and one phenotype with $N = 405{,}088$ and 458,464 genotyped markers for model fitting ($89,177$ genotyped markers for SAIGE, see Methods). Running times and costs are computed using the same hardware for all methods (mem1_ssd1_v2_x36, 72 GB of RAM, 36-core processor).

Method	**Quantitative Traits**		**Binary Traits**	
	Time (hrs)	**Cost (£)**	**Time (hrs)**	**Cost (£)**
Quickdraws	71.92	57.86	141.11	109.5
Regenie	26.71	22.38	24.73	22.40
BOLT-LMM	168.0	150.0	-	-
FastGWA	2.6	2.29	5.08	4.536
SAIGE	-	-	240.48	214.72

Table B.15: Computational efficiency of Quickdraws for $N = 1{,}000{,}000$ samples. We compare the computational requirements of Quickdraws and Regenie when computing summary statistics for 458k, 13.3 million and 600 million variants, 50 quantitative, and 50 binary traits, using $N = 1{,}000{,}000$ and 458,464 genotyped markers for model fitting. The data set was generated by duplicating samples from the $N = 405k$ subset of white British individuals. * indicates that the cost of imputed variant association testing is extrapolated from the testing time for $458,464$ genotyped markers. Quickdraws was run using the low-memory option. Running times and costs were computed using the same hardware for both methods (mem1_ssd1_v2_x36, 72 GB of RAM, 36-core processor).

Method	**Quantitative Traits**		**Binary Traits**	
	Time (hrs)	**Cost (£)**	**Time (hrs)**	**Cost (£)**
Regenie (Step 1)	13.38	12.11	178.41	159.29
Regenie (Step 2, $M = 458$k)	2.13	1.90	32.96	29.43
Regenie* (Step 2, $M = 13.3$ mil.)	62.04	55.39	957.33	854.71
Regenie* (Step 2, $M = 600$ mil.)	2798.8	2498.8	43187.82	38558.34
Quickdraws (Step 1)	243.23	161.50	424.56	281.91
Quickdraws (Step 2, $M = 458$k)	1.51	1.36	23.4	21.19
Quickdraws* (Step 2, $M = 13.3$ mil.)	44.04	39.77	679.51	615.43
Quickdraws* (Step 2, $M = 600$ mil.)	1986.76	1794.13	30654.58	27763.75

References

[1] Francis S. Collins, Michael Morgan, and Ari Patrinos. "The Human Genome Project: lessons from large-scale biology". In: *Science* 300.5617 (2003), pp. 286–290.

[2] KA Wetterstrand. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. 2022. URL: `https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data`.

[3] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (2018), pp. 203–209.

[4] Howard M Cann et al. "A human genome diversity cell line panel". In: *Science* 296.5566 (2002), pp. 261–262.

[5] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), p. 68.

[6] Leo Speidel et al. "A method for genome-wide genealogy estimation for thousands of samples". In: *Nature genetics* 51.9 (2019), pp. 1321–1329.

[7] Jerome Kelleher et al. "Inferring whole-genome histories in large population datasets". In: *Nature Genetics* 51.9 (2019), pp. 1330–1338.

[8] Brian C Zhang et al. "Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits". In: *Nature Genetics* 55.5 (2023), pp. 768–776.

[9] Augustine Kong et al. "Rate of de novo mutations and the importance of father's age to disease risk". In: *Nature* 488.7412 (2012), pp. 471–475.

[10] Jack N Fenner. "Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies". In: *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 128.2 (2005), pp. 415–423.

[11] Richard J Wang et al. "Human generation times across the past 250,000 years". In: *Science Advances* 9.1 (2023), eabm7047.

[12] Kelley Harris and Jonathan K Pritchard. "Recent advances in understanding de novo mutation through sequencing whole genomes". In: *Nature Reviews Genetics* 20.10 (2019), pp. 551–564.

[13] Frédéric Baudat et al. "PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice". In: *Science* 327.5967 (2010), pp. 836–840.

[14] Noah A Rosenberg et al. "Genetic structure of human populations". In: *science* 298.5602 (2002), pp. 2381–2385.

[15] John Novembre et al. "Genes mirror geography within Europe". In: *Nature* 456.7218 (2008), pp. 98–101.

[16] Stephen Leslie et al. “The fine-scale genetic structure of the British population”. In: *Nature* 519.7543 (2015), pp. 309–314.

[17] Adam Auton et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), p. 68.

[18] Saioa López, Lucy Van Dorp, and Garrett Hellenthal. “Human dispersal out of Africa: a lasting debate”. In: *Evolutionary Bioinformatics* 11 (2015), EBO–S33489.

[19] David Reich et al. “Reconstructing native American population history”. In: *Nature* 488.7411 (2012), pp. 370–374.

[20] Iosif Lazaridis et al. “Ancient human genomes suggest three ancestral populations for present-day Europeans”. In: *Nature* 513.7518 (2014), pp. 409–413.

[21] Wolfgang Haak et al. “Massive migration from the steppe was a source for Indo-European languages in Europe”. In: *Nature* 522.7555 (2015), pp. 207–211.

[22] Michael W Smith et al. “A high-density admixture map for disease gene discovery in African Americans”. In: *The American Journal of Human Genetics* 74.5 (2004), pp. 1001–1013.

[23] Alkes L Price et al. “Sensitive detection of chromosomal segments of distinct ancestry in admixed populations”. In: *PLoS genetics* 5.6 (2009), e1000519.

[24] Sriram Sankararaman et al. “The genomic landscape of Neanderthal ancestry in present-day humans”. In: *Nature* 507.7492 (2014), pp. 354–357.

[25] David Reich et al. “Genetic history of an archaic hominin group from Denisova Cave in Siberia”. In: *Nature* 468.7327 (2010), pp. 1053–1060.

[26] Sewall Wright. “Evolution in Mendelian Populations”. In: *Genetics* 16.2 (1931), pp. 97–159.

[27] R.A. Fisher. “The Genetical Theory of Natural Selection”. In: *Oxford University Press* (1930).

[28] John Frank Charles Kingman. “The coalescent”. In: *Stochastic processes and their applications* 13.3 (1982), pp. 235–248.

[29] John FC Kingman. “Origins of the coalescent: 1974-1982”. In: *Genetics* 156.4 (2000), pp. 1461–1463.

[30] Richard R Hudson. “Testing the constant-rate neutral allele model with protein sequence data”. In: *Evolution* (1983), pp. 203–217.

[31] Richard R Hudson. “Properties of a neutral allele model with intragenic recombination”. In: *Theoretical population biology* 23.2 (1983), pp. 183–201.

[32] Fumio Tajima. “Evolutionary relationship of DNA sequences in finite populations”. In: *Genetics* 105.2 (1983), pp. 437–460.

[33] Richard R Hudson. “Properties of a neutral allele model with intragenic recombination”. In: *Theoretical Population Biology* 23.2 (1983), pp. 183–201.

[34] Robert C Griffiths and Paul Marjoram. “Ancestral inference from gene trees”. In: *Statistical Science* 13.1 (1997), pp. 1–16.

[35] Yan Wong et al. “A general and efficient representation of ancestral recombination graphs”. In: *BioRxiv* (2023).

[36] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2 (2000), pp. 945–959.

[37] David H Alexander, John Novembre, and Ken Lange. "Fast model-based estimation of ancestry in unrelated individuals". In: *Genome Research* 19.9 (2009), pp. 1655–1664.

[38] Daniel J Lawson et al. "Inference of population structure using dense haplotype data". In: *PLoS Genetics* 8.1 (2012), e1002453.

[39] Sewall Wright. "The genetical structure of populations". In: *Annals of Eugenics* 15.1 (1951), pp. 323–354.

[40] Gustave Malécot. *Les Mathématiques de l'Hérédité*. Masson et Cie, 1948.

[41] Rebecca L Cann, Mark Stoneking, and Allan C Wilson. "Mitochondrial DNA and human evolution". In: *Nature* 325.6099 (1987), pp. 31–36.

[42] Charles A Tilford et al. "A physical map of the human Y chromosome". In: *Nature* 409.6822 (2001), pp. 943–945.

[43] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[44] Anil Raj, Matthew Stephens, and Jonathan K Pritchard. "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets". In: *Genetics* 197.2 (2014), pp. 573–589.

[45] Paolo Menozzi, Alberto Piazza, and Luigi Cavalli-Sforza. "Synthetic Maps of Human Gene Frequencies in Europeans: These maps indicate that early farmers of the Near East spread to all of Europe in the Neolithic." In: *Science* 201.4358 (1978), pp. 786–792.

[46] Nick Patterson, Alkes L Price, and David Reich. "Population structure and eigenanalysis". In: *PLoS Genetics* 2.12 (2006), e190.

[47] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature genetics* 38.8 (2006), pp. 904–909.

[48] John Novembre and Matthew Stephens. "Interpreting principal component analyses of spatial population genetic variation". In: *Nature genetics* 40.5 (2008), pp. 646–649.

[49] Juba Nait Saada et al. "Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations". In: *Nature communications* 11.1 (2020), p. 6130.

[50] N. Li and M. Stephens. "Modelling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data". In: *Genetics* 165.4 (2003), pp. 2213–2233.

[51] Garrett Hellenthal et al. "A genetic atlas of human admixture history". In: *science* 343.6172 (2014), pp. 747–751.

[52] Simon Gravel. "Population genetics models of local ancestry". In: *Genetics* 191.2 (2012), pp. 607–619.

[53] Po-Ru Loh et al. "Inferring admixture histories of human populations using linkage disequilibrium". In: *Genetics* 193.4 (2013), pp. 1233–1254.

[54] Priya Moorjani et al. “The history of African gene flow into Southern Europeans, Levantines, and Jews”. In: *PLoS genetics* 7.4 (2011), e1001373.

[55] Nick Patterson et al. “Ancient admixture in human history”. In: *Genetics* 192.3 (2012), pp. 1065–1093.

[56] Brian K Maples et al. “RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference”. In: *The American Journal of Human Genetics* 93.2 (2013), pp. 278–288.

[57] Sharon R Browning, Ryan K Waples, and Brian L Browning. “Fast, accurate local ancestry inference with FLARE”. In: *The American Journal of Human Genetics* 110.2 (2023), pp. 326–335.

[58] Michael Salter-Townshend and Simon Myers. “Fine-scale inference of ancestry segments without prior knowledge of admixing groups”. In: *Genetics* 212.3 (2019), pp. 869–889.

[59] David Reich et al. “Reconstructing Indian population history”. In: *Nature* 461.7263 (2009), pp. 489–494.

[60] Eric Y Durand et al. “Testing for ancient admixture between closely related populations”. In: *Molecular biology and evolution* 28.8 (2011), pp. 2239–2252.

[61] Benjamin M Peter. “A geometric relationship of F 2, F 3 and F 4-statistics with principal component analysis”. In: *Philosophical Transactions of the Royal Society B* 377.1852 (2022), p. 20200413.

[62] Matthew D Rasmussen et al. “Genome-wide inference of ancestral recombination graphs”. In: *PLoS genetics* 10.5 (2014), e1004342.

[63] Gilean AT McVean and Niall J Cardin. “Approximating the coalescent with recombination”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1459 (2005), pp. 1387–1393.

[64] Anthony Wilder Wohns et al. “A unified genealogy of modern and ancient genomes”. In: *Science* 375.6583 (2022), eabi8264.

[65] Pier Francesco Palamara et al. “High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability”. In: *Nature Genetics* 50.9 (2018), pp. 1311–1317.

[66] Yun Deng, Rasmus Nielsen, and Yun S Song. “Robust and accurate bayesian inference of genome-wide genealogies for large samples”. In: *bioRxiv* (2024), pp. 2024–03.

[67] Débora YC Brandt et al. “Evaluation of methods for the inference of ancestral recombination graphs”. In: *Genetics* 221.1 (2022).

[68] Stephan Schiffels and Richard Durbin. “Inferring human population size and separation history from multiple genome sequences”. In: *Nature genetics* 46.8 (2014), pp. 919–925.

[69] Leo Speidel et al. “High-resolution genomic ancestry reveals mobility in early medieval Europe”. In: *bioRxiv* (2024), pp. 2024–03.

[70] Yair Field et al. “Detection of human adaptation during the past 2000 years”. In: *Science* 354.6313 (2016), pp. 760–764.

[71] Aaron J Stern, Peter R Wilton, and Rasmus Nielsen. "An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data". In: *PLoS Genetics* 15.9 (2019), e1008384.

[72] Andrew H Vaughn and Rasmus Nielsen. "Fast and accurate estimation of selection coefficients and allele histories from ancient and modern DNA". In: *Molecular Biology and Evolution* 41.8 (2024).

[73] Caoqi Fan et al. "A likelihood-based framework for demographic inference from genealogical trees". In: *bioRxiv* (2023).

[74] Débora YC Brandt et al. "The promise of inferring the past using the ancestral recombination graph". In: *Genome biology and evolution* 16.2 (2024), evae005.

[75] Rasmus Nielsen, Andrew H Vaughn, and Yun Deng. "Inference and applications of ancestral recombination graphs". In: *Nature Reviews Genetics* (2024), pp. 1–12.

[76] Kelley Harris. "Evidence for recent, population-specific evolution of the human mutation rate". In: *Proceedings of the National Academy of Sciences* 112.11 (2015), pp. 3439–3444.

[77] Charles Darwin and Tom Griffith. *The descent of man*. Vol. 4. Prometheus Books New York, 1874.

[78] Guy S Jacobs et al. "Multiple deeply divergent Denisovan ancestries in Papuans". In: *Cell* 177.4 (2019), pp. 1010–1021.

[79] Jeffrey P Severinghaus and Edward J Brook. "Abrupt climate change at the end of the last glacial period inferred from trapped air in polar ice". In: *Science* 286.5441 (1999), pp. 930–934.

[80] Pontus Skoglund and David Reich. "A genomic view of the peopling of the Americas". In: *Current opinion in genetics & development* 41 (2016), pp. 27–35.

[81] Swapan Mallick et al. "The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes". In: *Scientific Data* 11.1 (2024), p. 182.

[82] James F Gusella et al. "A polymorphic DNA marker genetically linked to Huntington's disease". In: *Nature* 306.5940 (1983), pp. 234–238.

[83] The Wellcome Trust Case Control Consortium. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: *Nature* 447.7145 (2007), pp. 661–678.

[84] Benjamin Neale. *Heritability 101: What is heritability?* Accessed: 2024-07-13. 2017. URL: http://www.nealelab.is/blog/2017/9/13/heritability-101-what-is-heritability#:~:text=For%20example%2C%20the%20heritability%20of,all%20the%20relevant%20genetic%20effects.

[85] Peter M Visscher, William G Hill, and Naomi R Wray. "Heritability in the genomics era–concepts and misconceptions". In: *Nature Reviews Genetics* 9.4 (2008), pp. 255–266.

[86] Alexander Gusev. *Heritability: Project Hsq*. Accessed: 2024-07-13. 2021. URL: http://gusevlab.org/projects/hsq/.

[87] NHGRI-EBI GWAS Catalog. *The NHGRI-EBI GWAS Catalog*. Accessed: 2024-07-13. 2024. URL: https://www.ebi.ac.uk/gwas/.

[88] Polina V Rusina et al. “Genetic support for FDA-approved drugs over the past decade.” In: *Nature reviews. Drug Discovery* (2023).

[89] David Ochoa et al. “Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs”. In: *Nat Rev Drug Discov* 21.8 (2022), p. 551.

[90] Abdel Abdellaoui et al. “15 years of GWAS discovery: realizing the promise”. In: *The American Journal of Human Genetics* 110.2 (2023), pp. 179–194.

[91] Alexander Teumer et al. “Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria”. In: *Nature communications* 10.1 (2019), p. 4130.

[92] George L Bakris et al. “Effect of finerenone on chronic kidney disease outcomes in type 2 diabetes”. In: *New England journal of medicine* 383.23 (2020), pp. 2219–2229.

[93] Ennio G Favalli et al. “Baricitinib for COVID-19: a suitable treatment?” In: *The Lancet Infectious Diseases* 20.9 (2020), pp. 1012–1013.

[94] Julie E Horowitz et al. “Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease”. In: *Nature genetics* 54.4 (2022), pp. 382–392.

[95] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. “The personal and clinical utility of polygenic risk scores”. In: *Nature Reviews Genetics* 19.9 (2018), pp. 581–590.

[96] Nadia Solovieff et al. “Pleiotropy in complex traits: challenges and strategies”. In: *Nature Reviews Genetics* 14.7 (2013), pp. 483–495.

[97] Yuval B Simons et al. “A population genetic interpretation of GWAS findings for human quantitative traits”. In: *PLoS biology* 16.3 (2018), e2002985.

[98] George Davey Smith and Gibran Hemani. “Mendelian randomization: genetic anchors for causal inference in epidemiological studies”. In: *Human molecular genetics* 23.R1 (2014), R89–R98.

[99] Jean-Baptiste Pingault et al. “Using genetic data to strengthen causal inference in observational research”. In: *Nature Reviews Genetics* 19.9 (2018), pp. 566–580.

[100] Peter M Visscher et al. “10 years of GWAS discovery: biology, function, and translation”. In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[101] Emil Uffelmann et al. “Genome-wide association studies”. In: *Nature Reviews Methods Primers* 1.1 (2021), p. 59.

[102] Gao Wang. *Fine-mapping example*. URL: `https://stephenslab.github.io/susieR/articles/finemapping.html`.

[103] Hannah Wand et al. “Improving reporting standards for polygenic scores in risk prediction studies”. In: *Nature* 591.7849 (2021), pp. 211–219.

[104] Jennifer Listgarten et al. “Improved linear mixed models for genome-wide association studies”. In: *Nature methods* 9.6 (2012), pp. 525–526.

[105] Jian Yang et al. “Advantages and pitfalls in the application of mixed-model association methods”. In: *Nature genetics* 46.2 (2014), pp. 100–106.

[106] Po-Ru Loh et al. “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. In: *Nature genetics* 47.3 (2015), pp. 284–290.

[107] Po-Ru Loh et al. “Mixed-model association for biobank-scale datasets”. In: *Nature genetics* 50.7 (2018), pp. 906–908.

[108] Longda Jiang et al. “A resource-efficient tool for mixed model association analysis of large-scale data”. In: *Nature genetics* 51.12 (2019), pp. 1749–1755.

[109] Yue Wu and Sriram Sankararaman. “A scalable estimator of SNP heritability for biobank-scale data”. In: *Bioinformatics* 34.13 (2018), pp. i187–i194.

[110] Ali Pazokitoroudi et al. “Efficient variance components analysis across millions of genomes”. In: *Nature communications* 11.1 (2020), pp. 1–10.

[111] Jiazheng Zhu et al. “Fast variance component analysis using large-scale ancestral recombination graphs”. In: *bioRxiv* (2024).

[112] Shaun Purcell et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses”. In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[113] Christopher C Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. In: *Gigascience* 4.1 (2015), s13742–015.

[114] Manuel A Rivas and Christopher Chang. “Efficient storage and regression computation for population-scale genome sequencing studies”. In: *bioRxiv* (2024).

[115] Hyun Min Kang et al. “Efficient control of population structure in model organism association mapping”. In: *Genetics* 178.3 (2008), pp. 1709–1723.

[116] Hyun Min Kang et al. “Variance component model to account for sample structure in genome-wide association studies”. In: *Nature genetics* 42.4 (2010), pp. 348–354.

[117] Christoph Lippert et al. “FaST linear mixed models for genome-wide association studies”. In: *Nature methods* 8.10 (2011), pp. 833–835.

[118] Jennifer Listgarten, Christoph Lippert, and David Heckerman. “FaST-LMM-Select for addressing confounding from spatial structure and rare variants”. In: *Nature genetics* 45.5 (2013), pp. 470–471.

[119] Wei Zhou et al. “Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies”. In: *Nature genetics* 50.9 (2018), pp. 1335–1341.

[120] Diego Kuonen. “Miscellanea. Saddlepoint approximations for distributions of quadratic forms in normal variables”. In: *Biometrika* 86.4 (1999), pp. 929–935.

[121] Rounak Dey et al. “A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS”. In: *The American Journal of Human Genetics* 101.1 (2017), pp. 37–49.

[122] Longda Jiang et al. “A generalized linear mixed model association tool for biobank-scale data”. In: *Nature genetics* 53.11 (2021), pp. 1616–1621.

[123] Joelle Mbatchou et al. “Computationally efficient whole-genome regression for quantitative and binary traits”. In: *Nature genetics* 53.7 (2021), pp. 1097–1103.

[124] Akiko Nagai et al. “Overview of the BioBank Japan Project: study design and profile”. In: *Journal of epidemiology* 27.Supplement_III (2017), S2–S8.

[125] Andrea H Ramirez et al. “The All of Us Research Program: data quality, utility, and diversity”. In: *Patterns* 3.8 (2022).

[126] Genevieve L Wojcik et al. “Genetic analyses of diverse populations improves discovery for complex traits”. In: *Nature* 570.7762 (2019), pp. 514–518.

[127] Zhengming Chen et al. “China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up”. In: *International journal of epidemiology* 40.6 (2011), pp. 1652–1666.

[128] Mitja I Kurki et al. “FinnGen provides genetic insights from a well-phenotyped isolated population”. In: *Nature* 613.7944 (2023), pp. 508–518.

[129] Wei Zhou et al. “Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease”. In: *Cell Genomics* 2.10 (2022), p. 100192.

[130] Jianming Yu et al. “A unified mixed-model method for association mapping that accounts for multiple levels of relatedness”. In: *Nature genetics* 38.2 (2006), pp. 203–208.

[131] Zhiwu Zhang et al. “Mixed linear model approach adapted for genome-wide association studies”. In: *Nature genetics* 42.4 (2010), pp. 355–360.

[132] Xiang Zhou and Matthew Stephens. “Genome-wide efficient mixed-model analysis for association studies”. In: *Nature genetics* 44.7 (2012), pp. 821–824.

[133] Vincent Segura et al. “An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations”. In: *Nature genetics* 44.7 (2012), pp. 825–830.

[134] Loıc Yengo et al. “A saturated map of common genetic variants associated with human height”. In: *Nature* 610.7933 (2022), pp. 704–712.

[135] Quanli Wang et al. “Rare variant contribution to human disease in 281,104 UK Biobank exomes”. In: *Nature* 597.7877 (2021), pp. 527–532.

[136] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[137] Sriram Sankararaman et al. “The date of interbreeding between Neandertals and modern humans”. In: (2012).

[138] UCSC Genome Browser. *UCSC Genome Browser FTP Help*. Accessed: 2025-02-03. 2025. URL: `https://genome.ucsc.edu/goldenPath/help/ftp.html`.

[139] Anders Bergström et al. “Insights into human genetic variation and population history from 929 diverse genomes”. In: *Science* 367.6484 (2020), eaay5012.

[140] Swapan Mallick et al. “The Simons genome diversity project: 300 genomes from 142 diverse populations”. In: *Nature* 538.7624 (2016), pp. 201–206.

[141] Peter H Sudmant et al. “An integrated map of structural variation in 2,504 human genomes”. In: *Nature* 526.7571 (2015), pp. 75–81.

[142] Zan Koenig et al. “A harmonized public resource of deeply sequenced diverse human genomes”. In: *Genome Research* (2024).

[143] gnomAD Consortium. *HGDP and 1000 Genomes Phased Haplotypes v2*. Accessed: 2025-02-03. 2025. URL: `gs://gcp-public-data--gnomad/resources/hgdp_1kg/phased_haplotypes_v2/`.

[144] Reich Lab. *SGDP Phased Data (Known Bugs, Not Recommended - Please Use Newer Dataset Instead)*. Accessed: 2025-02-03. 2025. URL: https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data.knownbugs.not_recommended.please_use_newer_dataset_instead/.

[145] 1000 Genomes Project Consortium. *1000 Genomes Project Data Release (May 2013)*. Accessed: 2025-02-03. 2013. URL: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

[146] Ensembl Genome Browser. *Homo sapiens Ancestral Alleles - GRCh37 (Ensembl Release 74)*. Accessed: 2025-02-03. 2014. URL: ftp://ftp.ensembl.org/pub/release-74/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh37_e71.tar.bz2.

[147] Ensembl Genome Browser. *Homo sapiens Ancestral Alleles - GRCh38 (Ensembl Release 86)*. Accessed: 2025-02-03. 2016. URL: http://ftp.ensembl.org/pub/release-86/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh38_e86.tar.gz.

[148] Max Planck Institute for Evolutionary Anthropology. *Neandertal Genome Data Repository*. Accessed: 2025-02-03. 2025. URL: http://ftp.eva.mpg.de/neandertal/.

[149] Olivier Delaneau. *SHAPEIT5 Genetic Maps Repository*. Accessed: 2025-02-03. 2025. URL: https://github.com/odelaneau/shapeit5/tree/main/resources/maps/.

[150] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. "Efficient coalescent simulation and genealogical analysis for large sample sizes". In: *PLoS computational biology* 12.5 (2016), e1004842.

[151] Laurits Skov et al. "Detecting archaic introgression using an unadmixed outgroup". In: *PLoS genetics* 14.9 (2018), e1007641.

[152] Iosif Lazaridis et al. "The Genetic Origin of the Indo-Europeans". In: *bioRxiv* (2024).

[153] MH Day. "Early Homo sapiens remains from the Omo River region of south-west Ethiopia: Omo human skeletal remains". In: *Nature* 222.5199 (1969), pp. 1135–1138.

[154] Jean-Jacques Hublin et al. "New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens". In: *Nature* 546.7657 (2017), pp. 289–292.

[155] Anders Bergström et al. "Origins of modern human ancestry". In: *Nature* 590.7845 (2021), pp. 229–237.

[156] Ning Yu et al. "Larger genetic differences within Africans than between Africans and Eurasians". In: *Genetics* 161.1 (2002), pp. 269–274.

[157] Sarah A Tishkoff et al. "The genetic structure and history of Africans and African Americans". In: *science* 324.5930 (2009), pp. 1035–1044.

[158] Steven J Micheletti et al. "Genetic consequences of the transatlantic slave trade in the Americas". In: *The American Journal of Human Genetics* 107.2 (2020), pp. 265–277.

[159] Denis Pierron et al. “Genome-wide evidence of Austronesian–Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar”. In: *Proceedings of the National Academy of Sciences* 111.3 (2014), pp. 936–941.

[160] Denis Pierron et al. “Genomic landscape of human diversity across Madagascar”. In: *Proceedings of the National Academy of Sciences* 114.32 (2017), E6498–E6506.

[161] Esther S Brielle et al. “Entwined African and Asian genetic roots of medieval peoples of the Swahili coast”. In: *Nature* 615.7954 (2023), pp. 866–873.

[162] Cesare De Filippo et al. “Bringing together linguistic and genetic evidence to test the Bantu expansion”. In: *Proceedings of the Royal Society B: Biological Sciences* 279.1741 (2012), pp. 3256–3263.

[163] Sen Li, Carina Schlebusch, and Mattias Jakobsson. “Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples”. In: *Proceedings of the Royal Society B: Biological Sciences* 281.1793 (2014), p. 20141448.

[164] Etienne Patin et al. “Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America”. In: *Science* 356.6337 (2017), pp. 543–546.

[165] Cesar A Fortes-Lima et al. “The genetic legacy of the expansion of Bantu-speaking peoples in Africa”. In: *Nature* 625.7995 (2024), pp. 540–547.

[166] Jessica E Tierney, Francesco SR Pausata, and Peter B deMenocal. “Rainfall regimes of the Green Sahara”. In: *Science advances* 3.1 (2017), e1601503.

[167] Juan C Larrasoaña, Andrew P Roberts, and Eelco J Rohling. “Dynamics of green Sahara periods and their role in hominin evolution”. In: *PloS one* 8.10 (2013), e76514.

[168] Joseph K Pickrell et al. “The genetic prehistory of southern Africa”. In: *Nature communications* 3.1 (2012), p. 1143.

[169] Joseph K Pickrell et al. “Ancient west Eurasian ancestry in southern and eastern Africa”. In: *Proceedings of the National Academy of Sciences* 111.7 (2014), pp. 2632–2637.

[170] M Gallego Llorente et al. “Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa”. In: *Science* 350.6262 (2015), pp. 820–822.

[171] “Erratum for the Report “Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa” (previously titled “Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent”) by M. Gallego Llorente, E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, J. W. Arthur, M. C. Curtis, J. T. Stock, M. Coltorti, P. Pieruccini, S. Stretton, F. Brock, T. Higham, Y. Park, M. Hofreiter, D. G. Bradley, J. Bhak, R. Pinhasi, A. Manica”. en. In: *Science* 351.6275 (Feb. 2016), aaf3945–aaf3945.

[172] Lu Chen et al. “Identifying and interpreting apparent Neanderthal ancestry in African individuals”. In: *Cell* 180.4 (2020), pp. 677–687.

[173] Benjamin Vernot et al. “Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals”. In: *Science* 352.6282 (2016), pp. 235–239.

[174] Marieke Van de Loosdrecht et al. “Pleistocene North African genomes link near Eastern and sub-Saharan African human populations”. In: *Science* 360.6388 (2018), pp. 548–552.

[175] Rosa Fregel et al. “Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe”. In: *Proceedings of the National Academy of Sciences* 115.26 (2018), pp. 6774–6779.

[176] Luciana G Simões et al. “Northwest African Neolithic initiated by migrants from Iberia and Levant”. In: *Nature* 618.7965 (2023), pp. 550–556.

[177] Eleanor ML Scerri, Lounès Chikhi, and Mark G Thomas. “Beyond multiregional and simple out-of-Africa models of human evolution”. In: *Nature ecology & evolution* 3.10 (2019), pp. 1370–1372.

[178] Sally McBrearty and Alison S Brooks. “The revolution that wasn’t: a new interpretation of the origin of modern human behavior”. In: *Journal of human evolution* 39.5 (2000), pp. 453–563.

[179] Peter B deMenocal. “Climate and human evolution”. In: *science* 331.6017 (2011), pp. 540–542.

[180] Pontus Skoglund et al. “Reconstructing prehistoric African population structure”. In: *Cell* 171.1 (2017), pp. 59–71.

[181] Aaron P Ragsdale and Simon Gravel. “Models of archaic admixture and recent history from two-locus statistics”. In: *PLoS genetics* 15.6 (2019), e1008204.

[182] Michael F Hammer et al. “Genetic evidence for archaic admixture in Africa”. In: *Proceedings of the National Academy of Sciences* 108.37 (2011), pp. 15123–15128.

[183] Belen Lorente-Galdos et al. “Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations”. In: *Genome biology* 20 (2019), pp. 1–15.

[184] Arun Durvasula and Sriram Sankararaman. “Recovering signals of ghost archaic introgression in African populations”. In: *Science Advances* 6.7 (2020), eaax5097.

[185] Nina Hollfelder et al. “The deep population history in Africa”. In: *Human Molecular Genetics* 30.R1 (2021), R2–R10.

[186] Aaron P Ragsdale et al. “A weakly structured stem for human origins in Africa”. In: *Nature* 617.7962 (2023), pp. 755–763.

[187] Mark Lipson et al. “Ancient DNA and deep population structure in sub-Saharan African foragers”. In: *Nature* 603.7900 (2022), pp. 290–296.

[188] Trevor Cousins, Aylwyn Scally, and Richard Durbin. “A structured coalescent model reveals deep ancestral structure shared by all modern humans”. In: *bioRxiv* (2024), pp. 2024–03.

[189] Heng Li and Richard Durbin. “Inference of human population history from individual whole-genome sequences”. In: *Nature* 475.7357 (2011), pp. 493–496.

[190] Mark Lipson et al. “Ancient West African foragers in the context of African population history”. In: *Nature* 577.7792 (2020), pp. 665–670.

[191] Carina M Schlebusch et al. “Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago”. In: *Science* 358.6363 (2017), pp. 652–655.

[192] David A Murphy et al. “Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements”. In: *Elife* 12 (2022), e76065.

[193] Shuoguo Wang et al. “Apparent variation in Neanderthal admixture among African populations is consistent with gene flow from non-African populations”. In: *Genome biology and evolution* 5.11 (2013), pp. 2075–2081.

[194] Nina Hollfelder et al. “Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations”. In: *PLoS Genetics* 13.8 (2017), e1006976.

[195] Kay Prüfer et al. “The complete genome sequence of a Neanderthal from the Altai Mountains”. In: *Nature* 505.7481 (2014), pp. 43–49.

[196] Laurits Skov et al. “The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes”. In: *Nature* 582.7810 (2020), pp. 78–83.

[197] Matthias Meyer et al. “A high-coverage genome sequence from an archaic Denisovan individual”. In: *Science* 338.6104 (2012), pp. 222–226.

[198] Leonardo NM Iasi et al. “Neandertal ancestry through time: Insights from genomes of ancient and present-day humans”. In: *bioRxiv* (2024).

[199] Sriram Sankararaman et al. “The combined landscape of Denisovan and Neanderthal ancestry in present-day humans”. In: *Current Biology* 26.9 (2016), pp. 1241–1247.

[200] Kelley Harris and Jonathan K Pritchard. “Rapid evolution of the human mutation spectrum”. In: *Elife* 6 (2017), e24284.

[201] Laurent Duret and Nicolas Galtier. “Biased gene conversion and the evolution of mammalian genomic landscapes”. In: *Annual review of genomics and human genetics* 10.1 (2009), pp. 285–311.

[202] Leo Speidel et al. “Inferring population histories for ancient genomes using genome-wide genealogies”. In: *Molecular Biology and Evolution* 38.9 (2021), pp. 3497–3511.

[203] Graham Coop et al. “High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans”. In: *science* 319.5868 (2008), pp. 1395–1398.

[204] Simon Myers et al. “A common sequence motif associated with recombination hot spots and genome instability in humans”. In: *Nature genetics* 40.9 (2008), pp. 1124–1129.

[205] Simon Myers et al. “Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination”. In: *Science* 327.5967 (2010), pp. 876–879.

[206] Anjali G Hinch et al. “The landscape of recombination in African Americans”. In: *Nature* 476.7359 (2011), pp. 170–175.

[207] Daniel Wegmann et al. “Recombination rates in admixed individuals identified by ancestry-based inference”. In: *Nature genetics* 43.9 (2011), pp. 847–853.

[208] Benjamin Alleva et al. “Cataloging human PRDM9 allelic variation using long-read sequencing reveals PRDM9 population specificity and two distinct groupings of related alleles”. In: *Frontiers in cell and developmental biology* 9 (2021), p. 675286.

[209] Zachary Baker et al. “Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates”. In: *Elife* 6 (2017), e24133.

[210] Cathryn M Lewis and Evangelos Vassos. “Polygenic risk scores: from research tools to clinical instruments”. In: *Genome medicine* 12.1 (2020), p. 44.

[211] Alicia R Martin et al. “Clinical use of current polygenic risk scores may exacerbate health disparities”. In: *Nature genetics* 51.4 (2019), pp. 584–591.

[212] Yi Ding et al. “Polygenic scoring accuracy varies across the genetic ancestry continuum”. In: *Nature* 618.7966 (2023), pp. 774–781.

[213] Olivier Delaneau et al. “Accurate, scalable and integrative haplotype estimation”. In: *Nature communications* 10.1 (2019), p. 5436.

[214] Sile Hu et al. “Leveraging fine-scale population structure reveals conservation in genetic effect sizes between human populations across a range of human phenotypes”. In: *bioRxiv* (2023), pp. 2023–08.

[215] Dragana Vuckovic et al. “The polygenic and monogenic basis of blood traits and diseases”. In: *Cell* 182.5 (2020), pp. 1214–1231.

[216] Simone Rubinacci et al. “Efficient phasing and imputation of low-coverage sequencing data using large reference panels”. In: *Nature Genetics* 53.1 (2021), pp. 120–126.

[217] Bárbara Sousa da Mota et al. “Imputation of ancient human genomes”. In: *Nature Communications* 14.1 (2023), p. 3660.

[218] Niles Lehman et al. “Introgression of coyote mitochondrial DNA into sympatric North American gray wolf populations”. In: *Evolution* 45.1 (1991), pp. 104–119.

[219] Christine A Bozarth et al. “Coyote colonization of northern Virginia and admixture with Great Lakes wolves”. In: *Journal of Mammalogy* 92.5 (2011), pp. 1070–1080.

[220] Bridgett M vonHoldt et al. “Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf”. In: *Science Advances* 2.7 (2016), e1501714.

[221] Anders Bergström et al. “Grey wolf genomic history reveals a dual ancestry of dogs”. In: *Nature* 607.7918 (2022), pp. 313–320.

[222] Adam Auton et al. “Genetic recombination is targeted towards gene promoter regions in dogs”. In: *PLoS genetics* 9.12 (2013), e1003984.

[223] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. “Variational inference: A review for statisticians”. In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[224] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[225] Jeffrey Spence. “Flexible mean field variational inference using mixtures of non-overlapping exponential families”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 19642–19654.

[226] Matthew D Hoffman et al. “Stochastic variational inference”. In: *Journal of Machine Learning Research* (2013).

[227] Herbert Robbins and Sutton Monro. “A stochastic approximation method”. In: *The annals of mathematical statistics* (1951), pp. 400–407.

[228] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[229] Alex Graves. “Practical variational inference for neural networks”. In: *Advances in neural information processing systems* 24 (2011).

[230] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[231] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[232] Gulnara R Svishcheva et al. “Rapid variance components–based method for whole-genome association analysis”. In: *Nature genetics* 44.10 (2012), pp. 1166–1170.

[233] Johanna Jakobsdottir and Mary Sara McPeek. “MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals”. In: *The American Journal of Human Genetics* 92.5 (2013), pp. 652–666.

[234] Jian Yang et al. “Genomic inflation factors under polygenic inheritance”. In: *European journal of human genetics* 19.7 (2011), pp. 807–812.

[235] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. “Numba: A llvm-based python jit compiler”. In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. 2015, pp. 1–6.

[236] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. “The concrete distribution: A continuous relaxation of discrete random variables”. In: *arXiv preprint arXiv:1611.00712* (2016).

[237] Eric Jang, Shixiang Gu, and Ben Poole. “Categorical reparameterization with gumbel-softmax”. In: *arXiv preprint arXiv:1611.01144* (2016).

[238] Durk P Kingma, Tim Salimans, and Max Welling. “Variational dropout and the local reparameterization trick”. In: *Advances in neural information processing systems* 28 (2015).

[239] JM Hammersley and KW Morton. “A new Monte Carlo technique: antithetic variates”. In: *Mathematical proceedings of the Cambridge philosophical society*. Vol. 52. 3. Cambridge University Press. 1956, pp. 449–475.

[240] Charles Blundell et al. “Weight uncertainty in neural networks”. In: *arXiv preprint arXiv:1505.05424* (2015).

[241] Oran Shayer, Dan Levi, and Ethan Fetaya. “Learning discrete weights using the local reparameterization trick”. In: *arXiv preprint arXiv:1710.07739* (2017).

[242] Andrey Ziyatdinov et al. “Estimating the effective sample size in association studies of quantitative traits”. In: *G3* 11.6 (2021), jkab057.

[243] Ani Manichaikul et al. “Robust relationship inference in genome-wide association studies”. In: *Bioinformatics* 26.22 (2010), pp. 2867–2873.

[244] Adam Paszke et al. “Automatic differentiation in pytorch”. In: (2017).

[245] Brand Fortner. "HDF: The hierarchical data format". In: *Dr Dobb's J Software Tools Prof Program* 23.5 (1998), p. 42.

[246] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. URL: https://doi.org/10.1038/s41586-020-2649-2.

[247] The Haplotype Reference Consortium. "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature genetics* 48.10 (2016), pp. 1279–1283.

[248] The UK10K Consortium. "The UK10K project identifies rare variants in health and disease". In: *Nature* 526.7571 (2015), pp. 82–90.

[249] Gavin Band and Jonathan Marchini. "BGEN: a binary file format for imputed genotype and haplotype data". In: *BioRxiv* (2018), p. 308296.

[250] Benjamin B Sun et al. "Plasma proteomic associations with genetics and health in the UK Biobank". In: *Nature* 622.7982 (2023), pp. 329–338.

[251] Ryan S Dhindsa et al. "Rare variant associations with plasma protein levels in the UK Biobank". In: *Nature* 622.7982 (2023), pp. 339–347.

[252] Jian Zeng et al. "Signatures of negative selection in the genetic architecture of human complex traits". In: *Nature genetics* 50.5 (2018), pp. 746–753.

[253] Armin P Schoech et al. "Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection". In: *Nature communications* 10.1 (2019), p. 790.

[254] Yunfeng Ruan et al. "Improving polygenic prediction in ancestrally diverse populations". In: *Nature genetics* 54.5 (2022), pp. 573–580.

[255] Simon Haworth et al. "Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis". In: *Nature communications* 10.1 (2019), p. 333.

[256] Mashaal Sohail et al. "Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies". In: *Elife* 8 (2019), e39702.

[257] Jeremy J Berg et al. "Reduced signal for polygenic adaptation of height in UK Biobank". In: *Elife* 8 (2019), e39725.

[258] Eli A Stahl et al. "Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis". In: *Nature genetics* 44.5 (2012), pp. 483–489.

[259] Luke J O'Connor et al. "Extreme polygenicity of complex traits is explained by negative selection". In: *The American Journal of Human Genetics* 105.3 (2019), pp. 456–476.

[260] Steven Gazal et al. "Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection". In: *Nature genetics* 49.10 (2017), pp. 1421–1427.

[261] Hilary K Finucane et al. "Partitioning heritability by functional annotation using genome-wide association summary statistics". In: *Nature genetics* 47.11 (2015), pp. 1228–1235.

[262] Jian Zeng et al. “Widespread signatures of natural selection across human complex traits and functional genomic categories”. In: *Nature Communications* 12.1 (2021), p. 1164.

[263] Luke Jostins et al. “Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease”. In: *Nature* 491.7422 (2012), pp. 119–124.

[264] Meta-Analyses of Glucose et al. “Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes”. In: *Nature genetics* 44.9 (2012), pp. 981–990.

[265] Patrick CA Dubois et al. “Multiple common variants for celiac disease influencing immune gene expression”. In: *Nature genetics* 42.4 (2010), pp. 295–302.

[266] Mats Nagel et al. “Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways”. In: *Nature genetics* 50.7 (2018), pp. 920–927.

[267] Katrina M De Lange et al. “Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease”. In: *Nature genetics* 49.2 (2017), pp. 256–261.

[268] Qin Qin Huang et al. “Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals”. In: *Nature Communications* 13.1 (2022), p. 4664.

[269] Bjarni J Vilhjálmsson et al. “Modeling linkage disequilibrium increases accuracy of polygenic risk scores”. In: *The american journal of human genetics* 97.4 (2015), pp. 576–592.

[270] Luke R Lloyd-Jones et al. “Improved polygenic prediction by Bayesian multiple regression on summary statistics”. In: *Nature communications* 10.1 (2019), pp. 1–11.

[271] Florian Privé, Julyan Arbel, and Bjarni J Vilhjálmsson. “LDpred2: better, faster, stronger”. In: *Bioinformatics* 36.22-23 (2020), pp. 5424–5431.

[272] Jack Euesden, Cathryn M Lewis, and Paul F O’reilly. “PRSice: polygenic risk score software”. In: *Bioinformatics* 31.9 (2015), pp. 1466–1468.

[273] Tian Ge et al. “Polygenic prediction via Bayesian regression and continuous shrinkage priors”. In: *Nature communications* 10.1 (2019), p. 1776.

[274] Omer Weissbrod et al. “Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores”. In: *Nature Genetics* 54.4 (2022), pp. 450–458.

[275] Declan Bennett et al. “Controlling for background genetic effects using polygenic scores improves the power of genome-wide association studies”. In: *Scientific Reports* 11.1 (2021), p. 19571.

[276] Adrian I Campos et al. “Boosting the power of genome-wide association studies within and across ancestries by using polygenic scores”. In: *Nature Genetics* 55.10 (2023), pp. 1769–1776.

[277] Sean J Jurgens et al. “Adjusting for common variant polygenic scores improves yield in rare variant association analyses”. In: *Nature Genetics* 55.4 (2023), pp. 544–548.

[278] Tabea Schoeler et al. “Participation bias in the UK Biobank distorts genetic associations and downstream analyses”. In: *Nature Human Behaviour* 7.7 (2023), pp. 1216–1227.

[279] David Firth. “Bias reduction of maximum likelihood estimates”. In: *Biometrika* 80.1 (1993), pp. 27–38.

[280] Department of Statistics, University of Oxford. *Quickdraws Dataset Repository*. Accessed: 2025-02-03. 2025. URL: `https://www.stats.ox.ac.uk/publication-data/sge/ppg/quickdraws/`.

[281] Johannes Kettunen et al. “Genome-wide association study identifies multiple loci influencing human serum metabolite levels”. In: *Nature genetics* 44.3 (2012), pp. 269–276.

[282] Omer Weissbrod et al. “Functionally informed fine-mapping and polygenic localization of complex trait heritability”. In: *Nature genetics* 52.12 (2020), pp. 1355–1363.

[283] Arthur Korte et al. “A mixed-model approach for genome-wide association studies of correlated traits in structured populations”. In: *Nature genetics* 44.9 (2012), pp. 1066–1071.

[284] Xiang Zhou and Matthew Stephens. “Efficient multivariate linear mixed model algorithms for genome-wide association studies”. In: *Nature methods* 11.4 (2014), pp. 407–409.

[285] Nicola Pirastu et al. “Genetic analyses identify widespread sex-differential participation bias”. In: *Nature Genetics* 53.5 (2021), pp. 663–671.

[286] Stefania Benonisdottir and Augustine Kong. “Studying the genetics of participation using footprints left on the ascertained genotypes”. In: *Nature Genetics* 55.8 (2023), pp. 1413–1420.

[287] Haley M Amemiya, Anshul Kundaje, and Alan P Boyle. “The ENCODE blacklist: identification of problematic regions of the genome”. In: *Scientific reports* 9.1 (2019), p. 9354.