UNIVERSITY OF CALIFORNIA
SANTA CRUZ

NEW TOOLS FOR LOCALIZING ANCESTRY ACROSS GENOMES OF HYBRID INDIVIDUALS PROVIDE INSIGHT INTO THE GENETICS OF SPECIATION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Nathan K. Schaefer

March 2019

The Dissertation of Nathan K. Schaefer
is approved:

Professor Richard E. Green, Chair

Professor Beth Shapiro

Professor Grant Pogson

Lori Kletzer
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures **v**

List of Tables **xxxiii**

Abstract **xli**

Acknowledgments **xliii**

1 Introduction **1**

1.0.1 Detecting admixture without archaic genomes 3
1.0.2 Detecting admixture with archaic genomes 7
1.0.3 Local methods 7
1.0.4 Global methods 10
1.0.5 D-statistic 13
1.0.6 Detecting admixture with archaic hominins 16
1.0.7 Inferring modern human migrations 21
1.0.8 Detecting ancient admixture in other species 24
1.0.9 Conclusion 25

2 Inferring ancestry across low-coverage, downsampled genomes **30**

2.0.1 Background 31
2.0.2 Results 33
2.1 Discussion 56
2.2 Conclusion 59
2.3 Methods 60
2.3.1 Model description 60
2.3.2 Transition probabilities 61
2.3.3 Emission probabilities 64
2.3.4 Potential pitfalls 65
2.3.5 Simulations 66
2.3.6 Human and Neanderthal data 70
2.3.7 Bear data preparation 72

2.3.8 Low-coverage tests 75
2.3.9 Shared polar bear ancestry 76
2.3.10 Functional analyses 77

3 A method for fast, heuristic ancestral recombination graph inference 87
3.1 Background 88
3.2 Results 90
3.3 Discussion 97
3.4 Methods 98
3.4.1 Ancestral recombination graph inference 101
3.4.2 Inferring branch movements between two trees 103
3.4.3 General case algorithm 104
3.4.4 Simulations 109
3.4.5 Plotting tree articulation against mutation/recombination rate ratio . . . 110

4 Insights into human history from ancestral recombination graph inference 113
4.1 Background 113
4.2 Results 115
4.2.1 Discussion 172
4.3 Methods 175
4.3.1 Data processing 175
4.3.2 Simulations 175
4.3.3 Admixture scans 176
4.3.4 Testing for overlap 179
4.3.5 Analysis of introgressed haplotypes 179
4.3.6 Determining size of deserts 180
4.3.7 Analysis of human-specific alleles in deserts 181

5 Conclusion 185

A Supplementary methods for Chapter 2 188
A.1 HMM transition probabilities: computing z 188
A.2 HMM transition probabilities 191
A.3 Emission probability distributions: expectation 197

B Supplementary methods for Chapter 3 200
B.1 Adding nodes to the ARG 200
B.2 Solving ancestral recombination events 205
B.3 Finalizing ARG node ranges 215
B.4 Collapsing to trees 217

Bibliography 218

List of Figures

1.1 Library molecules for high-throughput sequencing (HTS) consist of target DNA fragments with adapter sequences ligated on either end. Adapters, with known sequence complementary to primer sequences, allow a single primer pair to amplify a diversity of DNA fragments, and another to be used for the sequencing reaction, where labeled nucleotides are incorporated [125]. For ancient DNA studies, HTS technology has allowed researchers to observe damage patterns at ends of molecules and amplify a large variety of genomic DNA fragments of unknown sequence. HTS size limitations are inconsequential, as ancient DNA is usually highly fragmented. 3

1.2 Overview of popular techniques for studying archaic admixture. a: Archaic genome free methods are test statistics that can be used to infer archaic introgression into modern individuals without archaic sequence data. Each is computed on real data, then data simulated under various demographic models, and compared. These are prone to errors in model specification and can produce false positives. b: Local methods can be used to find specific genes or genomic regions admixed individuals derive from one or another ancestral population. These are tuned to detect detect long introgressed haplotypes but have reduced power to detect old admixture events. c: Global methods consider individual sites across the genome. Many are formal tests for admixture and/or can be used to estimate admixture proportion. In each box, X" means true and S" means true in some cases. *" indicates methods applied to haplotype sequences, to which the concept of phasing does not apply. Note that, if sufficiently high-coverage genome-wide sequence data are available, these can be transformed into SNP calls if necessary. Also note that a method working on population-level data requires reference population data by default, as all inputs are population-level. 4

1.3 Adapted from [147]. Expected value of $f_3(C;A,B)$ under various tree topologies. Red lines trace genetic drift between populations C and A; blue lines trace genetic drift between C and B. f_3 measures drift between C and A that is also shared between C and B. Drift is shared along branches where arrows going in the same direction overlap. a and b: expected value of $f_3(C;A,B)$ with no admixture. If C is not a product of admixture between A and B, f_3 is expected to be positive. In the case where C is an outgroup to A and B (a), the value of f_3 is proportional to the distance separating C from A and B, which can also be thought of as the amount of shared history between A and B. c-f: expected value of when C is a product of admixture between A and B. α is the percent ancestry population C derives from A, and β is the percent derived from B. Distance j represents genetic drift between extant population A its ancestral population that admixed to form the population ancestral to C in the past; distance k is proportional to drift between extant population B and its admixing ancestral population. Computation of $f_3(C;A,B)$ in this case requires tracing multiple paths through the tree, since population C can share drift with population B that it received through admixture with population A and vice versa. The expectation is the sum of all shared drift: $E[f_3(C;A,B)] = \alpha\beta i + \alpha^2(i+j) + \beta^2(i+k) + \alpha\beta(i-p-q)$. This has the potential to be negative, although it can also be positive. Given that negative values are impossible if C is not a result of admixture (a and b), a negative result can be taken as evidence of admixture; a positive result, however, cannot be used to reject admixture. 27

1.4 Adapted from [181]. Visual explanation of expected values of $f_4(A,B;C,D)$ under various tree topologies. Red lines trace genetic drift from A to B; blue lines trace drift from C to D. f_4 measures drift shared between A and B that is also shared between C and D. Drift is shared along branches where arrows overlap going in the same direction. a-c: Positive, negative, and zero values of f_4 give support for different tree topologies relating the four populations. d, e: visual explanation of f_4 ratio method for inferring admixture proportion. Population X is a mixture of populations related to B and C; population D is an outgroup. The quantity of interest, α, is the proportion of ancestry population X has received from B. If the expected value of $f_4(A,D;B,C) = z$ (d), then the expected value of $f_4(A,D;X,C) = \alpha z$(e). It follows that $\alpha = f_4(A,D;B,C)/f_4(A,D;X,C)$ (Patterson et al. 2012). 28

1.5 Explanation of D statistic [56, 35]. Individuals are numbered according to the D-statistic notation: $D(P_1,P_2,P_3,P_4)$ and examples of individuals that could be used to yield a positive D-statistic result when testing for Neanderthal ancestry are given (D would be negative in this case if there had been gene flow between the Yoruba and Neanderthal instead). a: genome-wide tree relating the four individuals, based on prior knowledge. b: trees at ABBA and BABA sites used to compute D. In both, blue is used to represent a derived allele (does not match chimpanzee); red represents an ancestral allele (matches chimpanzee). To calculate D on sequence data, the number of sites with the topology of the left tree is N_{ABBA} and the number of sites with the topology of the right tree is N_{BABA}. Then $D = (N_{ABBA} - N_{BABA})/(N_{ABBA} + N_{BABA})$. 29

2.1 AD-LIBS accuracy on simulated data, using incorrect population parameters. Simulations here used the single-pulse" admixture model (see Figure 2.2 A, C, and E) except where otherwise noted, with 10kb windows, which were automatically adjusted by AD-LIBS as necessary. Asterisks denote accuracy significantly lower ($p < 0.001$) than that obtained using correct parameters. A: AD-LIBS accuracy using different prior population size estimates (true N = 3000) and correct number of generations since admixture (1000). B: AD-LIBS accuracy using different estimates for the number of generations since admixture (true g = 1000) and correct population size (3000). C: AD-LIBS accuracy using prior estimates of polar bear admixture proportion that differed from the true value, rounded to the nearest 10%. D: Same as C, but using the migration" admixture model (see Figure 2.2 B, D, and F), which produced a wider range of true admixture proportions. 35

2.2 AD-LIBS accuracy using simulated data. A, C, and E refer to simulations with a single polar-brown bear admixture event 12,000 years ago, followed by isolation (single-pulse model); B, D, and F refer to simulations in which a brown-polar bear admixture event 12,000 years ago is followed by continual breeding with unadmixed brown bears (migration model). A and B: percent of AD-LIBS inferences correct and percent of true ancestry recovered in each ancestry state (homozygous polar bear, heterozygous, and homozygous brown bear) for each individual. C and D: Effect of using different numbers of reference ancestral individuals (1, 2, 3, 4, or 5 from each population) on overall accuracy, using 10kb windows. Asterisks denote a distribution mean significantly lower ($p < 0.001$) than the best distribution (5 individuals from both populations), according to t-test. E and F: Effect of using different window sizes (5kb, 10kb, and 25kb), with 5 reference bears from each ancestral population. Asterisks denote a distribution mean significantly lower ($p < 0.001$) than the best distribution (10kb windows), according to t-test. 38

2.3 Accuracy of AD-LIBS estimates of the overall extent of polar bear ancestry, using simulated data. A and C refer to simulations with a single polar-brown bear admixture event 12,000 years ago, followed by isolation (single-pulse model); B and D refer to simulations in which a brown-polar bear admixture event 12,000 years ago is followed by continual breeding with unadmixed brown bears (migration model). All AD-LIBS runs in this figure used 5 reference individuals per ancestral population and 10kb windows. A and B: Inferred percent polar bear ancestry using AD-LIBS and $\hat{f}$ versus true percent polar bear ancestry. C and D: inferred percent polar bear ancestry of each type, according to AD-LIBS, versus true percent polar bear ancestry of each type. Each point represents the percent of a single simulated hybrid bear genome with a specific type of ancestry. E: overall accuracy of AD-LIBS inferences versus true percent polar bear ancestry, including both types of simulations. The line of best fit by least-squares regression is also shown. Accuracy decreases slightly as polar bear ancestry increases, probably due to the tendency of AD-LIBS to overestimate the extent of heterozygous ancestry (C and D). 79

2.4 Results from downsampling four ABC Islands brown bears, three Scandinavian brown bears, and four polar bears to 0.5x, 1x, 2x, 5x, and 10x coverage along the longest genomic scaffold, running HAPMIX [160] and AD-LIBS on the four ABC Islands bears at each coverage depth, and comparing these runs to results obtained from running both programs on the full-coverage versions of the same individuals. Each line represents an individual ABC Islands bear and each color represents a specific low coverage/full coverage comparison. A: percent of full coverage calls recovered by running each program at low coverage. Values given are averages across the three ancestry states (homozygous polar bear, heterozygous, and homozygous brown bear). B: percent of low coverage calls that were correct, according to full-coverage calls. Values given are averages across the three ancestry states. Some points are missing because HAPMIX was unable to detect any polar bear ancestry at 0.5x coverage. 80

2.5 Results from downsampling four ABC Islands brown bears, three Scandinavian brown bears, and four polar bears to 0.5x, 1x, 2x, 5x, and 10x coverage along the longest genomic scaffold, running HAPMIX and AD-LIBS on the four ABC Islands bears at each coverage depth, and comparing these runs to results obtained from running both programs on the full-coverage versions of the same individuals. Each line represents an individual ABC Islands bear and each color represents a specific low coverage/full coverage comparison. A and B assess homozygous polar bear (AA) calls, C and D assess heterozygous (AB) calls, and E and F assess homozygous brown bear (BB) calls. A, C, and E measure the percent of low-coverage calls that were correct" according to the high-coverage runs, while B, D, and F measure the percent of the high-coverage runs calls that were also detected by the low-coverage runs. In almost every case, AD-LIBS is more consistent with itself than other comparisons. We also note that low-coverage AD-LIBS inferences of homozygous polar and brown bear ancestry are more often correct, according to HAPMIX run at full coverage, than HAPMIX run at low coverage (A and E). AD-LIBS may, however, erroneously call more windows heterozygous than HAPMIX does (C), leading to its missing some windows of homozygous polar (B) and brown bear (F) ancestry. 81

2.6 Comparing overlap of regions of different types of ancestry among hybrid bears. For every combination of 2 or more American brown bears, we measured the number of bases that AD-LIBS labeled with the same type of ancestry (homozygous polar, homozygous/heterozygous polar, or homozygous brown) in each bear. We also performed one random trial per real comparison, in which coordinates comprising random regions were sampled from the reference genome, producing sets of genome regions of the same size and number as the regions of ancestry produced by AD-LIBS for each bear, but randomly scattered across the genome. We then measured the overlap between these random ancestry regions for the sake of comparing to the true overlap. Averages of every comparison of each number of bears are shown as solid lines, and averages of every comparison of randomized versions of those same bears are shows as dashed lines. 82

2.7 Geographic signal recovered in vectors of polar bear ancestry. A: principal components analysis (PCA) of polar bear ancestry state of 10kb genomic windows for 18 brown bears, using EIGENSOFT SmartPCA [148]. B: PCA of SNP data from a previous study [108], including a subset of the bears in A. Both plots show similar geographic patterns, with the Montana bear (GP01) falling close to the Admiralty Island bear(s), but only the SNP data separates Finnish (RF01) from Swedish brown bears. 83

2.8 Illustration of cases where either a local ancestry detection method (like AD-LIBS) or a global ancestry detection (like $\hat{f}$) might succeed, partially succeed, or fail. Each line represents a chromosome, with polar bear ancestry shown in blue and brown bear ancestry shown in brown. All five individuals needed for computation of are shown in each case. A: local and global methods both succeed in detecting all of the hybrid individuals polar bear ancestry. B: local and global methods both fail to detect the hybrid individuals polar bear ancestry. C: local methods successfully detect the hybrid individuals polar bear ancestry, since it is in a different part of the genome than the polar bear ancestry in the genome of the model "unadmixed" brown bear. Global methods fail to detect the hybrid individuals polar bear ancestry. Since global methods use genome-wide averages, the hybrid individual is not seen to possess any more polar bear ancestry than the model "unadmixed" brown bear. D: Both local and global methods will detect the hybrids first segment of polar bear ancestry but fail to detect the second segment, resulting in both types of methods underestimating the hybrid individuals true percent polar bear ancestry. 84

2.9 The state space of AD-LIBSs hidden Markov model. The three round states (AA, AB, and BB) are ancestry states that can emit scores. AA represents regions where both homologous chromosomes derive ancestry from ancestral population A, AB represents regions of heterozygous ancestry, and BB represents regions homozygous for population B ancestry. The three square states (sAA, sAB, and sBB) are skip states, each associated with one of the three ancestry states. Skip states can only emit scores representing windows of the genome in which data are too sparse to infer ancestry. Each skip state is more likely to transition back to its associated ancestry state than to one of the others. Arrow colors represent different types of transition probabilities. Green arrows are starting probabilities and are related to the pre-estimated percent ancestry derived from each ancestral population (A and B). Blue arrows represent recombination events; their probability is related to the probability of a recombination event having happened at a given site in the time since admixture, as well as the probability of sampling a base from population A or B. Black arrows are related to the probability of skipping a given window, computed from the number of "N" bases encountered. Red arrows are transitions to the end state, with probabilities related to the number of windows on the chromosome or scaffold being scanned. Gold arrows represent probabilities that are computed after other probabilities, by subtracting from 1 the sum of all other transition probabilities out of a given state. 85

2.10 All possible combinations of events leading to transitions between the three ancestry states of the hidden Markov model. “A” and “B” denote chromosomes derived from ancestral populations A and B, and the three states AA, AB, and BB model regions where ancestry is homozygous from population A, heterozygous, and homozygous from population B, respectively (AB and BA are represented by same state, but are shown separately here to clarify that the ancestry of both separate chromosomes must be considered when computing probabilities). The other columns denote possible recombination-related events on the two parental homologues of a given chromosome (henceforth “homologue 1” and “homologue 2”). A “Y” in the R1 column signifies that recombination took place at a given base on chromosome homologue 1 in the time since admixture, and a “Y” in the R2 column signifies that recombination took place at this base on chromosome homologue 2. C1a and C1b refer to the ancestry of the bases on chromosome homologue 1 immediately before and after the recombination event, if it happened; C2a and C2b refer to the ancestry of the bases on chromosome homologue 2 before and after recombination. Z indicates that the same ancestral recombination event, which happened in the time since admixture, was sampled twice in the same individual (once on chromosome homologue 1 and once on chromosome homologue 2). The parameter *g* is the number of generations since admixture, *r* is the recombination probability per site per generation, assumed to be 1 cM/Mb or 10^{-8} per site, and *z* is the probability of resampling the same ancestral recombination event twice in one individual, according to genetic drift approximated by the Wright-Fisher model. 86

3.1 A: example showing ARG inferred by parsimony. Columns are variable sites; shaded cells denote shared derived mutations. Red-highlighted sites fail the four haplotype test with other red- 5 highlighted sites (shown by brackets), vertical dotted red lines mark ancestral recombination breakpoints. Site numbers mark clades they tag in the trees, and red arrows show ancestral recombination events. B: example of a recombination event joining together the blue haplotype upstream of the break point (dotted line) and the red haplotype downstream of the breakpoint. Location of the blue and red haplotypes in a consensus tree are shown above. C: how the daughter haplotype produced in B will appear in an ARG: within the blue haplotypes clade upstream of the recombination breakpoint and in the red haplotypes clade downstream of it. 89

3.2 Example of algorithm for inferring branch movements between to trees known *a priori*. For more information on terminology, see Methods. A: Two trees, which differ by one branch movement. B: Clades from the two trees that fail the four haplotype test. Left column shows clades from the first (upstream) tree and right column shows clades from the second (downstream) tree; arrows indicate four haplotype test failures. C: Graph showing all possible branch movements that could explain the four haplotype test failures shown in B. The left and right columns are "tree" nodes, while the center column lists candidate γ clades. Colors indicate types of four haplotype test failures: red paths are conditional on a failure being the α/α type, green on it being α/β, and blue on it being β/β. In this case, a single candidate γ clade (C) has the most edges and can explain all four haplotype test failures. This is interpreted as the clade C moving from the smallest observed α clade in the first tree (CD) to the smallest observed β clade in the second tree (CH). If no β clades from the second tree are observed, the branch movement goes upward to a clade containing the union of all clades failing the four haplotype test. If no α clades from the first tree are observed, the branch movement goes downward from a clade containing the union of all clades failing the four haplotype test. 92

3.3 Schematic of data structure. For more information on terminology, see Methods. Top: rectangles are "tree nodes" representing clades in trees. Each has a set of haplotypes (represented by letters A-G), and a start and end coordinate (blue numbers in brackets) determined by coordinates of SNP sites tagging the clade (yellow numbers in braces), along with a propagation distance parameter (100 in this example). Parent/child edges (vertical arrows) also have start and end coordinates determined by the nodes. Ovals are candidate nodes that can explain four gamete test failures; colored edges indicate potential paths between tree nodes through candidate nodes that could explain four gamete test failures (colors indicate types of paths). The candidate node with the most edges is eventually chosen as the most parsimonious branch movement, allowing for the inference of new nodes; the two trees at the bottom show the "solved" ancestral recombination event with the branch movement marked in red and all clades inferred without SNP data marked with yellow stars. The coordinates of the recombination event (blue numbers in brackets) are taken to be midway between the furthest-downstream upstream site and the furthest-upstream downstream site involved in recombination. 93

3.4 A: Accuracy of SARGE on simulated data (defined as percent of all clades correct according to the true ARG in the simulation), with increasing numbers of human-like haplotypes from an unstructured population. Error bars are one standard deviation across 5 replicates. B: Number of nodes per tree with increasing number of haplotypes in simulated data. 95

3.5 Properties of SARGE performance on simulated data with a sub-Saharan African-like level of heterozygosity, constant population size history, and no structure. Points are means; error bars show one standard deviation. A: Tree articulation as a percent of all nodes possible (given the number of haplotypes), with increasing number of haplotypes. B: Percent of all clades (across all trees) inferred from solving recombination events (rather than shared mutations). C: Number of trees across the chromosome with two children of the root node (no root-level polytomies). D: Execution time as a function of the number of input haplotypes. Real data, where SNPs and recombination events cluster in the genome, is likely to increase execution time. 96

3.6 Comparisons of SARGE to two other ARG inference programs, using a simulated data with sub-Saharan African-like heterozosity, constant population size, and no structure. In each comparison, error bars represent one standard deviation across 5 replicates and ARGs were inferred across data sets with increasing numbers of haplotypes. A: percent of nodes correct, given the true ARG from the simulation. B: Number of nodes per tree. C: Kendall-Coljin distance [85] from true trees, without considering branch lengths. This metric was used in another study [83] and is negatively affected by polytomies. D: Execution time. Since tsinfer is a python module, its execution time does not include writing data to disk, while the other two execution times do. E: Same as D, but with Rent+ excluded. 111

3.7 A: Genome-wide UPGMA trees from SNP data (top and bottom) against similarity matrix from shared recombination events inferred by SARGE. Dark red (similar) groups are Native Americans and Papuans. Outliers (dark lines) are S_Naxi haplotypes; another study reported likely improper phasing in this sample [83]. B: Number of nodes per tree against the ratio of mutation rate to recombination rate in 50kb windows. 112

3.8 Different types of four haplotype test failures. In each, the γ clade is highlighted in purple, α in red, and β in blue. A: Lateral branch movement. Four haplotype test failures of type α/α, α/β, and β/β are observed. B: Upward branch movement. Only α/α four haplotype test failures are observed. C: Downward branch movement. Only β/β four haplotype test failures are observed. 112

4.1 Separating admixed haplotypes from incomplete lineage sorting (ILS). A: cartoon of properties expected of ILS vs. admixed haplotypes. B: Using the true ARG from a simulation involving Neanderthal admixture into modern humans 50kya, ILS is separable from admixture by considering both the TMRCA to admixer and the length of the haplotype. C: In real data, these two distributions are not separable. We computed an admixture p-value for each haplotype and binned haplotypes into admixture and ILS based on varying p-value cutoffs, and for each bin plotted the coefficient of variation (standard deviation divided by mean) in the overall extent of ILS and admixture across SGDP populations. We expect the true coefficient of variation in ILS to be low across populations, so we chose a p-value cutoff that minimized this value (p = 0.16). D: Real data from the SGDP data set binned using the p-value determined in C. 117

4.2 A: Genome-wide percent Neanderthal, Denisovan, and ambiguous (either Neanderthal or Denisovan) across SGDP populations, using the ARG and the D-statistic. D-statistic calculations considered only one archaic population at a time as introgressor and thus does not detect ambiguous ancestry and also might count some Denisovan ancestry as Neanderthal, and vice-versa. B: Times to most recent common ancestor of confidently introgressed ($p < 0.001$) segments across SGDP populations. Conversions assume 6.5 mya human-chimpanzee divergence time and branch shortening values from [163]. C: Mean frequency in all humans of each confidently introgressed ($p < 0.001$) segment across SGDP populations (top panel) and mean within-population frequency of each segment across SGDP populations (bottom panel). D: Sharing (Jaccard statistic) of Neanderthal-introgressed haplotypes (matrix). Haplotypes are ordered by a UPGMA tree using input SNP data (top and left). Populations are colored the same way as in B and C, and self-self comparisons are set to 0 similarity in order to not saturate the color scale. E: Sharing (Jaccard statistic) of Denisovan-introgressed haplotypes (matrix). Haplotypes are ordered by a UPGMA tree using input SNP data (top and left). 118

4.3 Worldwide distribution of times to most recent common ancestor (TMRCA) to the closest Neanderthal haplotype of Neanderthal-like haplotypes in modern humans. Haplotypes are from the confident set ($p < 0.001$), and points are averages across all haplotypes within all genomes from each location. Numbers are corrected for branch shortening, using the mean of the values given for the two Neanderthal genomes in [163]. 120

4.4 Worldwide distribution of frequencies of individual Neanderthal-like haplotypes in modern humans. For each introgressed haplotype, its frequency in all humans worldwide was computed, and these values were averaged across all haplotypes within all human genomes from each geographic location. 121

4.5 Worldwide distribution of times to most recent common ancestor (TMRCA) to the closest Denisovan haplotype of Denisovan-like haplotypes introgressed in modern humans. Haplotypes are from the confident set ($p < 0.001$), and points are averages across all haplotypes within all genomes from each location. Numbers are corrected for branch shortening, using the mean of the values given for the Denisovan genome in [163]. 122

4.6 Worldwide distribution of frequencies of individual Denisovan-like haplotypes in modern humans. For each introgressed haplotype, its frequency in all humans worldwide was computed, and these values were averaged across all haplotypes within all human genomes from each geographic location. 123

4.7 Percent of archaic-introgressed haplotypes in Africa (excluding Biaka, Mbuti, and Khomani-San) shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency. 123

4.8 Percent of archaic-introgressed haplotypes in basal African lineages used as an outgroup (Biaka, Mbuti, and Khomani-San) shared with other SGDP populations. 124

4.9 Percent of archaic-introgressed haplotypes in America shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency. 124

4.10 Percent of archaic-introgressed haplotypes in CentralAsiaSiberia shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency. 125

4.11 Percent of archaic-introgressed haplotypes in EastAsia shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency. 125

4.12 Percent of archaic-introgressed haplotypes in Oceania shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency. 126

4.13 Percent of archaic-introgressed haplotypes in SouthAsia shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency. 126

4.14 Percent of archaic-introgressed haplotypes in WestEurasia shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency. 127

4.15 TMRCAs to Neanderthal for confident ($p < 0.001$) Neanderthal-like haplotypes (left panel) and TMRCAs to Denisovan for confident ($p < 0.001$) Denisovan-like haplotypes (right panel), corrected for branch shortening using values in [163] and plotted by whether unique to a specific SGDP population or shared among multiple populations. 127

4.16 A: Confident ($p < 0.001$) Neanderthal-like haplotypes that were also detected using an ARG inferred over a subset of the samples. The top five outliers (ordered by length times frequency in all humans) are marked with the genes they contain (one contained no genes). B: The top outlier Neanderthal haplotype (by length times frequency), showing 250kb upstream and downstream and genes contained within. C: Confident ($p < 0.001$) Denisovan-like haplotypes that were also detected using an ARG inferred over a subset of the samples. The top five outliers (ordered by length times frequency in all humans) are marked with the genes they contain (one contained no genes). B: The top outlier Denisovan haplotype (by length times frequency), showing 250kb upstream and downstream and genes contained within. This region also coincides with a Neanderthal haplotype detected in some populations. 132

4.17 Properties of desert regions (free of admixture and ILS) and human-specific regions (all humans share a derived lineage). A: Sizes of deserts and human-specific regions detected in real data, with different numbers of randomly subsampled human haplotypes. Number of haplotypes refers to modern human haplotypes only; all archaic hominin haplotypes were included in each run. B: Proportions of the autosomal genome containing archaic hominin admixture, ILS, and in deserts and human-specific regions, in real data and data simulated under a popular demographic model, with a single pulse of Neanderthal admixture 50 kya. C: Lengths of deserts and human-specific regions (bp). D: Mean TMRCA of all humans across each desert and human-specific region. E: For loci where all humans share a derived lineage in a regulatory element binding site, genes regulated by elements that bind at that site are considered, and all pairs of genes in this set whose products interact are considered. For each interacting pair of genes in this set, the TMRCA of all humans at the regulatory element mutation for each gene in the pair are shown. Selected time bins with the most genes are 33-42 kya, 145-185 kya, and 400-500 kya. 183

4.18 Screenshot of the region around the three protocadherin gene clusters on chromosome 5. Human-specific regions are shown in green, human derived allele frequencies are shown in red, and archaic derived allele frequencies are shown in black. Humans have little allelic diversity in human-specific regions, with most derived alleles being fixed or high frequency. 184

B.1 How node indices are adjusted when four haplotype test failures are encountered. Black letters represent clades, yellow numbers in curly braces represent site indices, and blue numbers in brackets represent start and end coordinates (inclusive). Red text indicates an adjusted value. Red arrows show four haplotype test failures, and black arrows represent changes made to nodes. A: a simple case where the furthest donwnstream site owned by node 1 (50) is upstream of the furthest upstream site owned by node 2 (150). In this case, node 1's end coordinate is set to its furthest downstream site, and node 2's start coordinate is set to its furthest upstream site. B: Node 4 interrupts the range of node 3. Node 3 must be split into two nodes, and all three resulting nodes must have their ranges adjusted. 204

B.2 Tree-compatibility in three different situations. Gray squares are nodes, black letters are clades, yellow numbers in curly braces are genomic positions, and blue numbers in brackets are start/end coordinates. Red arrows indicate four haplotype test failures, and green ovals denote tree compatibility. A: three pairs of nodes are compatible (can belong to the same trees as each other). B: only two pairs of nodes are tree-compatible. C: Only one pair of nodes is tree-compatible. 206

B.3 An example case in which multiple ancestral recombination events may be considered as one. A: The true ARG across three adjacent genomic regions. Clades involved in recombination are marked α and β; subscripts denote the recombination event (first or second) to which they correspond. Clades observed in SNP data appear below each tree in the order in which they are observed; colors mark the true tree to which each clade belongs. Purple branches are true γ clades, and purple arrows show ancestral recombination events. B: The correct grouping of nodes path through them in a recombination graph. First, (J) moved downward from (ABCDEFGHIJ) to (EFGHIJ). Then, (E) moved from (EFGHI) to (ABCE). C: A likely incorrect inference made, if nodes are not grouped correctly into trees. It appears most parsimonious to say that (J) moved down from (ABCDEFGHIJ) in the first tree to (FGHIJ) in the third tree, skipping the middle tree altogether. If this choice is made, genomic positions for the ancestral recombination event will also be wrong, as it chooses the narrowest possible interval, which would place it between the first and second tree. Note that observing the clade (ABCE) in the third tree might help avoid this problem. . . . 208

List of Tables

2.1 The accuracy of AD-LIBS ancestry inferences using simulated genomes. Two demographic models representative of the ABC Islands bears history were used: one in which a single admixture event between polar and brown bears takes place 12,000 years ago, followed by isolation of the hybrid population (Single-pulse model), and one in which admixture takes place at the same time but is followed by continuous brown bear migration from the mainland (Migration model). Overall accuracy is the percent of all bases for which true ancestry matched AD-LIBS-inferred ancestry. Since this number is weighted toward more common ancestry states, the average across all three ancestry states is also given. 37

2.2 Sample details. All sequence data were published in previous studies and downloaded as reads from the NCBI SRA. Coverage levels shown were estimated from numbers of raw reads before mapping to the reference genome. All samples were aligned to the polar bear reference genome, then subjected to base and map quality filtering, indel realignment, and duplicate removal. 40

2.3 Results of running AD-LIBS on autosomal sequences of five randomly chosen European (CEU) individuals from the 1000 Genomes Project, with the Altai Neanderthal and five randomly chosen Yoruba (YRI) individuals from the 1000 Genomes Project as reference individuals from admixing populations. Using AD-LIBS with a window size (10kb) lower than the recommended minimum of 14kb gives bad results, while using a window size above this threshold (15kb) gives much more reasonable results. 41

2.4 Percent ancestry of each type, as called by AD-LIBS and HAPMIX, in the four bears sequenced to sufficient coverage depth for variant calling. AD-LIBS calls more heterozygous ancestry than HAPMIX and probably overestimates heterozygous ancestry genome-wide. 45

2.5 Numbers of variants obtained from four polar bears (PB7, PB12, PB68, and PB105), three brown bears (OFS01, RF01, and SJS01), and four ABC Islands bears (ABC01, ABC05, Adm2, and Bar) at different levels of coverage along the longest polar bear genomic scaffold. At low coverage, marker density is too low for tools like HAPMIX to be useful or accurate. 46

2.6 Percent polar bear for each brown bear in this study, calculated via , AD-LIBS, and HAPMIX, if available. The asterisk indicates that Swe was used as a model unadmixed brown bear in $'\hat{f}$ calculations, making polar bear ancestry undetectable. HAPMIX was only run on the four ABC Islands brown bears with minimum 20x coverage, to ensure that heterozygous variant calls were reliable. The"AD-LIBS conservative" column shows AD-LIBS estimates corrected according to the percent of homozygous and heterozygous polar bear ancestry calls that were incorrect in simulations under the single-pulse model. 47

2.7 Percent ancestry of each type called by AD-LIBS for all bears below 20x coverage, for which HAPMIX was not run. Heterozygous calls are probably overestimates. 48

2.8 Overlap of merged regions of polar bear ancestry among 11 brown bears, and regions free of polar bear ancestry in 18 brown bears, with protein-coding genes and exons. The distance p value is the Kolmogorov-Smirnov relative distance p-value and the projection p value is from the projection test, which measures overlap, both implemented in the Genometricorr R package [46]. 50

2.9 Enriched biological_process GO terms of whole genes intersecting high-frequency polar bear haplotypes in brown bears. Testing was done using the Wilcoxon rank-order test in FUNC [164], with genes sorted by mean frequency of polar bear haplotype (in all brown bears), with FUNC's refinement routine to account for hierarchical relationships among terms. Terms with $p < 0.01$ after refinement are shown. 53

2.10 Enriched biological_process GO terms of protein-coding gene exons intersecting high-frequency polar bear haplotypes in brown bears. Testing was done using the Wilcoxon rank-order test in FUNC [164], with genes sorted by mean frequency of polar bear haplotype (in all brown bears), with FUNC's refinement routine to account for hierarchical relationships among terms. Terms with $p < 0.01$ after refinement are shown. 54

2.11 Enriched biological_process GO terms of protein-coding genes intersecting regions where none of 18 hybrid brown bears have any polar bear ancestry. Testing was done using the hypergeometric test in FUNC [164], with FUNC's refinement routine to account for hierarchical relationships among terms. Terms with $p < 0.01$ after refinement are shown. 55

2.12 Enriched biological_process GO terms of protein-coding gene exons intersecting regions where none of 18 hybrid brown bears have any polar bear ancestry. Testing was done using the hypergeometric test in FUNC [164], with FUNC's refinement routine to account for hierarchical relationships among terms. Terms with $p < 0.01$ after refinement are shown. 56

4.1 Times to most recent common ancestor (TMRCA) of various given groups, averaged across all sites in the ARG. Values given in years assume 6.5 million years human-chimpanzee divergence (and thus 13 million years for mutations to accumulate). First TMRCA value is given as a percent of human-chimp divergence. Corrected values use approximate branch shortening values from [163] (minimum and maximum values given are based on minimum and maximum values in the paper). The TMRCA of all humans has only one value because there is no need to correct for branch shortening. Since more sites were included in this analysis than in previous studies focused on genome-wide statistics (i.e. no mappability filter was applied), archaic branch lengths might be slightly inflated from false singletons inferred from DNA damage. 115

4.2 Demographic parameters of Neanderthal and Denisovan admixture from ARG inference. Genome-wide percents given are the percent of the genome assayed (excluding sex chromosomes) classified as Neanderthal or Denisovan origin (or one of the two; ambiguous column), using a score cutoff chosen to minimize the cross-population variance in ILS (Appendix B, Fig. 4.1). Frequencies are means of frequencies of individual admixed segments. Frequency numbers reported are calculated using only confidently admixed haplotypes ($p < 0.001$) and are the frequencies across all human haplotypes in the Simons Genome Diversity Project Panel. Africa2 is a population of the most basal African lineages (Mbuti, Biaka, and Khomani-San) which were used as an outgroup in which no more than 10% of haplotypes were allowed to be admixed for any individuals in the data set were allowed to be called admixed. 119

4.3 Times to most recent common ancestor (TMRCAs) with admixers for confidently introgressed ($p < 0.001$) Neanderthal and Denisovan-like segments, as well as dates estimated using haplotype block lengths and assuming neutral evolution, with a 1 cM/Mb recombination rate and 25 year generation time. Since the TMRCA to admixer and haplotype length were both used to determine the scores of admixed segments and separate ILS from admixture, these values are affected by the score cutoff chosen ($p < 0.001$) and should be treated cautiously. Comparisons between populations should still be valid. 128

4.4 Overlap of confidently-called ($p < 0.001$) Neanderthal and Denisovan haplotypes with various genomic features. Genes are whole protein coding genes from Gencode [49], using Ensembl version 94 on human genome version GRCh38 lifted over to GRCh37 coordinates. Exons are for protein-coding genes from the same annotation. Regulatory element binding sites are from the filtered double-elite" set in the GeneHancer database [48], obtained from the UCSC Genome Browsers Table Browser utility [21]. Distance-based p-values are from the "relative distance" Kolmogorov-Smirnov test and projection p-values measure overlap, both implemented in the GenometricCorr R package [46]. Significant ($p < 0.01$ or $p > 0.99$) values are marked with asterisks. 131

4.5 Significantly enriched biological_process Gene Ontology (GO) terms, via a Wilcoxon rank-order test on genes overlapping high-confidence ($p < 0.001$) Neanderthal-introgressed haplotypes in modern humans, ranked by the frequency (in all humans) of the introgressed haplotype. 137

4.6 Significantly enriched biological_process Gene Ontology (GO) terms, via a Wilcoxon rank-order test on genes overlapping high-confidence ($p < 0.001$) Denisovan-introgressed haplotypes in modern humans, ranked by the frequency (in all humans) of the introgressed haplotype. 138

4.7 Significant GWAS hits that coincide with Neanderthal-introgressed variants in high-confidence ($p < 0.001$) haplotypes, sorted by decreasing Neanderthal haplotype frequency. Only high-frequency ($> 1\%$) variants are shown; there were no such variants in Denisovan-introgressed haplotypes. 144

4.9 Enriched biological_process Gene Ontology [1] terms in desert regions. Enrichment testing was done using the hypergeometric function in FUNC [164], with refinement and a cutoff of p = 0.01. 158

4.10 Enriched biological_process Gene Ontology [1] terms in human-specific regions. Enrichment testing was done using the hypergeometric function in FUNC [164], with refinement and a cutoff of p = 0.01. 164

4.8 Overlap of archaic hominin ancestry deserts and human-specific regions with various genomic features. Genes are whole protein coding genes from Gencode [49], using Ensembl version 94 on human genome version GRCh38 lifted over to GRCh37 coordinates. Exons are for protein-coding genes from the same annotation. Regulatory elements are from the filtered "double-elite" set in the GeneHancer database [48], obtained from the UCSC Genome Browsers Table Browser utility [21]. Distance correlation p-value comes from the relative distance Kolmogorov-Smirnov test and projection p-value, which measures overlap, were both computed using the GenometricCorr R package [46]. All significant ($p < 0.01$ or $p > 0.99$) values are marked with an asterisk. 165

4.11 Genes within the top 50 human-specific regions, by length. Coordinates given are for the boundaries of the human-specific region (hg19 coordinates). TMRCAs are means across sites within the human-specific regions and reported as a fraction of human/chimp divergence (approx. 13 my). 171

Abstract

New Tools for Localizing Ancestry Across Genomes of Hybrid Individuals Provide Insight Into the Genetics of Speciation

by

Nathan K. Schaefer

Understanding the processes by which new species arise has long been of interest to evolutionary geneticists. In some cases, abrupt changes in habitat or niche can spur on adaptive changes in one population that help lead to its genetic isolation from others. In cases where diverging species ranges overlap, speciation genes, or fast-evolving genes which can negatively affect an organism's fertility or viability when divergent alleles from separating populations are inherited together, are thought to spur on species divergence. Hybridization can be thought of as a natural experiment in which semi-incompatible, divergent genomes are brought together in a living organism. When this happens, natural selection should remove incompatible sets of alleles from the genome and increase the frequency of beneficial introgressed alleles (adaptive introgression). By sequencing the genomes of natural populations with hybrid ancestry and identifying the ancestral origin of each part of the genome, it is possible to identify genes involved in species divergence, as well as cases of adaptive introgression.

I review current techniques for studying hybridization using genomic data, as well as what has been learned about ancient DNA and human history using such techniques. I then present a new method for ancestry mapping using low-quality, low-coverage sequence data and demonstrate its application on a population of hybrid brown/polar bears from North America. I

also present a new heuristic ancestral recombination graph (ARG) inference algorithm, which can be used for fine-grained ancestry mapping, as well as a wide range of other population genomic applications. Finally, I use ARG inference to shed light on past human hybridization with Neanderthals and Denisovans, and to identify regions of the genome that define human genetic uniqueness.

Acknowledgments

I would like to thank my committee, Richard E. Green, Grant Pogson, and Beth Shapiro, for all help and guidance during my graduate education, as well as Ed and Beth for giving me opportunities to work on exciting projects.

I have also learned a lot from graduate students and postdocs in the lab, some of whom preceded me there and all of whom gave me valuable advice while I found my footing. These include James Cahill, Sam Vohr, Pete Heintzman, Dan Chang, Andre Soares, and Kelly Harkins.

I am fortunate to have had support from several funding organizations, without which the research described here would not have been possible. These sources of funding include the Gordon and Betty Moore Foundation, the National Institutes of Health, the National Science Foundation, and the Jack Baskin and Peggy Downes fellowship from UCSC.

I would also like to thank all others, academic and otherwise, who provided logistical, work-related, or emotional support, or some combination of all three, throughout my time at UCSC.

The text of this dissertation includes reprints of the following previously published material: Schaefer, NK, Shapiro, B, and Green, RE, Detecting Hybridization Using Ancient DNA, *Molecular Ecology* 25(11):2398-2412, 2016 and Schaefer, NK, Shapiro, B, and Green, RE, AD-LIBS: Inferring Ancestry Across Hybrid Genomes Using Low-Coverage Sequence Data, *BMC Bioinformatics* 18(1):203, 2017. The co-authors listed in this publication directed and supervised the research which forms the basis for the dissertation.

Chapter 1

Introduction

This section was published in the June 2016 journal *Molecular Ecology* under the title Detecting Hybridization using Ancient DNA, with coauthors Beth Shapiro and Richard E. Green. In it, I review existing methods for inferring admixture using genomic DNA, as well as their applications in the field of ancient DNA research.

For more than two decades after the first DNA sequences were isolated from ancient remains [70, 140], the field of ancient DNA was limited to cloning or PCR-based interrogation of one or a few genetic loci. Such data can be useful for studying some aspects of past demography such as population migrations and bottlenecks [68, 219]. For detecting subtle signals of admixture, however, genome-wide data sets are necessary. These data are becoming routinely available from ancient remains via high-throughput sequencing [125] of DNA. Beginning with the retrieval of 13 Mb of the mammoth genome [155] and portions of the Neanderthal genome [57, 135], a variety of approaches have been developed to extract DNA and make it available for direct sequencing, ushering in the new era of paleogenomics [194].

The field of ancient DNA has realized enormous benefits from the gains in efficiency of high-throughput sequencing (HTS). First, HTS libraries and the machines used to read them typically can accommodate a limited size fragment of DNA (up to several hundred nucleotides for currently-popular platforms; [81]). Because DNA molecules retrieved from ancient remains tend to be much smaller, this library and machine limitation is inconsequential. Second, to amplify library molecules during sequencing e.g. during bridge amplification or emulsion PCR a common set of adapters must be ligated onto each molecule. These adapters provide a convenient means to amplify the entire library before sequencing, effectively turning the library itself into a semi-renewable resource (limited by the diversity of DNA fragments present in the sample) Figure 1.1. This is an important consideration for libraries derived from rare and precious ancient samples. Third, library construction and sequencing is set up so that the natural ends of each molecule are read from the sequencer. This has enabled observation of the patterns of DNA base damage in ancient DNA molecules at their ends [55, 14], whereas efforts to characterize damage in molecules amplified by primers specific to sequence within them [142, 14, 16] were unable to do so. Finally, the sheer scale of data collection depending on the machine, up to billions of reads allows a means to retrieve genome-scale data sets from DNA extracts that are often mostly microbial DNA.

Driven by the accumulation of genome-scale data from ancient remains, a spate of methods for detecting admixture has been recently described. An overview of these methods and their requirements, strengths, and weaknesses is given in Figure 1.2; they will be described in detail in the following sections. Paleogenomic data and these methods have revealed many surprises in the evolutionary history of numerous species. Perhaps chief amongst these is that

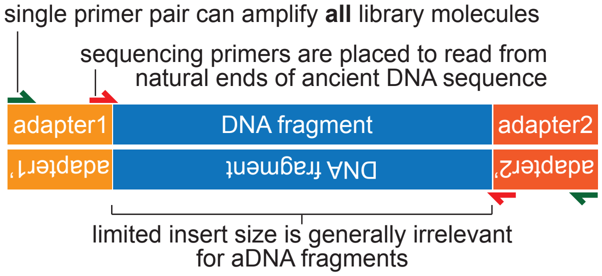
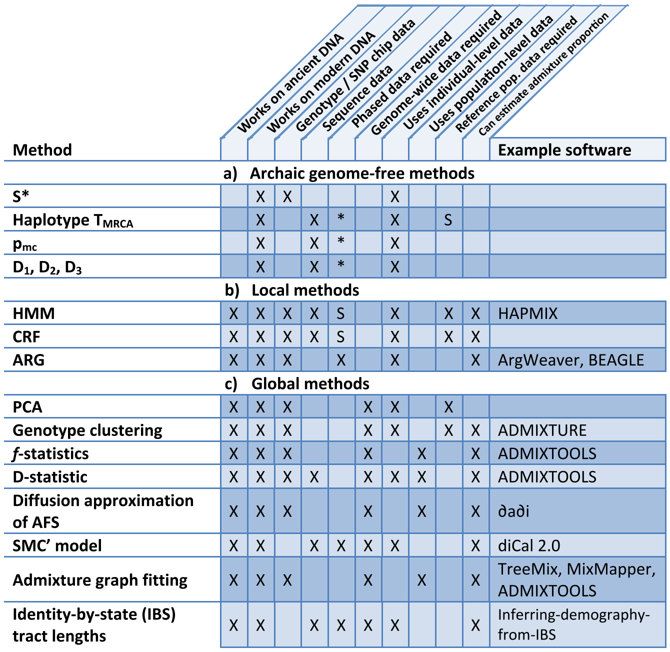

single primer pair can amplify **all** library molecules
sequencing primers are placed to read from natural ends of ancient DNA sequence
adapter1
DNA fragment
adapter2
adapter1'
DNA fragment
adapter2'
limited insert size is generally irrelevant for aDNA fragments

Figure 1.1: Library molecules for high-throughput sequencing (HTS) consist of target DNA fragments with adapter sequences ligated on either end. Adapters, with known sequence complementary to primer sequences, allow a single primer pair to amplify a diversity of DNA fragments, and another to be used for the sequencing reaction, where labeled nucleotides are incorporated [125]. For ancient DNA studies, HTS technology has allowed researchers to observe damage patterns at ends of molecules and amplify a large variety of genomic DNA fragments of unknown sequence. HTS size limitations are inconsequential, as ancient DNA is usually highly fragmented.

hybridization is extensive within the evolutionary history of many vertebrate species, including our own.

1.0.1 Detecting admixture without archaic genomes

Before the first paleogenomes had been assembled, approaches to detecting ancient admixture focused on analyzing data from present-day genomes, and in particular human genomes. Part of the reason for this is that single-locus data from ancient hominins, namely Neanderthals, were available for years before the first paleogenomic data that enabled definitive tests for admixture between Neanderthals and humans. By 2006, mitochondrial genomes were available from several Neanderthals, and the genetic divergence between Neanderthal and modern human mitochondrial genomes led to the prevailing view that humans and Neanderthals had not admixed [193, 58]. Others argued, however, that the data were not incompatible with admixture,



Method	Works on ancient DNA	Works on modern DNA	Genotype / SNP chip data	Sequence data	Phased data required	Genome-wide data required	Uses individual-level data	Uses population-level data	Reference pop. data required	Can estimate admixture proportion	Example software
a) Archaic genome-free methods											
S*		X	X				X				
Haplotype T_{MRCA}		X		X	*		X		S		
p_{mc}		X		X	*		X				
D_1, D_2, D_3		X		X	*		X				
b) Local methods											
HMM	X	X	X	X	S		X		X	X	HAPMIX
CRF	X	X	X	X	S		X		X	X	
ARG	X	X	X		X		X			X	ArgWeaver, BEAGLE
c) Global methods											
PCA	X	X	X			X	X		X		
Genotype clustering	X	X	X			X	X		X	X	ADMIXTURE
f-statistics	X	X	X			X		X		X	ADMIXTOOLS
D-statistic	X	X	X	X		X	X	X		X	ADMIXTOOLS
Diffusion approximation of AFS	X	X	X			X		X		X	∂a∂i
SMC’ model	X	X		X	X	X	X			X	diCal 2.0
Admixture graph fitting	X	X	X			X		X		X	TreeMix, MixMapper, ADMIXTOOLS
Identity-by-state (IBS) tract lengths	X	X		X	X	X	X			X	Inferring-demography-from-IBS

Figure 1.2: Overview of popular techniques for studying archaic admixture. a: Archaic genome free methods are test statistics that can be used to infer archaic introgression into modern individuals without archaic sequence data. Each is computed on real data, then data simulated under various demographic models, and compared. These are prone to errors in model specification and can produce false positives. b: Local methods can be used to find specific genes or genomic regions admixed individuals derive from one or another ancestral population. These are tuned to detect detect long introgressed haplotypes but have reduced power to detect old admixture events. c: Global methods consider individual sites across the genome. Many are formal tests for admixture and/or can be used to estimate admixture proportion. In each box, X” means true and S” means true in some cases. *” indicates methods applied to haplotype sequences, to which the concept of phasing does not apply. Note that, if sufficiently high-coverage genome-wide sequence data are available, these can be transformed into SNP calls if necessary. Also note that a method working on population-level data requires reference population data by default, as all inputs are population-level.

for example if gene flow were unidirectional and came only from males, or if enough time had elapsed for genetic drift to remove Neanderthal mitochondrial variants from modern humans [136, 57]. In the absence of a Neanderthal genome sequence, some sought to inform this debate by analyzing patterns within genomes of present-day humans.

Single-locus studies sought to find archaic alleles in present-day humans via a phylogenetic approach. Given sequence data from various human populations, researchers identified haplotypes showing unusually high divergence from other haplotypes, meaning that their time to most recent common ancestor (TMRCA) is much older than the genome-wide average. Data about geographic distribution of alleles and even archaic sequence data, when available, are incorporated to strengthen findings. This type of approach was used to detect a handful of potentially introgressed haplotypes without ancient sequence data: one specific to present-day Asians, at an X-linked pseudogene called *RRM2P4* [54], which was later found in the Neanderthal genome [62], as well as other two other haplotypes at clinically significant loci [63, 44], which were not found in the Neanderthal genome and are thus may have been false positives [122]. More recent single-locus studies have incorporated sequence data from ancient hominins and used similar techniques to discover archaic haplotypes of genes involved in the immune response [4, 123, 122].

Plagnol and Wall [154] tested for Neanderthal-human admixture using linkage patterns in modern human genomes. They reasoned that if humans had recently (e.g. 40,000 years ago) admixed with an archaic lineage, any introgressed variants should be tightly linked and occur in long (e.g. 40kb) blocks, since recombination would have had insufficient time to further erode the lengths of the archaic haplotypes. They defined a statistic called S*, which seeks to identify sets of SNPs that span long distances and show strong pairwise correlation between genotypes but are not necessarily adjacent, and computed S* over a data set of European and West African individuals. Assessing significance by comparison with simulated data, the authors concluded that European and West African genomes probably both carried genomic segments from separate ancient admixture events [154]. A follow-up study suggested that the admixture events involved multiple archaic hominin species, and inferred a low level of introgression into East Asians [217].

Other investigators have used a variety of techniques to infer archaic admixture from modern sequencing data alone. As in the Plagnol and Wall study, such efforts rely on summary statistics sensitive to admixture. These statistics are used to compare observed data to data simulated under a variety of demographic models, some of which include admixture. S* expanded upon earlier statistics by Wall designed to quantify numbers of tightly correlated genotypes and test demographic models [216]. Another group developed a summary statistic called pmc, which identifies basal gene tree clades containing a large proportion of non-African haplotypes, and used it to support the case for the archaic origin of the Asian-specific RRM2P4 haplotype [27]. Another study that used S* to infer archaic introgression also devised three summary statistics D1, D2, and D3 designed to measure time of admixture, split time between admixing lineages, and extent of admixture, after placing all individuals under study into two groups based on sequence similarity [62]. S* has also recently been used to infer archaic admixture in modern African lineages, using whole-genome data [95].

Methods to detect admixture without archaic genomes suffer from several shortcomings that can be avoided by the presence of sequence data from ancient individuals. Many techniques rely, for example, on assumptions about the demographic history of the species under investigation. Demographic model misspecification can thus bias results, as can misspecification of model parameters like mutation and recombination rates. This has led to several cases in which gene haplotypes inferred to have introgressed into modern humans from Neanderthals were not found in the Neanderthal genome [122]. For this reason, ancient sequence data have proven useful.

1.0.2 Detecting admixture with archaic genomes

The availability of sequence data directly from ancient genomes has led many to use as well as develop techniques for inferring admixture from genomic data. Although described here for their utility in ancient DNA studies, these statistical approaches are general-purpose and are used to study admixture in modern organisms as well. They can enable, for example, the inference of ancestry for specific segments of an admixed individuals genome (local methods), and genome-wide tests for admixture (global methods) that summarize the degree of ancestry components in an admixed individual. Local methods have reduced power to detect old admixture events compared to global methods [147], since they seek to identify long stretches of common ancestry, which recombination will degrade over time. Nonetheless, both categories of methods have developed considerably over the last several years, and both have provided novel insights into species evolutionary trajectories.

1.0.3 Local methods

Local methods for ancestry detection are of use to researchers interested in uncovering specific genes or genomic regions that an admixed individual derives from one or another ancestral population. Although they were generally developed without ancient DNA in mind, they have proven useful in recent attempts to investigate specific archaic variants that have been lost or fixed in modern individuals after archaic admixture. They have also been used to reduce noise in data by uncovering variants that individuals have received via gene flow from populations that are not of interest to investigators.

Local methods model an admixed individuals genome as a series of haplotype blocks, each of which originated in a specific ancestral population. As this requires considering blocks of linked polymorphisms rather than individual SNPs, hidden Markov models (HMMs) are popular local ancestry tools. HMMs are computational models in which sequences of observations are treated as emissions from a set of predefined "states;" in this case, observations are drawn from genotype or sequence data and states correspond to different sources of ancestry. The Viterbi algorithm can then be used to determine the most likely path through states given a sequence of observations [169, 39] and thus assign ancestry to regions of the genome. Early attempts at this strategy were used for admixture mapping in disease studies [45, 74, 73, 146, 231]. Another generation of HMM-based local ancestry methods built upon the same concept but sought to improve parameter estimation by using a more complex model, improving efficiency, or calculating different statistics to use as input observations [208, 206, 160, 10, 15]. A popular example, HAPMIX, uses unphased genotype data from admixed individuals to simultaneously determine phase and infer ancestry. Since errors in phasing techniques can cause local ancestry tools to mistake regions of heterozygous ancestry for transitions between ancestral haplotypes, HAPMIX incorporates phasing into the process of inferring ancestry. This is done by representing phase as well as ancestry in the HMM state space and determining the most likely ancestry of each genomic position over all possible phase configurations [160]. In addition to locating introgressed regions, techniques like HAPMIX have been used to find and mask" regions of European ancestry in Native Americans to improve inference of older population movements

[180, 174].

Conditional random fields (CRFs) are another, similar tool for local ancestry inference. CRFs can be thought of as generalized hidden Markov models. Where HMMs require each observation in a sequence to be a single data point, CRFs allow each observation to have an arbitrary number of features; this allows a CRF to train on and classify multiple types of data simultaneously [96]. This approach is useful when authors are uncertain which summary statistics will be most useful for inferring ancestry. However, unlike HMMs, CRFs require training data [169, 96], which usually comes from simulations with known ancestry. A CRF was used in a recent effort to map Neanderthal ancestry in modern human populations [186]. The features used for ancestry inference had to do with allele sharing patterns, sequence similarity to Neanderthals, and linkage disequilibrium [186].

Given current computational resources and available reference data, ancestral recombination graph (ARG) inference may soon become a feasible approach for local ancestry detection [196]. The ARG is a representation of all coalescence and recombination events, which join and split lineages going back in time, across all individuals and variable sites in a data set; it is thus a complete description of the relationships between individuals in a population panel, across their genomes [196]. ARG inference is computationally challenging, but at least two heuristic implementations currently exist. ArgWeaver [176] constructs the ARG one individual at a time, and uses Markov chain Monte Carlo (MCMC) sampling to draw from the distribution of all possible ARGs when a new individual is added. Song & Heins Beagle [200], not to be confused with popular haplotype-phasing software of the same name, conceptualizes the ARG as a sequence of trees describing non-recombined haplotype blocks separated by recombination events. Beagle, which was not designed for genome-scale data sets, computes the most parsimonious path between trees along the genome via dynamic programming. An accurate ARG could be used, for example, to determine where in the genome individuals and populations fall in clades with archaic hominins. Current implementations require high-quality, phased genotypes [200, 176].

1.0.4 Global methods

Global methods for ancestry detection consider individual sites throughout the genome. In this section, we will first describe the most commonly used global methods used to detect ancient admixture in paleogenomic data sets. We will then highlight some of the key discoveries facilitated by these methods. We focus on admixture between humans and archaic hominins, as this is the field in which the majority of the work using these statistics has been performed.

Several global methods arose from other areas of research before large numbers of complete genome sequences were available, and all have limitations. Principal Components Analysis (PCA), in which vectors of genotype data at many loci are projected onto the axes that capture the most variation within them, has a long history and is famous for recapitulating the geographic distribution of humans [124, 137]. Despite the visually interpretable results, however, PCA is not a formal test [147] and an individuals intermediacy between two groups in principal component space does not prove admixture [227]. EIGENSTRAT [159], which relies on PCA to infer ancestry of individuals, thus may wrongly infer admixture in some problematic cases. Structure [161] and ADMIXTURE [6] are common model-based clustering methods for inferring population structure. These methods attempt to learn local genotype frequencies for a user-defined number of groups across the genome. Then, individuals are described as being mixtures of one or more of these groups. ADMIXTURE provides an estimate of the extent of admixture between groups. Neither of these tests explicitly for significance.

1.0.4.1 *f*-statistics

With the advent of paleogenomics came the need for a new set of statistics that could describe tree topologies relating individuals and populations, formally test for admixture, and estimate the percent ancestry that admixed individuals and populations derive from ancestral groups. The f-statistics, which are included in the software package ADMIXTOOLS [181, 147], are popular for this purpose. The f-statistics work on population-level data, and each describes or tests a phylogenetic relationship by measuring genetic drift conceptualized as variance in allele frequencies along tree branches that is shared between populations. To avoid bias, f-statistics must be computed on sites ascertained in an outgroup to the populations being compared [147].

The f_3 statistic is a simple test for whether a population C is a product of admixture between populations A and B. At a single site, $f_3(C;A,B) = (c-a)(c-b)$, where a, b, and c are allele frequencies in populations A, B, and C. When calculated genome-wide, f_3 is usually positive because of genetic drift in the C lineage that is not shared with A or B (Figure 1.3 a,b). When C is the product of admixture between A and B, however, f_3 can be negative (Figure 1.3 c-f). Negative f_3 is strong evidence for admixture, although a positive f_3 does not necessarily disprove admixture [181, 147]. $f_3(C;A,B)$ can also be used to approximate the relatedness of populations A and B when C is a known outgroup to both (Figure 1.3 a); this is called an outgroup f_3 statistic [174].

The f_4 statistic is used to estimate the correct phylogenetic relationship between four populations. At a single site, $f_4(A,B;C,D) = (a-b)(c-d)$, where a, b, c, and d are allele frequencies in populations A, B, C, and D. Positive, negative, and zero genome-wide values support different tree topologies (Figure 1.4 a-c). A technique called f_4 ratio estimation can also be used to estimate the percent ancestry an admixed population derives from an ancestral population [146]. If data exist from admixing populations B and C, admixed population X, population A (which is more closely related to B than C), and outgroup population D, f_4 ratio estimation can approximate the percent ancestry α that X derives from B. The estimate for α is given by $f_4(a,D;X,C)/f_4(A,D;B,C)$ (Figure 1.4 d,e) [147].

Haak et al [61] used the f_4 statistic in a more exploratory way, to identify populations that may have contributed DNA to an admixed population of interest, and to estimate the amount of ancestry contributed by each of the admixing populations. The authors defined a set of candidate admixing populations Ref_1, Ref_2, ... Ref_N that may have contributed ancestry to the population of interest Test in unknown proportions α_1, α_2, ... α_N. They then chose three outgroup populations A, B, and C, none of which share recent gene flow with Test or $Ref_1 ... Ref_N$. They observed that $f_4(Test,A;B,C) = \sum_{i=1}^{N} \alpha_i f_4(Ref_i,A;B,C)$. After calculating f_4 for each candidate reference population and every possible permutation of available outgroups, the authors were able to calculate the α_i admixture coefficients for each candidate admixing population via linear regression [61].

1.0.5 D-statistic

Another popular genome-wide test for admixture is the D-statistic [56, 35]. D can be computed using either individual genomes or population allele frequency data [35]. In the case of individual genomes, D requires sequence from two potentially admixed individuals, P_1 and P_2, a candidate admixing individual, P_3, and an outgroup P_4. D always falls between -1 and 1; it is positive if P_1 shares more derived alleles with P_3 than P_2 shares with P_3. D is negative if P_2 shares more derived alleles with P_3 than P_1 shares with P_3. The idea behind D is that, if there has been gene flow from the population of which P_3 is a member, then any admixed individual (P_1 or P_2) will share more derived alleles with P_3 than an unadmixed individual. To calculate D, one scans the genome for sites where P_2 shares a derived allele with a P_3, termed ABBA sites. To compensate for incomplete lineage sorting (ILS), this is subtracted from this the number of sites at which P_1 shares a derived allele with P_3, termed BABA sites. Then $D = (N_{ABBA} - N_{BABA})/(N_{ABBA} + N_{BABA})$, where N_{ABBA} is the total number of ABBA sites and N_{BABA} is the number of BABA sites (Figure 1.5) [56]. Random processes like ILS and recurrent mutation can produce ABBA and BABA sites, but should produce an equal number of both. Admixture, if it occurs, will only increase ABBA or BABA counts in the admixed individual. D is robust to fluctuating ancestral population sizes but can be confounded by ancestral population structure [35]. One recent study, seeking to minimize the noise resulting from ancestral population structure, restricted D to sites where individuals from a population believed to be free of admixture matched the outgroup P_4 and thus carried the ancestral allele. This technique is called an "enhanced D-statistic" and can improve power to detect admixture, but it can also introduce bias. If analysis is restricted to sites where individuals from unadmixed population P_0 match the outgroup P_4, and populations P_1 and P_2 are equally related to P_3 but not equally related to P_0, $D_{enhanced}(P_1, P_2, P_3, P_4)$ can deviate from zero, although the expectation of $D(P_1, P_2, P_3, P_4)$ is zero [126].

D can be used in other ways as well. Like the f statistics, D can be calculated on population genotype data by replacing N_{ABBA} and N_{BABA} with products of allele frequencies in the four populations [35]. Another statistic $\hat{f}$ [56, 35] uses D to estimate admixture proportion: if P_{3a} and P_{3b} are two individuals from population P_3, then $\hat{f} = D(P_1, P_2, P_3, P_4)/D(P - 1, P_{3a}, P_{3b}, P_4)$, and it can be understood as a ratio of D calculated on the admixed individual to D calculated on an individual from the admixing population. D can also be calculated without a candidate admixing individual P_3, if a different outgroup P_0 to P_1 and P_2 is available: $E[D(P_2, P_1, P_0, P_4)] \propto E[D(P_1, P_2, P_3, P_4)]$, with the value changing slightly due to this statistics dependence on the split time of P_0 and the P_1/P_2 lineage, rather than the time of admixture [35]. Finally, Eaton and Ree introduced a variation on the D statistic, which they call the partitioned D statistic [38] and used it to analyze RADseq data collected from a genus of flowering plants within the broomrape family. This method is designed to remove the effect of shared ancestry amongst multiple candidate admixing populations by quantifying the number of derived alleles that are common in both and found in the admixed population.

1.0.5.1 Weighted block jackknife

A weighted block jackknife approach [93] can be used to assess significance of f and D statistics. To overcome bias introduced by linkage disequilibrium (LD), the block jackknife technique divides the genome into M blocks, each of which must be long enough to overcome LD between adjacent blocks. Appropriate block size can be determined by performing the block jackknife repeatedly with increasing block sizes until standard error estimates converge [181, 56]. Each block is then removed from the genome in turn, and the test statistic is computed over the rest of the genome. In the case of D, a single jackknife computation is D_i for $i = 1, 2, ...M$, the mean $D_\mu = (1/M)\sum_{i=1}^{M} D_i$, and the weight of jackknife block i is $W_i = (N_i)/(\sum_{j=1}^{M} N_j)$ where N_i is the number of informative sites in the block and $\sum_{j=1}^{M} N_j$ is the number of informative sites in the genome. The weighted variance of D in an individual is then given by $\sum_{i=1}^{M} W_i(D_i - D_\mu)^2$ and standard error is $SE_D = \sqrt{M\sum_{i=1}^{M} W_i(D_i - D_\mu)^2}$ [56]. Since the expectation of D is zero, Z scores can then be computed from D scores as $Z = D/SE_D$.

1.0.5.2 Other approaches

Other approaches to detecting archaic admixture use information about specific demographic and evolutionary parameters, such as split times between populations, population structure, and natural selection. The program $\delta a \delta i$ [60] considers the derived allele frequency in multiple populations at sites throughout the genome, termed the multi-population allele frequency spectrum (AFS). The expected AFS under a model that can include selection and migration is computed by solving a diffusion equation that approximates AFS evolution over time. Model parameters including extent of migration are then adjusted via (composite) maximum likelihood estimation to fit the observed AFS [60]. diCal 2.0 builds on the theory of the sequentially Markov conditional sampling distribution [149], using a hidden Markov model that trains on observed haplotypes and has states corresponding to discretized time points in the past. This HMM can be used to estimate parameters for demographic models that include population structure and migration [202]. TreeMix [153], MixMapper [105], and qpGraph from ADMIXTOOLS [147] all build on the concept of fitting graphs rather than trees to genotype data, allowing for migration between nodes.

Another set of methods seek to infer demographic parameters like admixture extent from linkage disequilibrium patterns [157, 147, 64]. In a popular implementation of this approach, pairs of phased haplotypes are drawn from populations of interest, and the distribution of lengths of identity state (IBS) tracts, or runs of identical sequence flanked by variable sites, is computed [64]. This distribution is then compared to one expected under a demographic model and used to optimize model parameters, which can include population growth rates, divergence times, and rates of admixture [64].

1.0.6 Detecting admixture with archaic hominins

One of the most visible contributions of paleogenomic studies to current understanding of admixture is the detection of gene flow between archaic hominins and modern humans. The first direct genetic evidence of admixture between Neanderthals and anatomically modern humans was from the 2010 publication of a draft Neanderthal genome sequence [56], which expanded upon an earlier analysis of 1 megabase of the Neanderthal genome that hinted at possible Neanderthal-human admixture [57]. Using the D-statistic and sequences from modern humans, Green et al. inferred Neanderthal gene flow into all non-Africans, and estimated the Neanderthal proportion of non-Africans ancestry to be 1-4% [56]. A subsequent study using a higher-quality Neanderthal genome revised this to 1.5-2.1% and concluded that the Neanderthal that admixed with modern Eurasians was more closely related to a Neanderthal from the Caucasus than to Neanderthals from the Altai Mountains and Croatia, suggesting a possible location for admixture [165].

Although the D-statistic can be confounded by ancestral population structure [35], and some studies have suggested that such structure did exist in early humans [54], other lines of evidence support Neanderthal-human admixture. First, patterns of linkage disequilibrium (LD) in present-day humans suggest admixture occurred 47-65 kya, more recently than would be expected if Neanderthal-like haplotypes were the result of ancestral population structure [188]. Second, a comparison of the site frequency spectrum of real data with that simulated under models of ancestral population structure and recent admixture also supported the recent admixture scenario [226]. The most convincing evidence came, however, from a more recent analysis of a previously unknown archaic hominin called the Denisovan. Denisovan DNA was extracted from a 30-50,000 year old finger bone found in Denisova cave in southern Siberia and was found to belong to a previously undiscovered hominin lineage [92, 179]. Phylogenies inferred from Denisovan mitochondrial and nuclear DNA are discordant: mitochondrial DNA suggests a deep, 1 mya divergence between the Denisovan lineage and a clade containing both human and Neanderthal lineages [92], while nuclear loci place the Denisovan closer to Neanderthals (650 kya diverged) than to modern humans (800 kya diverged) [179]. This discordance suggests either incomplete lineage sorting in a small population descended from a much larger one or admixture with an as-yet unknown archaic hominin with a more ancient divergence from humans and Neanderthals [179]. A subsequent study that included demographic simulations supported the admixture hypothesis, while also detecting a small amount of gene flow from

Neanderthals into the Denisovan [165].

Like Neanderthals, the Denisovan appears to have contributed to the modern human gene pool. Using the D-statistic, about 3-6% of the genomes of present-day Australian aborigines and Melanesians are of Denisovan-like origin [179, 126], as opposed to 0.2% of East Asian and Native American genomes and little to none of the genomes of other groups [165]. A possible explanation for this pattern is admixture with the ancestors of Australians and Melanesians followed by migration of admixed Oceanians to East Asia [165]. Another study suggests that New Guineans were the source for Denisovan ancestry detected in all other groups, including Australian aborigines [166].

This discovery that Denisovans admixed with modern humans has had two consequences. First, it bolsters the case for Neanderthal-human admixture. If the signal of Neanderthal-human admixture resulted from structure in the ancestral African population, then the Denisovan should exhibit excess allele sharing with all non-Africans and not just Australians and Melanesians, because of the phylogenetic proximity of the Denisovan to Neanderthals [126]. Second, it creates a geographic mystery. Although the range of the Denisovan population is not known, it is unclear how a Siberian population could have admixed with the ancestors of Australians and Melanesians. This mystery is compounded by the recent discovery of a 400,000 year old hominin bone from Sima de los Huesos in Spain, which has Neanderthal-like morphological features and mitochondrial DNA that is very similar to the Denisovan [127]. Given that the Denisovan mitochondrial haplotype may have originated within another, unknown hominin lineage [179, 165], this creates a connection between hominin lineages in western Europe, southern Siberia, and Oceania that is yet to be fully understood [127].

Ancient remains of modern humans have also helped inform the study of Neanderthal-human admixture. In 2014, the genome of a 45,000 year old human male from the UstIshim site in Siberia was sequenced [51]. Computational analysis, which included D-statistics to detect gene flow and f_4 ratio estimation to quantify the level of gene flow, determined that the individual came from a population ancestral to both modern Europeans and Asians, and had tracts of Neanderthal ancestry that were longer than those found in modern humans [51]. The length distribution of Neanderthal haplotypes was used to estimate that the UstIshim individuals Neanderthal ancestor lived between 50 and 60 kya [51]. In addition to UstIshim, two other ancient human genomes were found to have longer tracts of Neanderthal ancestry than modern humans: a 36-39,000 year old individual from western Russia [192], and a 37-42,000 year old human from Petera cu Oase in Romania [50]. In an analysis similar to the UstIshim study [51], the latter was found to have a substantially larger Neanderthal component than present-day humans, with longer un-recombined Neanderthal haplotype blocks [50]. Fu et al. concluded that the Petera cu Oase individual was only 4-6 generations removed from a Neanderthal ancestor and may have had one or more other Neanderthal ancestors. This finding weakens the case for a single human-Neanderthal admixture event and suggests that at least one admixture event may have taken place in Europe.

The idea of multiple admixture events has been upheld by computational studies. Contrary to initial reports, recent studies have detected more Neanderthal ancestry in East Asians compared to Europeans [218, 186, 214]. One proposed explanation for this is that Neanderthal alleles are generally deleterious and thus were able to drift to higher frequency in the historically smaller East Asian population than in the historically larger European population, where purifying selection would have been more powerful [186]. Another explanation is a two-pulse" model of admixture, in which the ancestors of East Asians admix with Neanderthals a second time, after the population split from western Eurasians [214]. Simulations under different demographic models have upheld either the latter scenario or a more complex scenario involving admixture with other groups, as more likely than the former [88, 215]. These studies are leading to a new view of hominin history in which barriers between divergent taxa are porous and rapid adaptation to new environments may have been facilitated in part by gene flow [141].

Many studies have moved beyond population genetics and sought to identify selective consequences of Neanderthal and Denisovan alleles present in modern humans. In some cases, there appears to have been adaptive introgression, as with several non-African human leukocyte antigen (HLA) haplotypes that may have originated in Neanderthals and Denisovans, where they probably arose under selective pressure from local pathogens long before modern humans migrated to the same areas [4]. In other cases, deleterious alleles introgressed from an archaic hominin and then went to high frequency in modern human populations, as with a set of disease-related variants discovered by a whole-genome scan [186] and a Neanderthal-origin haplotype across the gene SLC16A11 that poses high diabetes risk [222]. The diabetes risk allele could, however, have originally conferred a selective advantage to ancient humans upon entering a new habitat and adopting a new diet [171]. Other studies, reviewed in Racimo et al [171], have discovered cases in which selection has apparently spread archaic alleles of genes involved in immune defense, altitude adaptation, skin and hair phenotypes, and lipid metabolism. In addition to uncovering many cases of adaptive introgression, two recent studies that mapped out Neanderthal ancestry in present-day humans found depletion of Neanderthal sequence in and around coding regions, suggesting that natural selection may have acted to eliminate many Neanderthal variants [186, 214].

1.0.7 Inferring modern human migrations

Beyond Neanderthals and Denisovans, ancient DNA and statistics for detecting admixture can be used to infer the movement of genes, and therefore people, between locations. In addition to D and f-statistics, approaches to infer patterns of migration and admixture include but are not limited to admixture graph fitting, demographic model fitting to the sequentially Markovian conditional sampling distribution (diCal 2.0), and characterization of identity by state (IBS) tract length distributions. Together, these statistical approaches have reframed the existing view about the timing and nature of human movements across the globe.

In reconstructing the history of the peopling of Europe, for example, two early observations from paleogenomes demanded a context. First, the genome of tzi, a 5,300 year old man from the Italian alps, was found to resemble the genomes of present-day Sardinians [84]. Second, the genome of a 24,000 year old boy from Malta in south-central Siberia was found to share ancestry with both present-day European and Native American genomes [173]. A larger study followed up on these findings, adding many present-day genomes as well as several from ancient European farmers and hunter gatherers [98]. This study inferred that modern Europeans descend from three genetic sources: western European hunter-gatherers, early farmers from the Middle East, and a mystery population related to ancient Siberians and Native Americans [98]. This study also showed that tzis affinity to modern-day Sardinians was a trait shared with other Neolithic farmers [98]. Two more recent studies, one with 69 [61] and another with 101 ancient genomes [7], provided greater detail about past human migrations. In particular, these studies suggested that the mystery population identified earlier was probably a mixture of Eastern European hunter-gatherers, which were related to the ancient Siberian samples and Native Americans and to herders from the Eurasian steppe. This population was estimated to have invaded Europe during the Late Neolithic, after which they contributed genes to all populations, were a source for wheeled cart technology and Indo-European languages, and led to the rise of the Corded Ware culture throughout Copper Age Europe [7, 61]. This same group, known as Yamnaya, also spread east to create the Andronovo culture in the Altai region in Siberia, which later changed as it received migrants from East Asia in the Iron Age [7].

Admixture-based analyses of ancient human genomes have also shed light on the ongoing debate about the peopling of the Americas, in particular about whether Native Americans are descendants of a single group that migrated across the Bering Strait in the Late Pleistocene or a more complex mixture of groups. To date, Native American paleogenomes have shown strong continuity with present-day Native Americans, challenging hypotheses about ancient admixture that were based on analyses of skeletal morphology [177, 178]. One recent study divided Native Americans into three lineages:“First Americans,” Eskimo-Aleut speakers, and Na Dene speakers, and concluded that each of these could have represented a separate migration from Asia, with subsequent admixture and some possible back-migration from First Americans to Asia [180]. In contrast, a subsequent larger study concluded that“First Americans” and Na Dene speakers more likely diverged within the Americas, while the Inuit may represent a separate migration [174].

Several studies have also attempted to address the possible gene flow from Oceanians into Native American populations, which was first detected by the observation of low levels of Denisovan DNA in the New World [165, 166]. One study found a weak signal of differential Oceanian ancestry in New World populations, and concluded that a small amount of Oceanian ancestry made its way to different parts of the Americas via admixture first with East Asians and later with Aleutian Islanders [174]. Another detects Oceanian admixture in several Amazonian groups and argues for a larger Melanesian presence among New World populations [199].

One feature that distinguishes several of these recent ancient DNA investigations of human migration and demography from past ones is an increase in both the number of samples and the variety of analysis techniques used. In contrast to previous studies in which one or several paleogenomes were analyzed, e.g. [179, 56, 165], several recent studies have used dozens of samples [172, 174, 7, 61, 199].

Owing to the lack of well-preserved hominin remains, some global regions, like Africa, have thus far been difficult to study using ancient DNA [194]. Using patterns of Neanderthal ancestry, however, researchers have detected possible back-migrations from Eurasia to eastern Africa [4, 165]. More recently, ancient human remains with high endogenous DNA content were discovered in Ethiopia and yielded the first ancient African genome, called Mota [53]. Furthermore, several groups have sought to expand upon the original discovery of possible archaic introgression into African groups based on S* [154, 217]. For example, one study of noncoding autosomal loci inferred archaic gene flow into a variety of central and southern African populations within the last 70,000 years, to the exclusion of a West African agriculturalist population [62]. Another group calculated S* across genomes of African hunter-gatherer populations and concluded that there had been multiple instances of archaic introgression, first into the common ancestors of this group and later as regional admixture events [95]. Follow-up studies will be needed to assess whether this signal might be the result of ancestral population structure rather than admixture.

1.0.8 Detecting ancient admixture in other species

Although hominins remain the most popular lineage for ancient admixture studies, advances have also been made in understanding the history of gene flow in other species. Within mammals, a recent study investigating the relationship between modern cattle and aurochs, their extinct wild ancestor, used the D-statistic to detect a low level of gene flow from aurochs into British and Irish cattle breeds in the period since domestication [144]. While lacking nuclear sequence data, another study using ancient DNA from mammoths analyzed mitochondrial genomes from the morphologically divergent Columbian mammoth and wooly mammoth species. The authors found that the Columbian mammoths mtDNA fell within the diversity of that of the wooly mammoth, and thus that the two species may have hybridized at some point in time; this suggests a follow-up study involving nuclear data [43].

Admixture studies using ancient DNA have been applied to plants and fungi as well. A group interested in maize, specifically its arrival in and adaptation to the US Southwest about 4,000 years ago, sought to settle a debate about its route of diffusion using ancient DNA. By sequencing 32 ancient maize samples spanning much of the history of maize domestication and geographic spread, the authors found, using the D-statistic, TreeMix, and a genotype clustering method, that maize in the US Southwest likely spread from highland Mexico, with subsequent gene flow from coastal varieties [29]. Another study sought to clarify interspecific relationships and centers of origin within the fungal genus Phytophthora, which includes the pathogen responsible for late blight, the cause of the Irish potato famine. They found, using the same methods, that P. andina, a species native to the Andes, appears to have arisen through hybridization between a species closely related to that which caused the potato famine, and an as-yet unknown outgroup to the other species examined [117].

It is worth noting that high-coverage ancient genomes from non-hominin species are just now becoming available (*e.g.* [114, 143]). Just as studies of archaic hominin admixture have been enabled by the growing diversity of genomic data from humans and their close relatives, future progress in other taxa should enable detection and characterization of ancient admixture events in lineages further removed from our own. These studies will no doubt provide important insights into the effects of hybridization and gene flow on speciation and environmental adaptation [3].

1.0.9 Conclusion

Recent advances in extraction, sequencing, and analysis of ancient DNA have led the field away from studies of single loci and into the field of paleogenomics, where more ambitious studies and detection of admixture and inter-population migration are now possible. Such studies have both co-opted existing techniques and mandated the development of new tools for detecting and quantifying admixture. With these, they have shed light on past admixture events, in both the recent and distant past, that have changed our understanding of who we are as a species. As reference data become more available, and ancient DNA studies become more ambitious in sequencing a larger portion of genomes of an expanding number of ancient taxa, innovative new computational analysis techniques will follow. The result will be a wider perspective on the complex web of interactions between species past and present that defines Earths recent biological history.

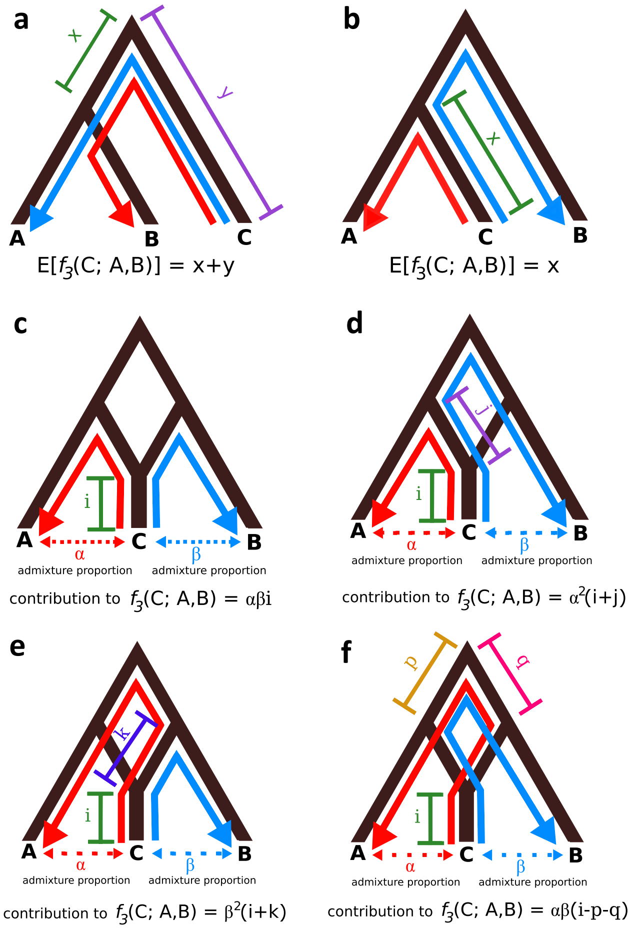


Figure 1.3: Adapted from [147]. Expected value of $f_3(C;A,B)$ under various tree topologies. Red lines trace genetic drift between populations C and A; blue lines trace genetic drift between C and B. f_3 measures drift between C and A that is also shared between C and B. Drift is shared along branches where arrows going in the same direction overlap. a and b: expected value of $f_3(C;A,B)$ with no admixture. If C is not a product of admixture between A and B, f_3 is expected to be positive. In the case where C is an outgroup to A and B (a), the value of f_3 is proportional to the distance separating C from A and B, which can also be thought of as the amount of shared history between A and B. c-f: expected value of when C is a product of admixture between A and B. α is the percent ancestry population C derives from A, and β is the percent derived from B. Distance j represents genetic drift between extant population A its ancestral population that admixed to form the population ancestral to C in the past; distance k is proportional to drift between extant population B and its admixing ancestral population. Computation of $f_3(C;A,B)$ in this case requires tracing multiple paths through the tree, since population C can share drift with population B that it received through admixture with population A and vice versa. The expectation is the sum of all shared drift: $E[f_3(C;A,B)] = \alpha\beta i + \alpha^2(i+j) + \beta^2(i+k) + \alpha\beta(i-p-q)$. This has the potential to be negative, although it can also be positive. Given that negative values are impossible if C is not a result of admixture (a and b), a negative result can be taken as evidence of admixture; a positive result, however, cannot be used to reject admixture.

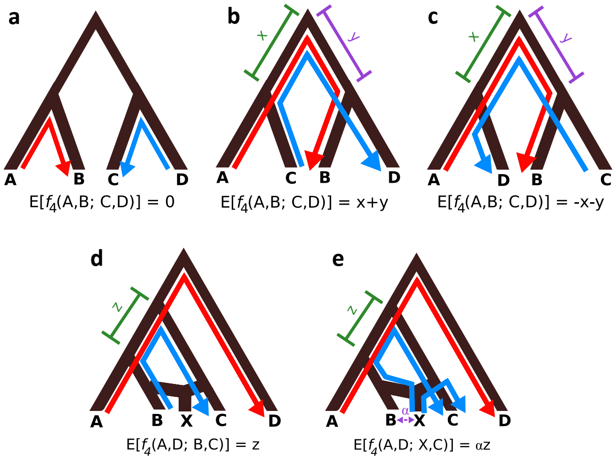


Figure 1.4: Adapted from [181]. Visual explanation of expected values of $f_4(A,B;C,D)$ under various tree topologies. Red lines trace genetic drift from A to B; blue lines trace drift from C to D. f_4 measures drift shared between A and B that is also shared between C and D. Drift is shared along branches where arrows overlap going in the same direction. a-c: Positive, negative, and zero values of f_4 give support for different tree topologies relating the four populations. d, e: visual explanation of f_4 ratio method for inferring admixture proportion. Population X is a mixture of populations related to B and C; population D is an outgroup. The quantity of interest, α, is the proportion of ancestry population X has received from B. If the expected value of $f_4(A,D;B,C) = z$ (d), then the expected value of $f_4(A,D;X,C) = \alpha z$(e). It follows that $\alpha = f_4(A,D;B,C)/f_4(A,D;X,C)$ (Patterson et al. 2012).

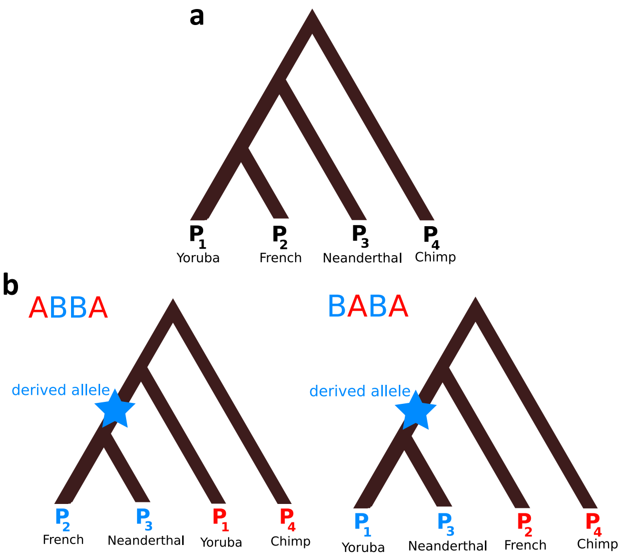


Figure 1.5: Explanation of D statistic [56, 35]. Individuals are numbered according to the D-statistic notation: $D(P_1, P_2, P_3, P_4)$ and examples of individuals that could be used to yield a positive D-statistic result when testing for Neanderthal ancestry are given (D would be negative in this case if there had been gene flow between the Yoruba and Neanderthal instead). a: genome-wide tree relating the four individuals, based on prior knowledge. b: trees at ABBA and BABA sites used to compute D. In both, blue is used to represent a derived allele (does not match chimpanzee); red represents an ancestral allele (matches chimpanzee). To calculate D on sequence data, the number of sites with the topology of the left tree is N_{ABBA} and the number of sites with the topology of the right tree is N_{BABA}. Then $D = (N_{ABBA} - N_{BABA})/(N_{ABBA} + N_{BABA})$.

Chapter 2

Inferring ancestry across low-coverage, downsampled genomes

In this chapter, I describe a computational tool for locally mapping ancestry across genomes of hybrid animals, using low-coverage and/or low-quality data, and I demonstrate its use on a population of brown bears with known polar bear ancestry, as well as down-sampled, pseudo-haploid modern human and Neanderthal genomes. This tool is well-suited to studies of non-model organisms, in which panels of phased reference data are unavailable, and ancient DNA studies, in which high-coverage sequencing data are unavailable. This was originally published in April 2017 in *BMC Bioinformatics* under the title AD-LIBS: Inferring ancestry across hybrid genomes using low-coverage sequence data, with coauthors Beth Shapiro and Richard E. Green. I have added a small amount of text to what was originally published (the "Functional consequences" section in Results, along with a paragraph about this section in the Discussion, and a paragraph describing methods used for it in Methods).

2.0.1 Background

Inferring the ancestry of different parts of admixed diploid individuals genomes has been a goal of fields as diverse as disease gene mapping [45] and paleogenomics [186, 214]. Several computational approaches have been developed for ancestry detection. Among these, global methods calculate genome-wide amounts of admixture but do not attempt to localize admixture segments in the genome. In contrast, local methods for ancestry detection scan across admixed individuals genomes to search for haplotype blocks originating from specific ancestral populations [189]. Because haplotype blocks are broken down by recombination over time, local methods sacrifice power to detect very old admixture events in exchange for the ability to make specific, local statements about ancestry [189].

Many of the techniques for local ancestry inference were developed to investigate human ancestry and therefore incorporate assumptions that may not be valid for analyses of non-human data. For example, methods that compute on genotype calls rely on accurate calling. Genotype calling from sequence data as implemented in applications such as GATK [121] rely on pre-existing knowledge of polymorphic sites to make high-quality variant calls. This information is often unavailable for non-model organisms. In addition, some fields, such as paleogenomics [194], are limited by the amount of data that can possibly be recovered. Specifically, the degraded nature of recovered ancient DNA, and the upper limit imposed by the endogenous DNA content of source material [194], often results in coverage well below the threshold of 20X that has is considered necessary for reliable genotype calling [134]. Additionally, population genomic analyses may benefit more from sequencing many individuals to low coverage than sequencing fewer individuals more deeply [52], meaning that data collected for other types of analyses may not be suitable for ancestry inference techniques that rely on genotype calling. As an example, a recent study used ancient DNA from aurochs, the extinct wild ancestor of domestic cattle, to produce a 6X coverage genome and infer gene flow into British and Irish cattle breeds post-domestication [144]. These data would be unsuited to current local ancestry inference techniques.

To address these challenges, we present AD-LIBS (Ancestry Detection through Length of Identity By State tracts), which is a software application that performs local ancestry inference by analyzing genetic data across genomic windows rather than at individual sites. AD-LIBS is designed for low-coverage shotgun resequencing data, and bypasses the need for variant calling and phasing. Input data for AD-LIBS is a single haploid sequence for each individual where every base is a random sample from one or the other chromosome, as has been done in other studies to mitigate genotyping errors [19, 20]. AD-LIBS uses a hidden Markov model to infer the most likely ancestral origin of each piece of the genome.

We test AD-LIBS using simulated data and find that it correctly infers 87-91% of ancestry, with a true positive rate of 89% for identifying admixed, homozygous regions and 82-85% for identifying admixed heterozygous regions.

We then use AD-LIBS to assign ancestry in two real data sets: one comprising five European humans with known Neanderthal ancestry and five West African individuals without Neanderthal ancestry, and another consisting of 18 brown bears from North America and Scandinavia with varying amounts of polar bear ancestry. In humans, we find that AD-LIBS produces maps of Neanderthal ancestry in Europeans that overlap significantly with published maps [186, 214] and global Neanderthal ancestry estimates that fall within 0-2% of what is expected from prior studies [56, 165]. In the bear data set, AD-LIBS identifies polar bear ancestry in all brown bear populations, including those believed previously to be unadmixed, and recovers a geographic signal in patterns of polar bear ancestry. We also test AD-LIBS on down-sampled, artificially low-coverage data from bears and find that it produces consistent results down to about 2X genome-wide coverage, outperforming HAPMIX [160], an existing local ancestry inference tool, at coverage levels below this. In summary, AD-LIBS is an effective tool for producing local ancestry maps for genomes of hybrid individuals when only low-coverage sequence data are available and/or reference data are scarce.

2.0.2 Results

2.0.2.1 Overview of AD-LIBS

AD-LIBS (Ancestry Detection through Length of Identity-By-State tracts) uses a hidden Markov model to determine the ancestry of specific regions of hybrid individuals genomes inferred from low-coverage shotgun sequence data. To circumvent problems inherent in genotyping and phasing individuals sequenced to low coverage, AD-LIBS uses non-overlapping windows to scan pseudo-haploid sequence data, allowing all nucleotide positions in a given window to“vote” on the correct ancestry of that window. AD-LIBS does not require phased sequences from reference individuals nor does it require prior knowledge of polymorphic sites. AD-LIBS does require prior estimates of the population size, the number of generations since admixture, and the proportion of ancestry that the admixed population derives from each ancestral population. Although population size is best taken from census data, a rough estimate may be obtained from nucleotide diversity in the ancestral populations [23]. Admixture proportion may be estimated using the $\hat{f}$ statistic [56, 35] if an outgroup genome is available, and time of admixture can be roughly inferred from the admixture proportion estimate together with prior knowledge about the ancestral species historical ranges and demography. AD-LIBS includes programs to calculate both average nucleotide diversity and $\hat{f}$, and in practice, incorrect estimates for these parameters do not have a large effect on results (Figure 2.1).

AD-LIBS scans across each hybrid individuals genome in windows of a fixed width. In each window, AD-LIBS calculates a score based on average identity-by-state (IBS) tract lengths between the admixed individual and each individual from each ancestral population. AD-LIBS considers three possible types of ancestry in each window: homozygous for ancestry from one of the two ancestral populations, or heterozygous. Thus, AD-LIBS works as an ancestry genotyper for genomic segments and determines the most likely sequence of ancestry states across each chromosome or scaffold, given expected score distributions under each type of ancestry, computed from nucleotide diversity values. The probability of transitioning between ancestry states is related to the probability of recombination having occurred at specific genomic loci in the time since admixture, as well as the overall prevalence of alleles from each ancestral population in the admixed population.

AD-LIBS is designed to be efficient: its genome scanning and scoring components are written in C and its hidden Markov model component uses a Cython package (`https://github.com/jmschrei/pomegranate`). AD-LIBS requires that the system running it possesses enough memory to hold the longest chromosome or genomic scaffold sequence for each reference ancestral individual and a single hybrid in memory at once; for humans, this would

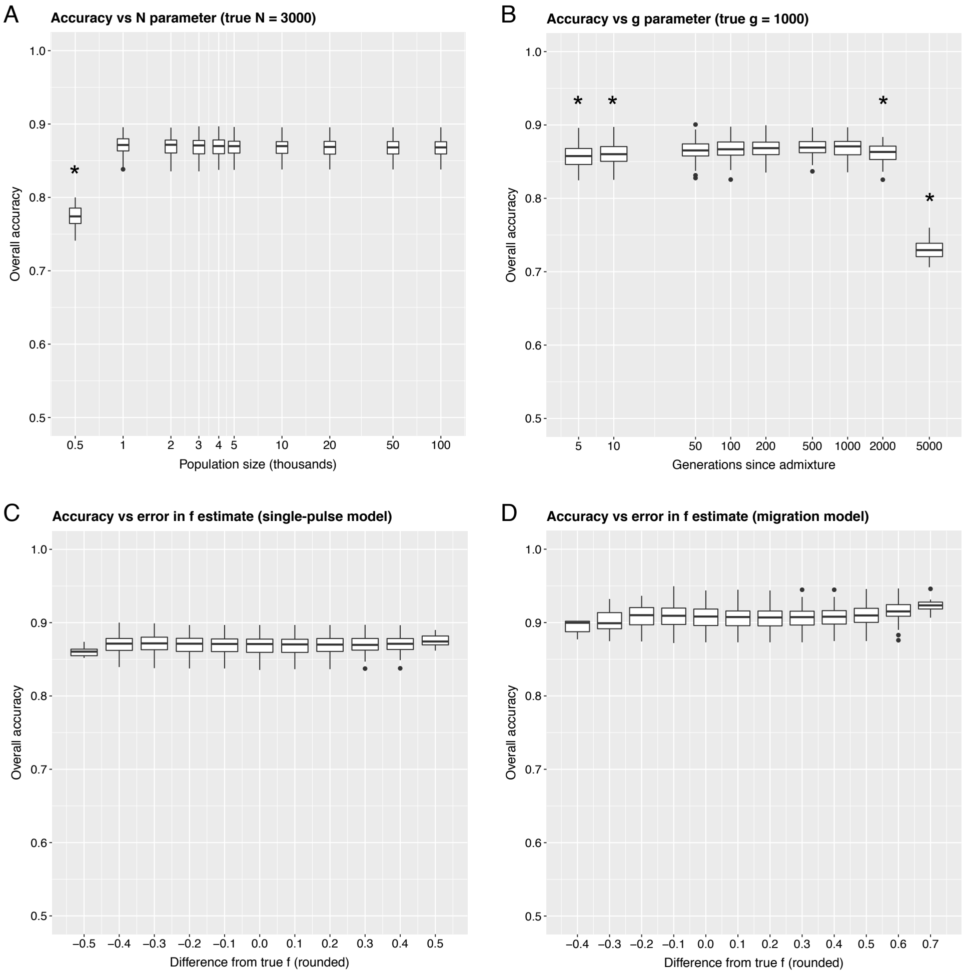


Figure 2.1: AD-LIBS accuracy on simulated data, using incorrect population parameters. Simulations here used the single-pulse" admixture model (see Figure 2.2 A, C, and E) except where otherwise noted, with 10kb windows, which were automatically adjusted by AD-LIBS as necessary. Asterisks denote accuracy significantly lower ($p < 0.001$) than that obtained using correct parameters. A: AD-LIBS accuracy using different prior population size estimates (true N = 3000) and correct number of generations since admixture (1000). B: AD-LIBS accuracy using different estimates for the number of generations since admixture (true g = 1000) and correct population size (3000). C: AD-LIBS accuracy using prior estimates of polar bear admixture proportion that differed from the true value, rounded to the nearest 10%. D: Same as C, but using the migration" admixture model (see Figure 2.2 B, D, and F), which produced a wider range of true admixture proportions.

comprise approximately 250 MB plus 250 MB RAM for each reference ancestral individual. On an Intel Xeon 2.7 GHz processor with 377 GB RAM, we ran AD-LIBS on a single 2.3 Gb hybrid brown bear genome, using ten ancestral reference genomes with numeric parameters pre-computed, in under 7.5 minutes. The same operation took approximately 9 minutes on a comparable machine with 70 GB RAM. When scanning multiple hybrid genomes, AD-LIBS can use multiple processes simultaneously to reduce execution time.

2.0.2.2 Simulations

To assess the accuracy of AD-LIBS, we generated 100 simulated hybrid genomes, each consisting of ten, one-megabase (1 Mb) chromosomes. We assumed a demographic history resembling that of the ABC Islands brown bears, a well-studied population of brown bears known to have polar bear ancestry [19, 20, 108, 129, 40]. We used two demographic models, one with a single polar-brown bear admixture event 12,000 years ago (single-pulse model), and another incorporating continuous brown bear dispersal to the ABC Islands from the initial admixture event until the present (migration model) (see Methods). We compared AD-LIBS ancestry calls to the known ancestry of each simulated chromosome. AD-LIBS performed well, with overall accuracy of 87% for the single-pulse model and 91% for the continuous migration model, and accurately recovered polar bear ancestry (82-85% true positive rate for heterozygous and 89% true positive rate for homozygous polar bear ancestry) (Table 2.1, Figure 2.2). While choice of window size and number of reference individuals from each ancestral population had a small effect on overall accuracy, simulations show that suboptimal choices for both e.g. one reference individual per ancestral population, or large windows of 25kb reduce overall accuracy only by several percent (Figure 2.2). Additionally, we found that inaccurate prior estimates of admixed population size, number of generations since admixture, and polar bear ancestry proportion for individual hybrid bears had a similarly small effect on overall accuracy (Figure 2.1).

Model	Ancestry state	Prop. calls correct	Prop. truth detected
Migration	AA	0.8858	0.931
Migration	AB	0.8447	0.9633
Migration	BB	0.9762	0.8459
Migration	Average	0.9022	0.9134
Migration	Overall accuracy	0.9069	
Single-pulse	AA	0.8933	0.9245
Single-pulse	AB	0.817	0.9513
Single-pulse	BB	0.9501	0.7134
Single-pulse	Average	0.8868	0.8631
Single-pulse	Overall accuracy	0.8694	

Table 2.1: The accuracy of AD-LIBS ancestry inferences using simulated genomes. Two demographic models representative of the ABC Islands bears history were used: one in which a single admixture event between polar and brown bears takes place 12,000 years ago, followed by isolation of the hybrid population (Single-pulse model), and one in which admixture takes place at the same time but is followed by continuous brown bear migration from the mainland (Migration model). Overall accuracy is the percent of all bases for which true ancestry matched AD-LIBS-inferred ancestry. Since this number is weighted toward more common ancestry states, the average across all three ancestry states is also given.

AD-LIBS was about as good at estimating each individuals extent of polar bear ancestry as $\hat{f}$, a widely-used statistic that estimates admixture proportion by comparing genome-wide frequencies of sites supporting tree topologies compatible and incompatible with admixture [56, 35]. AD-LIBS tends to overestimate the amount of heterozygous ancestry by several percent, however (Figure 2.3). This might explain why AD-LIBS was more accurate in identifying ancestry under the migration model than the single-pulse model (Table 2.1, Figure 2.2 A, B). Genomes simulated under the migration model tend to have a lower overall extent of polar

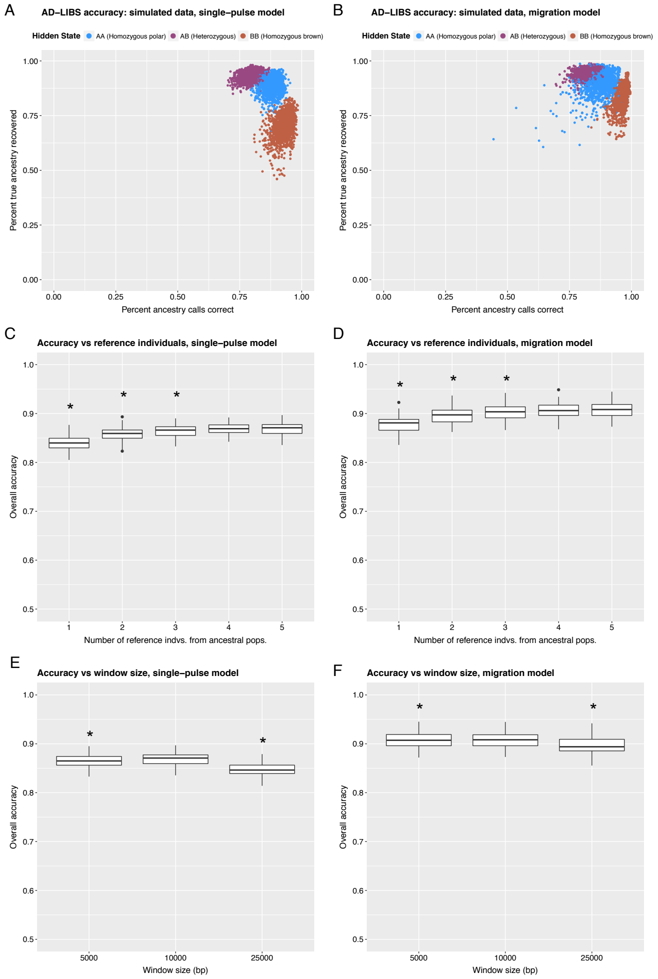


Figure 2.2: AD-LIBS accuracy using simulated data. A, C, and E refer to simulations with a single polar-brown bear admixture event 12,000 years ago, followed by isolation (single-pulse model); B, D, and F refer to simulations in which a brown-polar bear admixture event 12,000 years ago is followed by continual breeding with unadmixed brown bears (migration model). A and B: percent of AD-LIBS inferences correct and percent of true ancestry recovered in each ancestry state (homozygous polar bear, heterozygous, and homozygous brown bear) for each individual. C and D: Effect of using different numbers of reference ancestral individuals (1, 2, 3, 4, or 5 from each population) on overall accuracy, using 10kb windows. Asterisks denote a distribution mean significantly lower ($p < 0.001$) than the best distribution (5 individuals from both populations), according to t-test. E and F: Effect of using different window sizes (5kb, 10kb, and 25kb), with 5 reference bears from each ancestral population. Asterisks denote a distribution mean significantly lower ($p < 0.001$) than the best distribution (10kb windows), according to t-test.

bear ancestry (Figure 2.3 A, B), giving AD-LIBS less opportunity to overestimate heterozygous ancestry. This causes the overall accuracy of AD-LIBS to fall by a rate of approximately 0.2 percent per percent polar bear ancestry (Figure 2.3 E; slope of best fit line by least squares regression = -0.206; adjusted r^2 = 0.687; F-statisic p-value $< 2.2e-16$, via linear model function in R [168]), although this effect may level off as higher levels of polar bear ancestry will lead to greater amounts of homozygous polar bear ancestry, which AD-LIBS detects more accurately.

2.0.2.3 Real data

We collected two data sets for our study. First, we obtained five CEPH European (CEU) human genomes with Neanderthal ancestry (Green et al. 2010b) and five Yoruban (YRI) human genomes with little to no Neanderthal ancestry [56] from the 1000 Genomes Project [36], along with a single high-quality Neanderthal genome [165]. We used these data to map Neanderthal ancestry in Europeans using AD-LIBS and compare the results to previously-published local ancestry maps [186, 214] and global estimates of Neanderthal ancestry in Europeans [56, 165]. We also collected previously published shotgun sequence data from four polar bears, eighteen North American brown bears, and one American black bear [108, 129, 19, 20]. For a full list of bear samples used in this study, see Table 2.2. All reads were aligned to the polar bear reference genome [108] before pseudo-haploidization and variant calling (Methods). The black bear was used as an outgroup to perform $\hat{f}$ [56, 35] calculations for comparison with our findings.

Sample	Species	Location	Sex	Coverage	Study
PB7	U. maritimus	Spitsbergen, Svalbard	F	176.2X	Miller et al. 2012 [129]
PB12	U. maritimus	Qaanaq, Greenland	F	26.0X	Liu et al. 2014 [108]
PB68	U. maritimus	Qaanaq, Greenland	F	26.1X	Liu et al. 2014 [108]
PB105	U. maritimus	Disko West, Greenland	F	26.2X	Liu et al. 2014 [108]
OFS01	U. arctos	stanvik, Sweden	F	22.8X	Liu et al. 2014 [108]
RF01	U. arctos	Ruokolahti, Finland	F	20.9X	Liu et al. 2014 [108]
SJS01	U. arctos	Slakka, Sweden	F	15.2X	Liu et al. 2014 [108]
Swe	U. arctos	Dalarna, Sweden	F	11.0X	Cahill et al. 2015 [20]
Den	U. arctos	Denali Natl. Park, AK	F	12.1X	Cahill et al. 2013 [19]
GP01	U. arctos	Glacier Park, Montana	M	16.8X	Liu et al. 2014 [108]
GRZ	U. arctos	Kenai Peninsula, AK	F	83.6X	Miller et al. 2012 [129]
ABC01	U. arctos	Baranof Island, AK	M	20.0X	Liu et al. 2014 [108]
ABC02	U. arctos	Baranof Island, AK	F	18.4X	Liu et al. 2014 [108]
ABC03	U. arctos	Chichagof Island, AK	M	19.6X	Liu et al. 2014 [108]
ABC04	U. arctos	Chichagof Island, AK	M	18.8X	Liu et al. 2014 [108]
ABC05	U. arctos	Chichagof Island, AK	F	22.4X	Liu et al. 2014 [108]
ABC06	U. arctos	Admiralty Island, AK	F	19.5X	Liu et al. 2014 [108]
Adm1	U. arctos	Admiralty Island, AK	F	12.1X	Cahill et al. 2013 [19]
Adm2	U. arctos	Admirality Island, AK	F	76.5X	Miller et al. 2012 [129]
Bar	U. arctos	Baranof Island, AK	M	49.1X	Miller et al. 2012 [129]
Chi1	U. arctos	Chichagof Island, AK	F	9.2X	Cahill et al. 2015 [20]
Chi2	U. arctos	Chichagof Island, AK	F	10.2X	Cahill et al. 2015 [20]
Uam	U. americanus	Pennsylvania	M	11.6X	Cahill et al. 2013 [19]

Table 2.2: Sample details. All sequence data were published in previous studies and downloaded as reads from the NCBI SRA. Coverage levels shown were estimated from numbers of raw reads before mapping to the reference genome. All samples were aligned to the polar bear reference genome, then subjected to base and map quality filtering, indel realignment, and duplicate removal.

2.0.2.4 Neanderthal ancestry in humans

AD-LIBS produced maps of Neanderthal ancestry in modern Europeans that agreed with published data, including global estimates of Neanderthal ancestry proportion [56, 165] as well as population-specific local ancestry maps [186, 214]. We prepared pseudo-haploid genome sequences from five randomly selected admixed CEPH European (CEU) and five unadmixed Yoruban (YRI) individuals from the 1000 Genomes Project [36], as well as two "haplotype" sequences from the Altai Neanderthal [165] (variants were randomly assigned to one or the other haplotype at heterozygous sites; see Methods). We then ran AD-LIBS to infer Neanderthal ancestry in each European, using Neanderthal and Yoruban sequences as reference ancestral populations. For comparison, we also calculated each European individuals Neanderthal ancestry via $\hat{f}$. Although choice of window size affects results, the estimate of each individuals Neanderthal ancestry proportion from AD-LIBS, using appropriate parameters, is 0.80-1.90% greater than the estimate (Table 2.3). AD-LIBS estimates are also 0.22-1.92% greater than the published estimate of 1.5-2.1% in all modern humans [165]. We note that back-migration of Europeans to West Africa has also contributed some Neanderthal ancestry to Yoruban individuals [4, 165], biasing estimates downward. We also compared the maps of Neanderthal ancestry from AD-LIBS to those published for CEPH Europeans [186, 214]. The AD-LIBS map overlapped significantly (p of greater overlap = 0 in 500 random trials) with the two previously published maps, although each map also finds Neanderthal ancestry in regions of the genome where the other maps do not.

Individual	AD-LIBS, 10kb	AD-LIBS, 15kb	$\hat{f}$
NA11832	11.50%	2.35%	1.41%
NA11840	11.50%	2.32%	1.52%
NA12340	13.80%	3.42%	1.52%
NA12383	13.40%	3.39%	1.56%
NA12814	13.60%	3.29%	1.45%

Table 2.3: Results of running AD-LIBS on autosomal sequences of five randomly chosen European (CEU) individuals from the 1000 Genomes Project, with the Altai Neanderthal and five randomly chosen Yoruba (YRI) individuals from the 1000 Genomes Project as reference individuals from admixing populations. Using AD-LIBS with a window size (10kb) lower than the recommended minimum of 14kb gives bad results, while using a window size above this threshold (15kb) gives much more reasonable results.

Due to the nature of the emission probability distributions that AD-LIBS uses to distinguish between regions with different types of ancestry (Methods), AD-LIBS produces inaccurate results when the window size is too small and/or too much genetic variation within the ancestral populations is also shared between them. This was the case when using 10kb windows to scan for Neanderthal ancestry in Europeans (Table 2.3). When ancestral populations share a large amount of genetic variation between ancestral populations, the distributions that AD-LIBS uses to distinguish between different types of ancestry tend to overlap. Using larger window sizes can reduce the variance of these distributions, and AD-LIBS can suggest an appropriate window size automatically (Methods). Using a window size of 15kb, above the threshold recommended by AD-LIBS, produced more accurate estimates of Neanderthal ancestry in Europeans (Table 2.3). When too much genetic variation within ancestral populations is also shared between them, however, AD-LIBS is unlikely to be accurate no matter what window size is chosen. This is likely to be the case when both admixing populations consist of modern humans, and this problem can be avoided by choosing an alternative to AD-LIBS when genetic differentiation between ancestral populations, as measured by statistics such as F_{ST} [75], is low.

2.0.2.5 Polar bear ancestry in brown bears

We next used AD-LIBS to scan for polar bear ancestry in brown bears. We explored the effect of low sequence coverage depth on AD-LIBS inferences, compared global polar bear ancestry estimates from AD-LIBS to those produced using other techniques, and looked for geographic patterns in the distribution of polar bear ancestry across brown bear genomes.

Determining the necessary level of coverage

We sought to determine the effect of low sequence coverage depth on the accuracy of AD-LIBS by downsampling reads to produce artificial low-coverage genomes. We selected four admixed ABC Islands bears that were sequenced to at least 20X coverage (ABC01, ABC05, Adm2, and Bar), four polar bears over 20X coverage (PB7, PB12, PB68, and PB105) and three Scandinavian brown bears over 10X coverage (OFS01, RF01, and SJS01). The Scandinavian brown bears were hypothesized to be unadmixed with polar bears [108]. We obtained a set of variant calls and a pseudo-haploid genome sequence for each bear (Methods), and downsampled every bear to 0.5X, 1X, 2X, 5X, and 10X coverage along the longest genomic scaffold (scaffold1) to produce a set of variant calls and a pseudo-haploid sequence for this scaffold at these different coverage levels. We ran AD-LIBS on each of the four admixed ABC Islands bears at full coverage and at each downsampled coverage level, using the three Scandinavian brown bears and four polar bears as unadmixed reference sequences. For comparison to AD-LIBS, we then ran HAPMIX [160], a commonly-used tool for local ancestry inference, on the same data. At each depth, we compared inferences from HAPMIX and AD-LIBS to the output of both programs run on the full-coverage data.

Using the high coverage data, we found that AD-LIBS and HAPMIX produce comparable results, although AD-LIBS tends to label regions heterozygous that HAPMIX labels homozygous polar bear (Table 2.4). At the lowest coverage levels, marker density after variant calling was too low for HAPMIX to produce interpretable results (Table 2.5). AD-LIBS more consistently infers the same ancestry at low and high coverage than does HAPMIX. Additionally, at coverage below 2X, ancestry calls made by AD-LIBS are more similar to the full-coverage ancestry calls from HAPMIX than are the low-coverage ancestry calls from HAP-

MIX (Figure 2.4). When grouping results by ancestry state, low-coverage homozygous ancestry calls from AD-LIBS are more likely to match high-coverage calls than those from HAPMIX, although AD-LIBS labels some regions as heterozygous that are homozygous according to HAPMIX (Figure 2.5). We infer from this experiment that AD-LIBS is consistent with itself down to about 2X coverage, and that inferences of homozygous ancestry from AD-LIBS are more reliable than those from HAPMIX at low coverage, although AD-LIBS erroneously labels some regions of homozygous ancestry heterozygous. By avoiding the need for variant calling, AD-LIBS also outperforms HAPMIX in cases of very low (0.5X or 1X) coverage, when there are not enough called variants to detect any polar bear ancestry. We note that genotype imputation could help improve marker density when running HAPMIX on low-coverage data, but this is only possible when studying species for which variant catalogs from large panels of reference individuals are available, such as humans.

Measuring admixture proportion

We next sought to compare estimates of the genome-wide extent of polar bear ancestry in brown bears from AD-LIBS to those produced using other techniques. For each of eighteen brown bears, we ran AD-LIBS using four polar bears and four Scandinavian brown bears, the latter as potentially unadmixed models of ancestral populations (individual Scandinavian brown bears were excluded from the unadmixed reference brown bear set when treated as hybrid bears). For each bear, we also estimated genome-wide polar bear ancestry using the $\hat{f}$ statistic [56, 35], which was used in prior studies of polar bear ancestry in brown bears [19, 20]. For this analysis, we used PB7 and PB12 as model polar bears, Swe as a model brown bear, and the American black bear as the outgroup. We also ran HAPMIX on all brown bears sequenced

Bear	Ancestry state	AD-LIBS	HAPMIX	Agreement
ABC01	Hom. Polar	4.30%	7.30%	43.60%
ABC01	Heterozygous	28.70%	14.20%	46.70%
ABC01	Hom. Brown	67.10%	78.50%	84.20%
ABC01	Total	100%	100%	72.70%
ABC05	Hom. Polar	4.90%	8.39%	47.50%
ABC05	Heterozygous	28.30%	12.80%	44.20%
ABC05	Hom. Brown	66.80%	78.80%	84.00%
ABC05	Total	100%	100%	72.30%
Adm2	Hom. Polar	3.60%	6.52%	43.60%
Adm2	Heterozygous	26.80%	12.20%	42.60%
Adm2	Hom. Brown	69.60%	81.30%	84.70%
Adm2	Total	100%	100%	73.20%
Bar	Hom. Polar	4.56%	7.68%	46.60%
Bar	Heterozygous	27.80%	13.20%	44.60%
Bar	Hom. brown	67.60%	79.10%	84.30%
Bar	Total	100%	100%	72.80%

Table 2.4: Percent ancestry of each type, as called by AD-LIBS and HAPMIX, in the four bears sequenced to sufficient coverage depth for variant calling. AD-LIBS calls more heterozygous ancestry than HAPMIX and probably overestimates heterozygous ancestry genome-wide.

to at least 20X coverage, with all polar bears and the three Scandinavian brown bears above 20X coverage used as reference ancestral populations. For details about choices of parameters, see Methods.

The admixture proportions detected with AD-LIBS were higher than our estimates using $\hat{f}$ (Table 2.6). AD-LIBS-inferred admixture proportions are also higher than estimates from HAPMIX for the four ABC Islands bears of greater than 20x coverage (Table 2.6). We note that $\hat{f}$ is considered a lower bound on admixture proportion, since it can only detect mutations that arose in the hybridizing lineages between the time of speciation and admixture [35]. This was not the case in our simulations, however, in which AD-LIBS and both consistently overestimated the polar bear admixture proportion by several percent (Figure 2.3 A,B).

Coverage level	Num. Raw variants	Num. Filtered	Num. Phased
10	490,705	410,174	410,174
5	443,470	288,761	288.821
2	353,433	44,706	44,706
1	268,595	1,494	1,494
0.5	172,214	14	14

Table 2.5: Numbers of variants obtained from four polar bears (PB7, PB12, PB68, and PB105), three brown bears (OFS01, RF01, and SJS01), and four ABC Islands bears (ABC01, ABC05, Adm2, and Bar) at different levels of coverage along the longest polar bear genomic scaffold. At low coverage, marker density is too low for tools like HAPMIX to be useful or accurate.

Although overestimation of heterozygous ancestry could explain why AD-LIBS produces erroneously high polar bear admixture proportions, an alternative explanation is needed to explain its discrepancy with $\hat{f}$. One possibility is that in real data, purifying selection in the brown bear lineage could reduce nucleotide diversity below the level typical of neutrally-evolving regions of the brown bear genome, but not below the level typical across the entire polar bear genome. This could cause windows of the genome in which brown bear-specific selection has taken place subsequent to the brown-polar bear split to appear erroneously heterozygous. It is also possible that polar bear ancestry in the Scandinavian brown bears, which were assumed to be unadmixed in calculations and in prior studies [19, 20], may also explain why $\hat{f}$ estimates were lower than estimates using both AD-LIBS and HAPMIX.

To be conservative, we recalculated all of our admixture proportions from AD-LIBS, this time multiplying the numbers of bases called homozygous and heterozygous polar bear by the rate at which these types of ancestry calls were correct in our simulations under the “single-pulse” model (Table 2.1) (Table 2.6, “AD-LIBS conservative” column). Since bases mis-called as heterozygous might actually be of either homozgyous polar bear or homozygous

Bear	Origin	f	AD-LIBS	AD-LIBS conservative	HAPMIX
ABC01	Baranof Island, AK	8.63%	18.60%	15.50%	14.40%
ABC02	Baranof Island, AK	8.87%	18.80%	15.70%	14.80%
ABC03	Chichagof Island, AK	9.63%	19.40%	16.20%	N/A
ABC04	Chichagof Island, AK	9.03%	19.00%	15.80%	N/A
ABC05	Chichagof Island, AK	8.93%	19.10%	15.90%	N/A
ABC06	Admiralty Island, AK	6.56%	17.10%	14.20%	N/A
Adm1	Admiralty Island, AK	6.12%	16.60%	13.80%	N/A
Adm2	Admirality Island, AK	6.05%	17.00%	14.20%	12.60%
Bar	Baranof Island, AK	8.14%	18.50%	15.40%	14.30%
Chi1	Chichagof Island, AK	8.57%	18.60%	15.50%	N/A
Chi2	Chichagof Island, AK	8.69%	18.70%	15.60%	N/A
Den	Denali Natl. Park, AK	7.02%	14.50%	11.90%	N/A
GP01	Glacier Park, Montana	4.37%	17.20%	14.30%	N/A
GRZ	Kenai Peninsula, AK	3.30%	13.00%	10.70%	N/A
OFS01	stanvik, Sweden	0.46%	5.35%	4.41%	N/A
RF01	Ruokolahti, Finland	0.32%	6.90%	5.67%	N/A
SJS01	Slakka, Sweden	0.21%	5.27%	4.33%	N/A
Swe	Dalarna, Sweden	0%*	4.89%	4.02%	N/A

Table 2.6: Percent polar bear for each brown bear in this study, calculated via , AD-LIBS, and HAPMIX, if available. The asterisk indicates that Swe was used as a model unadmixed brown bear in $'\hat{f}$ calculations, making polar bear ancestry undetectable. HAPMIX was only run on the four ABC Islands brown bears with minimum 20x coverage, to ensure that heterozygous variant calls were reliable. The“AD-LIBS conservative” column shows AD-LIBS estimates corrected according to the percent of homozygous and heterozygous polar bear ancestry calls that were incorrect in simulations under the single-pulse model.

brown bear ancestry, treating them all as homozygous brown bear this way should produce an under-estimate of polar bear ancestry. The observation that AD-LIBS tends to find less homozygous polar bear ancestry than HAPMIX does (Table 2.4) also suggests that some of the mis-called heterozygous ancestry should be treated as homozygous polar bear ancestry. Regardless, we find using this technique that AD-LIBS still predicts more polar bear ancestry than $\hat{f}$.

Comparing specific ancestry calls (homozygous polar bear, homozygous brown bear, and heterozygous) shows 72-74% overall agreement between AD-LIBS and HAPMIX, with most discrepancy resulting from AD-LIBS overestimating the extent of heterozygous ancestry (Table 2.4 and Table 2.7). It is possible that HAPMIX underestimates homozygous polar bear ancestry as well, and that the problems described earlier with variant calling and phasing may lower the reliability of inferences from HAPMIX.

Bear	Hom. polar	Heterozygous	Hom. brown
ABC02	4.56%	28.50%	66.90%
ABC03	4.82%	29.10%	66.10%
ABC04	4.67%	28.60%	66.80%
ABC06	3.71%	26.70%	69.60%
Adm1	2.92%	27.40%	69.70%
Chi1	4.24%	28.70%	67.00%
Chi2	4.03%	29.30%	66.70%
Den	1.20%	26.60%	72.20%
GP01	3.43%	27.50%	69.10%
GRZ	1.72%	22.50%	75.80%
OFS01	0.44%	9.82%	89.70%
RF01	0.44%	12.90%	86.60%
SJS01	0.33%	9.88%	89.80%
Swe	0.34%	9.10%	90.60%

Table 2.7: Percent ancestry of each type called by AD-LIBS for all bears below 20x coverage, for which HAPMIX was not run. Heterozygous calls are probably overestimates.

Shared patterns of ancestry

We next investigated the extent to which the same regions of the genome had the same type of ancestry in multiple bears. For each possible combination of two or more bears, we computed the number of bases in the genome for which all bears were inferred to have the same type of ancestry. Considering each type of ancestry separately, we then created random ancestry maps for each bear by sampling genomic coordinates comprising randomly-drawn regions of the same number and size from the reference genome. Computing the overlap among these random ancestry maps for all bears in the set gave us a null model against which to compare the extent of overlap among our true ancestry maps. For each group of bears and each type of ancestry, overlap is greater than for random samples (Figure 2.6). This suggests that polar bear introgression took place within the shared demographic history of all of the brown bears in this study, as hypothesized by others [129, 40, 19, 20, 108].

As another way to visualize sharing of polar bear-derived haplotypes among brown bears, we used principal components analysis (PCA), to test whether the ancestry data from AD-LIBS contain a similar geographic signal of admixture to that which has been observed previously from SNP data [108]. Using EIGENSOFT SmartPCA [148], we created vectors of ancestry across 10kb genomic windows and performed PCA on these vectors for all 18 brown bears. Principal components place individuals into groups based on geography, with the first component corresponding to polar bear ancestry proportion. The ancestry results are largely similar to those from SNP data (Figure 2.7). For example, the Montana bear clusters with the Admiralty Island individual(s), to the exclusion of the Baranof and Chichagof Island bears. The SNP PCA distinguishes bears from Finland and Sweden, however, while the ancestry PCA does not, suggesting that polar bear ancestry in these individuals might stem from the same historical event, despite different recent evolutionary histories.

2.0.2.6 Functional consequences

Overall, we find that polar bear ancestry extends over roughly 65% of the queryable part of the genome (scaffolds of length 500kb or greater). We tested the set of all regions of polar bear introgression in brown bears, and the set of regions free of polar bear introgression (the

Map	Features	Dist. P	Proj. P
polar	genes	0.3363769	0.9834
polar	exons	0.471	0.32
nopolar	genes	0.028	1
nopolar	exons	0.322	0.0000573

Table 2.8: Overlap of merged regions of polar bear ancestry among 11 brown bears, and regions free of polar bear ancestry in 18 brown bears, with protein-coding genes and exons. The distance p value is the Kolmogorov-Smirnov relative distance p-value and the projection p value is from the projection test, which measures overlap, both implemented in the Genometricorr R package [46].

complement of this set) for correlation and intersection with protein-coding genes and exons of protein coding genes. We find no significant enrichment or depletion of overlap between, or positional correlation between, polar bear-introgressed regions and protein-coding genes or exons (Table 2.8).

To look for evidence of adaptive introgression, we took all protein coding genes and exons that intersected regions of polar bear ancestry (in any bear), sorted them by the mean frequency of polar bear ancestry across each gene, and performed a Wilcoxon-rank test for enrichment of Gene Ontology terms using FUNC [164]. We find many terms related to immune response, as has been found in many archaic-introgressed haplotypes in modern humans [171]. We also find the terms DNA repair, chromatin organization, and spermatid development (for whole genes) (Table 2.9) and DNA repair (for exons) (Table 2.10). This is surprising, as we would expect such terms to be related to genes often involved in hybrid incompatibility, rather than adaptive introgression. Elucidating what these genes are, and why they appear to be in relatively high-frequency introgressed haplotypes, merits further study.

p	GO ID	GO term
9.34E-05	GO:0000122	negative regulation of transcription by RNA polymerase II
0.00825521	GO:0002828	regulation of type 2 immune response
0.00462537	GO:0005979	regulation of glycogen biosynthetic process
0.00129983	GO:0006281	DNA repair
0.00483142	GO:0006325	chromatin organization
0.00719853	GO:0006691	leukotriene metabolic process
0.00370973	GO:0006909	phagocytosis
0.00327674	GO:0007286	spermatid development
0.00521529	GO:0007565	female pregnancy
0.00666644	GO:0009124	nucleoside monophosphate biosynthetic process
0.00277393	GO:0010922	positive regulation of phosphatase activity
0.00706058	GO:0016188	synaptic vesicle maturation
0.00758499	GO:0030857	negative regulation of epithelial cell differentiation
0.00687972	GO:0032436	positive regulation of proteasomal ubiquitin-dependent protein catabolic process
0.00897533	GO:0032507	maintenance of protein location in cell
0.00999749	GO:0032570	response to progesterone
0.00719853	GO:0032753	positive regulation of interleukin-4 production
0.0066798	GO:0032784	regulation of DNA-templated transcription, elongation

0.00377917	GO:0042752	regulation of circadian rhythm
0.00840516	GO:0043044	ATP-dependent chromatin remodeling
0.00737313	GO:0043470	regulation of carbohydrate catabolic process
0.00785181	GO:0045581	negative regulation of T cell differentiation
0.00719853	GO:0045589	regulation of regulatory T cell differentiation
0.00686625	GO:0045683	negative regulation of epidermis development
0.00706058	GO:0045792	negative regulation of cell size
0.000941951	GO:0045892	negative regulation of transcription, DNA-templated
0.00534869	GO:0045893	positive regulation of transcription, DNA-templated
0.00303331	GO:0060850	regulation of transcription involved in cell fate commitment
0.0075634	GO:0070372	regulation of ERK1 and ERK2 cascade
0.00719853	GO:0098930	axonal transport
0.00824451	GO:2001023	regulation of response to drug

Table 2.9: Enriched biological_process GO terms of whole genes intersecting high-frequency polar bear haplotypes in brown bears. Testing was done using the Wilcoxon rank-order test in FUNC [164], with genes sorted by mean frequency of polar bear haplotype (in all brown bears), with FUNC's refinement routine to account for hierarchical relationships among terms. Terms with $p < 0.01$ after refinement are shown.

p	GO ID	GO term
0.00156607	GO:0000122	negative regulation of transcription by RNA polymerase II
0.00592948	GO:0005979	regulation of glycogen biosynthetic process
0.00178018	GO:0006281	DNA repair
0.00946091	GO:0006590	thyroid hormone generation
0.00245489	GO:0010922	positive regulation of phosphatase activity
0.00664922	GO:0023058	adaptation of signaling pathway
0.00946091	GO:0032329	serine transport
0.004439	GO:0032436	positive regulation of proteasomal ubiquitin-dependent protein catabolic process
0.00442608	GO:0032507	maintenance of protein location in cell

0.00736052	GO:0032784	regulation of DNA-templated transcription, elongation
0.00857897	GO:0045066	regulatory T cell differentiation
0.00702846	GO:0045581	negative regulation of T cell differentiation
0.00429937	GO:0045792	negative regulation of cell size
0.00269204	GO:0045892	negative regulation of transcription, DNA-templated
0.00347511	GO:0050795	regulation of behavior
0.0089133	GO:0051006	positive regulation of lipoprotein lipase activity
0.00923071	GO:0051602	response to electrical stimulus
0.00994172	GO:0060338	regulation of type I interferon-mediated signaling pathway
0.00538082	GO:0060712	spongiotrophoblast layer development
0.00360331	GO:0086091	regulation of heart rate by cardiac conduction
0.00946091	GO:0099622	cardiac muscle cell membrane repolarization

Table 2.10: Enriched biological_process GO terms of protein-coding gene exons intersecting high-frequency polar bear haplotypes in brown bears. Testing was done using the Wilcoxon rank-order test in FUNC [164], with genes sorted by mean frequency of polar bear haplotype (in all brown bears), with FUNC's refinement routine to account for hierarchical relationships among terms. Terms with $p < 0.01$ after refinement are shown.

p	GO ID	term
0.00221323	GO:0000122	negative regulation of transcription by RNA polymerase II
0.0026105	GO:0043009	chordate embryonic development
0.00280615	GO:0016050	vesicle organization
0.00477465	GO:0003229	ventricular cardiac muscle tissue development
0.00490197	GO:0010922	positive regulation of phosphatase activity
0.00591571	GO:0021954	central nervous system neuron development
0.00731015	GO:0045620	negative regulation of lymphocyte differentiation
0.00795445	GO:0048745	smooth muscle tissue development
0.00806575	GO:0032436	positive regulation of proteasomal ubiquitin-dependent protein catabolic process

Table 2.11: Enriched biological_process GO terms of protein-coding genes intersecting regions where none of 18 hybrid brown bears have any polar bear ancestry. Testing was done using the hypergeometric test in FUNC [164], with FUNC's refinement routine to account for hierarchical relationships among terms. Terms with $p < 0.01$ after refinement are shown.

We then checked the possible functional significance of regions free from polar bear introgression. To this end, we performed Gene Ontology enrichment analyses of both protein-coding genes and protein-coding gene exons intersecting these regions. We found among our enriched terms some related to neurodevelopment, as well as embryonic development (Table 2.11 and Table 2.12). Although catalogs of genes for which positive selection likely played a role in the differentiation of brown and polar bears already exist [108], these new catalogs likely contain the most important genes for driving the species apart, as hybridization and subsequent backcrossing in these bears serves as a natural experiment in which divergent sets of alleles were "tested" together in nature, and the most compatible combinations were likely to be those passed on. Furthermore, where commonly used tests like $\frac{dN}{dS}$ focus exclusively on non-synonymous substitutions, our catalog is likely to include regulatory and splicing-related mutations as well. This merits follow-up study of the specific genes in our sets.

p	GO ID	term
0.000174025	GO:0001701	in utero embryonic development
0.00141049	GO:0000122	negative regulation of transcription by RNA polymerase II
0.00578795	GO:0072384	organelle transport along microtubule
0.00776506	GO:0034968	histone lysine methylation
0.0087984	GO:0014033	neural crest cell differentiation
0.00994697	GO:0021955	central nervous system neuron axonogenesis
0.00994697	GO:0051055	negative regulation of lipid biosynthetic process

Table 2.12: Enriched biological_process GO terms of protein-coding gene exons intersecting regions where none of 18 hybrid brown bears have any polar bear ancestry. Testing was done using the hypergeometric test in FUNC [164], with FUNC's refinement routine to account for hierarchical relationships among terms. Terms with $p < 0.01$ after refinement are shown.

2.1 Discussion

AD-LIBS is a new technique for the detection and analysis of ancestry in admixed individuals, designed for use with low-coverage shotgun sequence data from non-model organisms. The technique works well on both simulated and real data, requires only several reference individuals from each ancestral population (Figure 2.2), and is accurate at coverage depths as low as 2X (Figure 2.4).

AD-LIBS is unlikely to perform as well as other local ancestry inference techniques when high-confidence genotype calls and phased data from reference populations are available. Moreover, AD-LIBS overestimates heterozygous ancestry (Figure 2.3), although it has a lower false positive rate for identifying regions homozygous for one or the other type of ancestry (Figure 2.2) and infers the correct amount of homozygous, introgressed ancestry genome-wide (Figure 2.3). Therefore, one can be confident in results from AD-LIBS when analyzing genomic regions labeled as homozygous for ancestry from one or the other reference population but should use caution when describing regions heterozygous for ancestry.

Although AD-LIBS is robust to suboptimal choices of most parameters, window size must be chosen carefully, and F_{ST} between ancestral populations [75] must also be considered. The latter is important because overlap between the three emission probability distributions that AD-LIBS uses to determine which type of ancestry produced the set of IBS tract lengths in each window (see Methods) depends to a large extent on F_{ST} between the two ancestral populations. If nucleotide diversity between populations is large relative to nucleotide diversity within populations, then the means of the emission probability distributions will lie further apart than distributions for ancestral populations with low F_{ST} (see Appendix A for expected emission probability distributions). While increasing the window size can help mitigate this problem by decreasing the variances of the distributions, it may be impossible to get accurate results when dealing with populations with low F_{ST} between them. As an example, AD-LIBS is not expected to give accurate results for human populations, in which within-population nucleotide diversity is often very similar to between-population nucleotide diversity. With populations of sufficiently high F_{ST}, such as polar and brown bears, users should either allow AD-LIBS to determine an appropriate window size by measuring the overlap among emission probability distributions (see Methods) or carefully evaluate results to ensure they are realistic. Using too small a window size to distinguish populations that are closely related can result in error (Table 2.3).

Using AD-LIBS, we detected a greater amount of polar bear ancestry in 18 brown bear genomes than has been previously reported using other methods [19, 19, 108, 129]. It is possible that these polar bear ancestry estimates are inflated by several percent due to AD-LIBS overestimating the extent of heterozygous ancestry (Figure 2.3). If still valid, however, this finding illustrates an advantage of using local ancestry inference techniques like AD-LIBS over global techniques in admixture studies. If AD-LIBS is correctly inferring that Scandinavian brown bears have some polar bear ancestry, then prior studies that used these bears as model "unadmixed" brown bears may have underreported polar bear ancestry in all bears. The reason for this underreporting is that the global ancestry inference techniques used in prior studies, such as $\hat{f}$, are genome-wide averages relative to a genome presumed to be unaffected by past admixture. As such, global ancestry inference methods cannot detect polar bear ancestry in an individual brown bear as long as the model "unadmixed" brown bear to which it is compared has the same amount of polar bear ancestry anywhere else in its genome. Local methods like AD-LIBS and HAPMIX, conversely, can detect polar bear ancestry at a particular genomic locus within an individual, as long as the model brown bear genome to which it is being compared is free of polar bear ancestry at that same locus (Figure 2.8).

AD-LIBS maps of polar bear ancestry in brown bears also provide a look into the geographic history of polar-brown bear admixture. Given that principal components analysis (PCA) of polar bear ancestry groups bears geographically largely the same way as PCA of SNP data (Figure 2.7), polar-brown bear admixture may have taken place before the present day North American brown bear populations formed. The placement of the Montana brown bear near bears from Admiralty Island in principal component (PC) space also suggests that ABC Islands brown bears could have been the source of polar bear ancestry in mainland brown bears, as previously hypothesized [20]. The existence of a small amount of polar bear ancestry in Scandinavian brown bears, which is similarly observed in Finnish and Swedish bears in

PC space despite these bears clear difference in genotype PC space, suggests that there may have been a single, older polar-brown bear introgression event in Europe, independent from the source of polar bear ancestry in North American brown bears. If true, this result is evidence that hybridization between brown and polar bears may have been common in their evolutionary history, and may be the expected outcome of shifting habitat boundaries in times of global climate change.

Genes located in high-frequency introgressed regions, as well as those located in regions free of admixture, are also likely to tell an interesting story about speciation and the consequences of admixture. It is possible, for example, that our set of genes within regions devoid of polar bear ancestry might include genes involved in hybrid incompatibility. It is also noteworthy that we find evidence of adaptive introgression of some immune system-related alleles, as was found to be the case in archaic introgression into modern humans [171].

2.2 Conclusion

AD-LIBS expands the potential range of admixture analyses both to non-model organisms and to data sets in which only low-coverage genomes are available. While AD-LIBS should not replace existing approaches for high-coverage data or where phased reference panels are available, AD-LIBS accurately identifies genomic regions in hybrids that are homozygous for ancestry from a specific ancestral population, even with low coverage data. By thus reducing the quantity and quality of data needed, AD-LIBS can make admixture mapping a viable tool in a wider range of studies than was previously possible.

2.3 Methods

2.3.1 Model description

AD-LIBS (Ancestry Detection through Length of Identity By State tracts) is designed for use with low-coverage sequence data from diploid organisms. The insight behind AD-LIBS is to consider windows of a genome, rather than individual SNP sites, when determining ancestry. This allows groups of variants to vote together on the ancestry of windows in the genome, decreasing the influence of individual sites that might be prone to genotyping or sequencing error. AD-LIBS takes as input pseudo-haploid FASTA sequences, in which every base is randomly sampled from one or the other homologous chromosome, rather than sets of genotype calls at variable sites. This eliminates the need for variant calling, which can be problematic without prior knowledge of polymorphic sites as in non-model organisms. It also avoids problems inherent in identifying heterozygous sites using low-coverage sequencing data [19, 19].

If an individual has ancestors from both population A and B, each window of that individuals genome can be classified as a sample of two chromosomes from population A, two from population B, or one of each. The state space of the hidden Markov Model (HMM) used by AD-LIBS therefore includes three ancestry states: AA, which models genomic windows in which both of an individuals chromosomes descend from population A; AB, which models windows in which an individual derives one chromosome from each ancestral population; and BB, which models windows in which an individual is homozygous for ancestry from population B. Note that no attempt is made to phase" variants when ancestry is heterozygous: AD-LIBS does not attempt to determine which of the two homologous chromosomes is of population A or B ancestry in the heterozygous state. In our model, we always designate the ancestral population with lower genetic diversity as population A and the other as population B. Figure 2.9 describes a cartoon of the HMM state space, including states not yet described. AD-LIBS uses the Python Pomegranate library for hidden Markov model operations, available at https://github.com/jmschrei/pomegranate.

2.3.2 Transition probabilities

The transitions between states are related to the probability of recombination having occurred since admixture between the two ancestral populations. For this, AD-LIBS requires an estimate of the genome-wide extent of admixture and the number of generations since admixture. Given that the number of generations since admixture is g, and the per-nucleotide recombination probability per generation is r, the probability of a recombination event having taken place at a single nucleotide position in the time since admixture is gr. AD-LIBS assumes r to be a flat rate of 1 centimorgan per megabase, or 10^{-8} per site. If p, the extent of ancestry from population A in the admixed population, is known, then the probabilities of switching between state AA (homozygous population A ancestry), AB, (heterozygous ancestry), and BB (homozygous population B ancestry) can be determined. This requires considering the per-site probabilities of recombination events having happened or not in the time since admixture on both homologous chromosomes, along with the probabilities of the next base on each homologous chromosome being derived from population A or B. Additionally, AD-LIBS accounts for the effect of genetic drift: considering recombination events as alleles in the classic Wright-Fisher model, it derives the probability of resampling the same ancestral recombination event twice in a single individual, hereafter referred to as z. The probabilities of transitions between the three ancestry states are given in Figure 2.10.

As an example, two possible sets of events can lead to a transition from a region of homozygous population A ancestry (AA) to a region of homozygous population B ancestry (BB). One is that there has been a recombination event at the same site on both chromosome homologues in the time since admixture, the probability of which is $(gr)^2$, and that the base immediately after the recombination event is of population B ancestry on both chromosome homologues, the probability of which is $(1-p)^2$. This set of events thus has probability $(gr)^2(1-p)^2$. Conversely, if the two chromosome homologues have a recent common ancestor at the site of interest, it is possible that a historical recombination event between a region of population A ancestry and a region of population B ancestry happened once, but was inherited by both parents of the individual of interest. The probability of the individual inheriting the same historical recombination event from both parents is z, and the probability of the base immediately after this recombination event deriving from ancestral population B is $(1-p)$, making the probability of this set of events $z(1-p)$ (see Table 2.10).

Whereas r is a hard-coded approximation and g is a model parameter inferred from prior knowledge, the parameters p and z can be calculated. A popular method for estimating the admixture proportion p from sequence data is the statistic $\hat{f}$, an extension of the D statistic used to estimate the extent of Neanderthal ancestry in modern humans [56, 35]. D is a genome-wide measure of excess derived allele sharing between an admixed individual and candidate introgressor; it compares numbers of sites, genome wide, that support alternative tree topologies. The statistic $\hat{f}$ is a ratio of D computed on an admixed individual to D computed on an individual from the admixing population of interest. $\hat{f}$ can be used to obtain a lower bound on admixture proportion. When p is not supplied by the user, AD-LIBS requires a genome from an outgroup individual and at least two individuals from admixing population A; these are used to compute $\hat{f}$ as an approximation of p. Sometimes, for example when the test individual derives less of its genome from the introgressor than the individual hypothesized to be unadmixed, $\hat{f}$ can yield negative values. In this case, and in every other case where $p \leq 0$, we set p to a minimum value of 0.001. This allows AD-LIBS to detect regions of population A ancestry even when they were not originally expected, if the signal is strong enough.

The parameter z, or the probability of resampling the same ancestral recombination event twice in an individual, is less straightforward to compute. Conceptualizing recombination events of interest as alleles that arise within the admixed population during the time since admixture, with the per-site, per generation probability r, a Markov chain can be used to compute the probability of such a recombination event drifting to any frequency between 0 and $(2N)/(2N)$ where N is the number of individuals in the population, over the course of g generations [66]. This probability distribution can then be used to compute the probability of resampling the same recombination event twice in a single individual. Since the transition probability matrix for this Markov chain can become very large with large population sizes, making computation difficult, we implemented the solution to the diffusion approximation of this problem presented by McKane and Waxman [120] in AD-LIBS. For a detailed explanation of how the value of z is computed in AD-LIBS, see Appendix A.

In addition to the three ancestry states AA, AB, and BB, we defined three skip states, sAA, sAB, and sBB, which each model windows in which there is insufficient data to make an inference about ancestry (Figure 2.9). Each of these states is only capable of emitting a designated score representative of low-quality windows. Each is also much more likely to transition back to its associated ancestry state than to one of the others: the transition probability $P(AB|sAA) = p(AB|AA)$, $p(BB|sAB) = p(BB|AB)$, and so on. This allows the HMM to have memory of the state in which a sequence was before encountering windows of sparse data: the probability of transitioning to a new ancestry state is the same, whether or not windows of sparse data are encountered. The probabilities of transitioning from ancestry states to skip states can only be calculated after scanning a sequence: windows with a percentage of ambiguous or "N" bases above a user-specified threshold are designated "skipped," and the skip probability s is the number of skipped windows divided by the total number of windows in an input DNA sequence. The transition probability from each ancestry state to its associated skip state, as well as the probability of remaining in a skip state once there, is s. Since the emission probability distributions of skip states are very different from those of ancestry states, in practice the magnitude of s does not matter: windows intended to be skipped will be skipped whether s is high or low. For a more detailed explanation of other transition probabilities, and how transition probabilities are set on sequences belonging to the X chromosome (or Z chromosome for species using the ZW sex determination system), see Appendix A.

2.3.3 Emission probabilities

Rather than considering individual genotypes at known variable sites, AD-LIBS divides genomic sequences into windows and computes a score based on identity-by-state (IBS) tract lengths in each window. Identity-by-state tract lengths have proven useful in quantifying parameters of demography and admixture and underlie some popular methods for demographic inference [157, 64]. They are also easy to compute, can be measured without a set of high-confidence genotype calls, and have a clearly defined expected distribution, which should not be affected by the fact that our input data are pseudo-haploidized sequences rather than phased haplotypes.

AD-LIBS computes scores based on IBS tract lengths in fixed-width genomic windows. In each window, the "query" sequence from the hybrid individual is compared to all available sequences from ancestral populations A and B. The score x in a given window is $log((1/(aw))\sum_{i=1}^{a}[(1/n)\sum_{j=1}^{n}IBS_{i,j}]) - log((1/(bw))\sum_{i=1}^{b}[(1/n)\sum_{j=1}^{n}IBS_{i,j}])$ where a is the number of sampled individuals from population A, b is the number of sampled individuals from population b, n is the total number of IBS tracts found in a given window between the hybrid and another individual, $IBS_{i,j}$ is the length of the jth IBS tract with individual i sampled from either population A or B, and w is the window size in base pairs. In simpler terms, x is the ratio of the log transformed mean IBS tract length between the hybrid and individuals from population A, and between the hybrid and individuals from population B. For its emission probability distributions, AD-LIBS computes the expected distribution of x for each of the three ancestry states, with slight adjustments for scores in windows along the X (or Z) chromosome. For details, see Appendix A.

2.3.4 Potential pitfalls

One parameter that must be chosen carefully is the window size. Apart from upper and lower bounds set on window size by mathematical limitations (see Appendix A), users have the ability to choose window sizes for their analyses. AD-LIBS can recommend a window size by testing the amount of overlap among the three emission probability distributions. Since overlap among emission probability distributions can hinder the ability of AD-LIBS to distinguish among different types of ancestry, and since the variance of all three distributions will decrease as window size increases (see Appendix A), increasing window size can improve discriminative power while risking failure to detect short ancestral haplotypes. AD-LIBS recommends a window size by computing the emission probability distributions for a range of window sizes beginning at the minimum bound. At each window size, it integrates a function returning the minimum value of each pair of distributions over those distributions full range, which gives a measure of overlap [79]. The smallest window size for which the maximum pairwise distribution overlap is 0.5 or lower is recommended. If the chosen window size causes the maximum overlap of any two of the three distributions to exceed 0.5, AD-LIBS iteratively multiplies the standard deviations of all three emission probability distributions by 0.5 and recomputes the overlap until it falls below 0.5. While this makes the model less realistic, it has the potential to improve discriminative power.

2.3.5 Simulations

To test AD-LIBS, we used Hudsons coalescent simulator, ms [77], to simulate haplotypes under a demographic model representative of brown bears, polar bears, ABC Islands brown bears, and American black bears. We performed 20 trials in which ten 1 Mb pseudo-haploid chromosomes were generated for each of five polar bears, five mainland brown bears, five ABC Islands brown bears, and one black bear. Our demographic model is similar to that proposed by Cahill et al [19], in which hybridization between brown and polar bears takes place on Alaskas Admiralty, Baranof, and Chichagof (ABC) islands at the end of the Pleistocene epoch (the initial hybrid bear population consists of 50% polar bears and 50% brown bears). We chose for our model 0.94 Mya for the split time between the American black bear and brown and polar bears [94], 411 kya for the split between brown and polar bears [108], and 12 kya, the approximate end of the Pleistocene epoch, as the time of hybridization between mainland brown bears and the polar bears of the ABC islands [19]. Furthermore, we chose a generation time of 11.35 years and a per-site, per-generation mutation rate of 1.82572810^{8} [108], as well as a default recombination rate of 1 centimorgan per megabase, or per site. For nucleotide diversity values, we used and estimated by Cahill et al [19], along with estimated by Kutschera et al [94]; we converted these into effective population size values by dividing by four times the mutation rate. Our full ms command, which produced two haplotypes for each simulated individual, was ms 32 10 -t 1700.0 -r 931.135415571 1000000 -I 4 10 10 10 2 -n 1 0.235294117647 -n 4 1.23529411765 -es 0.0113546182949 2 0.5 -ej 0.0113546182949 2 1 -ej 0.0113546182949 5 3 -ej 0.3888956766 1 3 -ej 8.89445099767 3 4 T.

In each of the 20 simulations, ms generated two black bear haplotypes and 10 each of polar bear, mainland brown bear, and ABC Islands brown bear haplotypes, with ten repetitions. After splitting the ms output files into individual repetitions, we then used Seq-Gen [175] with the Hasegawa, Kishino, and Yano (HKY) nucleotide substitution model [67] and a 4:1 transition:transversion ratio to convert each haplotype from each repetition into a DNA sequence. The full Seq-Gen command was `seq-gen -mHKY -t 4 -l 1000000 -s 0.0017 -p [number of trees in ms output file] [ms output file]`. We then sampled two haplotypes per individual and, using a Python program, randomly choose the base from one or the other haplotype at each position to generate 1 Mb pseudo-haploid chromosome sequences. Finally, we concatenated the 1 Mb haploid sequences for each individual across the ten repetitions to yield 10 Mb simulated genomes for 5 polar bears, 5 mainland brown bears, 5 ABC Islands brown bears, and one black bear.

We used the trees from ms to produce maps of "true" ancestry for each hybrid bear in order to validate AD-LIBS results for each trial. We output trees with the `t` parameter of ms and used these to produce BED files of the true ancestry of each segment of the simulated chromosomes for all five ABC Islands bears. To do this, we used a Python program to parse the trees describing the relationship of all simulated haplotypes at each segment of the simulated chromosome. For each ABC Islands brown bear haplotype at each segment, we computed the time to most recent common ancestor (TMRCA) with all polar bear haplotypes and with all brown bear haplotypes. In order to distinguish admixture from incomplete lineage sorting, we designated an ABC Islands bear haplotype as having polar bear ancestry only if its TMRCA with all polar bear haplotypes was more recent than its TMRCA with all brown bear haplotypes, and if its TMRCA with all polar bear haplotypes postdated the polar-brown bear split. If both haplotypes comprising a pseudo-haploid ABC Islands bear chromosome have polar bear ancestry in a given region, that region is designated "AA" for homozygous polar bear ancestry; if both have brown bear ancestry, it is designated "BB;" if there is one haplotype with each type of ancestry, it is designated "AB;" and if none of these is the case, no ancestry call is made. These are used as maps of "true" ancestry across the simulated chromosomes.

We then ran AD-LIBS on each simulated hybrid bear and assessed its accuracy using its map of "true" ancestry. For our initial estimate of polar bear ancestry in each hybrid bear, we calculated using the first two polar bear sequences, the first mainland brown bear sequence, and the black bear sequence as an outgroup. We then ran AD-LIBS with an admixed population size of 3000 and 1000 generations since admixture, and nucleotide diversity values that were calculated directly from the generated sequences. We note that only having 10 megabases of sequence for each bear may have hurt the accuracy of our calculations, since $\hat{f}$ is a genome-wide average that requires a large number of single-site observations to disentangle true admixture from incomplete lineage sorting [56, 35]. For each simulated ABC Islands brown bear chromosome, we tried running AD-LIBS with one polar bear and one brown bear sequence to use as reference data, then two of each, three of each, four of each, and five of each, to determine whether the number of reference sequences affected output. We also used three different window sizes – 5 kb (slightly above the minimum threshold set by AD-LIBS, given the nucleotide diversity in the sequences), 10 kb, and 25 kb – for the same reason. After generating the results, we compared the output of AD-LIBS to the BED files of "true" ancestry by compiling the intersection of AD-LIBS ancestry features with the true ancestry features using BEDTools intersect [167] and determining the true ancestry across each window by majority vote of true ancestry regions contained within. For each state, then, we calculated the percent of bases for which the HMMs classification was correct, as well as the percent of bases for which the true ancestry was recovered by the HMM. Admixture proportion was calculated as two times the number of bases in homozygous polar bear windows, plus the number of bases in heterozygous windows, all divided by two times the total number of windows for which AD-LIBS produced a label.

To compare performance, we repeated this experiment, but this time used a demographic model in which polar bears on the ABC islands are gradually converted into brown bears by continuous gene flow from mainland American brown bears. In this model, we allowed mainland bears to migrate to the ABC islands population at a rate of 0.001 (0.1% of each generation of the ABC Islands population is composed of brown bear migrants), beginning 12,000 years ago and continuing until the present. This model produces ABC Islands bears more varied in their polar bear ancestry proportion, and possibly more similar to the true ABC Islands bears. The full procedure for simulations with this model was the same as above, but using the ms command ms 32 10 -t 1700.0 -r 931.135415571 1000000 -I 4 10 10 10 2 -n 1 0.235294117647 -n 4 1.23529411765 -m 2 3 93.1135415571 -em 0.0113546182949 2 3 0 -ej 0.0113546182949 2 1 -ej 0.3888956766 1 3 -ej 8.89445099767 3 4 T. We refer to the former model, with a single hybridization event, as the "single-pulse" model and the latter model, with continuous gene flow, as the "migration model."

2.3.6 Human and Neanderthal data

We ran AD-LIBS on human and Neanderthal data as a further test of AD-LIBSs ability to correctly calculate admixture proportion, since high-coverage human and Neanderthal sequencing data are readily available, many studies have already sought to identify Neanderthal admixture proportions in modern humans and F_{ST} between humans and Neanderthals is reasonably high. We chose to scan European genomes for Neanderthal ancestry, because Neanderthal-human admixture is well studied, and we chose European over East Asian individuals because the history of Neanderthal-European gene flow may be simpler and involve fewer admixture events than that of Neanderthal-East Asian gene flow [214, 88, 215]. We randomly selected five European (CEU) individuals and five Yoruba (YRI) individuals from phase 3 of the 1000 Genomes Project [36], downloaded BAM files mapped to reference genome hg19 for each, and created a haploidized genomic sequence for each individual using the samtools mpileup utility [104] with map and base quality cutoffs of 20, along with a program that chooses a random base from the set that passed filters at every position, filtering out bases where coverage was greater than the 97.5th percentile of coverage genome-wide. The European individuals used were NA11832, NA11840, NA12340, NA12383, and NA12814; the Yoruba were NA18504, NA18870, NA18934, NA19099, and NA19238. We then downloaded variant calls for the high-coverage Altai Neanderthal [165] and generated two "haplotype" sequences in hg19 coordinates using a program that transforms VCF to FASTA format, randomly assigning each variant at heterozygous sites to one or the other haplotype. Treating YRI and Altai as the two reference populations, we then calculated $\pi_{Altai} = 0.000303$, $\pi_{YRI} = 0.001525$, and $\pi_{Altai-YRI} = 0.001763$ from these sequences and chose a population size of 10,000, based on prior estimates [182], and 2,000 generations since admixture, roughly based on inferences drawn from Neanderthal haplotype block lengths in ancient human genomes [51, 50]. Neanderthal admixture proportions were estimated by calculating using both Neanderthal haplotype sequences, the Yoruba individual NA18504, and the reads from chimpanzee genome release PanTro4 [25], mapped to hg19 coordinates by the UCSC Genome Browser team [87]. After running AD-LIBS on each individual, we computed admixture proportion using the same technique as described in the Testing with simulated data section.

2.3.7 Bear data preparation

Our bear sequence data were all published as part of previous studies [129, 19, 20, 108]; sample details are given in 2.2. We selected for study 11 hybrid brown/polar bears from Alaskas Admiralty, Baranof, and Chichagof (ABC) islands (ABC01, ABC02, ABC03, ABC04, ABC05, ABC06, Adm1, Adm2, Bar, Chi1, and Chi2), one brown bear from Montana known to have polar bear ancestry (GP01), two brown bears with some polar bear ancestry from the Alaskan mainland (Den and GRZ), four Scandinavian brown bears hypothesized to be free of polar bear ancestry (OFS01, RF01, SJS01, and Swe), and four polar bears selected for high coverage depth (PB7, PB12, PB68, and PB105). Most data were downloaded as reads from the NCBI SRA, subjected to adapter removal and read merging using Seq-Prep (`https://github.com/jstjohn/SeqPrep`) and mapped to the polar bear reference genome [108] using BWA MEM [102], sorted and indexed with samtools [104], and subjected to indel realignment via GATK, followed by duplicate removal via PicardTools [121]. The Denali park brown bear (Den), Swedish brown bear (Swe), Admiralty Island brown bear (Adm1), and American black bear (Uam), however, were processed as published in previous studies [19, 20]: adapter trimming using Trimmomatic (Bolger, Lohse, and Usadel 2014), mapping using BWA aln [103], and duplicate removal using samtools rmdup [104] followed by GATKs indel realignment [121]. Following this, we selected four polar bears (PB7, PB12, PB68, and PB105), two Scandinavian brown bears (OFS01 and RF01), and four ABC Islands brown bears (ABC01, ABC05, Adm2, and Bar), each of which had a minimum of 20x genome-wide average coverage, and performed variant calling on these using GATKs Unified Genotyper. We set a minimum base and map quality of 30, and then discarded variants with a genotype quality lower than 30 or a variant quality lower than 50. We also filtered to exclude sites for which coverage was lower than 4 or greater than the 97.5th genome-wide percentile for any individual bear; this yielded 16,635,425 SNPs and 3,054,975 indels. In addition to using these variant calls for downstream analysis, we used BEAGLE [17] with no reference panel, no imputation, and five iterations to phase our SNPs, resulting in a panel of 15,637,657 (94% of the original SNPs) phased polymorphic sites.

To compensate for our inability to reliably identify heterozygous sites in low-coverage (¡ 20x) individuals, and to format our data for use with AD-LIBS, we generated pseudo-haploid sequences in reference genome coordinates for all bears by choosing a random base with minimum map and base quality of 30 at every site, skipping sites where coverage was greater than the 97.5th percentile of genome-wide coverage [19, 20]. This was done using samtools mpileup with the polar bear reference genome and map and base quality filters, then piping to an in-house program that selects and outputs a random high-quality base at each position, yielding a FASTA file. Genome-wide coverage was computed using bedtools genomecov [167]. We then filtered these sequences to only scaffolds with a minimum length of 500kb in the reference genome and calculated $\pi_{polar} = 0.000615$, $\pi_{brown} = 0.00233$, and $\pi_{polar-brown} = 0.003564$ using a utility included with AD-LIBS on these sequences. We then ran AD-LIBS on all bears, assuming an admixed population size of 3,000 and 2,000 generations since admixture, using PB105, PB12, OFS01, and Uam to estimate each admixed bears admixture proportion via $\hat{f}$. For the Scandinavian bears OFS01, RF01, SJS01, and Swe, assumed to be free of polar bear admixture [20, 108], we specified an admixture proportion of 0.001 in order to allow the model to detect polar bear ancestry if it existed. We inferred ancestry for each of our brown bears using a window size of 10kb (the size that worked best using simulations), a skip threshold of 0.25 (which gave very similar results to runs with skip thresholds of 0.1, 0.5, and 0.75), and using an X chromosome model for scaffolds determined belong to the X chromosome in a previous study [19]. We set the time since admixture to 1000 generations ago, the approximate end of the Pleistocene epoch assuming a generation time of 11.35 years [108] and an admixed population size of 3,000 individuals.

We also used our panel of phased SNPs to infer ancestry for our four ABC Islands bears that had at least a 20x average depth of coverage (ABC01, ABC05, Adm2, and Bar), using HAPMIX [160], with GENOTYPE=1, OUTPUT_SITES=1, THETA=0.08, LAMBDA=900.0, RECOMBINATION_VALS=600 600, MUTATION_VALS=0.2 0.2 0.01, and MISCOPYING_VALS=0.05 0.05. This gave us an independent map of polar bear ancestry for these four bears against which to compare AD-LIBSs results. We note that our HAPMIX results are not as reliable as those for human data, since our reference panel was phased computationally and thus subject to switch errors. After running HAPMIX, we converted output to BED files that could be compared to AD-LIBS results using an in-house program. This program assigns an ancestry state (homozygous population A, heterozygous, or homozygous population B) to each site by choosing the highest ancestry probability output by HAPMIX, or skipping sites where two or more probabilities are equal. It then merges runs of sites with the same ancestry into contiguous regions of ancestry and prints results in BED format.

2.3.8 Low-coverage tests

To test AD-LIBSs performance on low-coverage data, we used the same alignments as our full-coverage data. We chose to limit analysis, however, to the four hybrid ABC Islands bears for which we were able to run HAPMIX at full coverage (ABC01, ABC05, Adm2, and Bar), owing to the fact that these bears were sequenced to minimum 20X coverage and thus yielded reliable genotype calls. For unadmixed "reference" bears, we included all four polar bears (PB7, PB12, PB68, and PB105), as well as the three Scandinavian brown bears sequenced to at least 10X coverage (OFS01, RF01, and SJS01). We note that our full-coverage HAPMIX runs used only the Scandinavian bears over 20X coverage (OFS01 and RF01), and so our low-coverage HAPMIX runs actually had one more reference individual available than our high-coverage runs. For computational efficiency, we limited analysis to the longest scaffold of the polar bear reference genome (scaffold1, 67.4 Mb). For each bear, we compiled a random set of properly-paired reads that mapped to scaffold1 with minimum map quality 30 using samtools view, samtools bamshuf, and samtools bam2fq [104]. We then calculated, for each bear, the number of reads from these random sets required to obtain 0.5X, 1X, 2X, 5X, and 10X coverage across scaffold1. We then took subsets of our sets of high-quality mapping reads and, for each bear at each coverage level, mapped these reads to polar bear scaffold1 using BWA MEM [102], then performed GATKs [121] indel realignment on the resulting BAM files. We did not remove duplicates, since the BAM files were already deduped prior to downsampling. We used our previously described strategy for preparing pseudo-haploid FASTA sequences (samtools mpileup and a program that randomly chooses a base at each position that passes quality filters), with map and base quality cutoffs of 20, to prepare data for use with AD-LIBS. We also used GATKs UnifiedGenotyper to call SNPs along scaffold1 for every bear at each coverage level, removing sites with map or base quality below 20, as well as indel or non-biallelic variants. Following this, we phased variants using BEAGLE [17] with no reference panel, no imputation, and five iterations at each coverage level.

We ran HAPMIX on each bear for each coverage level using the same parameters as for full-coverage data (see Data preparation section), and its results were converted to BED files for easy comparison to AD-LIBSs results. AD-LIBS was then run on each hybrid bear at each coverage level with the same parameters as the full-genome runs, with the exception that nucleotide diversity values were computed from the haploidized FASTA files rather than using the previously-calculated values, the skip threshold was set to 0.75 to accommodate more missing data, and prior estimates of polar bear ancestry proportion were all set to 0.08, as they were in all HAPMIX runs. To compare output of HAPMIX and AD-LIBS runs to each other, we used the same technique as we did when comparing AD-LIBS results for simulated data to the BED files of true ancestry, described in the Testing with simulated data section.

2.3.9 Shared polar bear ancestry

To test for sharing of the same types of ancestry across the same regions of the genomes in multiple bears, we used a custom Python program, along with several existing tools. We first created merged BED files of each specific type of ancestry for each bear using BEDTools [167], grouping heterozygous and homozygous ancestry together for one ancestry type. We also used BEDTools intersect to compute the size (in base pairs) of the intersection of each group of bears for each ancestry type, and random samples were taken from the polar bear genome using BEDTools shuffle, limited to the polar bear genomic scaffolds that were at least 500kb long the same set of scaffolds on which AD-LIBS was run.

In order to run EIGENSOFT SmartPCA [148] on the bear ancestry data, we used a custom script to convert AD-LIBSs BED files into EIGENSTRAT format, using the starting coordinate of each window as the position of each "variant," dropping "scaffold" from scaffold names, setting genetic distance to 0 (the default) for each "variant" so that a flat recombination rate can be assumed across each scaffold, and coding each homozygous polar bear window as 2, heterozygous windows as 1, and homozygous brown bear windows as 0.

For the SNP-based PCA run, we downloaded the set of polar and brown bear SNPs published as part of a recent study [108], excluded all polar bears, converted to EIGENSTRAT format, and ran EIGENSOFT SmartPCA [148] the same way as with our ancestry data. Only the first two principal components were considered.

2.3.10 Functional analyses

We performed all tests of overlap and correlation of genomic intervals using the GenometricCorr R package [46]. We note that our tests might be slightly biased, since we do not have a map of centromeric and telomeric regions to exclude from analyses; such regions should be devoid of genes. We performed all gene ontology enrichment tests using FUNC [164], using the Wilcoxon rank order test for introgressed segments and the hypergeometric test for polar bear-free genomic regions. In both cases, we used the refinement script to account for the hierarchy of terms, with a p-value cutoff of 0.01. For Gene Ontology terms, we used the October 29, 2018 version of the Gene Ontology tables [1]. To compute frequencies of introgressed haplotypes, we first created BED files for two pseudo-“haplotypes” for each hybrid brown bear, assigning each homozygous polar (AA) segment to both haplotypes and each heterozygous (AB) ancestry segment to one of the two haplotype files. We then compiled all files using bedtools multiinter [167] and divided by the number of haplotypes possessing each feature by 36 (2 times the number of bears) to obtain frequencies. Means of these frequencies were used to rank genes when performing Wilcoxon rank-order Gene Ontology enrichment tests.

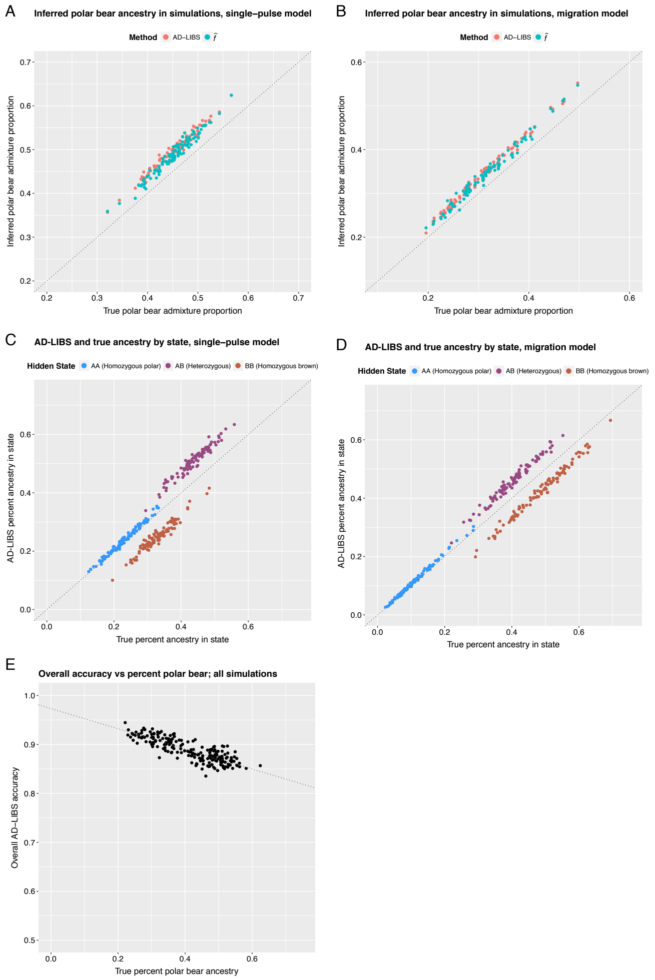


Figure 2.3: Accuracy of AD-LIBS estimates of the overall extent of polar bear ancestry, using simulated data. A and C refer to simulations with a single polar-brown bear admixture event 12,000 years ago, followed by isolation (single-pulse model); B and D refer to simulations in which a brown-polar bear admixture event 12,000 years ago is followed by continual breeding with unadmixed brown bears (migration model). All AD-LIBS runs in this figure used 5 reference individuals per ancestral population and 10kb windows. A and B: Inferred percent polar bear ancestry using AD-LIBS and $\hat{f}$ versus true percent polar bear ancestry. C and D: inferred percent polar bear ancestry of each type, according to AD-LIBS, versus true percent polar bear ancestry of each type. Each point represents the percent of a single simulated hybrid bear genome with a specific type of ancestry. E: overall accuracy of AD-LIBS inferences versus true percent polar bear ancestry, including both types of simulations. The line of best fit by least-squares regression is also shown. Accuracy decreases slightly as polar bear ancestry increases, probably due to the tendency of AD-LIBS to overestimate the extent of heterozygous ancestry (C and D).

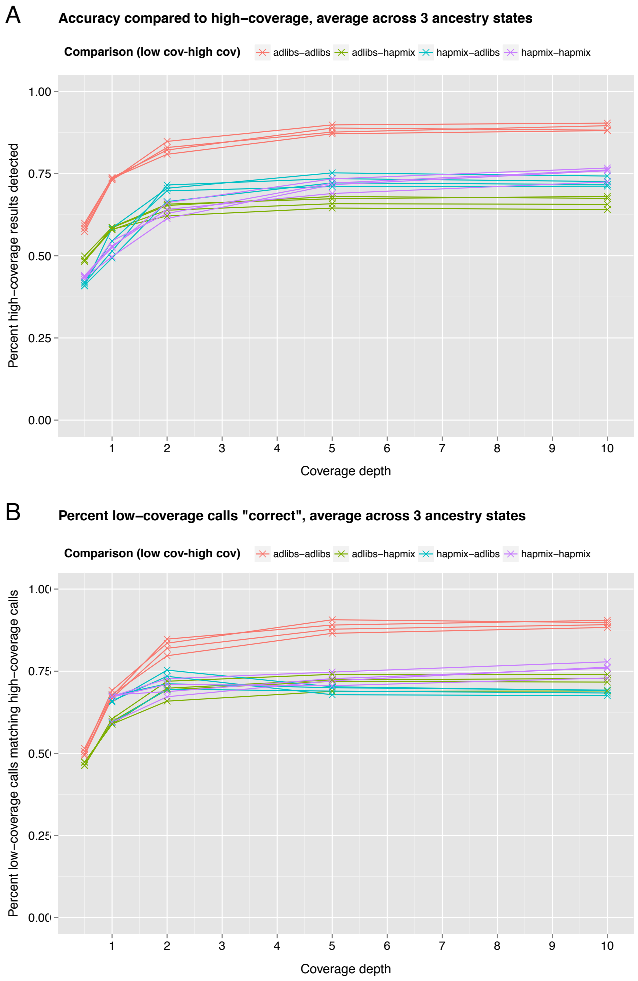


Figure 2.4: Results from downsampling four ABC Islands brown bears, three Scandinavian brown bears, and four polar bears to 0.5x, 1x, 2x, 5x, and 10x coverage along the longest genomic scaffold, running HAPMIX [160] and AD-LIBS on the four ABC Islands bears at each coverage depth, and comparing these runs to results obtained from running both programs on the full-coverage versions of the same individuals. Each line represents an individual ABC Islands bear and each color represents a specific low coverage/full coverage comparison. A: percent of full coverage calls recovered by running each program at low coverage. Values given are averages across the three ancestry states (homozygous polar bear, heterozygous, and homozygous brown bear). B: percent of low coverage calls that were correct, according to full-coverage calls. Values given are averages across the three ancestry states. Some points are missing because HAPMIX was unable to detect any polar bear ancestry at 0.5x coverage.

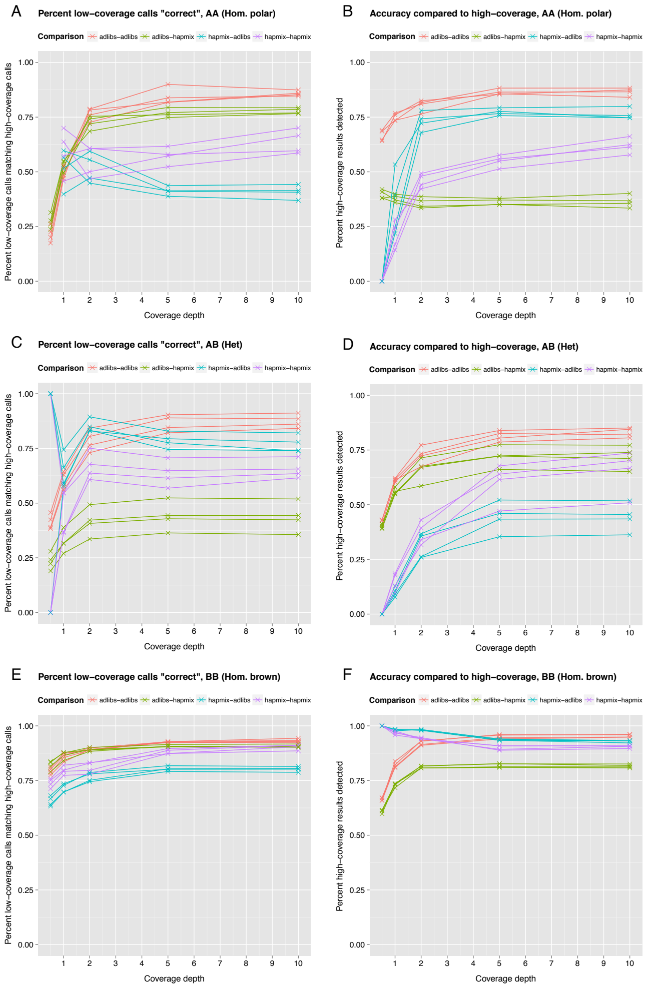


Figure 2.5: Results from downsampling four ABC Islands brown bears, three Scandinavian brown bears, and four polar bears to 0.5x, 1x, 2x, 5x, and 10x coverage along the longest genomic scaffold, running HAPMIX and AD-LIBS on the four ABC Islands bears at each coverage depth, and comparing these runs to results obtained from running both programs on the full-coverage versions of the same individuals. Each line represents an individual ABC Islands bear and each color represents a specific low coverage/full coverage comparison. A and B assess homozygous polar bear (AA) calls, C and D assess heterozygous (AB) calls, and E and F assess homozygous brown bear (BB) calls. A, C, and E measure the percent of low-coverage calls that were correct" according to the high-coverage runs, while B, D, and F measure the percent of the high-coverage runs calls that were also detected by the low-coverage runs. In almost every case, AD-LIBS is more consistent with itself than other comparisons. We also note that low-coverage AD-LIBS inferences of homozygous polar and brown bear ancestry are more often correct, according to HAPMIX run at full coverage, than HAPMIX run at low coverage (A and E). AD-LIBS may, however, erroneously call more windows heterozygous than HAPMIX does (C), leading to its missing some windows of homozygous polar (B) and brown bear (F) ancestry.

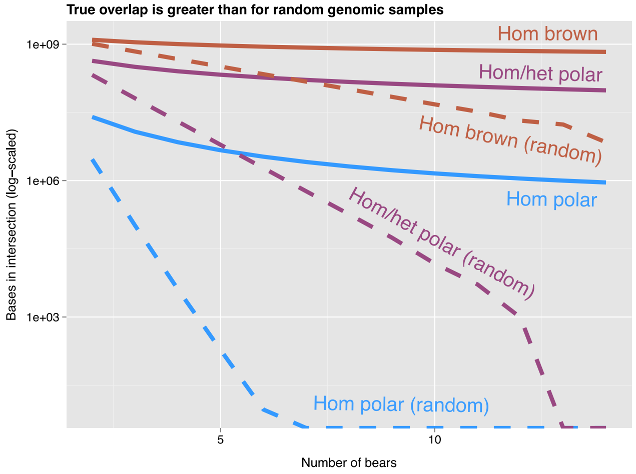


Figure 2.6: Comparing overlap of regions of different types of ancestry among hybrid bears. For every combination of 2 or more American brown bears, we measured the number of bases that AD-LIBS labeled with the same type of ancestry (homozygous polar, homozygous/heterozygous polar, or homozygous brown) in each bear. We also performed one random trial per real comparison, in which coordinates comprising random regions were sampled from the reference genome, producing sets of genome regions of the same size and number as the regions of ancestry produced by AD-LIBS for each bear, but randomly scattered across the genome. We then measured the overlap between these random ancestry regions for the sake of comparing to the true overlap. Averages of every comparison of each number of bears are shown as solid lines, and averages of every comparison of randomized versions of those same bears are shows as dashed lines.

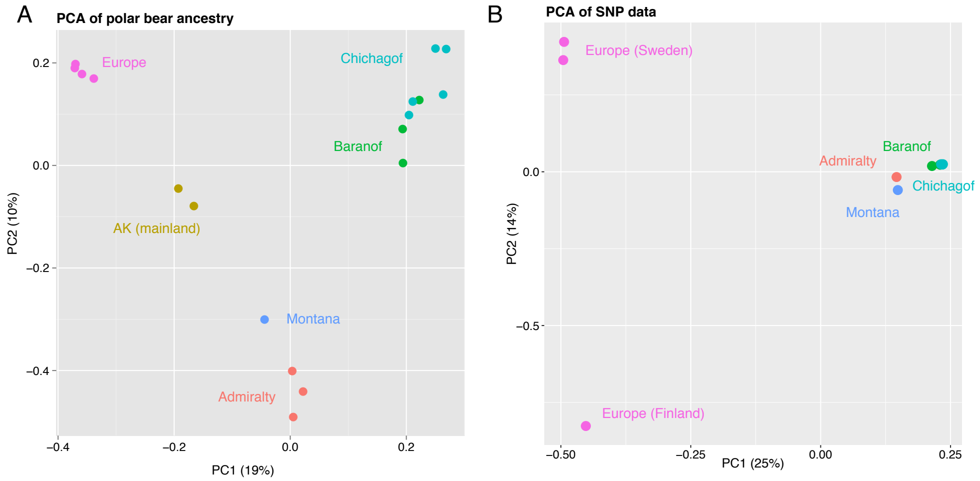


Figure 2.7: Geographic signal recovered in vectors of polar bear ancestry. A: principal components analysis (PCA) of polar bear ancestry state of 10kb genomic windows for 18 brown bears, using EIGENSOFT SmartPCA [148]. B: PCA of SNP data from a previous study [108], including a subset of the bears in A. Both plots show similar geographic patterns, with the Montana bear (GP01) falling close to the Admiralty Island bear(s), but only the SNP data separates Finnish (RF01) from Swedish brown bears.

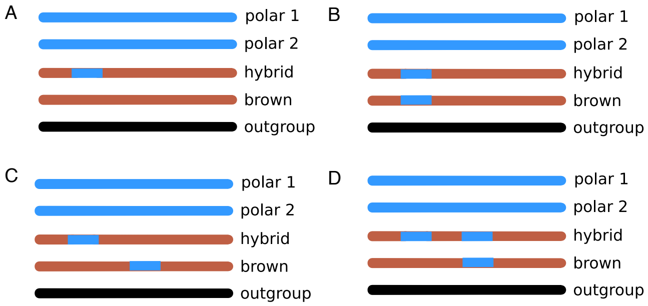


Figure 2.8: Illustration of cases where either a local ancestry detection method (like AD-LIBS) or a global ancestry detection (like $\hat{f}$) might succeed, partially succeed, or fail. Each line represents a chromosome, with polar bear ancestry shown in blue and brown bear ancestry shown in brown. All five individuals needed for computation of are shown in each case. A: local and global methods both succeed in detecting all of the hybrid individuals polar bear ancestry. B: local and global methods both fail to detect the hybrid individuals polar bear ancestry. C: local methods successfully detect the hybrid individuals polar bear ancestry, since it is in a different part of the genome than the polar bear ancestry in the genome of the model "unadmixed" brown bear. Global methods fail to detect the hybrid individuals polar bear ancestry. Since global methods use genome-wide averages, the hybrid individual is not seen to possess any more polar bear ancestry than the model "unadmixed" brown bear. D: Both local and global methods will detect the hybrids first segment of polar bear ancestry but fail to detect the second segment, resulting in both types of methods underestimating the hybrid individuals true percent polar bear ancestry.

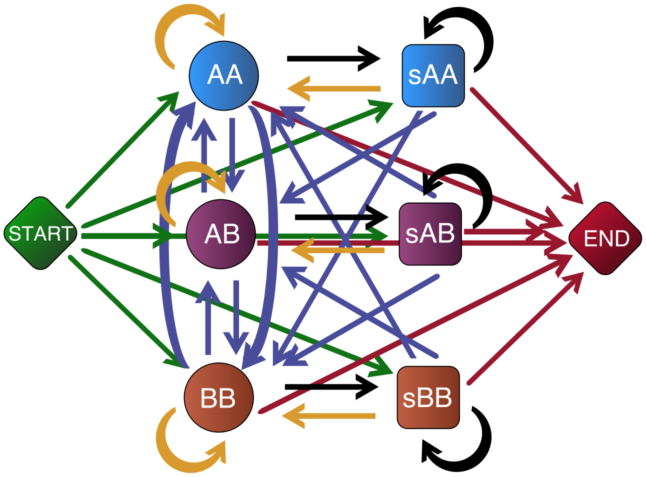
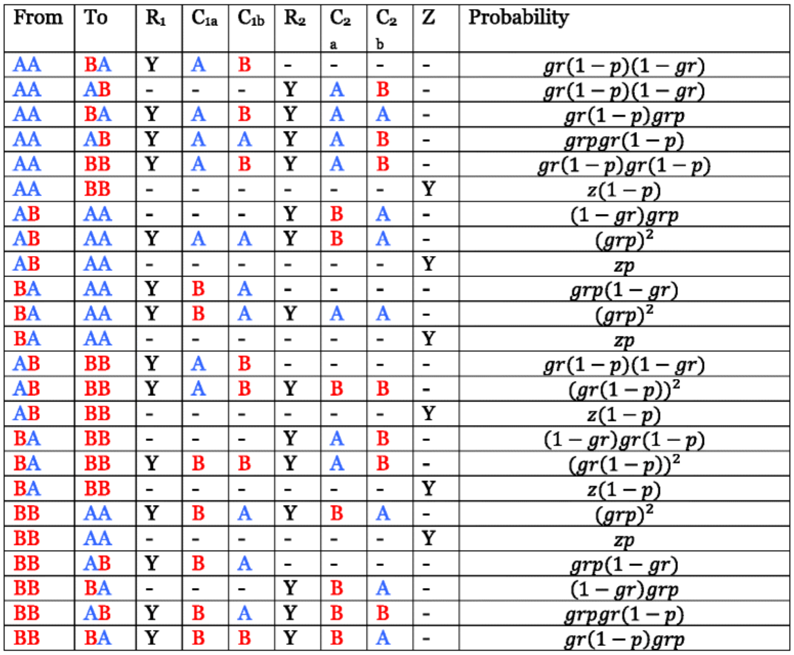


Figure 2.9: The state space of AD-LIBSs hidden Markov model. The three round states (AA, AB, and BB) are ancestry states that can emit scores. AA represents regions where both homologous chromosomes derive ancestry from ancestral population A, AB represents regions of heterozygous ancestry, and BB represents regions homozygous for population B ancestry. The three square states (sAA, sAB, and sBB) are skip states, each associated with one of the three ancestry states. Skip states can only emit scores representing windows of the genome in which data are too sparse to infer ancestry. Each skip state is more likely to transition back to its associated ancestry state than to one of the others. Arrow colors represent different types of transition probabilities. Green arrows are starting probabilities and are related to the pre-estimated percent ancestry derived from each ancestral population (A and B). Blue arrows represent recombination events; their probability is related to the probability of a recombination event having happened at a given site in the time since admixture, as well as the probability of sampling a base from population A or B. Black arrows are related to the probability of skipping a given window, computed from the number of "N" bases encountered. Red arrows are transitions to the end state, with probabilities related to the number of windows on the chromosome or scaffold being scanned. Gold arrows represent probabilities that are computed after other probabilities, by subtracting from 1 the sum of all other transition probabilities out of a given state.



From	To	R_1	C_{1a}	C_{1b}	R_2	C_2 a	C_2 b	Z	Probability
AA	BA	Y	A	B	-	-	-	-	$gr(1-p)(1-gr)$
AA	AB	-	-	-	Y	A	B	-	$gr(1-p)(1-gr)$
AA	BA	Y	A	B	Y	A	A	-	$gr(1-p)grp$
AA	AB	Y	A	A	Y	A	B	-	$grpgr(1-p)$
AA	BB	Y	A	B	Y	A	B	-	$gr(1-p)gr(1-p)$
AA	BB	-	-	-	-	-	-	Y	$z(1-p)$
AB	AA	-	-	-	Y	B	A	-	$(1-gr)grp$
AB	AA	Y	A	A	Y	B	A	-	$(grp)^2$
AB	AA	-	-	-	-	-	-	Y	zp
BA	AA	Y	B	A	-	-	-	-	$grp(1-gr)$
BA	AA	Y	B	A	Y	A	A	-	$(grp)^2$
BA	AA	-	-	-	-	-	-	Y	zp
AB	BB	Y	A	B	-	-	-	-	$gr(1-p)(1-gr)$
AB	BB	Y	A	B	Y	B	B	-	$(gr(1-p))^2$
AB	BB	-	-	-	-	-	-	Y	$z(1-p)$
BA	BB	-	-	-	Y	A	B	-	$(1-gr)gr(1-p)$
BA	BB	Y	B	B	Y	A	B	-	$(gr(1-p))^2$
BA	BB	-	-	-	-	-	-	Y	$z(1-p)$
BB	AA	Y	B	A	Y	B	A	-	$(grp)^2$
BB	AA	-	-	-	-	-	-	Y	zp
BB	AB	Y	B	A	-	-	-	-	$grp(1-gr)$
BB	BA	-	-	-	Y	B	A	-	$(1-gr)grp$
BB	AB	Y	B	A	Y	B	B	-	$grpgr(1-p)$
BB	BA	Y	B	B	Y	B	A	-	$gr(1-p)grp$

Figure 2.10: All possible combinations of events leading to transitions between the three ancestry states of the hidden Markov model. "A" and "B" denote chromosomes derived from ancestral populations A and B, and the three states AA, AB, and BB model regions where ancestry is homozygous from population A, heterozygous, and homozygous from population B, respectively (AB and BA are represented by same state, but are shown separately here to clarify that the ancestry of both separate chromosomes must be considered when computing probabilities). The other columns denote possible recombination-related events on the two parental homologues of a given chromosome (henceforth "homologue 1" and "homologue 2"). A "Y" in the R1 column signifies that recombination took place at a given base on chromosome homologue 1 in the time since admixture, and a "Y" in the R2 column signifies that recombination took place at this base on chromosome homologue 2. C1a and C1b refer to the ancestry of the bases on chromosome homologue 1 immediately before and after the recombination event, if it happened; C2a and C2b refer to the ancestry of the bases on chromosome homologue 2 before and after recombination. Z indicates that the same ancestral recombination event, which happened in the time since admixture, was sampled twice in the same individual (once on chromosome homologue 1 and once on chromosome homologue 2). The parameter g is the number of generations since admixture, r is the recombination probability per site per generation, assumed to be 1 cM/Mb or 10^{-8} per site, and z is the probability of resampling the same ancestral recombination event twice in one individual, according to genetic drift approximated by the Wright-Fisher model.

Chapter 3

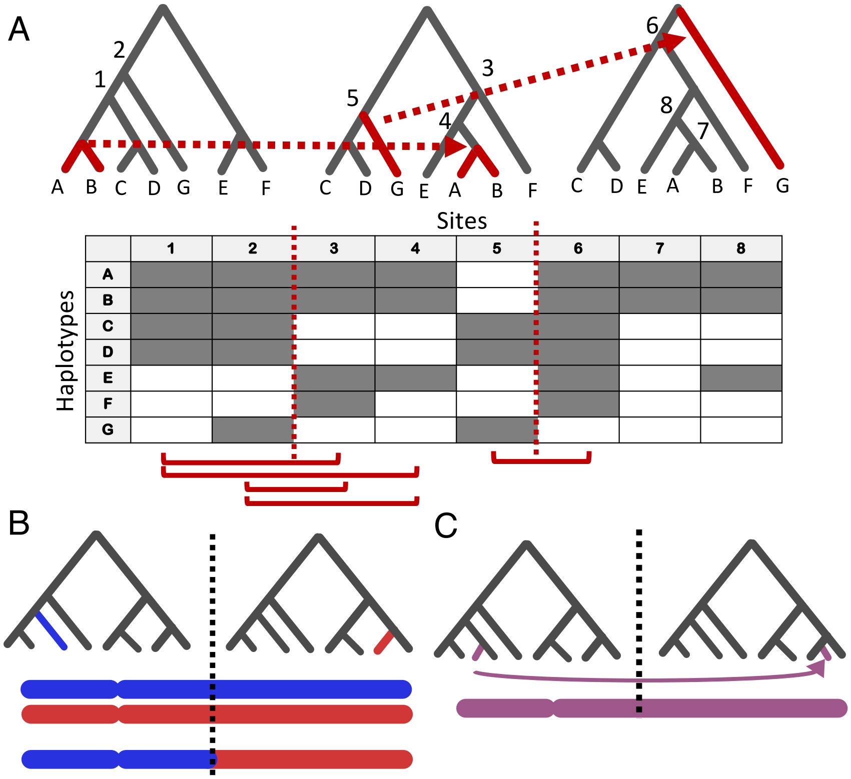
A method for fast, heuristic ancestral recombination graph inference

In this section, I describe a new algorithm for quickly inferring an ancestral recombination graph (ARG) over a set of phased haplotypes. In addition to many other applications, an ARG can be used to map ancestry across hybrid or admixed genomes. The algorithm I present here is demographic model-free and uses only parsimony (in mutations as well as ancestral recombination events) as a guiding principle. I then demonstrate its use on simulated genomes and find that it produces trees that are 80-90% accurate on average. It scales to run on thousands of haplotypes, although more trees contain ambiguities (polytomies) as more haplotypes are added.

3.1 Background

Many methods have been developed for mapping ancestry across hybrid or admixed genomes. S*, a statistic summarizing linkage between SNPs, has been used to detect [154], map [214], and test the existence of [211] recent archaic hominin ancestry in modern humans, without the need for an archaic hominin reference genome. Gene tree topologies and branch lengths have been used for the same purpose [62, 4], although such approaches have produced false positive results [97, 122]. Other studies have taken an ensemble approach, combining multiple locus-specific statistics using techniques like linear regression [37] or conditional random fields [186].

Ancestral recombination graph (ARG) inference [59] provides an appealing alternative method for ancestry mapping, with higher resolution and fewer built-in assumptions. An ARG is a series of trees, mapped to individual sites, over all phased haplotypes in a genomic data set. Ancestral recombination events, or sites at which chromosome segments with different histories were joined together by historical recombination, form boundaries between trees. Each ancestral recombination event manifests as a clade of haplotypes, all of which descend from the first ancestral haplotype to possess it, moving from one position in the tree upstream of the event to a new position in the downstream tree [200] (Fig. 3.1). ARGs are complete descriptions of phylogenomic data sets and present for recombining genomes what simple trees present for nonrecombining ones. As prior techniques for ancestry mapping can be thought of as summaries of the ancestral recombination graph, higher resolution ancestry maps could be produced if the ARG were known.



	1	2	3	4	5	6	7	8
A	▨	▨	▨	▨		▨	▨	▨
B	▨	▨	▨	▨		▨	▨	▨
C	▨	▨			▨	▨		
D	▨	▨			▨	▨		
E			▨	▨		▨		▨
F			▨			▨		
G		▨			▨			

Figure 3.1: A: example showing ARG inferred by parsimony. Columns are variable sites; shaded cells denote shared derived mutations. Red-highlighted sites fail the four haplotype test with other red- 5 highlighted sites (shown by brackets), vertical dotted red lines mark ancestral recombination breakpoints. Site numbers mark clades they tag in the trees, and red arrows show ancestral recombination events. B: example of a recombination event joining together the blue haplotype upstream of the break point (dotted line) and the red haplotype downstream of the breakpoint. Location of the blue and red haplotypes in a consensus tree are shown above. C: how the daughter haplotype produced in B will appear in an ARG: within the blue haplotypes clade upstream of the recombination breakpoint and in the red haplotypes clade downstream of it.

Although software exists to infer ARGs, Existing ARG inference techniques often do not scale well to genome-wide data sets with many samples. These include BEAGLE, which was designed for smaller data sets [200], ArgWeaver, which uses powerful statistical techniques that limit scalability [176], and Rent+ [131]. Furthermore, some techniques produce fully articulated trees and therefore describe relationships not knowable from the data [176, 131]. Another technique, Margarita, randomly samples histories at ancestral recombination event boundaries and does not seek to produce parsimonious recombination histories [130]. A recently described approach, tsinfer, overcomes many other techniques problems with scalability but does not infer branch lengths and assumes that each mutation's frequency is correlated with its age [83]. Since this assumption is violated at loci undergoing either admixture or selection, this technique is poorly suited for mapping archaic ancestry in modern humans.

We present a new heuristic, parsimonious ARG inference algorithm called SARGE (Speedy Ancestral Recombination Graph Estimator) that can run on thousands of phased genomes, makes no prior assumptions other than parsimony, estimates branch lengths, and represents uncertainty due to missing mutations as polytomies in output trees. We validate SARGE using simulated data and demonstrate that it is as accurate as existing methods. We also demonstrate that it is suited to the analysis of real data, using a set of 279 human genomes from the Simons Genome Diversity Project panel [116] together with several genomes from archaic hominins [165, 126, 163].

3.2 Results

We developed a parsimony-based ARG inference technique, SARGE, that uses ancestral recombination events to help articulate trees. Our method relies on the four-gamete test (or four-haplotype test) [76], a simple test that identifies discordant clades between a pair of loci. The crux of our technique is a simple algorithm for choosing the branch movement(s) capable of explaining the highest number of discordant clades. In short, if the full topology of two trees adjacent in the genome is known ahead of time, one can compile the list of four-gamete test failures between the two trees. Each four-gamete test failure then implies one of three possible edits (or branch movements [200]) to transform the upstream into the downstream tree. One can create a graph, where each four gamete test failure becomes two nodes (the upstream clade and the downstream clade failing the test), and each possible tree edit is another type of node. For each pair of nodes failing the four-gamete test, an edge is drawn from the upstream node, through each candidate "edit" node, to the downstream node. This can result in up to 3 possible paths between each pair of nodes (any candidate edit node that fails the four gamete test with a node in either tree is omitted from the graph). Once the graph is drawn, the edit node with the most edges represents the most parsimonious way to edit the first tree to transform it into the second, or the edit that can explain the most four-gamete test failures. Under parsimony, this edit serves as a solution to the problem of which ancestral recombination events happened between two adjacent trees (Fig. 3.2).

To expand this algorithm to more general use, we created a graph data structure where every node is a clade over a contiguous genomic region. Nodes keep track of genomic sites that tag them and have genomic start and end coordinates, as do their edges to parent and child nodes. Each node is only allowed to be compared with other nodes within a set distance of their furthest upstream and downstream sites. When a node's clade fails the four gamete test with another node, candidate "edit" or moving nodes are created, with potential edges through them, as described in the above algorithm (Fig. 3.3). The way our data structure is currently designed, we require input data to be phased in advance.

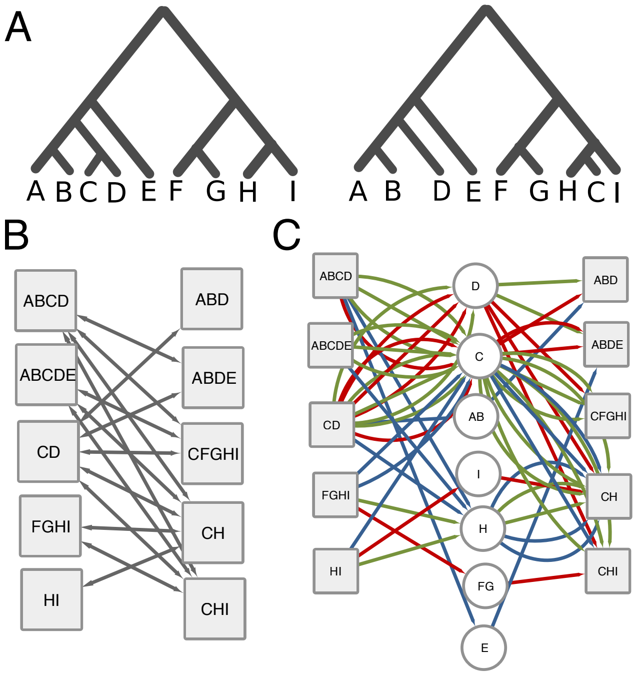


Figure 3.2: Example of algorithm for inferring branch movements between to trees known *a priori*. For more information on terminology, see Methods. A: Two trees, which differ by one branch movement. B: Clades from the two trees that fail the four haplotype test. Left column shows clades from the first (upstream) tree and right column shows clades from the second (downstream) tree; arrows indicate four haplotype test failures. C: Graph showing all possible branch movements that could explain the four haplotype test failures shown in B. The left and right columns are "tree" nodes, while the center column lists candidate γ clades. Colors indicate types of four haplotype test failures: red paths are conditional on a failure being the α/α type, green on it being α/β, and blue on it being β/β. In this case, a single candidate γ clade (C) has the most edges and can explain all four haplotype test failures. This is interpreted as the clade C moving from the smallest observed α clade in the first tree (CD) to the smallest observed β clade in the second tree (CH). If no β clades from the second tree are observed, the branch movement goes upward to a clade containing the union of all clades failing the four haplotype test. If no α clades from the first tree are observed, the branch movement goes downward from a clade containing the union of all clades failing the four haplotype test.

One benefit of our method is that it can infer clades in trees from shared ancestral recombination events. This helps circumvent a foundational problem in ARG inference: unrecombined segments of chromosomes often do not contain enough genetic diversity to fully articulate trees. Therefore, even if the boundaries of all ancestral recombination events were

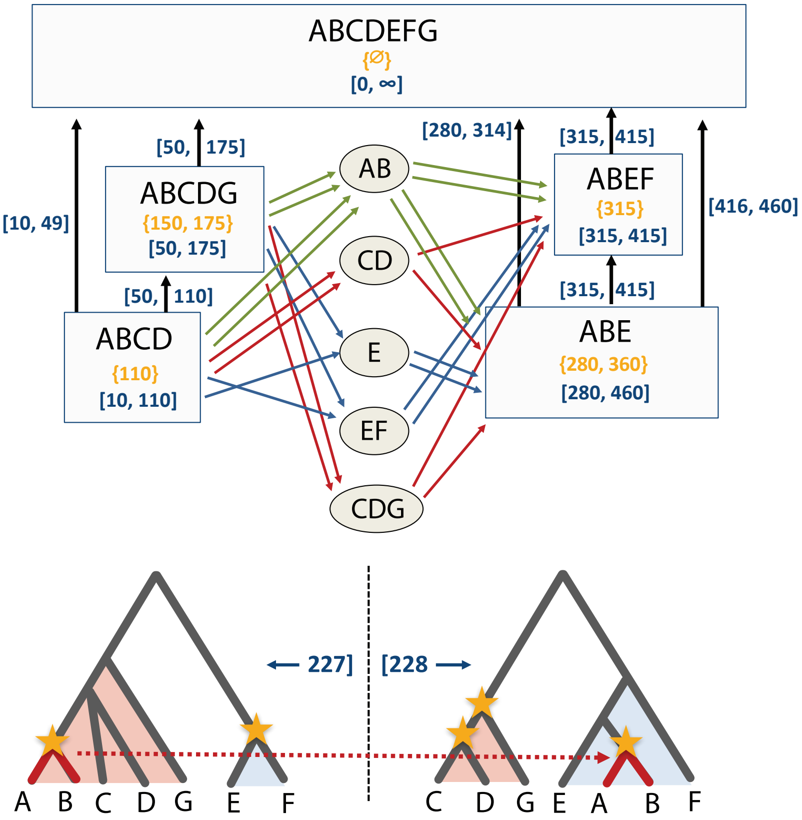


Figure 3.3: Schematic of data structure. For more information on terminology, see Methods. Top: rectangles are "tree nodes" representing clades in trees. Each has a set of haplotypes (represented by letters A-G), and a start and end coordinate (blue numbers in brackets) determined by coordinates of SNP sites tagging the clade (yellow numbers in braces), along with a propagation distance parameter (100 in this example). Parent/child edges (vertical arrows) also have start and end coordinates determined by the nodes. Ovals are candidate nodes that can explain four gamete test failures; colored edges indicate potential paths between tree nodes through candidate nodes that could explain four gamete test failures (colors indicate types of paths). The candidate node with the most edges is eventually chosen as the most parsimonious branch movement, allowing for the inference of new nodes; the two trees at the bottom show the "solved" ancestral recombination event with the branch movement marked in red and all clades inferred without SNP data marked with yellow stars. The coordinates of the recombination event (blue numbers in brackets) are taken to be midway between the furthest-downstream upstream site and the furthest-upstream downstream site involved in recombination.

known *a priori*, many organisms would lack the level of nucleotide diversity necessary to observe enough clades within each segment to build a complete tree. In the case of high-heterozygosity organisms, one would observe more ancestral recombination events than in low-heterozygosity organisms, thus making the lengths of un-recombined segments smaller and leading to the same problem. In our method, when an ancestral recombination event is “solved,” we can often infer the existence of clades implied by the recombination event, without observing them in the input SNP data (Fig. 3.3). In addition to helping articulate the trees, this process also allows us to use sites upstream and downstream of the ancestral recombination event, but not affected by it, to articulate the trees.

One other feature of our method is that it produces results that are always consistent with the input data. This means both that it is fairly accurate and that it often avoids making a statement when there is no evidence for any possible inference. This manifests in output data as polytomies (clades with more than two children).

For the sake of quality control, we created a simulated data set consisting of haplotypes drawn from a constant-size, unstructured population with a level of heterozygosity similar to that of modern-day sub-Saharan Africans. Our algorithm’s performance on this data set should be a lower bound on actual performance, because we expect population structure to result in more derived alleles and ancestral recombination events tagging groups, thus making ARG trees easier to articulate. We sampled increasing numbers of haplotypes from our simulated data set, inferring an ARG on each set of haplotypes and assessing performance. We find that our algorithm is on average 84.76% correct (95% C.I. 84.75-84.77%), with both accuracy and the number of polytomies increasing with the number of input haplotypes (Fig. 3.4). We also find that there is an asymptote to how articulated trees can become and how many clades can be inferred from ancestral recombination events, given increasing numbers of input haplotypes (3.5A,B). It is also reasonably fast, requiring approximately 1 hour to infer an ARG over 1 Mb of sequence with 5000 input haplotypes (3.5D).

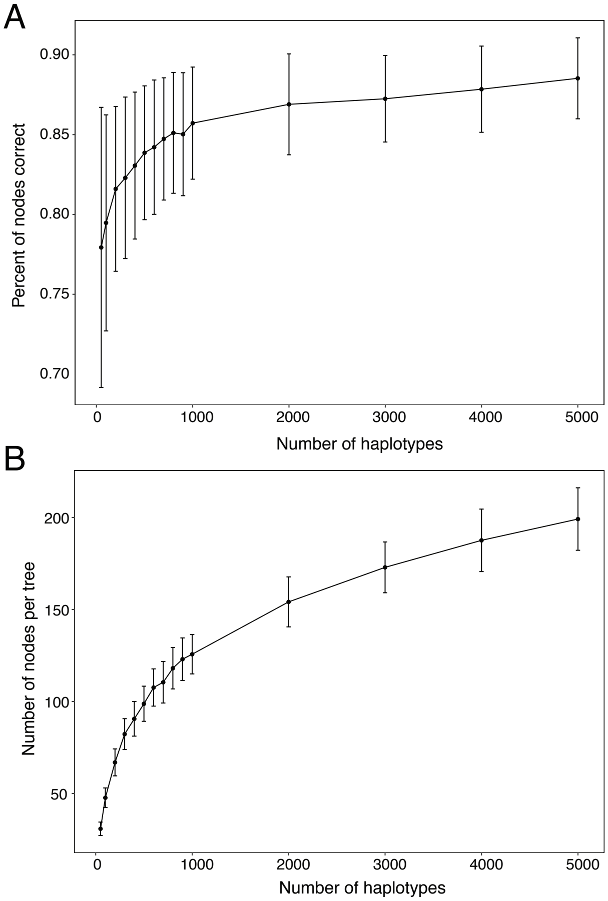


Figure 3.4: A: Accuracy of SARGE on simulated data (defined as percent of all clades correct according to the true ARG in the simulation), with increasing numbers of human-like haplotypes from an unstructured population. Error bars are one standard deviation across 5 replicates. B: Number of nodes per tree with increasing number of haplotypes in simulated data.

Next, we compared the performance of our ARG inference algorithm to two other scalable, recently-published alternatives. We omitted ArgWeaver [176], often considered the

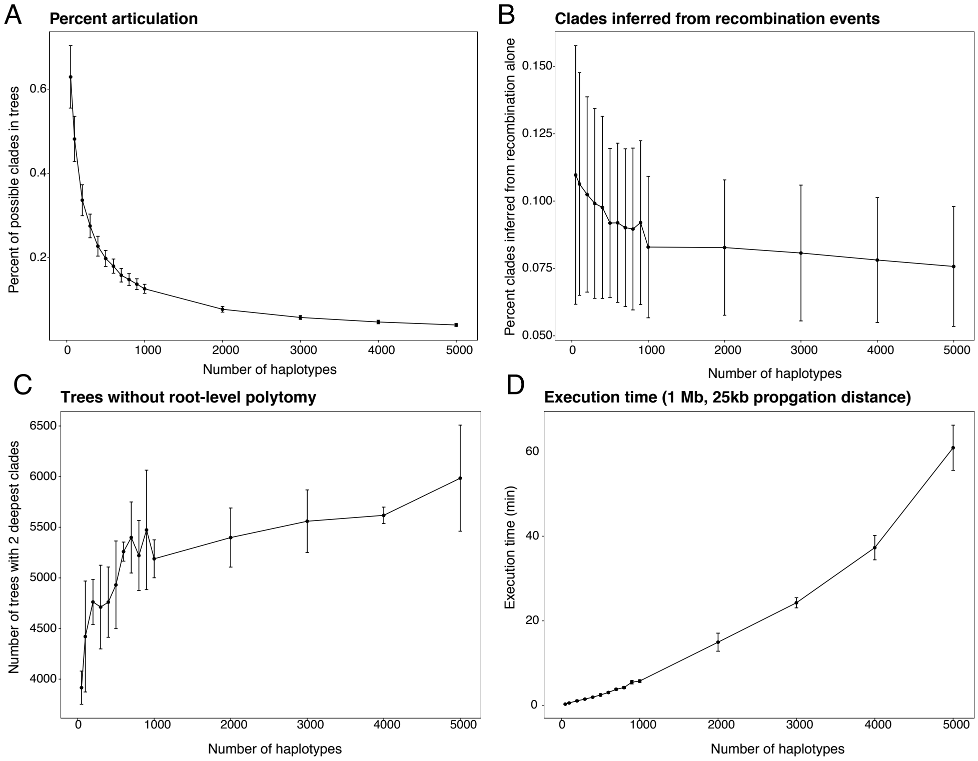


Figure 3.5: Properties of SARGE performance on simulated data with a sub-Saharan African-like level of heterozygosity, constant population size history, and no structure. Points are means; error bars show one standard deviation. A: Tree articulation as a percent of all nodes possible (given the number of haplotypes), with increasing number of haplotypes. B: Percent of all clades (across all trees) inferred from solving recombination events (rather than shared mutations). C: Number of trees across the chromosome with two children of the root node (no root-level polytomies). D: Execution time as a function of the number of input haplotypes. Real data, where SNPs and recombination events cluster in the genome, is likely to increase execution time.

state of the art, from this analysis because it does not run efficiently on data sets of the size we considered. We ran both tsinfer [83] and RENT+ [131] on our simulated data and compared the performance of SARGE to both. We find that SARGE produces the least-articulated but most accurate trees of the three programs. Although it is slower than tsinfer, it is orders of magnitude faster than Rent+ (Fig. 3.6). Given that tsinfer does not infer branch lengths and makes an assumption that is violated in cases of admixture and selection, we believe SARGE to be a useful alternative method for exploring data sets of similar size.

Next, we assessed the performance of SARGE on a real data set consisting of 279 human genomes [116] together with three archaic hominin genomes [165, 126, 163]. We were able to infer an ARG over all haplotypes in approximately 5 days, using 24 cores and never over 10 GB of RAM total. In this data set, an average 13.2% of clades SARGE inferred were learned solely from ancestral recombination events and are not observed in the input SNP data. Additionally, a genome wide tree inferred from shared ancestral recombination events agrees well with one produced from SNP data (Fig. 3.7A), suggesting that its inferences are reliable. We also observe that tree articulation is positively correlated with the mutation rate to recombination rate ratio, as expected (Fig. 3.7B; Spearmans rho = 0.40; $p < 2.2e-16$).

3.3 Discussion

We have presented a new ancestral recombination graph inference algorithm that scales to run on hundreds (or thousands) of input haplotypes and produces results that are always consistent with input data. Our algorithm is slower than tsinfer, but it provides data (branch lengths) unavailable using that technique, and it achieves greater accuracy by being more conservative.

In addition to many other uses, ARGs present a promising new way to map ancestry in hybrid genomes, as well as scan for evidence of selection. These two concepts could be combined to look for cases of adaptive introgression. One could also look for genomic regions where admixture is absent across a panel of genomes to infer the existence of hybrid incompatibility loci.

One current shortcoming of our method is that it requires phased data. Although experimental phasing techniques are becoming more widely available, another version of our algorithm that incorporates phasing would be useful. This would require an expansion of the data structure such that each clade would have multiple potential versions, each including or excluding specific haplotypes based on potential phase configurations. One could then solve ancestral recombination events in a way that minimizes their overall number. This would be a large undertaking, but it would be useful.

3.4 Methods

We downloaded data from the Simons Genome Diversity Project (SGDP) panel [116], along with two Neanderthal [165, 163] genomes and one Denisovan [126] genome. The Simons data were downloaded in pre-phased form from `https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data`; phasing was done using SHAPEIT2 [33]. We note that the hosts of the data state that the genotypes they provide at sites lacking a homologous chimpanzee are unreliable; we discarded all such sites from analysis.

Existing variant call sets for the ancient samples were either created using a genotype caller that did not account for ancient DNA damage [165, 126] or were subjected to a mappability filter that discarded many sites in the genome [163]. Because our method is sensitive to genotype errors and seeks to make inferences at every possible site in the genome, we chose to re-call variants in these three genomes using the ancient DNA-aware genotype caller snpAD version 0.3.0 [162]. For all snpAD runs, we required a minimum base quality of 25 and treated different types of libraries separately, separating UDG-treated and non-UDG treated libraries in the case of the Vindija Neanderthal, and separating single-stranded and double-stranded library data for the Altai Neanderthal and Denisovan.

Although the SGDP data were already phased, phasing posed a challenge for the Neanderthal and Denisovan data, for which there is no reference panel and for which DNA is fragmented into short segments. Fortunately, the comparatively low nucleotide diversity in these archaic hominins results in the presence of long runs of homozygosity, which are phased by definition. As an unbiased first step, we performed read-backed phasing using WhatsHap version 0.16 [145] (with default parameters, plus `ignore-read-groups`). Before filtering SNPs for quality and coverage, this phased 722,828 of 11,746,838 heterozygous sites (6.2%) in the Altai Neanderthal, 346,992 of 48,083,469 heterozygous sites (0.7%) in the Vindija Neanderthal, and 514,575 of 33,951,346 heterozygous sites (1.5%) in the Denisovan. Many of the remaining, unconfidently phased heterozygous sites were later removed from analysis, however: after filtering our data, we were left with only 1,677,774 of 49,876,210 total SNPs (3.4%) for which at least one archaic hominin individual was heterozygous and not phased by read-backed phasing.

Following read-backed phasing, we merged archaic hominin VCF files (using bcftools merge from bcftools version 1.8 [101]) and then phased the merged files using Eagle2.4 [110], with the 1000 Genomes Project data [2] as a reference panel. We used Eagle2s default parameters, but specified that it should not impute missing data (–noImpMissing) and that it should output alleles that it could not phase (–outputUnphased). After this, we randomly assigned both alleles at every unphased heterozygous site to one or the other haplotype. Although this decision, along with the use of a modern human reference panel, undoubtedly introduced haplotype switch errors, we deemed this preferable to excluding sites that were not confidently phased (which would require us to exclude data from all of the Simons Genome Diversity Project individuals at the same sites). To mitigate problems arising from this decision, we avoided performing any haplotype-specific analyses on the archaic hominin genomes. When creating maps of archaic hominin ancestry in modern humans, for example, we track only whether a modern human haplotype is in a clade with one or more archaic hominin haplotypes at each site, but not which specific archaic hominin haplotype is in the clade.

We merged the phased archaic hominin files with the SGDP data, using bcftools merge with the `missing-to-ref` option, and then used bcftools norm to remove duplicate alleles (`-d`). To avoid mis-identifying all SGDP samples as homozygous reference at sites that were originally excluded from the SGDP data set, we limited the variant call set for each chromosome to the sites between the first and last site in the SGDP data on that chromosome. To mitigate the same problem, we also removed any site for which all non-reference alleles in our SGDP data were private to archaic hominins, but for which non-reference alleles were present in modern humans within the 1000 Genomes data set [2]. We then discarded all sites for which any individual had a missing genotype or genotype quality below 25 or for which any archaic sample fell within the upper or lower tail of its genome-wide coverage distribution (extracted from the VCF file). The allowed coverage ranges (determined by eye) were 23-70X for the Altai Neanderthal, 10-43X for the Denisovan, and 10-47X for the Vindija33.19 Neanderthal.

Finally, we polarized our variant call set into ancestral and derived alleles, using the chimpanzee reference genome panTro4 [25] (mapped to hg19 by the UCSC Genome Browser team [21] and downloaded in AXT format) as an ancestral sequence, discarding any variant that was an indel, had more than two alleles, or lacked a known chimpanzee homolog. We chose panTro4 as an ancestral sequence rather than a composite ancestral sequence as some other studies have done (e.g. [36]) because it allowed us to more easily estimate branch lengths, at the cost of discarding some sites. Additionally, because our approach assumes the infinite sites model of mutation, we excluded all CpG dinucleotide sites from analysis, as methylated cytosines in CpG dinucleotides are highly mutable and are thus more likely than other nucleotides to undergo repeated mutations [41].

3.4.1 Ancestral recombination graph inference

We developed an ancestral recombination graph inference program called SARGE (available at `https://github.com/nkschaefer/sarge`), which is optimized for speed and low memory usage, in addition to making minimal model assumptions. SARGE assumes parsimony and the infinite sites model and uses the four gamete test [76] as a central insight. SARGE avoids using statistical techniques to smooth branch lengths or infer clades, opting instead to describe only that which can be inferred directly from the input data. The result is a set of trees that contain polytomies and have relatively low-resolution branch lengths.

Our algorithm centers on the observation that a single tree cannot contain two clades that share members unless one is a superset of the other. We assume that every shared derived allele in our data set defines a clade. It has been shown that, under this assumption, pairs of sites for which the inferred clades share members, but for which neither is a superset of the other, mark ancestral recombination events, or breakpoints between different trees. This is referred to as the “four haplotype test” or “four gamete test” [76, 200]. One could use this technique to map ancestral recombination events, which mark boundaries between trees, articulate trees using the sites within these boundaries. In practice, however, this can only produce minimally articulated trees. In the case of organisms with low nucleotide diversity, this is because there will not often be enough polymorphic sites between ancestral recombination breakpoints to observe many of the possible clades per tree (Fig. 3.1). In the case of organisms with high nucleotide diversity, however, it will be possible to detect far more ancestral recombination events, thus making the size of “bins” between ancestral recombination breakpoints smaller and leading to the same problem.

Our algorithm therefore seeks to infer all relevant information about each ancestral recombination event. An ancestral recombination event can be conceptualized as a branch movement [200], and so each consists of a set of haplotypes moving from one clade in an upstream tree into a new clade in a downstream tree. Given two clades that share members, but for which neither is a superset of the other (henceforth described as a failure of the four haplotype test), and assuming that this four haplotype test failure describes only one ancestral recombination event, there are then three possible branch movements than can explain it (Fig. 3.8). We refer to the clade in the upstream tree from which a subclade moved as α, the clade in the downstream tree into which a subclade moved as β, and the subclade that moved positions as γ. Four haplotype test failures are possible between the following sets of clades (with the clade in the upstream tree listed first and the clade in the downstream tree listed second): α/α, α/β, and β/β. In the case of an upward branch movement, all four haplotype test failures are α/α, and all four haplotype test failures are of the type β/β in the case of downward branch movements. The members of the moving clade γ can then be inferred once the type of four haplotype test failure is known. Denoting the members of the upstream clade as U and the members of the downstream clade as D, γ contains $U \setminus D$ in the α/α case, $U \cap D$ in the α/β case, or $D \setminus U$ in the β/β case.

3.4.2 Inferring branch movements between two trees

With this insight, we developed a simple algorithm to infer the most parsimonious ancestral recombination event (branch movement) between two trees, if the trees are known *a priori* and fully articulated. First, all clades in the upstream tree are compared to all clades in the downstream tree to collect four haplotype test failures. We then create a graph with a node for every upstream or downstream clade involved in a four haplotype test failure. Then, for each pair of nodes (U, D) failing the four haplotype test, where U is the set of haplotypes belonging to the upstream clade and D is the set of haplotypes belonging to the downstream clade, we create three nodes representing candidate γ clades: $U \setminus D$, $U \cap D$, and $D \setminus U$. Each of these candidate γ nodes is added to the graph only if it does not fail the four haplotype test with any clade in the upstream or downstream tree. Additionally, if the node already exists, it is retrieved from the graph rather than created anew. For each four haplotype test failure, we then create edges connecting the upstream and downstream node through an intermediate candidate γ node; these edges store the type of four haplotype test failure (α/α, α/β, or β/ β), conditional on the candidate γ node in the path.

Once all paths have been added, the candidate γ node with the most edges is chosen as the most parsimonious branch movement explaining the data (Fig. 3.2). If the chosen γ node is connected to both α nodes in the upstream tree and β nodes in the downstream tree, then the branch movement is inferred to be lateral; the node moved from the smallest upstream α to the smallest downstream β. If the γ node is not connected to any downstream β nodes, the branch movement is inferred to be an upward movement from the smallest upstream α clade to a clade containing the union of all clades connected to the chosen γ node. If the γ node is not connected to any upstream α nodes, then the branch movement is inferred to be a downward movement from a clade containing the union of all clades connected to the chosen γ node to the smallest downstream β node. If multiple γ are tied, then there are multiple equivalent ways to describe the same branch movement.

After a given γ is chosen, the set of four haplotype test failures is revisited in case multiple branch movements are required to explain the data. If for a given set of clades U and D, the chosen γ equals $U \setminus D$, $U \cap D$, or $D \setminus U$, then U and D are removed from consideration. Otherwise, new candidate γ nodes are created from the clades $U \setminus D \setminus\gamma$, $(U \cap D) \setminus\gamma$, and $D \setminus U \setminus\gamma$. The graph is then rebuilt using remaining four haplotype test failures, and another most parsimonious γ is chosen. This process is repeated until there are no four haplotype test failures remaining.

3.4.3 General case algorithm

Extrapolating this approach to ARG inference poses several problems. First, it cannot be known *a priori* which clades belong together in trees. Grouping clades together into upstream and downstream sets is therefore a difficult problem that we solve by exploring many possible groupings and bound using heuristic assumptions (see Heuristic). Second, many of the clades that could inform ancestral recombination events will be unobserved, if they are not tagged by mutations at sites in the data set.

Knowing this, we infer ancestral recombination events using the available mutations and then use these inferred ancestral recombination events to infer clades that they imply (Fig. 3.5B). Namely, we assume that γ clades should exist as clades in the ARG, whether or not they are tagged by mutations, because the haplotypes in γ share at least one ancestral recombination event. All subclades within the upstream α clade, with the γ clade haplotypes removed, must also exist as clades in the downstream tree. Likewise, all subclades within the downstream β clade, with the addition of γ haplotypes, must also exist in the upstream tree. Finally, in the case of an upward or downward branch movement (inferred by the absence of any β clades or α clades in the four haplotype test failures, respectively), the union of all clades failing the four haplotype test should exist as a clade in the ARG.

The other key component of our algorithm is a "propagation distance" parameter, p. This parameter describes how far upstream and downstream (in physical distance) each sites clade is allowed to communicate its existence. Because the all-versus-all clade comparisons required by our algorithm would become very computationally expensive without knowing *a priori* which clades belong to adjacent trees, this parameter helps bound the number of comparisons and thus the execution time. It also allows us to avoid storing an entire ARG over a chromosome in memory at once. As we read new sites into memory, we can identify nodes sufficiently far away upstream to be unaffected by the new data. We can then "solve" ancestral recombination events for those upstream nodes, and other nodes even further upstream, whose ranges leave them unaffected by the newly-solved recombination events, can be written to disk and erased. Because errors and violations of the infinite sites model (such as back-mutations) invariably exist, this parameter has the extra benefit of limiting how far along a chromosome erroneous data can propagate (although a cascade of incorrect clades inferred by recombination could hypothetically propagate errors outside of the range of the original, erroneous node).

This leads us to define a graph containing two types of nodes: "tree nodes," which are part of the ARG, and "recombination nodes," which represent candidate γ clades for unsolved ancestral recombination events. Each tree node represents a given clade over a contiguous genomic span and has a start and end coordinate, a set of positions of SNPs that tag it, and a set of other sites at which it was inferred to exist as part of a recombination event. Tree nodes have parent/child edges, also with start and end coordinates, and there is a single root node that spans the entire chromosome. Node range coordinates are initially set to the furthest upstream site owned by the node minus the propagation distance, up to the furthest downstream site owned by the node, plus the propagation distance. When a node encounters another node with which it fails the four haplotype test, however, its coordinates are adjusted either its end coordinate is set to the furthest-downstream site at which it is known to exist, or its start coordinate is set to the furthest-upstream site at which it is known to exist. Nodes also can have recombination edges, which point to nodes with which they fail the four haplotype test, with paths through recombination nodes (Fig. 3.3). These edges are analogous to the edges described in the two-trees algorithm (Fig. 3.2). When a recombination event is solved, all nodes implied by the recombination event are created as tree nodes in the ARG (Fig. 3.3), with "solved" recombination edges describing the inferred recombination event, to avoid creating redundant recombination events in the future. Furthermore, when no possible γ node explaining a four haplotype test failure can exist (i.e. all three possible clades fail the four haplotype test with existing ARG nodes within the ranges over which they must exist), we add “unsolvable” recombination edges connecting the two nodes that fail the four haplotype test. These edges allow us to adjust start and end coordinates of the nodes without inferring the branch movement that separates them.

The propagation distance parameter p allows us to bin the ARG into regions 2*p bases wide, each of which undergoes a different process simultaneously. Denoting the coordinate (in base pairs) of the most recently-read site in the input file as c, ARG nodes whose range ends within the range [c - 2*p, c] are subject to gain new parents, children, and/or recombination edges from comparison with newly-read sites. ARG nodes whose range ends in the bin [c 4*p, c-2*p), however, can no longer be affected by new sites read from the input file and are thus candidates to have their ancestral recombination events solved via the ARG version of the two trees algorithm. ARG nodes whose ranges end in the bin [c 6*p, c-4*p) cannot share recombination edges with nodes in the bin [c 4*p, c-2*p) and thus are candidates to be written to disk as trees. Finally, nodes with ranges that end upstream of c 6*p cannot affect the topology of branch lengths of trees at sites in the bin [c 4*p, c-2*p) and thus can be deleted. We note that this scheme is designed primarily for the sake of memory and time efficiency and that it is not perfect; namely, any time a recombination event is solved,” new nodes can be created that will create new parent/child relationships and recombination edges outside of the bin [c 2*p, c]. It would be relatively straightforward to create a version of SARGE that keeps the entire ARG in memory. This would result in the inference of more ancestral recombination events and therefore potentially fewer polytomies in output trees at the cost of higher execution time and memory usage. For the sake of this study, however, we note that limiting the number of clades we infer by binning the ARG this way makes our inferences conservative and ensures that our results do not over-interpret the SNP data.

We determine branch lengths when writing trees. Since each tree is defined only at a single site, we determine a nodes branch length by counting the number of mutations it owns within the range defined by the edge to its parent at the current site. If this parent/child edge expands beyond the range $[s-p, s+p]$, where s is the current site and p is the propagation distance, we limit to mutations found only within that range. We then divide the number of mutations by the number of bases in the range over which they were collected. In the case where a parent/child edge is valid only at a single site, this will lead to the extremely large branch length of 1. To help compensate for this, when we load trees from an output file, we scale each branch length by dividing it by the total height of the tree, both above and below that branch length. This puts all branch lengths on a scale between 0 and 1. When all fixed differences between the ancestral sequence and the reference genome are included as sites that can contribute to the root branch length in the ARG (as in this study), these branch lengths can then be multiplied by two times the divergence time between the ancestral and reference genomes to get approximate (low resolution) branch lengths. We note that many clades in our ARG have branch lengths of zero, meaning that no mutations were observed on those lineages. We also note that the number of times a given node serves as a clade in an ancestral recombination event also provides a measure of age. Although we store these values, we do not use them when computing branch lengths in this study, since it is difficult to reconcile time measured using two different types of units (mutations and shared recombination events). Thus, clades inferred solely from ancestral recombination events will have branch lengths of zero.

3.4.4 Simulations

For the sake of assessing our and other ARG inference programs, we simulated sampling an increasing number of haplotypes from a single panmictic population with no history of growth or bottlenecks. We did this using msprime [82]. We chose a recombination rate of 1 centimorgan per megabase and a mutation rate of $1 * 10^{-9}$ per year with a 25-year generation time, giving a per-generation mutation rate of $2.5 * 10^{-8}$. Additionally, we chose a heterozygosity value of 10.1 per 10,000 bases, comparable to the rate in modern sub-Saharan Africans [165]. We simulated 1 megabase of sequence per run, running 5 replicates each of simulations with 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, and 5000 haplotypes. The complete command used was `mspms X 1 -t 1010.0 -r 404.0 1000000 --precision 6 T`, where X is the number of haplotypes. Whenever there were duplicate base positions in a simulated data set, we ignored the allele data at all but the first occurrence of each position. We then ran SARGE on each data set with a propagation distance of 25,000 bases, along with tsinfer [83] (converting its output to a sequence of trees linked to specific variable sites) and Rent+ [131] with the `t` option to infer branch lengths. For each inferred tree, we loaded the tree output by msprime for the same variable site and defined the percent of clades correct as the fraction of all clades in all inferred trees that existed as clades in the msprime tree at the same sites. Other metrics were straightforward to compute, including the Kendall-Coljin distance, which was calculated as described in [85].

3.4.5 Plotting tree articulation against mutation/recombination rate ratio

We binned the genome into 50kb blocks and measured tree articulation as the mean number of nodes per tree, across all trees within each window. We then measured the mutation rate by sampling the branch length of the root node of each ARG tree (this is the number of mutations separating all hominin lineages from the chimpanzee genome, collected over 2*(propagation distance) bases and reported in units of mutations per base). Assuming 6.5 mya for the hominin-chimpanzee split and a 25-year generation time, we transformed numbers of mutations into a per-site, per-generation mutation rate by dividing by 13,000,000 divided by 25 and taking means across windows. Finally, we took the mean recombination rate in cM/Mb from the sex-averaged Oxford map contained within Eagle2 [110] and converted it to Morgans per base, to get a value in the same units.

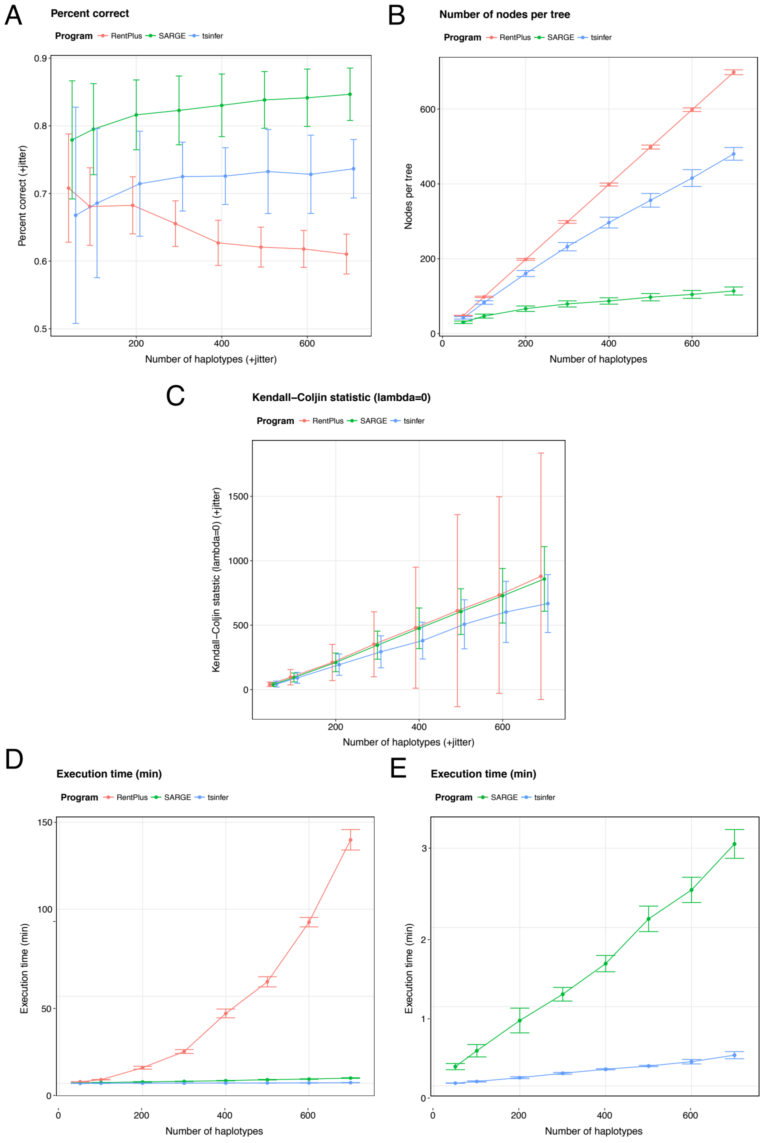


Figure 3.6: Comparisons of SARGE to two other ARG inference programs, using a simulated data with sub-Saharan African-like heterozosity, constant population size, and no structure. In each comparison, error bars represent one standard deviation across 5 replicates and ARGs were inferred across data sets with increasing numbers of haplotypes. A: percent of nodes correct, given the true ARG from the simulation. B: Number of nodes per tree. C: Kendall-Coljin distance [85] from true trees, without considering branch lengths. This metric was used in another study [83] and is negatively affected by polytomies. D: Execution time. Since tsinfer is a python module, its execution time does not include writing data to disk, while the other two execution times do. E: Same as D, but with Rent+ excluded.

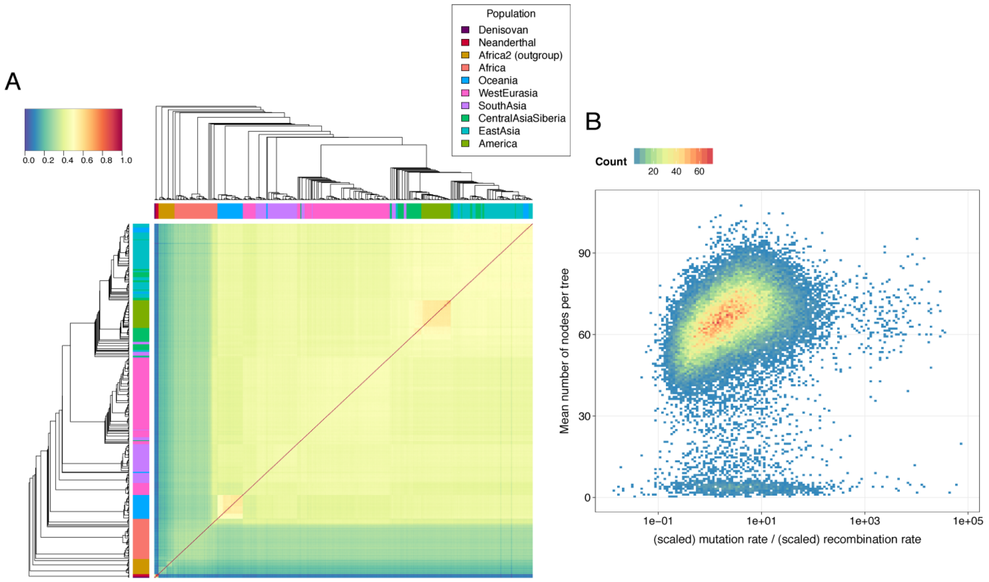


Figure 3.7: A: Genome-wide UPGMA trees from SNP data (top and bottom) against similarity matrix from shared recombination events inferred by SARGE. Dark red (similar) groups are Native Americans and Papuans. Outliers (dark lines) are S_Naxi haplotypes; another study reported likely improper phasing in this sample [83]. B: Number of nodes per tree against the ratio of mutation rate to recombination rate in 50kb windows.

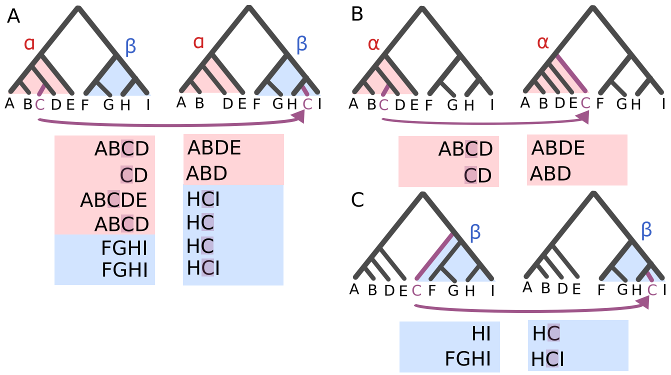


Figure 3.8: Different types of four haplotype test failures. In each, the γ clade is highlighted in purple, α in red, and β in blue. A: Lateral branch movement. Four haplotype test failures of type α/α, α/β, and β/β are observed. B: Upward branch movement. Only α/α four haplotype test failures are observed. C: Downward branch movement. Only β/β four haplotype test failures are observed.

Chapter 4

Insights into human history from ancestral recombination graph inference

In this section, I apply the algorithm described in the previous chapter to a panel of human genomes, along with several archaic hominin genomes. I use the results to describe the demography of admixture between humans and archaic hominins, and I also explore possible functional consequences of admixture. Finally, I describe regions of the human genome free of both admixture and incomplete lineage sorting with archaic hominins, which may contain some of the key genes involved in human speciation.

4.1 Background

Genomic studies have made clear that admixture between subpopulations and hybridization between species are common; this holds true in our own species [189]. One consequence of this is that species divergence must proceed in the face of periods of gene flow [3].

For this to happen, hybrids must have reduced fitness relative to non-hybrids. This can result from the accumulation of inter-species genetic incompatibilities; hybrids that inherit different species-specific alleles at these loci will have reduced fitness [224, 225]. Prior studies examining Neanderthal and Denisovan ancestry at specific loci in modern humans have seen evidence of this process. In general, genomic regions depleted for archaic ancestry in modern humans are enriched for genes [186, 214, 213, 187], and introgressed archaic alleles appear to be expressed at lower levels in the brain and testis [119]. Of all loci contributing to hybrid incompatibility, a subset should represent human-specific lineages that arose around the time of speciation. A catalog of these lineages would help illuminate not only where in the genome hybrid incompatibilities arose, but which specific mutations were important in distinguishing modern humans from our archaic relatives.

Although a variety of techniques have been used to map ancestry across genomes of hybrid individuals, ancestral recombination graph (ARG) inference provides the best opportunity to pinpoint the true boundaries of ancestry segments in the genome (rather than binning the genome into windows). Branch lengths in ARG trees also enable further study of genomic regions of interest, such as admixed regions. One could use lengths of haplotypes and branch lengths within them to pinpoint episodes of selection, for example.

We ran our ARG inference program, SARGE, on a set of 279 phased human genomes from around the world [116], with the addition of two high-coverage Neanderthal genomes [165, 163] and one high-coverage Denisovan genome [126]. We then used the results to create human haplotype-specific archaic ancestry maps and uncover instances of both adaptive introgression and likely selection against introgressed archaic lineages.

Group	TMRCA (%)	Uncorr. TMRCA	Min. TMRCA	Max. TMRCA
Denisovan	0.00511	66,500	137,000	157,000
Vindija33.19	0.00469	61,000	111,000	126,000
Altai	0.00498	64,700	185,000	195,000
Neanderthals	0.00985	128,000	213,000	226,000
Archaic hominins	0.029	377,000	455,000	471,000
Humans	0.0364	473,000	473,000	473,000
All hominins	0.0492	639,000	640,000	640,000

Table 4.1: Times to most recent common ancestor (TMRCA) of various given groups, averaged across all sites in the ARG. Values given in years assume 6.5 million years human-chimpanzee divergence (and thus 13 million years for mutations to accumulate). First TMRCA value is given as a percent of human-chimp divergence. Corrected values use approximate branch shortening values from [163] (minimum and maximum values given are based on minimum and maximum values in the paper). The TMRCA of all humans has only one value because there is no need to correct for branch shortening. Since more sites were included in this analysis than in previous studies focused on genome-wide statistics (i.e. no mappability filter was applied), archaic branch lengths might be slightly inflated from false singletons inferred from DNA damage.

4.2 Results

Having judged our results reliable (see Chapter 3), we first calculated the time to most recent common ancestor (TMRCA) of all groups within the ARG data set. We did this by taking the mean of the TMRCA of each group across all trees, genome-wide. Since these values have been estimated by others, this provided a simple test for the accuracy of our results. We find that the values we calculated agree with prior estimates (Table 4.1).

We used our ARG to produce haplotype-specific maps of archaic hominin ancestry in modern humans by first scanning for clades that grouped modern human haplotypes with one or more archaic hominin haplotypes, to the exclusion of other modern human haplotypes. To reduce false positives produced by large polytomies, we used an outgroup consisting of the most basal lineages of sub-Saharan Africans (Mbuti, Biaka, and Khomani-San) and required such clades to contain 10% or fewer outgroup haplotypes. We then disentangled admixture from incomplete lineage sorting (ILS) using a technique designed to minimize cross-population variance in the amount of incomplete lineage sorting (Appendix B, Fig. 4.1). This gave us genome-wide Neanderthal ancestry estimates close to, but lower than, those produced using an estimator based on the genome-wide D-statistic [56, 35] (Fig. 4.2A, Table 4.2). In the case of Denisovan ancestry, however, we underestimated the proportion in Oceanians by 2% relative to the D-statistic, with a relatively large amount (0.3%) detected in other populations (Fig. 4.2A, Table 4.2). One possible explanation for this is that many clades in which Oceanians have Denisovan-like ancestry also tend to include outgroup haplotypes. This could be because the outgroup haplotypes possess ancestry from other, as-yet-unsequenced archaic hominins (hereafter referred to as "super-archaic" ancestry) [62, 37], which are about as diverged from the sequenced Denisovan genome as the source of Denisovan-like ancestry in Papuans. Some of the haplotypes we call Denisovan might be of Neanderthal origin, due to incomplete lineage sorting and admixture among the archaic lineages [165].

The distribution of Neanderthal ancestry in modern humans largely agrees with prior studies. Considering only high-confidence ($p < 0.001$) Neanderthal-like haplotypes, the mean TMRCA to Neanderthal across each is consistent across populations, centered around 63 kya in all populations except those in Africa (Fig. 4.2B, Table 4.3, Fig. 4.3). This suggests that available Neanderthal genomes are good models for the introgressing Neanderthal(s) and that introgression probably took place around the time of the out-of-Africa migration. We also detect a number of Neanderthal-like haplotype blocks in sub-Saharan African populations, with the highest amount in Somali (0.8%) and the lowest in Mbuti (0.3%) genomes. TMRCAs to

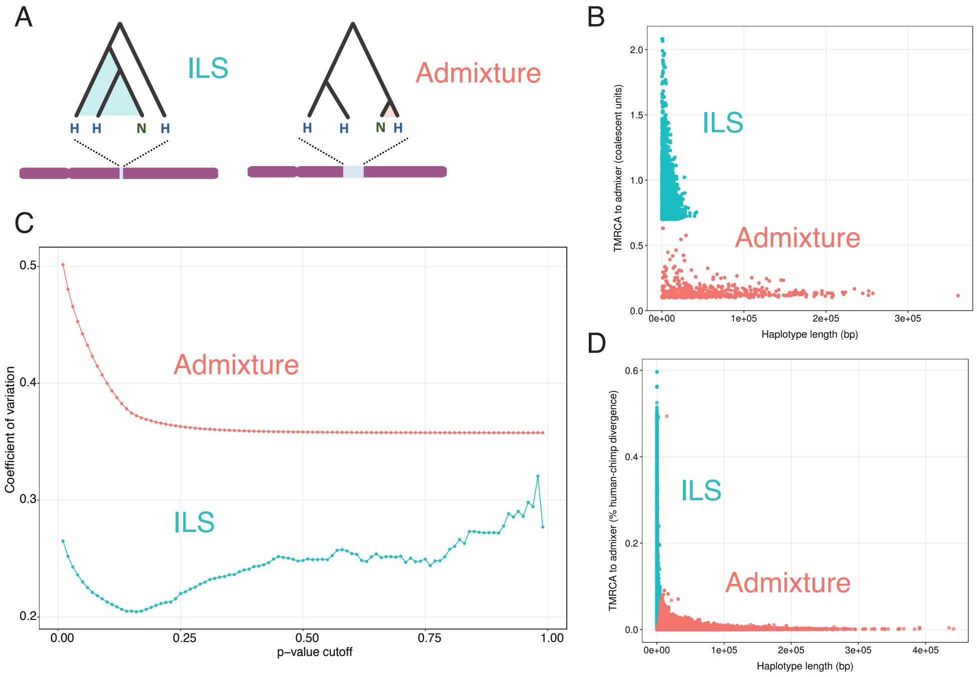


Figure 4.1: Separating admixed haplotypes from incomplete lineage sorting (ILS). A: cartoon of properties expected of ILS vs. admixed haplotypes. B: Using the true ARG from a simulation involving Neanderthal admixture into modern humans 50kya, ILS is separable from admixture by considering both the TMRCA to admixer and the length of the haplotype. C: In real data, these two distributions are not separable. We computed an admixture p-value for each haplotype and binned haplotypes into admixture and ILS based on varying p-value cutoffs, and for each bin plotted the coefficient of variation (standard deviation divided by mean) in the overall extent of ILS and admixture across SGDP populations. We expect the true coefficient of variation in ILS to be low across populations, so we chose a p-value cutoff that minimized this value (p = 0.16). D: Real data from the SGDP data set binned using the p-value determined in C.

Neanderthal appear to be nearly twice as old in the basal African lineages used as an outgroup as in non-African populations (113 kya), with TMRCAs to Neanderthal in the rest of sub-Saharan Africa (106 kya) intermediate between the two. This appears to be the result of two different components, one unique to Africa and fairly diverged from sequenced Neanderthal genomes, and the other shared predominantly with West Eurasian genomes and likely the result of ancient back-to-Africa migration [152]. This is further supported by our observation that the

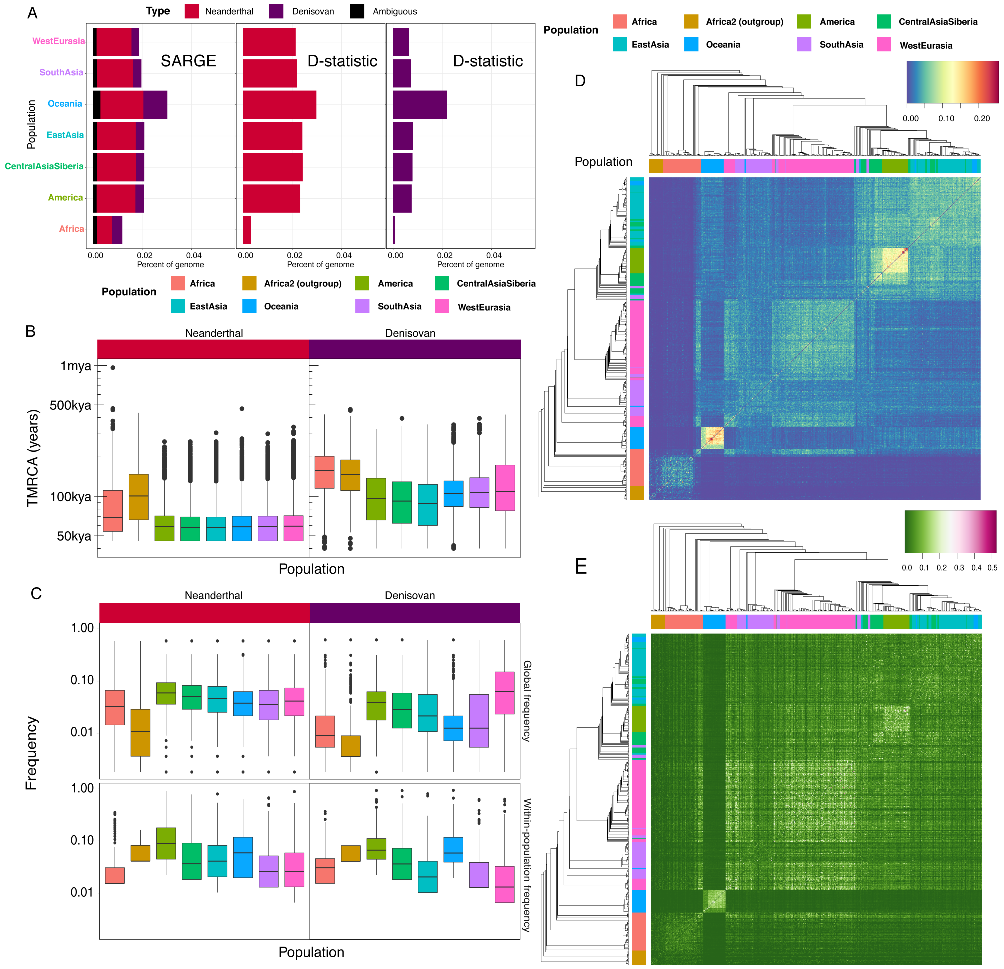


Figure 4.2: A: Genome-wide percent Neanderthal, Denisovan, and ambiguous (either Neanderthal or Denisovan) across SGDP populations, using the ARG and the D-statistic. D-statistic calculations considered only one archaic population at a time as introgressor and thus does not detect ambiguous ancestry and also might count some Denisovan ancestry as Neanderthal, and vice-versa. B: Times to most recent common ancestor of confidently introgressed ($p < 0.001$) segments across SGDP populations. Conversions assume 6.5 mya human-chimpanzee divergence time and branch shortening values from [163]. C: Mean frequency in all humans of each confidently introgressed ($p < 0.001$) segment across SGDP populations (top panel) and mean within-population frequency of each segment across SGDP populations (bottom panel). D: Sharing (Jaccard statistic) of Neanderthal-introgressed haplotypes (matrix). Haplotypes are ordered by a UPGMA tree using input SNP data (top and left). Populations are colored the same way as in B and C, and self-self comparisons are set to 0 similarity in order to not saturate the color scale. E: Sharing (Jaccard statistic) of Denisovan-introgressed haplotypes (matrix). Haplotypes are ordered by a UPGMA tree using input SNP data (top and left).

Population	Neanderthal	Denisovan	ambiguous	Nea. freq.	Den. freq.
Africa	0.62%	0.42%	0.16%	4.00%	4.40%
Africa2	0.30%	0.27%	0.09%	2.50%	1.80%
America	1.60%	0.34%	0.15%	6.90%	7.90%
CentralAsiaSiberia	1.60%	0.34%	0.16%	6.10%	7.30%
EastAsia	1.60%	0.35%	0.16%	5.80%	6.20%
Oceania	1.70%	1.00%	0.30%	4.90%	3.10%
SouthAsia	1.50%	0.34%	0.16%	5.00%	7.10%
WestEurasia	1.40%	0.30%	0.15%	5.40%	13%

Table 4.2: Demographic parameters of Neanderthal and Denisovan admixture from ARG inference. Genome-wide percents given are the percent of the genome assayed (excluding sex chromosomes) classified as Neanderthal or Denisovan origin (or one of the two; ambiguous column), using a score cutoff chosen to minimize the cross-population variance in ILS (Appendix B, Fig. 4.1). Frequencies are means of frequencies of individual admixed segments. Frequency numbers reported are calculated using only confidently admixed haplotypes ($p < 0.001$) and are the frequencies across all human haplotypes in the Simons Genome Diversity Project Panel. Africa2 is a population of the most basal African lineages (Mbuti, Biaka, and Khomani-San) which were used as an outgroup in which no more than 10% of haplotypes were allowed to be admixed for any individuals in the data set were allowed to be called admixed.

highest percent of Neanderthal haplotypes in non-basal sub-Saharan Africans are shared with West Eurasia (Fig. 4.7), that the highest percent of those in basal sub-Saharan Africans are shared with other sub-Saharan Africans (Fig. 4.8), and that the mean TMRCA to Neanderthal of African haplotypes shared with other populations is lower than that for haplotypes unique to Africa (Fig. 4.15). We note that gene flow between these basal and non-basal sub-Saharan African groups has already been documented [191], and therefore some of their Neanderthal-like haplotypes could be of Eurasian origin as well.

Next, we computed the frequency across all humans of each high-confidence ($p < 0.001$) Neanderthal-like haplotype block and mapped the means of these frequencies across human genomes in different parts of the world (Fig. 4.2C, Table 4.2, Fig. 4.4). We hereafter refer to this value as the “global frequency” of an introgressed haplotype. Because bottlenecks

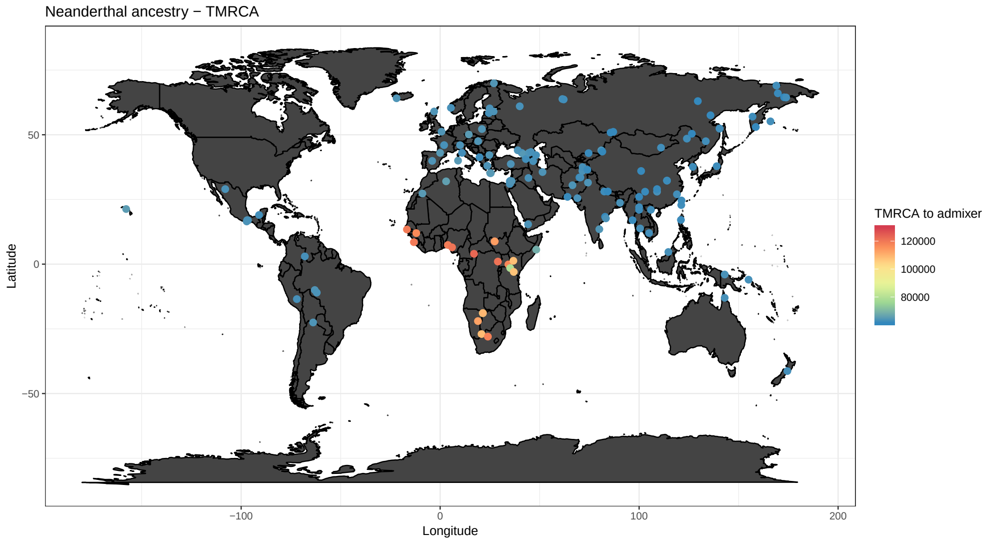


Figure 4.3: Worldwide distribution of times to most recent common ancestor (TMRCA) to the closest Neanderthal haplotype of Neanderthal-like haplotypes in modern humans. Haplotypes are from the confident set ($p < 0.001$), and points are averages across all haplotypes within all genomes from each location. Numbers are corrected for branch shortening, using the mean of the values given for the two Neanderthal genomes in [163].

from migration should increase these frequencies, we expected to see the lowest frequencies in places where introgression events took place. We observe low global frequencies in South Asia (Fig. 4.2C), with global frequency increasing in a gradient from South Asia through East Asia to the Americas (Fig. 4.4). A previous study similarly found Peruvians to carry the most high-frequency Neanderthal alleles of all studied populations [170]. We interpret this to mean that introgression likely took place near South or Southwest Asia, before most non-African populations had formed, with admixed individuals migrating to the east and undergoing successive bottlenecks. This is supported by the observation that sharing of Neanderthal haplotypes is relatively high within extant human populations, and especially so within the American and Oceanian populations, which have undergone multiple founder events (Fig. 4.2D). We also ob-

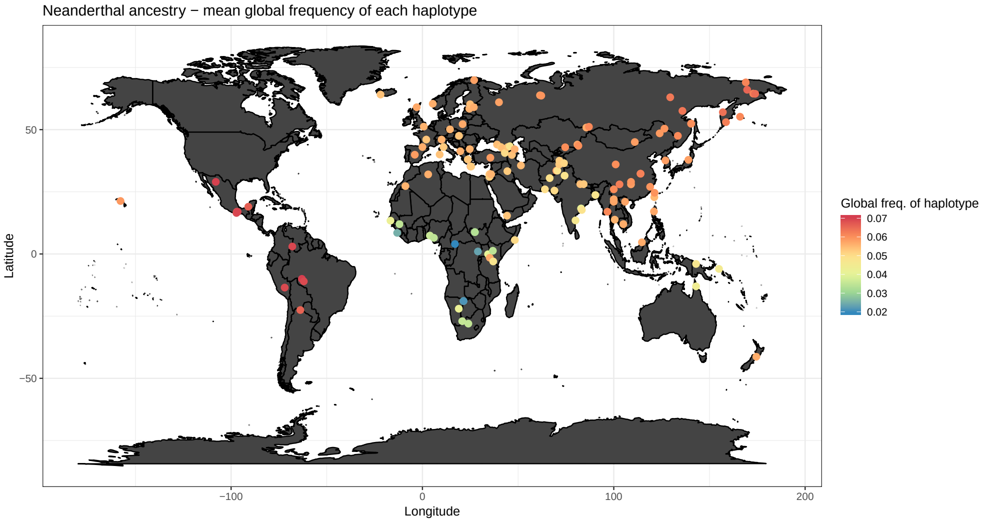

Neanderthal ancestry − mean global frequency of each haplotype
Latitude
50
0
−50
Longitude
−100
0
100
200
Global freq. of haplotype
0.07
0.06
0.05
0.04
0.03
0.02

Figure 4.4: Worldwide distribution of frequencies of individual Neanderthal-like haplotypes in modern humans. For each introgressed haplotype, its frequency in all humans worldwide was computed, and these values were averaged across all haplotypes within all human genomes from each geographic location.

serve low global frequency of Neanderthal-like haplotypes in sub-Saharan Africa; although we cannot rule out the possibility that these are segments of incomplete lineage sorting (ILS) that we falsely classified as admixture, it is possible that many of these haplotypes are the result of population-specific instances of super-archaic admixture. In Oceania, where we also see low global frequencies of Neanderthal-like segments, we suspect some Neanderthal-like haplotypes are actually a result of Denisovan-like introgression, which is largely specific to that population. This is supported by a slightly higher mean TMRCA to Neanderthal within Neanderthal-like haplotypes specific to Oceanians than within Neanderthal-like haplotypes shared between Oceanians and other populations (Fig. 4.15).

Denisovan-like ancestry segments appear more likely than Neanderthal-like segments to have multiple origins. As in other studies, we find most Denisovan-like haplotypes in Ocea-

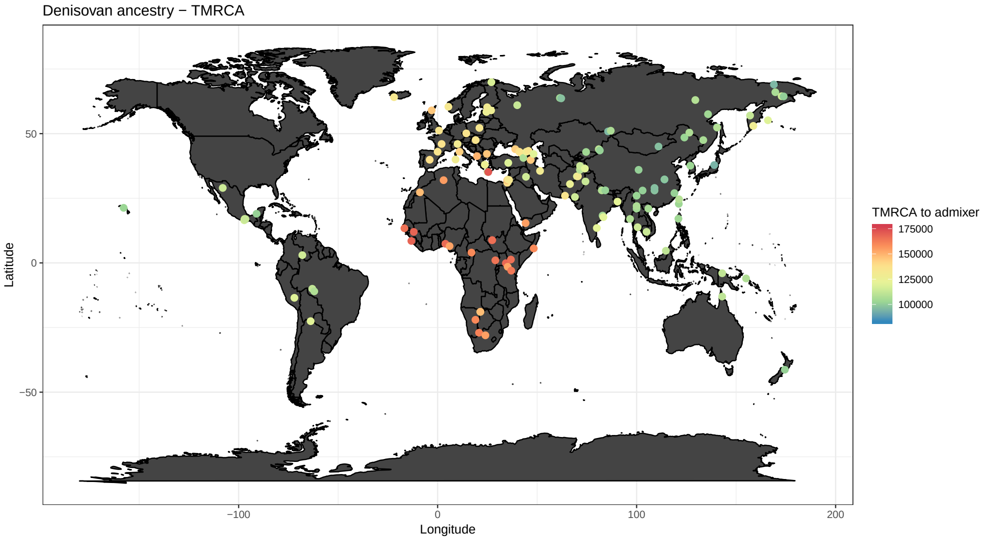


Figure 4.5: Worldwide distribution of times to most recent common ancestor (TMRCA) to the closest Denisovan haplotype of Denisovan-like haplotypes introgressed in modern humans. Haplotypes are from the confident set ($p < 0.001$), and points are averages across all haplotypes within all genomes from each location. Numbers are corrected for branch shortening, using the mean of the values given for the Denisovan genome in [163].

nia (1.0% average), primarily in Papuans (1.2% average), with most Oceanian Denisovan-like haplotypes unique to Oceanians (Fig. 4.12). We find that the mean TMRCA to Denisovan in Denisovan-like segments is much higher than for Neanderthal-like segments, however, with the highest value (in basal sub-Saharan African lineages) higher than the TMRCA of the two Denisovan haplotypes genome-wide (Fig. 4.2B, Table 4.3, Table 4.1). Additionally, admixture times estimated from haplotype lengths are less concordant with the TMRCAs to Denisovan than these two values computed for Neanderthal-like haplotypes (Table 4.3). We take this to mean that the available Denisovan genome is not a good model for the introgressor. We also observe that the distribution of TMRCAs to Denisovan in Denisovan-like haplotypes is less consistent across populations than for Neanderthal-like haplotypes (Fig. 4.2B). Notably, the

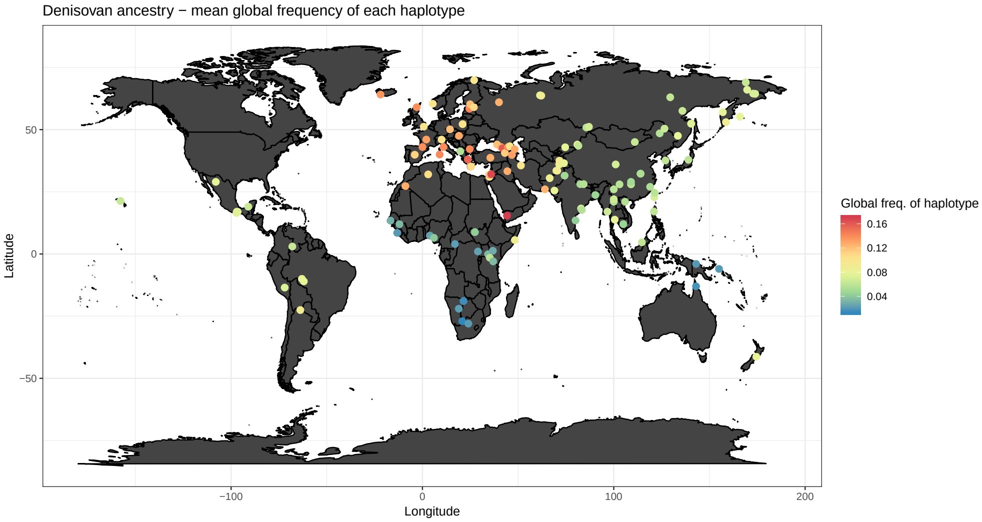

Denisovan ancestry − mean global frequency of each haplotype
Latitude
50
0
−50
Longitude
−100
0
100
200
Global freq. of haplotype
0.16
0.12
0.08
0.04

Figure 4.6: Worldwide distribution of frequencies of individual Denisovan-like haplotypes in modern humans. For each introgressed haplotype, its frequency in all humans worldwide was computed, and these values were averaged across all haplotypes within all human genomes from each geographic location.

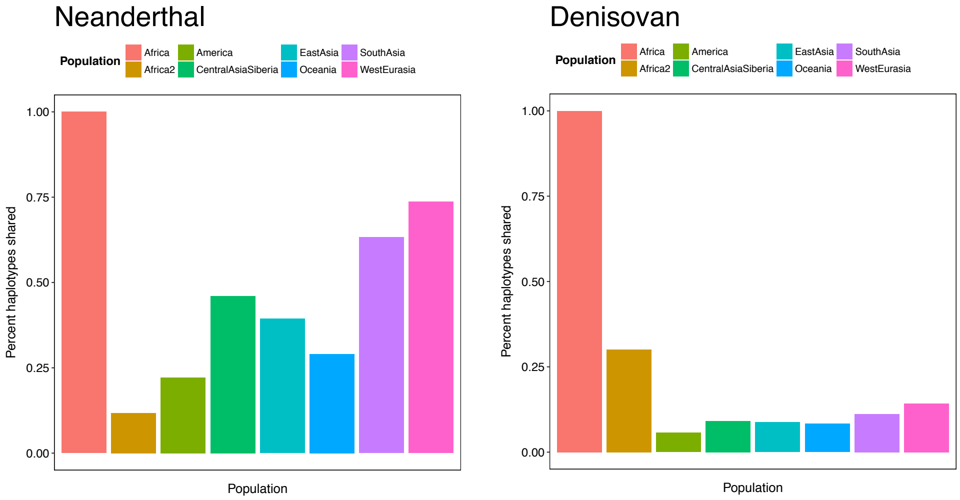


Figure 4.7: Percent of archaic-introgressed haplotypes in Africa (excluding Biaka, Mbuti, and Khomani-San) shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency.

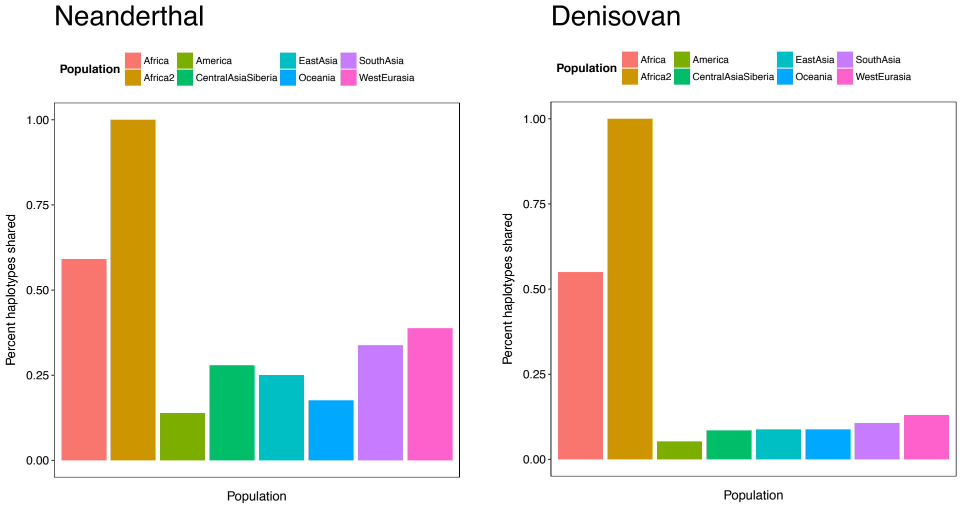


Figure 4.8: Percent of archaic-introgressed haplotypes in basal African lineages used as an outgroup (Biaka, Mbuti, and Khomani-San) shared with other SGDP populations.

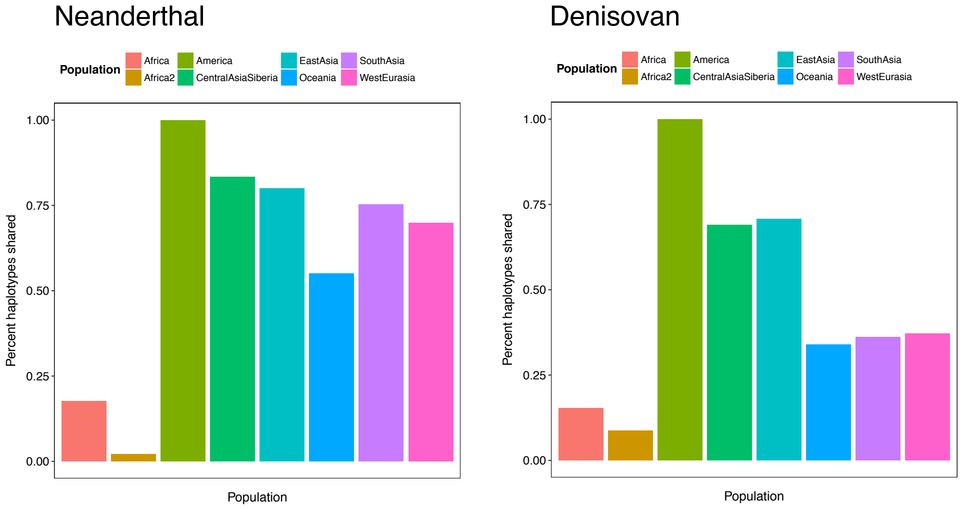


Figure 4.9: Percent of archaic-introgressed haplotypes in America shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency.

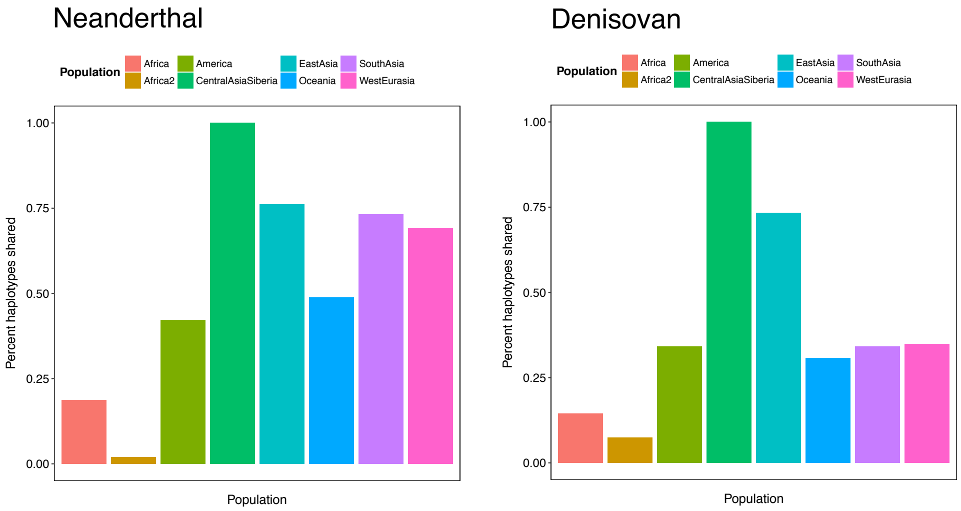


Figure 4.10: Percent of archaic-introgressed haplotypes in CentralAsiaSiberia shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency.

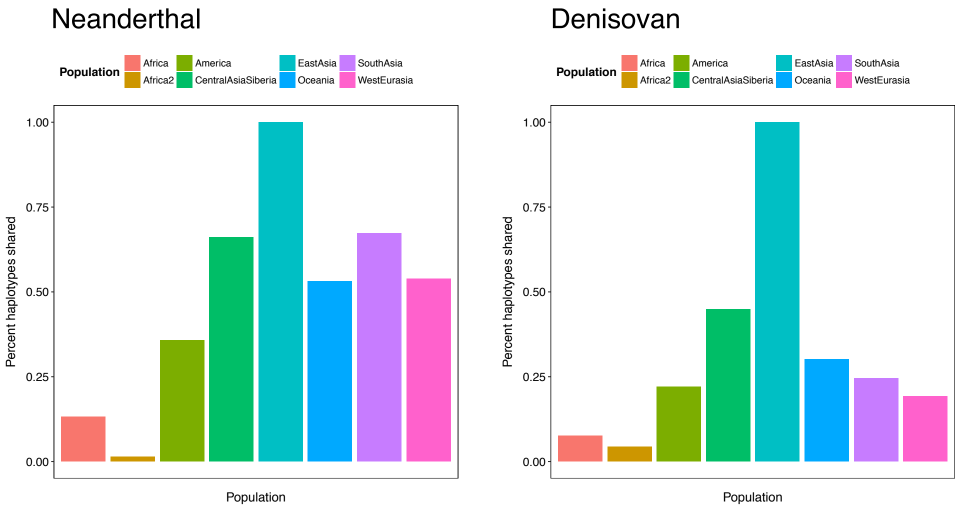


Figure 4.11: Percent of archaic-introgressed haplotypes in EastAsia shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency.

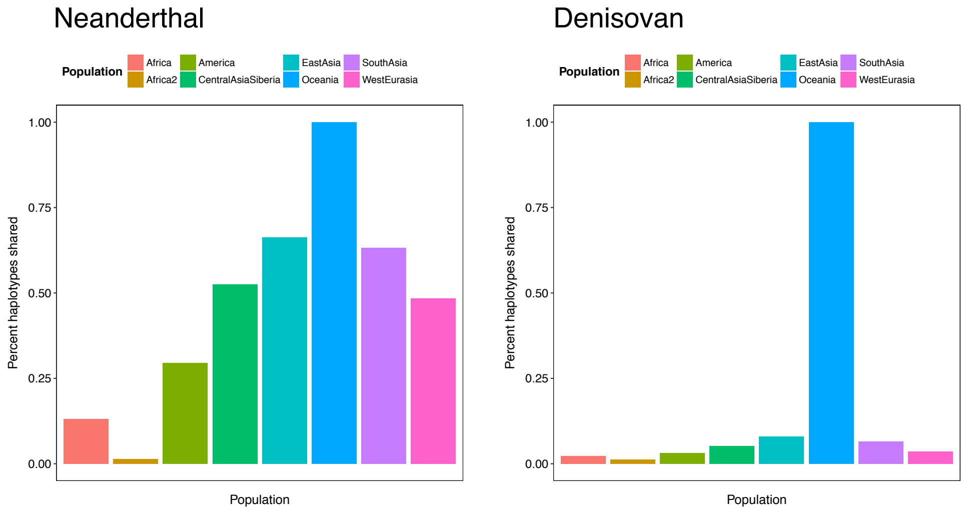


Figure 4.12: Percent of archaic-introgressed haplotypes in Oceania shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $<10\%$ frequency.

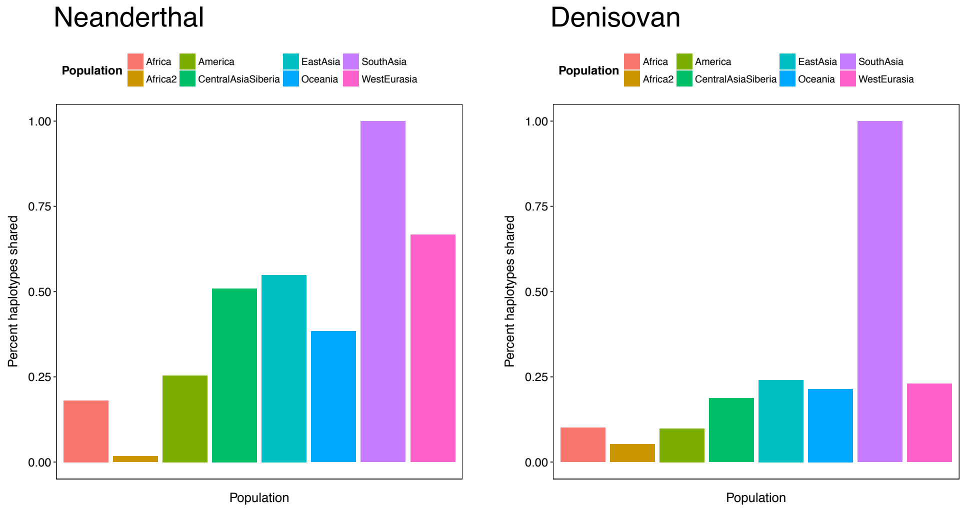


Figure 4.13: Percent of archaic-introgressed haplotypes in SouthAsia shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $<10\%$ frequency.

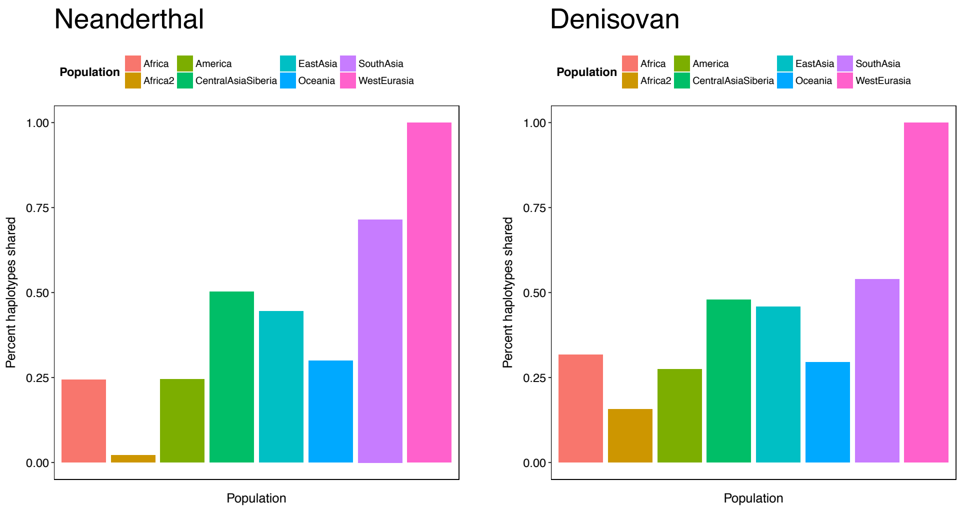


Figure 4.14: Percent of archaic-introgressed haplotypes in WestEurasia shared with other SGDP populations. Africa2 consists of most basal sub-Saharan African lineages (Mbuti, Biaka, and Khomani-San) used as an outgroup in which all introgressed haplotypes were required to be $< 10\%$ frequency.

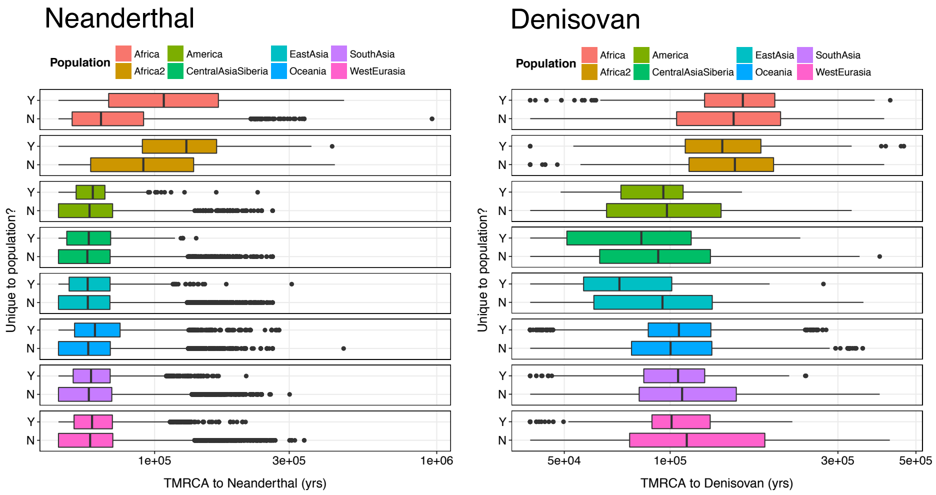


Figure 4.15: TMRCAs to Neanderthal for confident ($p < 0.001$) Neanderthal-like haplotypes (left panel) and TMRCAs to Denisovan for confident ($p < 0.001$) Denisovan-like haplotypes (right panel), corrected for branch shortening using values in [163] and plotted by whether unique to a specific SGDP population or shared among multiple populations.

Pop.	Nea. TMRCA	Den. TMRCA	Nea. date	Den. date
Africa	106 kya	161 kya	72 kya	82 kya
Africa2	113 kya	157 kya	80 kya	84 kya
America	63 kya	113 kya	68 kya	73 kya
CentralAsiaSiberia	62 kya	106 kya	67 kya	72 kya
EastAsia	62 kya	104 kya	67 kya	70 kya
Oceania	63 kya	109 kya	67 kya	72 kya
SouthAsia	63 kya	119 kya	64 kya	75 kya
WestEurasia	64 kya	133 kya	65 kya	83 kya

Table 4.3: Times to most recent common ancestor (TMRCAs) with admixers for confidently introgressed ($p < 0.001$) Neanderthal and Denisovan-like segments, as well as dates estimated using haplotype block lengths and assuming neutral evolution, with a 1 cM/Mb recombination rate and 25 year generation time. Since the TMRCA to admixer and haplotype length were both used to determine the scores of admixed segments and separate ILS from admixture, these values are affected by the score cutoff chosen ($p < 0.001$) and should be treated cautiously. Comparisons between populations should still be valid.

mean TMRCA to the Denisovan genome in Denisovan-like haplotypes is lower in genomes from East Asia, Central Asia, and the Americas than it is in Oceania. The source of these lower-TMRCA haplotypes appears to be East Asia: East Asia has more unique Denisovan-like haplotypes than America or Central Asia (Fig. 4.9, Fig. 4.10, Fig. 4.11) and the mean TMRCA of Denisovan-like haplotypes unique to East Asia is notably lower than that of haplotypes shared among multiple populations (Fig. 4.15; 82 kya vs. 108 kya). Although sampling bias caused by the low overall number of Denisovan haplotypes could affect our results, these observations contradict the model of Denisovan admixture into Oceanians, followed by migration of admixed individuals to mainland Asia [165, 166]. Our results agree, however, with a recent study that inferred two pulses of Denisovan-like ancestry into modern humans, with one component specific to mainland Asia and more closely related to the Denisovan genome than a second component specific to South Asia and Papua [18]. Additionally, another prior study has detected a potential ancient ancestry component of unknown origin in East Asians [132]; this may be related.

The mean global frequencies of individual Denisovan-like haplotypes are mostly concordant with the model of Denisovan gene flow into Papuans, followed by subsequent migration: we see low frequencies of Denisovan-like haplotypes in Oceania, with a gradient of increasing frequencies moving from East Asia to Central Asia and the Americas. We also see very low frequencies for Denisovan-like haplotypes in sub-Saharan Africans, as with Neanderthal-like haplotypes (Fig. 4.2C, Fig. 4.6). Furthermore, whereas population-specific frequencies of Neanderthal-like haplotypes are fairly consistent, population-specific frequencies of Denisovan-like haplotypes in mainland Asia are lower than in Oceania (Fig. 4.2C), suggesting that those populations may have already been established when gene flow happened (either by admixture with archaic introgressors or with a human population already carrying archaic ancestry). This is further supported by the observation that mainland Asian and South Asian genomes share Denisovan-like haplotypes much less than they share Neanderthal-like haplotypes (Fig. 4.2D-E). The fact that contradictory inferences can be drawn from analyses of the TMRCA to the Denisovan genome and the global frequencies of Denisovan-like haplotypes suggests that the Denisovan-like haplotypes we detect might stem from multiple introgression events with different archaic hominins, all of which are somewhat, but not closely, related to the sequenced Denisovan genome. Our observation of a small number of Denisovan-like haplotypes in West Eurasia, highly shared within that population, and with high global frequency (Fig. 4.2C,E, Fig. 4.6) also presents a mystery and might be the result of incompletely sorted haplotypes mislabeled as admixture.

We tested our maps of introgressed archaic hominin ancestry for overlap with various genomic features, as well as with each other. We find that our Neanderthal and Denisovan ancestry maps significantly overlap each other (Table 4.4), possibly due to previously-documented gene flow between Neanderthals and Denisovans [126]. We find depletion of Neanderthal ancestry at regulatory element binding sites, concordant with the observation that Neanderthal-introgressed alleles are preferentially downregulated in humans [119]. Otherwise, we do not find significant positional correlation or overlap between introgressed regions and genes or exons (Table 4.4), suggesting that much of the introgressed sequence might be randomly distributed. We do find nearly-significant enrichment of exons in Neanderthal-introgressed sequence, however, suggesting perhaps some cases of adaptive introgression. To locate the most important cases of adaptive introgression, we sorted our confidently-called Neanderthal-like and Denisovan-like haplotypes in decreasing order of length times global frequency. In order to minimize the effect of random choices made in ARG inference, we removed any haplotype that did not intersect a haplotype detected using an ARG inferred over a randomly chosen subset of 50 human haplotypes and all archaic hominin haplotypes. Because most long Denisovan-like haplotypes have low global frequency, most outlier Denisovan-like haplotypes are short and high-frequency (Fig. 4.16). Of the top five Neanderthal and top five Denisovan-like outlier haplotypes, three contain transmembrane proteins in the TMEM family, a group with many uncharacterized members which has also turned up in prior studies of archaic introgression [171]. The top Neanderthal haplotype contains TMEM236, a transmembrane protein of unknown function and MRC1L1 (MRC1), a mannose receptor that plays a role in the uptake of HIV-1 particles by macrophages [204]. Other genes in the top five outlier Neanderthal-like

Map	Features	Dist. P	Proj. P
Nea.	genes	0.884	0.745
Nea.	exons	0.229	0.977
Nea.	reg elt.	0.461	0.99736*
Nea.	Den.	6.01E-03*	0*
Den.	genes	0.682	2.51E-02
Den.	exons	0.721	0.676
Den.	reg elt.	0.781	0.782
Den.	Nea.	0*	0 *

Table 4.4: Overlap of confidently-called ($p < 0.001$) Neanderthal and Denisovan haplotypes with various genomic features. Genes are whole protein coding genes from Gencode [49], using Ensembl version 94 on human genome version GRCh38 lifted over to GRCh37 coordinates. Exons are for protein-coding genes from the same annotation. Regulatory element binding sites are from the filtered double-elite" set in the GeneHancer database [48], obtained from the UCSC Genome Browsers Table Browser utility [21]. Distance-based p-values are from the "relative distance" Kolmogorov-Smirnov test and projection p-values measure overlap, both implemented in the GenometricCorr R package [46]. Significant ($p < 0.01$ or $p > 0.99$) values are marked with asterisks.

haplotypes include ZNF605 and ZNF26, both C2H2-like zinc finger proteins that serve as transcription factors, BAZ2B, a transcription factor subunit [139], PPP2R5A, a component of a protein complex that binds kinetochores and plays a role in the control of mitotic cell division [158], and TMEM206, another transmembrane protein of unknown function. The top outlier Denisovan-like haplotype contains the uncharacterized transmembrane protein TMEM248, and other outlier Denisovan-like haplotypes include CNGA1, a cyclic GMP-activated cation channel involved in phototransduction by rod cells in the retina [228], HSD3B2, an enzyme involved in catalysis of steroid hormones [183], and FAHD2B, a relatively uncharacterized gene that with possible hydrolase activity.

To uncover biological processes that may have been acted upon by selection on introgressed variants, we performed a Wilcoxon rank-sum test on Gene Ontology terms [1] an-

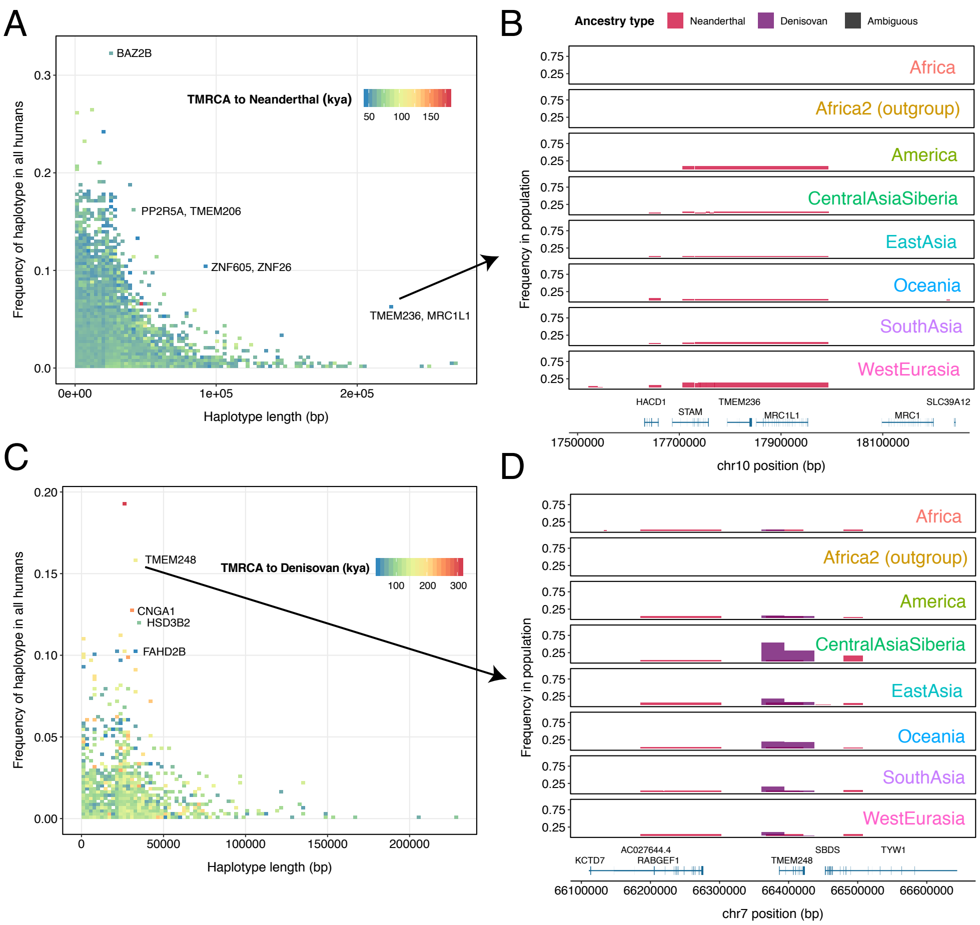


Figure 4.16: A: Confident ($p < 0.001$) Neanderthal-like haplotypes that were also detected using an ARG inferred over a subset of the samples. The top five outliers (ordered by length times frequency in all humans) are marked with the genes they contain (one contained no genes). B: The top outlier Neanderthal haplotype (by length times frequency), showing 250kb upstream and downstream and genes contained within. C: Confident ($p < 0.001$) Denisovan-like haplotypes that were also detected using an ARG inferred over a subset of the samples. The top five outliers (ordered by length times frequency in all humans) are marked with the genes they contain (one contained no genes). B: The top outlier Denisovan haplotype (by length times frequency), showing 250kb upstream and downstream and genes contained within. This region also coincides with a Neanderthal haplotype detected in some populations.

notating genes that overlapped introgressed haplotypes, ranking each term by the frequency in modern humans of its parent introgressed haplotype. Although this test did not consider specific variants and thus is prone to false positives, we detected terms related to G-protein coupled receptor signaling, keratin, metal ion homeostasis, and immune processes for both types of introgressed ancestry (Table 4.5, Table 4.6). In order to hone in on possible effects of specific introgressed variants, we also compiled all single-nucleotide variants tagging these introgressed haplotypes and intersected them with a catalog of significant GWAS hits from many studies [115] (Table 4.7). As GWAS studies are focused on medically and socially relevant traits, this is not an unbiased test, but it provides insight into phenotypic relevance of introgressed variants. We find high-frequency ($> 1\%$) Neanderthal variants implicated in traits related to immune dysfunction as well as nutrient levels, which agrees with the prevailing narrative that many adaptively introgressed Neanderthal variants largely had to do with dietary and immune adaptation to the Eurasian environment [171] but may be maladaptive to modern lifestyles [197]; many of our GWAS hits are the same as those found in other studies [187]. In the case of introgressed Denisovan variants, we do not find any significant GWAS hits at greater than 1% frequency in modern humans.

p-value	GO ID	GO term
4.19E-14	GO:0050911	detection of chemical stimulus involved in sensory perception of smell
7.49E-08	GO:0007186	G protein-coupled receptor signaling pathway
6.90E-05	GO:0060333	interferon-gamma-mediated signaling pathway
8.94E-05	GO:0061844	antimicrobial humoral immune response mediated by antimicrobial peptide
0.000146995	GO:0031424	keratinization

0.000167407	GO:0071280	cellular response to copper ion
0.000203636	GO:0007565	female pregnancy
0.000227352	GO:0010469	regulation of signaling receptor activity
0.000321707	GO:0006954	inflammatory response
0.000378327	GO:0071222	cellular response to lipopolysaccharide
0.000574185	GO:0070268	cornification
0.000739217	GO:0051281	positive regulation of release of sequestered calcium ion into cytosol
0.00123207	GO:0006882	cellular zinc ion homeostasis
0.00124776	GO:0002684	positive regulation of immune system process
0.0014972	GO:0006397	mRNA processing
0.00153823	GO:0048247	lymphocyte chemotaxis
0.00206373	GO:0071346	cellular response to interferon-gamma
0.00209769	GO:0010043	response to zinc ion
0.00210984	GO:0010273	detoxification of copper ion
0.00210984	GO:0051238	sequestering of metal ion
0.00213432	GO:0042742	defense response to bacterium
0.002301	GO:0006357	regulation of transcription by RNA polymerase II
0.00235922	GO:0006376	mRNA splice site selection
0.00239173	GO:0030593	neutrophil chemotaxis

0.00257379	GO:0046686	response to cadmium ion
0.00300734	GO:0002475	antigen processing and presentation via MHC class Ib
0.00363107	GO:0071356	cellular response to tumor necrosis factor
0.00410772	GO:0031638	zymogen activation
0.00424668	GO:0052697	xenobiotic glucuronidation
0.00425411	GO:0032823	regulation of natural killer cell differentiation
0.00483847	GO:0032495	response to muramyl dipeptide
0.00488235	GO:0050832	defense response to fungus
0.00500499	GO:0071549	cellular response to dexamethasone stimulus
0.00501819	GO:0043392	negative regulation of DNA binding
0.00505791	GO:0071347	cellular response to interleukin-1
0.00515977	GO:0051187	cofactor catabolic process
0.00516876	GO:0014003	oligodendrocyte development
0.00523049	GO:0032689	negative regulation of interferon-gamma production
0.00523293	GO:0032570	response to progesterone
0.00551232	GO:0007597	blood coagulation, intrinsic pathway
0.00601035	GO:0050776	regulation of immune response
0.00604952	GO:0006054	N-acetylneuraminate metabolic process
0.00604952	GO:0032966	negative regulation of collagen biosynthetic process
0.00604952	GO:0046007	negative regulation of activated T cell proliferation

0.00604952	GO:1904526	regulation of microtubule binding
0.0061107	GO:0071242	cellular response to ammonium ion
0.00636598	GO:0009749	response to glucose
0.00662589	GO:0031349	positive regulation of defense response
0.0068822	GO:1903707	negative regulation of hemopoiesis
0.00691491	GO:0008380	RNA splicing
0.00710093	GO:0001580	detection of chemical stimulus involved in sensory perception of bitter taste
0.00710981	GO:0098581	detection of external biotic stimulus
0.00731573	GO:0045907	positive regulation of vasoconstriction
0.00732096	GO:0009264	deoxyribonucleotide catabolic process
0.00755721	GO:0002467	germinal center formation
0.00772887	GO:0045687	positive regulation of glial cell differentiation
0.00775419	GO:0042772	DNA damage response, signal transduction resulting in transcription
0.00847483	GO:0045109	intermediate filament organization
0.00856996	GO:0003207	cardiac chamber formation
0.0086391	GO:0019221	cytokine-mediated signaling pathway
0.00864698	GO:0045869	negative regulation of single stranded viral RNA replication via double stranded DNA intermediate

0.00864698	GO:0052696	flavonoid glucuronidation
0.00864698	GO:0061004	pattern specification involved in kidney development
0.00880768	GO:0031640	killing of cells of other organism
0.00921992	GO:0044060	regulation of endocrine process
0.00945952	GO:0031343	positive regulation of cell killing
0.00948689	GO:0051412	response to corticosterone
0.00956698	GO:0071294	cellular response to zinc ion
0.00973362	GO:0021515	cell differentiation in spinal cord
0.00978883	GO:0010888	negative regulation of lipid storage
0.00985606	GO:0017001	antibiotic catabolic process
0.00997497	GO:0002637	regulation of immunoglobulin production

Table 4.5: Significantly enriched biological_process Gene Ontology (GO) terms, via a Wilcoxon rank-order test on genes overlapping high-confidence ($p < 0.001$) Neanderthal-introgressed haplotypes in modern humans, ranked by the frequency (in all humans) of the introgressed haplotype.

!

p-value	GO ID	GO term
8.31E-03	GO:0000245	spliceosomal complex assembly
1.82E-04	GO:0007186	G protein-coupled receptor signaling pathway
3.99E-03	GO:0007187	G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger
9.95E-03	GO:0007631	feeding behavior
0.00109767	GO:0009952	anterior/posterior pattern specification
0.00983203	GO:0010043	response to zinc ion
0.00149728	GO:0010469	regulation of signaling receptor activity
0.0091891	GO:0010677	negative regulation of cellular carbohydrate metabolic process
0.00641595	GO:0010830	regulation of myotube differentiation
0.00843255	GO:0019731	antibacterial humoral response
0.00100468	GO:0031424	keratinization
0.00680734	GO:0032355	response to estradiol
0.00981516	GO:0032677	regulation of interleukin-8 production
0.00789479	GO:0035821	modification of morphology or physiology of other organism
0.00876491	GO:0046916	cellular transition metal ion homeostasis
0.00816657	GO:0048706	embryonic skeletal system development
8.16E-09	GO:0050911	detection of chemical stimulus involved in sensory perception of smell
0.000818869	GO:0051179	localization
0.0091891	GO:0055069	zinc ion homeostasis
0.00545075	GO:0060337	type I interferon signaling pathway
0.00787755	GO:0061844	antimicrobial humoral immune response mediated by antimicrobial peptide
0.00334059	GO:0070268	cornification
0.00365757	GO:0072676	lymphocyte migration
0.0073354	GO:0098542	defense response to other organism
0.00311815	GO:0140053	mitochondrial gene expression

Table 4.6: Significantly enriched biological_process Gene Ontology (GO) terms, via a Wilcoxon rank-order test on genes overlapping high-confidence ($p < 0.001$) Denisovan-introgressed haplotypes in modern humans, ranked by the frequency (in all humans) of the introgressed haplotype.

chr	Position (hg38)	Allele	Frequency (modern humans)	Trait	Ref
7	129041008	T	0.102150538	Sjögren's syndrome	[100]
7	129043485	A	0.102150538	Primary biliary cirrhosis	[107]
7	129044262	C	0.102150538	Systemic sclerosis	[118]
6	121409576	A	0.086021505	Heart rate	[34]
12	40346421	C	0.082437276	Crohn's disease	[31]
12	40346421	C	0.082437276	Inflammatory bowel disease	[31]
19	32962753	T	0.082437276	Creatinine levels	[22]
2	118406151	C	0.077060932	Bone ultrasound measurement (broadband ultrasound attenuation)	[133]
2	69679438	C	0.073476703	Adolescent idiopathic scoliosis	[106]
3	154337010	T	0.069892473	Coronary artery disease	[212]
2	222184302	A	0.066308244	Vitamin D levels	[8]
1	92543881	C	0.064516129	Cholesterol, total	[210]
1	92543881	C	0.064516129	Cholesterol, total	[221]
3	119498262	C	0.044802867	Blood protein levels	[42]

1	150351808	T	0.043010753	Autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia (combined)	[28]
2	156055088	T	0.041218638	Cognitive performance (MTAG)	[99]
2	156055088	T	0.041218638	Cognitive performance	[99]
2	156055088	T	0.041218638	Intelligence (MTAG)	[71]
9	30915161	A	0.037634409	Post bronchodilator FEV1/FVC ratio	[113]
9	30916632	A	0.037634409	Post bronchodilator FEV1/FVC ratio	[113]
9	30926855	A	0.037634409	Post bronchodilator FEV1/FVC ratio	[113]
9	30928041	T	0.037634409	Post bronchodilator FEV1/FVC ratio	[113]
9	30934831	A	0.037634409	Post bronchodilator FEV1/FVC ratio	[113]

9	30955026	G	0.037634409	Post bronchodilator FEV1/FVC ratio	[113]
9	30957907	G	0.037634409	Post bronchodilator FEV1/FVC ratio	[113]
9	34656482	A	0.037634409	Blood protein levels	[205]
10	64725315	C	0.037634409	Central corneal thickness	[78]
10	94991513	C	0.037634409	Dehydroepiandrosterone sulphate levels	[230]
2	81441684	C	0.034050179	Aging traits	[112]
2	118563482	C	0.034050179	Erosive tooth wear (severe vs non-severe)	[5]
7	95826533	T	0.032258065	Dementia and core Alzheimer's disease neuropathologic changes	[11]
7	95826533	T	0.032258065	Dementia and core Alzheimer's disease neuropathologic changes	[11]
1	207843629	A	0.03046595	Reaction time	[30]
2	118078265	A	0.03046595	Cholesterol, total	[221]
2	118078265	A	0.03046595	LDL cholesterol	[221]

11	85186530	A	0.03046595	Self-reported math ability (MTAG)	[99]
9	89001928	A	0.028673835	Monocyte percentage of white cells	[9]
13	74421523	G	0.02688172	Immune reponse to smallpox (secreted IFN-alpha)	[86]
11	116568134	A	0.025089606	Clozapine-induced cytotoxicity	[32]
2	156270452	G	0.023297491	Menarche (age at onset)	[150]
11	63147874	G	0.023297491	Sex hormone levels	[185]
13	38041119	C	0.023297491	Alanine aminotransferase (ALT) levels after remission induction therapy in actute lymphoblastic leukemia (ALL)	[109]
4	105887024	C	0.021505376	Post bronchodilator FEV1	[113]
4	122478336	G	0.021505376	Rheumatoid arthritis	[138]
4	122532956	A	0.021505376	Allergic disease (asthma, hay fever or eczema)	[47]
1	159566423	A	0.019713262	Blood protein levels	[205]

2	156320222	T	0.019713262	Highest math class taken (MTAG)	[99]
1	149937602	A	0.017921147	HDL cholesterol	[91]
10	6749072	G	0.017921147	Hip circumference (psychosocial stress interaction)	[198]
11	25271868	G	0.017921147	Peripheral arterial disease (traffic-related air pollution interaction)	[220]
11	103145464	A	0.017921147	Diisocyanate-induced asthma	[229]
17	39897636	C	0.016129032	Subcutaneous adipose tissue	[26]
4	161890417	T	0.014336918	Risky sexual behaviors (alcohol dependence interaction)	[156]
11	103097123	C	0.014336918	Interleukin-10 levels	[209]
12	71105594	G	0.014336918	Adolescent idiopathic scoliosis	[106]
17	1672578	T	0.014336918	Blood protein levels	[42]
20	59644708	T	0.014336918	General cognitive ability	[30]

7	101642897	G	0.012544803	Response to serotonin reuptake inhibitors in major depressive disorder (plasma drug and metabolite levels)	[80]

Table 4.7: Significant GWAS hits that coincide with Neanderthal-introgressed variants in high-confidence ($p < 0.001$) haplotypes, sorted by decreasing Neanderthal haplotype frequency. Only high-frequency ($> 1\%$) variants are shown; there were no such variants in Denisovan-introgressed haplotypes.

In addition to cases of adaptive introgression, genomic regions free of, and possibly resistant to the incursion of, archaic hominin ancestry have been previously studied [186]. We expand upon this idea of archaic hominin ancestry “deserts” by searching for regions devoid of both archaic admixture and incomplete lineage sorting. “Deserts,” defined this way, denote regions of the genome in which modern humans comprise a distinct lineage from all other archaic hominins; they therefore should contain the alleles responsible for uniquely human phenotypic traits. We find that only about 10% of the autosomal genome lacks lineages that group together any modern human haplotypes with archaic hominins (henceforth called archaic hominin deserts) and 1.5% of the autosomal genome has a history in which all modern humans form a single clade (henceforth called human-specific regions). We are confident that we have sampled enough human genomes to find the correct extent of the deserts: an ARG inferred on a random subsample of 100 human haplotypes, with all archaic sequences, labeled approximately the same proportion of the genome as belonging to deserts and human-specific regions (Fig. 4.17A). In comparison, a (neutral) coalescent simulation with human demographic parameters inferred from data and a single pulse of Neanderthal admixture 50 kya (Appendix B) produced deserts over 44% and human-specific lineages over 42% of the genome (Fig. 4.17B). We note that although the percent of the genome in which we find archaic hominin admixture (60%) is higher than in our simulation, it slightly lower than what has been previously reported (70%; [214]). Additionally, the simulation did not include widely-hypothesized selection against weakly deleterious archaic alleles, which would increase the size of deserts in real relative to simulated data; this is the opposite of what we observe. One explanation for this discrepancy is further admixture with other as-yet-unknown archaic hominin lineages reducing the size of deserts and increasing the amount of the genome labeled as admixture and incomplete lineage sorting.

Human-specific regions appear to at least partly be the result of natural selection. Compared to deserts, where a human-specific mutation does not necessarily exist, human-specific regions are on average 275% larger (Fig. 4.17C). Additionally, the mean TMRCA of all humans across individual human-specific regions is much more variable than across individual deserts (Fig. 4.17D), suggesting cases of purifying selection (lower tail of the distribution) as well as neutral evolution or positive selection long ago in the past (middle and upper tail of the distribution). As deserts represent the widest possible span of true regions devoid of ILS and admixture (there could be unobserved clades marking ILS or admixture in these regions), testing them for enrichment or depletion of genomic features is a conservative test. Nonetheless, we find that positions of deserts are tightly correlated with positions of both exons and regulatory element binding sites, and that deserts tend to overlap both genes and regulatory element binding sites (Table 4.8). Human-specific regions, which represent very confident deserts of ILS and admixture, are correlated with positions of both genes and exons, and enriched for overlap with genes, exons, and regulatory element binding sites, but are not position-correlated with regulatory element binding sites (Table 4.8). We take this to mean that genes, exons, and regulatory element binding sites are highly enriched in true deserts, and that the positional correlation of human-specific regions with genes may even imply that they also often occur upstream and/or downstream of genes, at as-yet-unknown regulatory element binding sites. Gene Ontology enrichment analysis shows both desert and human-specific regions to be heavily enriched for biological process terms related to brain development, with homophilic cell-cell adhesion showing the highest enrichment (Table 4.9, Table 4.10). Because human-specific regions contain mutations shared by and specific to all humans, we focus our analyses of individual genes on those contained within human-specific regions.

p-value	GO ID	term
1.25E-11	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules
8.26E-07	GO:0070588	calcium ion transmembrane transport
4.35E-06	GO:0030335	positive regulation of cell migration
4.79E-06	GO:0051056	regulation of small GTPase mediated signal transduction
4.98E-06	GO:0030198	extracellular matrix organization

5.14E-06	GO:0007411	axon guidance
1.93E-05	GO:0035725	sodium ion transmembrane transport
2.54E-05	GO:0090630	activation of GTPase activity
3.40E-05	GO:0035249	synaptic transmission, glutamatergic
6.09E-05	GO:0007160	cell-matrix adhesion
8.82E-05	GO:0048813	dendrite morphogenesis
0.000122942	GO:0035556	intracellular signal transduction
0.000127814	GO:0043547	positive regulation of GTPase activity
0.000142609	GO:0045332	phospholipid translocation
0.000142941	GO:0007009	plasma membrane organization
0.000150135	GO:0050808	synapse organization
0.000160949	GO:0009967	positive regulation of signal transduction
0.000188395	GO:2000300	regulation of synaptic vesicle exocytosis
0.000193259	GO:0048008	platelet-derived growth factor receptor signaling pathway
0.000225242	GO:0034332	adherens junction organization
0.000226518	GO:0050900	leukocyte migration
0.000288477	GO:0006805	xenobiotic metabolic process
0.000400207	GO:0048488	synaptic vesicle endocytosis
0.000408619	GO:0007269	neurotransmitter secretion
0.00042998	GO:0030334	regulation of cell migration

0.000430334	GO:0072583	clathrin-dependent endocytosis
0.000432366	GO:0010976	positive regulation of neuron projection development
0.000457578	GO:1903530	regulation of secretion by cell
0.000465516	GO:0007155	cell adhesion
0.00051065	GO:0007215	glutamate receptor signaling pathway
0.000530815	GO:0030534	adult behavior
0.000610593	GO:0060078	regulation of postsynaptic membrane potential
0.000623651	GO:0008361	regulation of cell size
0.000700357	GO:0010885	regulation of cholesterol storage
0.000729067	GO:0060079	excitatory postsynaptic potential
0.000836436	GO:0007097	nuclear migration
0.00086671	GO:0015872	dopamine transport
0.000876401	GO:0043269	regulation of ion transport
0.000906691	GO:0021537	telencephalon development
0.000914601	GO:0035023	regulation of Rho protein signal transduction
0.0010915	GO:0030010	establishment of cell polarity
0.00110164	GO:0060045	positive regulation of cardiac muscle cell proliferation
0.00120304	GO:0016032	viral process
0.0012365	GO:0061001	regulation of dendritic spine morphogenesis
0.0013301	GO:1905039	carboxylic acid transmembrane transport

0.00134639	GO:0048013	ephrin receptor signaling pathway
0.00137139	GO:0003231	cardiac ventricle development
0.00142556	GO:0046854	phosphatidylinositol phosphorylation
0.00147866	GO:0061912	selective autophagy
0.00148203	GO:0014065	phosphatidylinositol 3-kinase signaling
0.00150962	GO:1902115	regulation of organelle assembly
0.00153611	GO:0001525	angiogenesis
0.00158847	GO:1903779	regulation of cardiac conduction
0.00158936	GO:0010811	positive regulation of cell-substrate adhesion
0.0015957	GO:0098659	inorganic cation import across plasma membrane
0.00167132	GO:0001666	response to hypoxia
0.00172151	GO:0015698	inorganic anion transport
0.00172479	GO:1903827	regulation of cellular protein localization
0.00174498	GO:0038083	peptidyl-tyrosine autophosphorylation
0.00174498	GO:0042558	pteridine-containing compound metabolic process
0.00174498	GO:0045773	positive regulation of axon extension
0.00181725	GO:0097484	dendrite extension
0.00184518	GO:0048729	tissue morphogenesis
0.00184571	GO:0015812	gamma-aminobutyric acid transport
0.00184571	GO:0097091	synaptic vesicle clustering

0.00199382	GO:0051336	regulation of hydrolase activity
0.00200355	GO:0007416	synapse assembly
0.0020374	GO:0007229	integrin-mediated signaling pathway
0.00210686	GO:0120035	regulation of plasma membrane bounded cell projection organization
0.00225378	GO:1903146	regulation of autophagy of mitochondrion
0.00227195	GO:0045807	positive regulation of endocytosis
0.00230044	GO:0007399	nervous system development
0.00237396	GO:0022603	regulation of anatomical structure morphogenesis
0.00238578	GO:0043122	regulation of I-kappaB kinase/NF-kappaB signaling
0.00245267	GO:0048048	embryonic eye morphogenesis
0.0024626	GO:0014912	negative regulation of smooth muscle cell migration
0.00252995	GO:0002791	regulation of peptide secretion
0.00256057	GO:0000904	cell morphogenesis involved in differentiation
0.00259392	GO:0030201	heparan sulfate proteoglycan metabolic process
0.0026916	GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway
0.00277874	GO:0035850	epithelial cell differentiation involved in kidney development

0.00282556	GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis
0.00283451	GO:0008360	regulation of cell shape
0.00286755	GO:0003179	heart valve morphogenesis
0.00286755	GO:0016339	calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules
0.00286755	GO:1903307	positive regulation of regulated secretory pathway
0.00300608	GO:1902476	chloride transmembrane transport
0.0030823	GO:0086014	atrial cardiac muscle cell action potential
0.00312522	GO:0032835	glomerulus development
0.00318401	GO:0090090	negative regulation of canonical Wnt signaling pathway
0.00343277	GO:0019228	neuronal action potential
0.00352809	GO:0003014	renal system process
0.00358537	GO:0018108	peptidyl-tyrosine phosphorylation
0.00363324	GO:0042493	response to drug
0.00366202	GO:0002063	chondrocyte development
0.00366202	GO:2000114	regulation of establishment of cell polarity
0.00369162	GO:0001667	ameboidal-type cell migration
0.00378185	GO:0045860	positive regulation of protein kinase activity
0.00378546	GO:0050919	negative chemotaxis

0.00380318	GO:0015721	bile acid and bile salt transport
0.00380318	GO:0048169	regulation of long-term neuronal synaptic plasticity
0.00403981	GO:0010977	negative regulation of neuron projection development
0.00415122	GO:0071495	cellular response to endogenous stimulus
0.00418701	GO:1901888	regulation of cell junction assembly
0.00431272	GO:0001952	regulation of cell-matrix adhesion
0.00433836	GO:0030516	regulation of axon extension
0.00444817	GO:0010107	potassium ion import
0.00444817	GO:0046638	positive regulation of alpha-beta T cell differentiation
0.00455429	GO:0045913	positive regulation of carbohydrate metabolic process
0.00467259	GO:0001657	ureteric bud development
0.00469898	GO:0051494	negative regulation of cytoskeleton organization
0.00471349	GO:0019216	regulation of lipid metabolic process
0.00473483	GO:0014866	skeletal myofibril assembly
0.00473483	GO:0071871	response to epinephrine
0.00473983	GO:2001257	regulation of cation channel activity
0.00478298	GO:0097120	receptor localization to synapse
0.00479548	GO:0003151	outflow tract morphogenesis
0.00482812	GO:0060041	retina development in camera-type eye
0.00482831	GO:0010594	regulation of endothelial cell migration

0.00482831	GO:0010611	regulation of cardiac muscle hypertrophy
0.00482831	GO:0071320	cellular response to cAMP
0.0048604	GO:0050807	regulation of synapse organization
0.0048613	GO:0051592	response to calcium ion
0.00486339	GO:0001845	phagolysosome assembly
0.00486339	GO:0006828	manganese ion transport
0.00486339	GO:0033539	fatty acid beta-oxidation using acyl-CoA dehydrogenase
0.00486339	GO:0036465	synaptic vesicle recycling
0.00486339	GO:0072178	nephric duct morphogenesis
0.00486339	GO:1903651	positive regulation of cytoplasmic transport
0.00493842	GO:0045446	endothelial cell differentiation
0.0050549	GO:0001676	long-chain fatty acid metabolic process
0.00514139	GO:0010975	regulation of neuron projection development
0.00514139	GO:0043647	inositol phosphate metabolic process
0.00524807	GO:0034220	ion transmembrane transport
0.00542368	GO:0003323	type B pancreatic cell development
0.00542368	GO:0032332	positive regulation of chondrocyte differentiation
0.00542368	GO:0035455	response to interferon-alpha
0.00542368	GO:0043046	DNA methylation involved in gamete generation
0.00548774	GO:0003176	aortic valve development

0.00548774	GO:0050855	regulation of B cell receptor signaling pathway
0.0054889	GO:0030100	regulation of endocytosis
0.00575316	GO:0090287	regulation of cellular response to growth factor stimulus
0.0057689	GO:0120161	regulation of cold-induced thermogenesis
0.0058732	GO:0046329	negative regulation of JNK cascade
0.0058732	GO:2001222	regulation of neuron migration
0.0059561	GO:0007212	dopamine receptor signaling pathway
0.00602533	GO:0060291	long-term synaptic potentiation
0.00627833	GO:0033036	macromolecule localization
0.00629847	GO:0019229	regulation of vasoconstriction
0.00629847	GO:0061098	positive regulation of protein tyrosine kinase activity
0.00638947	GO:0048593	camera-type eye morphogenesis
0.00642035	GO:0007010	cytoskeleton organization
0.00648286	GO:0030336	negative regulation of cell migration
0.00663261	GO:0050775	positive regulation of dendrite morphogenesis
0.006678	GO:1902904	negative regulation of supramolecular fiber organization
0.00672647	GO:0110020	regulation of actomyosin structure organization
0.00679646	GO:0034329	cell junction assembly
0.00683878	GO:0051147	regulation of muscle cell differentiation
0.00697793	GO:0030036	actin cytoskeleton organization

0.00709914	GO:0043149	stress fiber assembly
0.00724958	GO:0009812	flavonoid metabolic process
0.00724958	GO:0098911	regulation of ventricular cardiac muscle cell action potential
0.00724958	GO:0098969	neurotransmitter receptor transport to postsynaptic membrane
0.00727583	GO:0007162	negative regulation of cell adhesion
0.00740288	GO:0007041	lysosomal transport
0.00756096	GO:0016310	phosphorylation
0.00761249	GO:0072593	reactive oxygen species metabolic process
0.00764409	GO:0031623	receptor internalization
0.00772882	GO:0030111	regulation of Wnt signaling pathway
0.00774189	GO:0043393	regulation of protein binding
0.00782428	GO:0048041	focal adhesion assembly
0.00782915	GO:0045667	regulation of osteoblast differentiation
0.00784108	GO:0030155	regulation of cell adhesion
0.0078806	GO:0003433	chondrocyte development involved in endochondral bone morphogenesis
0.0078806	GO:0014898	cardiac muscle hypertrophy in response to stress
0.0078806	GO:0098661	inorganic anion transmembrane transport

0.0078806	GO:0098810	neurotransmitter reuptake
0.00789408	GO:0034145	positive regulation of toll-like receptor 4 signaling pathway
0.00789408	GO:0042415	norepinephrine metabolic process
0.00789408	GO:0061469	regulation of type B pancreatic cell proliferation
0.00789408	GO:1904322	cellular response to forskolin
0.00799742	GO:0003177	pulmonary valve development
0.00799742	GO:0007063	regulation of sister chromatid cohesion
0.00799742	GO:0032011	ARF protein signal transduction
0.00799742	GO:0035024	negative regulation of Rho protein signal transduction
0.00799742	GO:0036119	response to platelet-derived growth factor
0.00799742	GO:0071305	cellular response to vitamin D
0.00799742	GO:0097062	dendritic spine maintenance
0.00799742	GO:1901017	negative regulation of potassium ion transmembrane transporter activity
0.00799742	GO:2000310	regulation of NMDA receptor activity
0.00799742	GO:2000737	negative regulation of stem cell differentiation
0.00810685	GO:0062013	positive regulation of small molecule metabolic process
0.00811039	GO:0098657	import into cell
0.00811968	GO:0000902	cell morphogenesis

0.0082919	GO:0033627	cell adhesion mediated by integrin
0.00831191	GO:0051899	membrane depolarization
0.00842463	GO:0009887	animal organ morphogenesis
0.00869904	GO:0006942	regulation of striated muscle contraction
0.00870789	GO:1905037	autophagosome organization
0.00887433	GO:0010595	positive regulation of endothelial cell migration
0.00914977	GO:0048639	positive regulation of developmental growth
0.00915116	GO:0016242	negative regulation of macroautophagy
0.00915116	GO:0044091	membrane biogenesis
0.00915116	GO:0044319	wound healing, spreading of cells
0.00915116	GO:1900006	positive regulation of dendrite development
0.00918547	GO:0055074	calcium ion homeostasis
0.00931042	GO:0031076	embryonic camera-type eye development
0.00931042	GO:0051590	positive regulation of neurotransmitter transport
0.0093559	GO:0043200	response to amino acid
0.0093695	GO:0099536	synaptic signaling
0.00940916	GO:0060411	cardiac septum morphogenesis
0.00941286	GO:0030833	regulation of actin filament polymerization
0.00950095	GO:0007613	memory
0.00955764	GO:0050804	modulation of chemical synaptic transmission

0.00964434	GO:0032273	positive regulation of protein polymerization
0.00971488	GO:0030193	regulation of blood coagulation
0.00976581	GO:0035418	protein localization to synapse
0.00983252	GO:0043588	skin development

Table 4.9: Enriched biological_process Gene Ontology [1] terms in desert regions. Enrichment testing was done using the hypergeometric function in FUNC [164], with refinement and a cutoff of p = 0.01.

p-value	GO ID	term
8.92E-10	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules
1.45E-05	GO:0099072	regulation of postsynaptic membrane neurotransmitter receptor levels
3.26E-05	GO:0048813	dendrite morphogenesis
8.09E-05	GO:0071625	vocalization behavior
0.000106681	GO:0070588	calcium ion transmembrane transport
0.000111693	GO:0071417	cellular response to organonitrogen compound
0.000137589	GO:0050804	modulation of chemical synaptic transmission
0.000202768	GO:0007268	chemical synaptic transmission

0.00021313	GO:0071313	cellular response to caffeine
0.000269131	GO:0030010	establishment of cell polarity
0.000293738	GO:0007416	synapse assembly
0.00041081	GO:0051272	positive regulation of cellular component movement
0.000434341	GO:0007264	small GTPase mediated signal transduction
0.000441847	GO:0010976	positive regulation of neuron projection development
0.000458796	GO:0036465	synaptic vesicle recycling
0.000459866	GO:0016339	calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules
0.000464144	GO:0086014	atrial cardiac muscle cell action potential
0.000476936	GO:1904322	cellular response to forskolin
0.000481311	GO:0098698	postsynaptic specialization assembly
0.000485931	GO:0042297	vocal learning
0.000984275	GO:0007409	axonogenesis
0.00104939	GO:0021942	radial glia guided migration of Purkinje cell
0.00105734	GO:1905114	cell surface receptor signaling pathway involved in cell-cell signaling
0.00107765	GO:1903539	protein localization to postsynaptic membrane
0.00116814	GO:0015872	dopamine transport
0.00117966	GO:0000904	cell morphogenesis involved in differentiation

0.00119672	GO:0097120	receptor localization to synapse
0.00151518	GO:0035235	ionotropic glutamate receptor signaling pathway
0.00152567	GO:0061912	selective autophagy
0.00183832	GO:0034329	cell junction assembly
0.00188633	GO:0003192	mitral valve formation
0.00188633	GO:0034727	piecemeal microautophagy of the nucleus
0.00235145	GO:0044351	macropinocytosis
0.00246344	GO:0035735	intraciliary transport involved in cilium assembly
0.00281075	GO:0051494	negative regulation of cytoskeleton organization
0.0028388	GO:0010793	regulation of mRNA export from nucleus
0.0028388	GO:0051552	flavone metabolic process
0.00290296	GO:0030335	positive regulation of cell migration
0.00290771	GO:0071466	cellular response to xenobiotic stimulus
0.00314583	GO:1901021	positive regulation of calcium ion transmembrane transporter activity
0.00354249	GO:0035023	regulation of Rho protein signal transduction
0.00385041	GO:0051491	positive regulation of filopodium assembly
0.00395873	GO:0043547	positive regulation of GTPase activity
0.00422597	GO:0007256	activation of JNKK activity
0.00422597	GO:0061000	negative regulation of dendritic spine development

0.00473981	GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II
0.00480876	GO:0034332	adherens junction organization
0.00493758	GO:0060078	regulation of postsynaptic membrane potential
0.005337	GO:0048639	positive regulation of developmental growth
0.00553232	GO:1990138	neuron projection extension
0.00568608	GO:0031589	cell-substrate adhesion
0.00597595	GO:0019530	taurine metabolic process
0.00597595	GO:0060956	endocardial cell differentiation
0.00597595	GO:2000210	positive regulation of anoikis
0.00599516	GO:0055007	cardiac muscle cell differentiation
0.00621657	GO:0010975	regulation of neuron projection development
0.00632312	GO:0017156	calcium ion regulated exocytosis
0.00634401	GO:0010880	regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum
0.00634401	GO:0021680	cerebellar Purkinje cell layer development
0.00634401	GO:0032369	negative regulation of lipid transport
0.00634401	GO:0046058	cAMP metabolic process
0.00648969	GO:1903779	regulation of cardiac conduction
0.00654834	GO:0006298	mismatch repair

0.00656229	GO:0071380	cellular response to prostaglandin E stimulus
0.00665707	GO:0098609	cell-cell adhesion
0.00666753	GO:0001525	angiogenesis
0.00676885	GO:0007399	nervous system development
0.00678893	GO:0010507	negative regulation of autophagy
0.0068467	GO:0014878	response to electrical stimulus involved in regulation of muscle adaptation
0.0068467	GO:0018916	nitrobenzene metabolic process
0.0068467	GO:0033594	response to hydroxyisoflavone
0.0068467	GO:0035022	positive regulation of Rac protein signal transduction
0.0068467	GO:0042853	L-alanine catabolic process
0.0068467	GO:0046449	creatinine metabolic process
0.0068467	GO:0050923	regulation of negative chemotaxis
0.0068467	GO:0061669	spontaneous neurotransmitter secretion
0.0068467	GO:0071922	regulation of cohesin loading
0.0068467	GO:0072757	cellular response to camptothecin
0.0068467	GO:0090116	C-5 methylation of cytosine
0.0068467	GO:0097118	neuroligin clustering involved in postsynaptic membrane assembly

0.0068467	GO:1903377	negative regulation of oxidative stress-induced neuron intrinsic apoptotic signaling pathway
0.00690902	GO:0003300	cardiac muscle hypertrophy
0.0069647	GO:0016446	somatic hypermutation of immunoglobulin genes
0.0069647	GO:0038007	netrin-activated signaling pathway
0.0069647	GO:1903651	positive regulation of cytoplasmic transport
0.00706799	GO:0072659	protein localization to plasma membrane
0.00730382	GO:0006935	chemotaxis
0.00751327	GO:0007017	microtubule-based process
0.00766163	GO:0007610	behavior
0.0077158	GO:0010769	regulation of cell morphogenesis involved in differentiation
0.00783274	GO:0003283	atrial septum development
0.00825468	GO:0031646	positive regulation of neurological system process
0.00833621	GO:0099560	synaptic membrane adhesion
0.00833621	GO:0099590	neurotransmitter receptor internalization
0.008398	GO:0006928	movement of cell or subcellular component
0.00843191	GO:0048489	synaptic vesicle transport
0.00898072	GO:0051492	regulation of stress fiber assembly
0.00926198	GO:0001504	neurotransmitter uptake

0.0094164	GO:0006544	glycine metabolic process
0.0094164	GO:0071044	histone mRNA catabolic process
0.00967361	GO:0051965	positive regulation of synapse assembly
0.00999857	GO:0014909	smooth muscle cell migration

Table 4.10: Enriched biological_process Gene Ontology [1] terms in human-specific regions. Enrichment testing was done using the hypergeometric function in FUNC [164], with refinement and a cutoff of p = 0.01.

Long human-specific regions are more likely to be the result of natural selection and may contain genes important in differentiating humans from archaic hominins. We find the mean TMRCAs of all humans across human-specific regions to be negatively correlated with the length of the regions (Spearmans rho = -0.166; $p < 2.2e-16$), suggesting that longer regions are more subject to purifying selection than shorter regions. Furthermore, human-specific regions are enriched for interacting sets of genes (interaction score $>$ 700 in the STRING database [207]; permutation test p = 0.006 of more interactions in 1000 trials). The longest two human-specific regions contain zinc finger proteins with unknown function but likely to have transcription factor activity, ZNF626, ZNF726, and ZNF737. Taking the set of all genes that intersect the 50 longest human-specific regions, we find additional zinc finger proteins: ZNF561, ZNF562, ZNF675, ZNF707, and ZNF846. Of these, ZNF675 (TIZ) is known to have a likely role in osteoclast differentiation [195]; the others are less well understood. The 50 longest human-

Map	Features	Dist. P	Proj. P
Deserts	genes	0.14	0*
Deserts	exons	2.87E-09*	4.53E-01
Deserts	reg. elt.	2.12E-09*	0*
Human-specific	genes	4.57E-04*	5.88E-14*
Human-specific	exons	2.51E-03*	3.70E-10*
Human-specific	reg. elt.	9.47E-01	2.61E-06*

Table 4.8: Overlap of archaic hominin ancestry deserts and human-specific regions with various genomic features. Genes are whole protein coding genes from Gencode [49], using Ensembl version 94 on human genome version GRCh38 lifted over to GRCh37 coordinates. Exons are for protein-coding genes from the same annotation. Regulatory elements are from the filtered "double-elite" set in the GeneHancer database [48], obtained from the UCSC Genome Browsers Table Browser utility [21]. Distance correlation p-value comes from the relative distance Kolmogorov-Smirnov test and projection p-value, which measures overlap, were both computed using the GenometricCorr R package [46]. All significant ($p < 0.01$ or $p > 0.99$) values are marked with an asterisk.

specific regions also include one (chr12:11121063-11248363) containing salivary proteins as well as a cluster of seven bitter taste receptors, including 2 of 3 identified in previous studies as outliers for having broad specificity [128] (Table 4.11). This hints that dietary change may have been a potential factor in human speciation. Another long human-specific region encompasses a cluster of PRAMEF genes, which are cell surface antigens that probably play a role in cell proliferation in gonadal tissues and cancer [203], suggesting a possible role in hybrid incompatibility. Another noteworthy find is a set of 9 members of the β-protocadherin gene cluster. So-called clustered protocadherin genes undergo extensive alternative splicing, and the combinatorial expression of protocadherins on cell surfaces is thought to guide developing neural projections by allowing self-recognition and avoidance. Furthermore, genes in the Pcdhα cluster, which intersects with and extends downstream of another human-specific region, are probably involved in cleaning up cases of incorrect neural wiring [69]; we find a small human-specific region in the upstream-most portion of the γ-protocadherin cluster as well. For a visual representation of human-specific regions and allele frequencies within them in the region of all three protocadherin clusters, see Fig. 4.18. In all, we find 48 of 158 (30%) of autosomal genes with the GO term “homophilic cell adhesion via plasma membrane adhesion molecules” in our human-specific regions, suggesting that other cadherins were also important in differentiating the human lineage from archaic hominins. Slit/Robo signaling has been proposed as another mechanism for axon repulsion, among other processes important in neurodevelopment [12], and we find both ROBO1 and ROBO2 in human-specific regions, suggesting that regulation and/or splicing of self-recognition and repulsion in developing neurons underwent changes important in the human lineage.

gene	chrom	start	end	length	TMRCA
AC005488.1	7	72373953	72484431	110478	0
AC010760.1	15	22376613	22549285	172672	0
AC018630.1	12	11121063	11248363	127300	0.001835365
AC091057.6	15	30882290	30980523	98233	0
AC108868.1	2	107014665	107109876	95211	0
AC134980.3	15	22376613	22549285	172672	0
AC135068.3	15	22376613	22549285	172672	0
AC244517.10	5	140538160	140622981	84821	0.000171793
ALMS1	2	73705219	73780914	75695	0.000834066
ANKRD30A	10	37399546	37488981	89435	0.000490825
ARHGAP11B	15	30882290	30980523	98233	0

C1orf158	1	12819122	12945655	126533	0.000125513
CCDC166	8	144742887	144836457	93570	1.62E-05
CD8B2	2	107014665	107109876	95211	0
CLEC18B	16	74372357	74460377	88020	0.000461297
FAM83H	8	144742887	144836457	93570	1.62E-05
FLT4	5	180059136	180162268	103132	8.77E-07
GML	8	143862328	143946821	84493	8.81E-05
GOLGA8H	15	30882290	30980523	98233	0
H3.Y	5	17600702	17675089	74387	0.000157675
HLA-B	6	31304861	31378939	74078	0
HLA-DQA1	6	32589966	32689385	99419	0
HLA-DQB1	6	32589966	32689385	99419	0
HLA-DRB1	6	32495832	32573989	78157	0
HLA-DRB5	6	32495832	32573989	78157	0
HNRNPCL1	1	12819122	12945655	126533	0.000125513
HRH2	5	174990878	175090975	100097	0.000299065
IQANK1	8	144742887	144836457	93570	1.62E-05
LINC02203	15	22376613	22549285	172672	0
LY6D	8	143862328	143946821	84493	8.81E-05
MAPK15	8	144742887	144836457	93570	1.62E-05

METTL2A	17	60518577	60598285	79708	6.00E-06
MICA	6	31304861	31378939	74078	0
MTRNR2L6	7	142353927	142431384	77457	0
NPIPA1	16	15036996	15118351	81355	0.001525955
NPIPB15	16	74372357	74460377	88020	0.000461297
NSUN5	7	72651982	72730656	78674	0.007747081
OR4N4	15	22376613	22549285	172672	0
PCDHA1	5	140094294	140179492	85198	6.69E-06
PCDHA2	5	140094294	140179492	85198	6.69E-06
PCDHB10	5	140538160	140622981	84821	0.000171793
PCDHB11	5	140538160	140622981	84821	0.000171793
PCDHB12	5	140538160	140622981	84821	0.000171793
PCDHB13	5	140538160	140622981	84821	0.000171793
PCDHB14	5	140538160	140622981	84821	0.000171793
PCDHB16	5	140538160	140622981	84821	0.000171793
PCDHB7	5	140538160	140622981	84821	0.000171793
PCDHB8	5	140538160	140622981	84821	0.000171793
PCDHB9	5	140538160	140622981	84821	0.000171793
PDXDC1	16	15036996	15118351	81355	0.001525955
POM121	7	72373953	72484431	110478	0

POM121C	7	75049912	75152873	102961	0.000161375
PRAMEF1	1	12819122	12945655	126533	0.000125513
PRAMEF11	1	12819122	12945655	126533	0.000125513
PRAMEF12	1	12819122	12945655	126533	0.000125513
PRAMEF2	1	12819122	12945655	126533	0.000125513
PRAMEF4	1	12819122	12945655	126533	0.000125513
PRB1	12	11390844	11540562	149718	0
PRB3	12	11390844	11540562	149718	0
PRB4	12	11390844	11540562	149718	0
PRH1	12	11121063	11248363	127300	0.001835365
PRH1-PRR4	12	11121063	11248363	127300	0.001835365
PROP1	5	177333734	177425706	91972	0.000580695
RGPD3	2	107014665	107109876	95211	0
SLC6A2	16	55738787	55816560	77773	0.000965875
SPDYE5	7	75049912	75152873	102961	0.000161375
TAF11L12	5	17600702	17675089	74387	0.000157675
TAF11L13	5	17600702	17675089	74387	0.000157675
TAF11L14	5	17600702	17675089	74387	0.000157675
TAS2R14	12	11121063	11248363	127300	0.001835365
TAS2R19	12	11121063	11248363	127300	0.001835365

TAS2R20	12	11121063	11248363	127300	0.001835365
TAS2R31	12	11121063	11248363	127300	0.001835365
TAS2R43	12	11121063	11248363	127300	0.001835365
TAS2R46	12	11121063	11248363	127300	0.001835365
TAS2R50	12	11121063	11248363	127300	0.001835365
TLK2	17	60518577	60598285	79708	6.00E-06
TRIM50	7	72651982	72730656	78674	0.007747081
TRIM74	7	72373953	72484431	110478	0
UGT3A1	5	35946482	36069762	123280	0.001710388
UGT3A2	5	35946482	36069762	123280	0.001710388
ZNF561	19	9729793	9864085	134292	0.000109891
ZNF562	19	9729793	9864085	134292	0.000109891
ZNF626	19	20709939	20911175	201236	0
ZNF675	19	23782867	23892190	109323	0
ZNF707	8	144742887	144836457	93570	1.62E-05
ZNF726	19	23973277	24177630	204353	1.98E-06
ZNF737	19	20709939	20911175	201236	0
ZNF846	19	9729793	9864085	134292	0.000109891

Table 4.11: Genes within the top 50 human-specific regions, by length. Coordinates given are for the boundaries of the human-specific region (hg19 coordinates). TMRCAs are means across sites within the human-specific regions and reported as a fraction of human/chimp divergence (approx. 13 my).

Next, we sought to determine which mutations in human-specific regions may have affected gene regulation, which biological processes may have been affected by these mutations, and when in history these mutations may have arisen. Using the heavily-filtered double elite" set of regulatory element binding sites and regulatory targets from the GeneHancer database [48], we compiled a list of fixed derived alleles in modern humans intersecting these binding sites, along with the TMRCA of all humans in each. We then investigated which of these binding sites target genes interact with each other, and for each such interacting pair, compared the TMRCAs of the mutations in the genes regulators binding sites. We expected this to shed light on which mutations clustered together in time, with the hypothesis that mutations affecting regulation of genes involved in specific biological processes may have arisen around the same time. We find three clusters of TMRCAs this way: one from roughly 33-42 kya, one from 145-185 kya, and another from 300-400 kya (Fig. 4.17D). This appears to suggest three different "bursts" of adaptive changes, although the most recent (33-42 kya) is too young to predate the geographical spread of all humans this number is therefore probably biased downward by continual purifying selection reducing the amount of diversity in modern humans, or by an increased locus-specific mutation rate in humans compared to chimpanzees artificially inflating the human-specific branch length. Additionally, although many regulatory mutations affecting interacting genes have approximately the same TMRCA (on the diagonal in the plot), we find many off-diagonal effects as well, suggesting that different bursts of selection may have acted on some of the same biological processes. To locate which genes may have played the biggest role in the history of the human lineage, we counted the number of genes in our list of targets of regulatory elements with human-mutated binding sites with which each other gene in the list is known to interact. We then ordered our list of genes by this number times the number of regulatory element binding site mutations we found; top candidates in this list should be genes involved in multiple biological processes, with potentially large changes in human-specific expression from the ancestral state. The top two genes we find this way are SF1 and SRRM1, both important splicing factors [184]. The other top five genes are FYN, a protein tyrosine kinase that regulates neuronal migration [111], KAT5 (TIP60), a histone acetyltransferase subunit involved in chromatin remodeling and Notch1 signaling [89], and TRIM63 (MuRF1), which is involved in regulation of atrophy in skeletal muscle [13].

4.2.1 Discussion

Our findings pertaining to the demography of ancient admixture in modern humans largely agree with prior studies. We find a similar amount of Neanderthal ancestry in modern humans, but less Denisovan ancestry, than has been previously reported using genome-wide statistics [165]. Furthermore, our model of Neanderthal admixture (a main pulse after the out-of-Africa migration but before the diversification of modern Eurasian populations) agrees with what has been previously suggested [56]. In the case of Denisovan ancestry, the lower global amount of ancestry we recovered per individual is comparable to what has been detected in other studies using local ancestry scans [213]. The fact that we observe shorter Denisovan than Neanderthal haplotype block lengths, and that this contradicts previous inferences of a more recent Denisovan than Neanderthal introgression time [132] may hint that the Denisovan genome is not a good model for the true introgressor, which may have itself been a hybrid [223]. This is supported by much higher TMRCAs to Denisovan in Denisovan-like haplotypes than TMRCAs to Neanderthal in Neanderthal-like haplotypes. The possibility that a separate Denisovan-like component exists in mainland Asians than that in Oceanians merits further study.

Our detection of highly-diverged archaic-like haplotypes in sub-Saharan Africa using both Neanderthal and Denisovan haplotypes as the subjects of ancestry scans suggests that we may have located pieces of previously-described super-archaic admixture [62, 37], although this merits further study. Performing specific analyses of these haplotypes, and following them along the ARG after the archaic hominin haplotypes they contain recombine out of them, might allow us to uncover their full extent. Furthermore, scans based entirely on the TMRCA and genomic span of clades, rather than requiring the presence of archaic hominin haplotypes, might uncover more segments of super-archaic ancestry.

We find similar evidence as in previous studies (e.g. [186]) of introgressed alleles being slightly deleterious on average. Namely, we see enrichment of protein-coding genes and exons in human-specific genomic regions, as well as enriched intersection of human-specific regions with regulatory element binding sites. As in previous studies, we also find many of the same types of genes to be within these introgressed haplotypes, including genes related to diet, keratinogenesis, and immune function. We also find evidence that some previously-adaptive archaic alleles may contribute to modern maladaptive phenotypes.

According to our analysis, only about 1.5% of the genome (at minimum) and 10% of the genome (at maximum) is completely sorted and devoid of admixture between modern and archaic hominins, implying that this fraction of the genome alone makes modern humans a unique species. We find these regions to be enriched for genes, particularly those related to brain development and function. This finding is compatible with previous research suggesting that archaic-introgressed alleles cause downregulation of genes expressed in the brain [119]. In particular, the existence of completely sorted genomic regions intersecting all three protocadherin gene clusters on chromosome 5, as well as 48% of 158 total genes involved in homophilic cell-cell adhesion, suggests that these genes may prove to be particularly interesting targets of further analysis. The fact that exons of these genes are often spliced together in a unique manner to produce unique combinations of cell surface proteins to guide neurite growth [24] hints that changes in splicing, as well as amino acid changes and changes to gene regulation, may have been important in our evolutionary history. The finding that recurrent human-specific mutations happened in regions regulating multiple splicing factors, dating back to the oldest evolutionary "burst" we detected in human evolutionary history circa 400-500 kya, helps bolster this narrative. Furthermore, the abundance of C2H2-containing zinc finger proteins in the longest completely sorted genomic regions hints at their importance in human evolution; learning whether these proteins regulate the expression of other genes, and where they bind in the genome, could be illuminative.

We believe that ancestral recombination graph inference has much to bring to the study of archaic hominin admixture and human evolution and hope that future studies will expand upon our approach. Furthermore, assays that make use of gene-edited human stem cells will allow many of the human specific changes we find to be tested in the lab for phenotypic effects; this is a crucial next step in determining what makes us a unique species.

4.3 Methods

4.3.1 Data processing

All data were processed as described in the previous chapter. This study used phased human genomes from the Simons Genome Diversity Project [116], along with the high-coverage Altai [165] and Vindija33.19 Neanderthals [163], and the high-coverage Denisovan genome [126].

4.3.2 Simulations

Our demographic simulation was done using scrm [201] because it allows users to sample haplotypes from time points in the past, mimicking the branch shortening due to "missing evolution" when analyzing ancient genomes. For this simulation, we combined a popular, three population demographic model for modern humans [60] with populations meant to approximate the Altai [165] and Vindija [163] Neanderthals. We again assumed a 1 centimorgan per megabase recombination rate and a $1 * 10^{-9}$ per year mutation rate, along with a 25-year generation time, giving a per-generation mutation rate of $2.5 * 10^{-8}$. In addition to the demographic model parameters listed in [60], we modeled a Neanderthal/human split time of 575kya [165], an Altai/Vindija split time of 137.5kya, and modeled the heterozygosity in all Neanderthals as $1.6*10^{-4}$ [163]. Additionally, we chose 100kya as the divergence time between the Vindija and introgressing Neanderthal, and we modeled human/Neanderthal introgression as a single pulse 50kya, in the population ancestral to both Europeans and Asians, with a 5% admixture proportion. Finally, we assigned 57.kya of missing evolution to the Vindija haplotypes and 123kya of missing evolution to the Altai haplotypes [163]. Our simulated chromosome was 25 megabases long, and we sampled 2 haplotypes from each Neanderthal (but not the introgressing Neanderthal), as well as 20 haplotypes from each modern human population (African, Asian, and European). The full command was scrm 64 1 -t 17253.7128713 -r 7342.00547714 25000000 -T -I 6 20 20 20 0 0 0 -eI 0.0772268560362 0 0 0 0 0 2 -eI 0.167529158597 0 0 0 2 0 0 -n 1 1.68 -n 2 3.74 -n 3 7.29 -n 4 0.231834158238 -n 5 0.231834158238 -n 6 0.231834158238 -eg 0 2 116.010723 -eg 0 3 160.246047 -m 2 3 2.797460 -m 3 2 2.797460 -ej 0.028985 3 2 -en 0.028985 2 0.287184 -es 0.0681012839825 2 0.95 -ej 0.0681012839825 7 5 -ej 0.136202567965 6 5 -ej 0.197963 2 1 -en 0.303501 1 1 -ej 0.187278530952 5 4 -ej 0.783164765799 4 1. We discarded all but the first instance of every unique base position in the output file, and we converted the "true" trees into SARGE format for running analyses.

4.3.3 Admixture scans

The central challenge of creating admixture maps is to disentangle incomplete lineage sorting (ILS) from admixture. Both processes create local trees in the genome that group candidate admixed haplotypes with admixer haplotypes. Clades resulting from ILS are older than those resulting from admixture, however; they should therefore persist for shorter stretches along the genome and have older times to most recent common ancestor (TMRCAs) (Fig. 4.1A). In "true" ARG trees from our demographic simulation, clades resulting from admixture were indeed easily distinguishable from incompletely sorted lineages, using these two metrics (Fig. 4.1B). In the ARG we inferred using SARGE, however, the low-resolution nature of branch lengths inferred using mutations made the cutoff between ILS and admixture more difficult to determine (Fig. 4.1C,D). To map Neanderthal and Denisovan ancestry, we first scanned through ARG output for all clades that grouped some modern human haplotypes with one or more admixer haplotypes (Neanderthal and/or Denisovan) to the exclusion of some other modern human haplotypes. Since SARGE produces many polytomies, this carries the risk of observing a parent of one or more true admixed clades, but not the true admixed clade. This would manifest as a clade containing many modern humans, in addition to one or more archaic hominins, and would falsely be interpreted as a very high-frequency archaic-introgressed haplotype. To mitigate this problem, we defined the Mbuti, Biaka, and Khomani-San genomes as an outgroup and discarded any clade that contained more than 10% of the outgroup members. We also discarded one extremely long candidate haplotype that spanned a centromere. For each clade that passed our selection criteria, we visited each non-archaic hominin member and determined whether that member possessed candidate Neanderthal, Denisovan, or undetermined ancestry by assessing whether it was closer (by tree topology, ignoring branch lengths) to a Neanderthal or Denisovan haplotype, or equidistant to both. We then computed the mean time to most recent common ancestor (TMRCA) between each human member and the candidate archaic introgressor across each haplotype, not accounting for branch shortening.

This provided a set of haplotypes resulting from both admixture and incomplete lineage sorting. To separate these two categories of haplotypes, we computed a p-value (which we used as a score) designed to be low in cases of admixture and high in cases of ILS. We first computed the TMRCA of all lineages under study, using SARGE output; this value was approximately 639 kya. Assuming a 1 cM/Mb recombination rate, we consider an unrecombined haplotype in the ancestor of all lineages to be 100 Mb long. We assume that neutrally-evolving haplotype block lengths are exponentially distributed with a mean equal to 100 Mb divided by the number of generations that have passed. We then define, for each haplotype block length x, $p_{len} = e^{-\frac{639000}{1e8*25}x}$. Next, since TMRCAs in SARGE output trees appear to be exponentially distributed, we define a p-value that a given TMRCA y results from this lineage, as a percent of human-chimp divergence rather than in years, assuming 6.5 mya human-chimp divergence and using a pseudocount of 0.01: $p_{TMRCA} = 1 - e^{-(\frac{13e6}{639000}+0.001)(y+0.01)}$. We then compute a combined p-value, or score, as $p_{combined} = p_{len} * p_{TMRCA}$.

The central insight of the D-statistic, a popular genome-wide statistic used for detecting and quantifying admixture, is that admixture should affect certain human populations more than others, while ILS is expected to affect all approximately equally [56, 35]. We used this reasoning to determine a score cutoff for binning candidate admixed haplotypes into ILS and true admixture. Using score cutoffs ranging from 0 to 1 and offset by 0.01, we selected all haplotypes with combined p greater than or equal to the cutoff (ILS) and less than the cutoff (admixture). We then computed, for each haploid human genome, the percent of the queryable genome (excluding sex chromosomes and unplaced scaffolds) belonging to regions of ILS and regions of admixture. We then computed the coefficient of variation (standard deviation divided by mean) of this value across SGDP populations. We hypothesized that the cross-population coefficient of variation in ILS should be low around the ideal threshold, while the coefficient of variation in admixture should be higher but stabilize. We found that this is the case around a score cutoff of 0.16 (Fig. 4.1C). Haplotypes classified as admixture and ILS using this score cutoff are shown in Fig. 4.1D, and binning this way produced reasonable estimates of the global extent of admixture (Fig. 4.2A). For subsequent, haplotype-specific analyses, however, we limited to a more conservative set of admixed haplotypes, with a score cutoff of 0.001. We do not seek to interpret these p-values beyond their use as a score for separating admixture from ILS.

4.3.4 Testing for overlap

All overlap-related tests were done using the GenometricCorr R package [46]. We loaded each "reference" and "query" BED file, along with the lengths of all hg19 autosomes, We then limited all tests to a list of regions deemed queryable by the ARG. We defined these regions as 50kb windows of the genome that contained at least one site in our ARG input files; these "queryable" windows should therefore exclude non-genic regions of the genome such as centromeres and telomeres. We set the number of permutations in each test to zero, as we were only interested in two parametric tests implemented in the package.

4.3.5 Analysis of introgressed haplotypes

We merged and computed the frequencies of confidently introgressed (score ¡ 0.001) Neanderthal and Denisovan haplotypes using BEDTools multiinter [167] and dividing by the number of haploid human genomes under study. For GO enrichment analyses, we used the Wilcoxon test implemented in FUNC [164], with the October 29, 2018 release of the Gene

Ontology tables [1], ranking GO terms by the frequency of the Neanderthal or Denisovan haplotype containing them. We used FUNCs refinement routine to account for the hierarchy of GO terms, and we used a p-value cutoff of 0.01.

For our GWAS analysis, we obtained the GWAS catalog from the European Bioinformatics Institute [115] on hg38 coordinates, discarding any hits that were not on autosomes. We then extracted every derived allele in our data set shared between humans and archaic hominins whose frequency in modern humans equaled the frequency of its parent archaic-introgressed haplotype. We used liftOver [72] to port these alleles to hg38 coordinates and then compared them with the significant GWAS hits, considering only alleles within introgressed haplotypes with frequency greater than 1% in modern humans. There were no such GWAS hits coinciding with high-frequency Denisovan-introgressed alleles.

4.3.6 Determining size of deserts

After obtaining our set of deserts and human-specific regions using real data, we scanned our inferred ARG on data from the demographic simulation (see Simulations section) for deserts, human-specific regions, and admixture. Due to the small size of the simulated chromosome, the small number of simulated samples, and the lack of an unadmixed outgroup, we did not use the same technique as in real data to detect admixture. Instead, we first scanned for all clades that grouped one or more modern human haplotypes with archaic hominin haplotypes to the exclusion of other modern human haplotypes, without the use of an outgroup. This produced a set of clades representing both admixture and incomplete lineage sorting. Next, we repeated this process, using true ARG trees from scrm output and plotted the mean TMRCA to Neanderthal against the length of each candidate introgressed haplotype; this allowed us to easily separate admixture from ILS by eye, since branch lengths were exact rather than inferred from mutations (Fig. 4.1B). We created a BED file of each true admixed haplotype thus detected, and we labeled all candidate introgressed haplotypes from our inferred ARG data as truly introgressed if they intersected one of these true admixed haplotypes. Merging this file (using BEDTools merge [167]) and counting the number of bases gave us the overall amount of admixture in our simulations.

To determine if the deserts and human-specific regions we detected represent the full extent of those regions across all humans, or whether they are a superset that would decrease with the examination of more genomes, we randomly sampled (without regard to population or phylogenetic position) sets of 10, 50, and 100 human haplotypes from the SGDP data set and added all archaic hominin haplotypes (Altai, Vindija33.19, and Denisovan) to each set. We re-ran SARGE on each of these data sets, with the same parameters as the full run (excluding CpG sites, and with 25kb propagation distance) and scanned the results of each for deserts and human-specific regions.

4.3.7 Analysis of human-specific alleles in deserts

We conducted Gene Ontology enrichment analysis using the hypergeometric test implemented in FUNC [164], with the October 29, 2018 release of the Gene Ontology tables [1], using FUNCs refinement routine and a p-value cutoff of 0.01.

To study the age of candidate functional mutations in these regions, we first searched ARG output for clades containing all modern humans and noted the TMRCA of each clade.

We then intersected these sets of positions with positions of confidently-called binding sites for regulatory elements, obtained from the “double-elite” set within the GeneHancer database [48], downloaded from the UCSC Genome Browser [21] (“interactions” table on hg19). We then took the target genes for each regulatory element and searched for interactions with other genes in this set in the STRING database [207], with a minimum interaction score of 300. We then determined time bins containing the most genes by eye, after plotting.

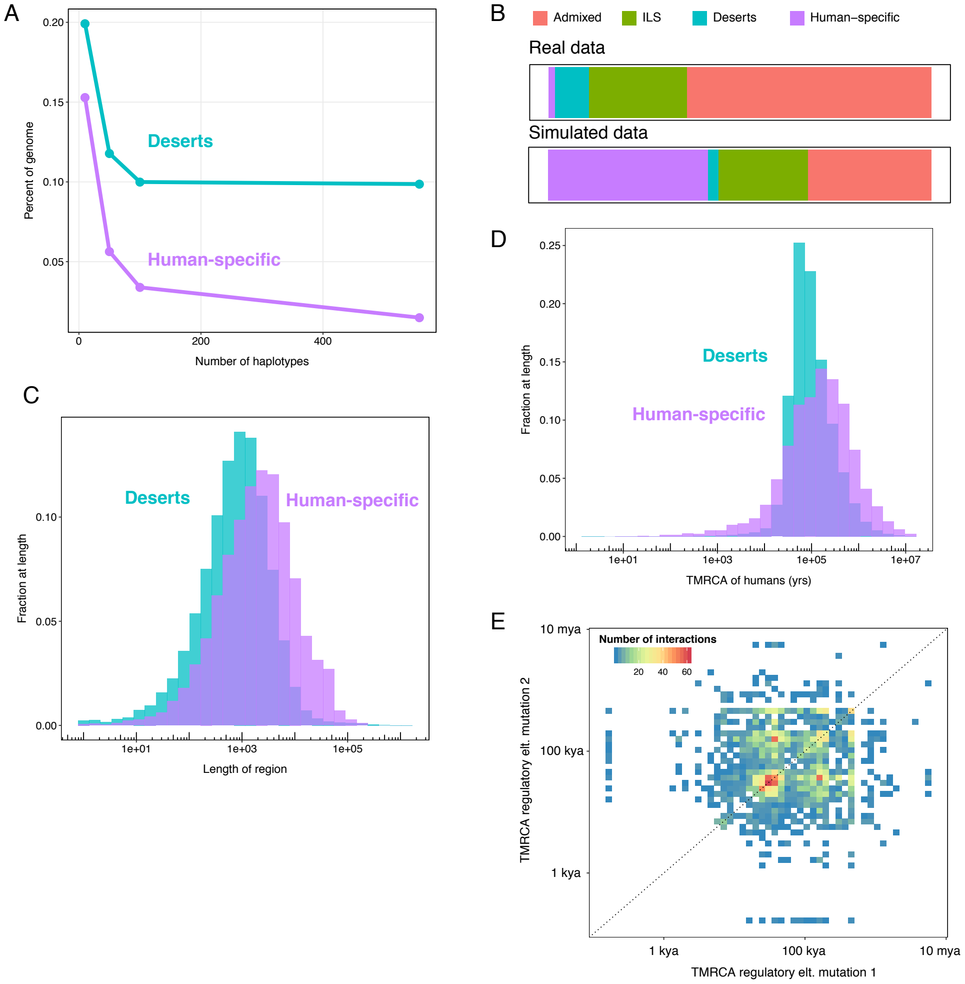


Figure 4.17: Properties of desert regions (free of admixture and ILS) and human-specific regions (all humans share a derived lineage). A: Sizes of deserts and human-specific regions detected in real data, with different numbers of randomly subsampled human haplotypes. Number of haplotypes refers to modern human haplotypes only; all archaic hominin haplotypes were included in each run. B: Proportions of the autosomal genome containing archaic hominin admixture, ILS, and in deserts and human-specific regions, in real data and data simulated under a popular demographic model, with a single pulse of Neanderthal admixture 50 kya. C: Lengths of deserts and human-specific regions (bp). D: Mean TMRCA of all humans across each desert and human-specific region. E: For loci where all humans share a derived lineage in a regulatory element binding site, genes regulated by elements that bind at that site are considered, and all pairs of genes in this set whose products interact are considered. For each interacting pair of genes in this set, the TMRCA of all humans at the regulatory element mutation for each gene in the pair are shown. Selected time bins with the most genes are 33-42 kya, 145-185 kya, and 400-500 kya.

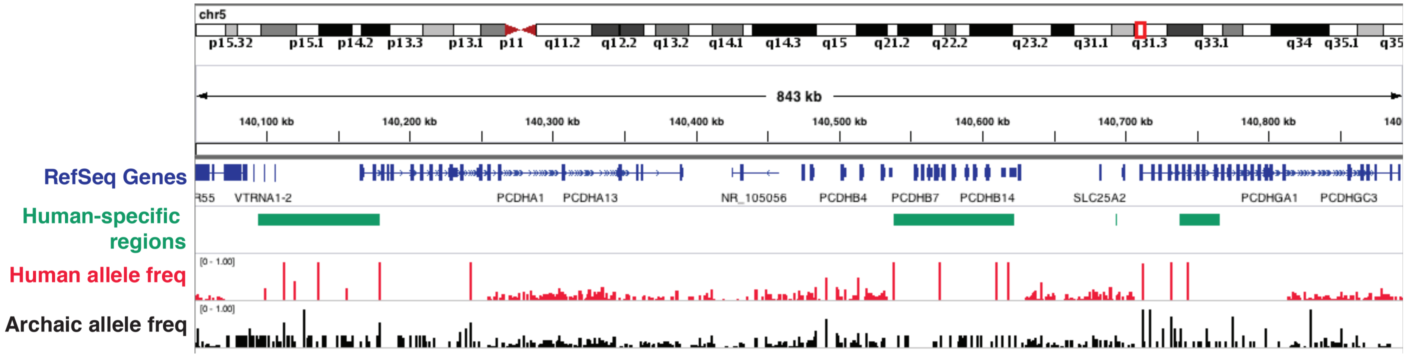


Figure 4.18: Screenshot of the region around the three protocadherin gene clusters on chromosome 5. Human-specific regions are shown in green, human derived allele frequencies are shown in red, and archaic derived allele frequencies are shown in black. Humans have little allelic diversity in human-specific regions, with most derived alleles being fixed or high frequency.

Chapter 5

Conclusion

I have introduced two new methods for localizing introgressed segments in the genomes of hybrid individuals, one suited to low-coverage data, especially from non-model organisms where large panels of phased reference data are unavailable, and another suited to large panels of phased data. Although AD-LIBS, the hidden Markov model-based approach, has a fairly specific use case, SARGE, the ancestral recombination graph inference program, can be used for a wide variety of analyses not limited to admixture mapping. Both programs are freely available to other researchers at `https://github.com/nkschaefer`.

I have used these two programs to analyze the inheritance of introgressed ancestry in two different species, brown bears and modern humans. These two cases have some qualities in common. First, both sets of hybridizing species diverged very recently (350-500 kya [108] for brown and polar bears and 500-600 kya for humans and archaic hominins [163]). Second, both involve gene flow from a relatively genetically homogenous population into one with higher nucleotide diversity, and gene flow in both cases appears to have been (at least mostly) unidirectional [56, 19]. Therefore, comparing and contrasting these two data sets can be illuminative.

The genomic distribution of hybrid ancestry in in both species generally follows the accepted wisdom that hybrid ancestry is somewhat deleterious [65]; introgression-free regions are enriched for functional elements in both species, although evidence for adaptive introgression exists in both species as well. Intriguingly, we find evidence of adaptively introgressed immune-related genes in both species. This could be due to the fact that diversity in MHC alleles is useful for defending against a variety of pathogens, and introgression is therefore a universal source of divergent MHC alleles. It could also be the case that both Neanderthals and polar bears were the first species to become specialized for life in a particular environment, and the defenses they evolved against local pathogens later proved beneficial to those they hybridized with as they colonized new environments. This has been suggested previously in the human-Neanderthal case [4]. The functions of other potentially adaptively introgressed regions provides an interesting topic for future study. These include genes involved in DNA repair, spermatid development, and circadian rhythm in bears, and zinc finger transcription factors (Neanderthal) and a gene involved in phototransduction (Denisovan), as well as transmembrane proteins of unknown function (both) in humans.

Genomic regions free of introgressed ancestry in both species (and also free of incomplete lineage sorting in humans) provide another interesting avenue for further study. These regions in both species contain genes involved in neurodevelopment, although they are likely to be different sets of genes. Looking at derived alleles fixed in both species within these regions, and those alleles' proximity to functionally consequential genome features, could shed light on the significance of these regions. This would be especially useful in the case of human-specific regions within the three protocadherin clusters, where alternative splicing regulation may have been important in the differentiation of modern from archaic hominins. Another avenue of analysis is to look for evidence of hybrid incompatibility in these regions, for example, by locating interacting sets of genes with members in separate introgression-free regions. This could help determine whether there are certain genes (or pathways) universally involved in hybrid incompatibility, or if hybrid incompatibility loci in humans and bears are fundamentally different from those already discovered (i.e. in *Drosophila* [151]).

Ancestral recombination graph inference using large panels of data also promises to enable many new types of population genetic analyses. As many existing anlyses act on statistics that are in effect summaries of the ARG [196], these could be re-created using ARG output, and perhaps become more accurate in the process. The ARG could also be used to disentangle complicated demographic questions, such as the true phylogenetic position and demographic history of Denisovans.

Appendix A

Supplementary methods for Chapter 2

A.1 HMM transition probabilities: computing *z*

Transition probabilities in AD-LIBS depend primarily on the values of g, the number of generations since admixture, r, the recombination probability per site per generation, and z, the probability of resampling the same ancestral recombination event twice in a single individual (see Materials and Methods and Figure 2.10). While g and r are simple to conceptualize and compute, z is more challenging, and we describe the computation of z in this section.

The Wright-Fisher model, in which populations are modeled as constant-sized unstructured groups of randomly mating individuals with non-overlapping generations, can be used to set up a calculation of z. Hartl and Clark [66] conceptualized allele frequencies under genetic drift in a Wright-Fisher population as a Markov chain problem: they described a state space of size 2N, where each state corresponds to an allele frequency in a diploid population of size N. Each entry in the transition probability matrix T, where is the probability of transitioning from i to j copies of an allele in a diploid population of size N in one generation, is given by $T_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(\frac{2N-i}{2N}\right)^{2N-j}$. Given that recombination events arise with probability r, then in this case $T_{0,j>0} = 2Nr^{j}$ and $T_{0,0} = 1 - \sum_{j=1}^{2N} T_{0,j}$ and the probability of a single recombination being resampled in an individual after allowing g generations of random drift is $\sum_{j=2}^{2N} T_{0,j}^{g}(\frac{j}{2N})^{2})$. While conceptually appealing, this formulation quickly becomes computationally intractable when there is a large population size N or number of generations since admixture g.

To improve computational efficiency, continuous approximations to the genetic drift problem, based on the mathematics of diffusion, have been proposed. [90] made an influential contribution, but his solution did not account for the so-called "absorbing" states of loss and fixation. We implemented the solution to the pure drift equation, without selection or migration, given by [120]. Given a population size N, a number of generations t, and an initial allele (or recombination) frequency of $1/2N$, the probability density of allele frequency x in the t^{th} generation is given by $\sum_{n=0}^{\infty}[(2n+3)(n+1)(n+2)_2F_1(-n,n+3;2;\frac{1}{2N})_2F_1(-n,n+3;2;x)e^{\frac{-(n+1)(n+2)t}{4N}}$, the probability of allele loss is a Dirac delta function with weight $(1-\frac{1}{2N})\sum_{n=0}^{\infty}[(\frac{1}{2N})(2n+3)_2F_1(-n1,n+3;2;\frac{1}{2N})e^{\frac{-(n+1)(n+2)t}{4N}}$, and the probability of allele fixation is a Dirac delta function with weight $(\frac{1}{2N})\sum_{n=0}^{\infty}[(1-\frac{1}{2N})(2n+3)-1)_{2}^{n+1}F_1(-n,n+3;2;\frac{1}{2N})e^{\frac{-(n+1)(n+2)t}{4N}}$, where ${}_2F_1(a,b;c;z)$ is the hypergeometric function.

In our implementation, we terminate the infinite sums when newly added terms are lower than 10^{-20} and we divide the allele frequency spectrum into 500 bins (or 2N-1 bins, if 2N-1 ¡ 500), plus one bin each for the probabilities of allele loss and fixation. For any given number of generations t, we use the above equations to compute probability density at the upper and lower limit of each bin, obtain probabilities from probability density via the trapezoid rule, and map the probability of each bin to the frequency value at its midpoint, giving a vector of probabilities V and a vector of allele frequencies F, each with 502 entries corresponding to bins indexed by b, where $b \sim [0, 501]$. We then iterate over each generation $0 \leq i < g$, and compute V and F for an allele that arose in generation i at frequency $\frac{1}{2N}$ and underwent $t = g - i$ generations of drift. For each of these vectors, then, we compute the per-site probability of resampling this allele (or recombination event) twice in an individual in generation g: if r is the recombination probability per site, $b \sim [0, 501]$ is the bin index, V_b is the probability of a given bin, and F_b is the mean allele frequency associated with that bin, then the probability of resampling the same allele or recombination event twice in an individual is $\sum_{b=0}^{501} rV_bF_b^2$. The overall probability of resampling the same ancestral recombination event twice in an individual is then $z = \sum_{i=0}^{g-1} [\sum_{b=0}^{501} rV_{i,b}F_{i,b}^2]$.

With these values, the transition probabilities between the three ancestry states can be computed per site as in Figure 2.10. To transform these into transition probabilities between windows, each can be multiplied by the window size w, so that the transition probabilities between ancestry states are as follows:

$$p(AB|AA) = 2grw(1-p)(gpr-gr+1)$$

$$p(BB|AA) = w(1-p)(-g^2pr^2+g^2r^2+z)$$

$$p(AA|AB) = 2pw(g^2pr^2-g^2r^2+gr+z)$$

$$p(BB|AB)+2w(1-p)(-g^2pr^2+gr+z)$$

$$p(AA|BB) = wp(g^2pr^2+z)$$

$$p(AB|BB) = 2gprw(1-gpr)$$

A.2 HMM transition probabilities

In addition to the three ancestry states and three skip states described in Materials and Methods, AD-LIBS also has start and end states. The probability of transitioning to the end state from any other state is $1/l$, where l is the number of windows in a genomic input sequence. The transition probabilities from the start state are based on the percent ancestry the admixed population derives from population A, p, and the distribution of population A-like bases under Hardy-Weinberg equilibrium (s is the skip probability):

$$p(AA|start) = p^2(1-s)$$

$$p(sAA|start) = p^2 s$$

$$p(AB|start) = 2p(1-p)(1-s)$$

$$p(sAB|start) = 2p(1-p)s$$

$$p(BB|start) = (1-p)^2(1-s)$$

$$p(sBB|start) = (1-p)^2 s$$

When all other probabilities have been determined, the probability of any ancestry state transitioning to itself is defined as one minus the sum of all other transition probabilities from that state. This is also how transition probabilities from skip states back to their associated ancestry states are determined. For example:

$$p(AA|AA) = 1 - p(AB|AA) - p(BB|AA) - p(sAA|AA) - p(end|AA)$$

$$p(AA|sAA) = 1 - p(sAA|sAA) - p(AB|sAA) - p(BB|sAA) - p(end|sAA)$$

AD-LIBS can also optionally account for the approximate general reduction in heterozygosity due to genetic drift, as formulated by Hartl and Clark [66]: given an initial level of heterozygosity H_0, a population of N diploid individuals, the level of heterozygosity after t generations of genetic drift is $H_t \approx H_0 e^{\frac{-t}{2N}}$. This is approximated in AD-LIBS using t =the number of generations since admixture g and population size N. We then reduce the probability of transitioning from any state to the heterozygous state according to this predicted reduction in heterozygosity, and compensate for this by increasing the probability of staying in or transitioning to one of the homozygous ancestry states:

$$p(AA|AA) = p(AA|AA) + (\frac{p(AA|AA)}{p(AA|AA) + p(BB|AA)})(p(AB|AA) - p(AB|AA) * (H_t/H_0))$$

$$p(AA|sAA) = p(AA|sAA) + (\frac{p(AA|sAA)}{p(AA|sAA) + p(BB|sAA)})(p(AB|sAA) - p(AB|sAA) * (H_t/H_0))$$

$$p(BB|AA) = p(BB|AA) + (\frac{p(BB|AA)}{p(AA|AA) + p(BB|AA)})(p(AB|AA) - p(AB|AA) * (H_t/H_0))$$

$$p(BB|sAA) = p(BB|sAA) + (\frac{p(BB|sAA)}{p(AA|sAA) + p(BB|sAA)})(p(AB|sAA) - p(AB|sAA) * (H_t/H_0))$$

$$p(AA|BB) = p(AA|BB) + (\frac{p(AA|BB)}{p(AA|BB) + p(BB|AA)})(p(AB|BB) - p(AB|BB) * (H_t/H_0))$$

$$p(AA|sBB) = p(AA|sBB) + (\frac{p(AA|sBB)}{p(AA|sBB) + p(BB|sAA)})(p(AB|sBB) - p(AB|sBB) * (H_t/H_0))$$

$$p(BB|BB) = p(BB|BB) + (\frac{p(BB|BB)}{p(AA|BB) + p(BB|BB)})(p(AB|BB) - p(AB|BB) * (H_t/H_0))$$

$$p(BB|sBB) = p(BB|sBB) + (\frac{p(BB|sBB)}{p(AA|sBB) + p(BB|sBB)})(p(AB|sBB) - p(AB|sBB) * (H_t/H_0))$$

$$p(AA|AB) = p(AA|AB) + (\frac{p(AA|AB)}{p(AA|AB) + p(BB|AB)})(p(AB|AB) - p(AB|AB) * (H_t/H_0))$$

$$p(AA|sAB) = p(AA|sAB) + (\frac{p(AA|sAB)}{p(AA|sAB) + p(BB|sAB)})(p(AB|sAB) - p(AB|sAB) * (H_t/H_0))$$

$$p(BB|AB) = p(BB|AB) + (\frac{p(BB|AB)}{p(AA|AB) + p(BB|AB)})(p(AB|AB) - p(AB|AB) * (H_t/H_0))$$

$$p(BB|sAB) = p(BB|sAB) + (\frac{p(BB|sAB)}{p(AA|sAB) + p(BB|sAB)})(p(AB|sAB) - p(AB|sAB) * (H_t/H_0))$$

$$p(AB|AA) = p(AB|AA) * (H_t/H_0)$$

$$p(AB|sAA) = p(AB|sAA) * (H_t/H_0)$$

$$p(AB|AB) = p(AB|AB) * (H_t/H_0)$$

$$p(AB|sAB) = p(AB|sAB) * (H_t/H_0)$$

$$p(AB|BB) = p(AB|BB) * (H_t/H_0)$$

$$P(AB|sBB) = p(AB|sBB) * (H_t/H_0)$$

Since many transition probabilities depend upon the window size w, we risk obtaining sums of transition probabilities from individual states that are greater than 1 when the window size is large. We take two steps to prevent this: first, we cap both the skip probability s and the probability of ending the sequence $1/l$ at maximum values of 0.05. In our experience, this number is high enough to still allow the model to detect skipped windows and to end sequences in the appropriate places. Second, given the other model parameters, we calculate the maximum possible window size that will allow the sum of transition probabilities out of each state to fall below 1. If the user chooses a window size that exceeds this threshold, the program notifies the user and exits.

Our implementation also considers some of the unique properties of the X chromosome [190]. Since there are fewer copies of the X chromosome than any autosome in a given population, the X chromosome only recombines approximately $2/3$ as often as the autosomes. We therefore build a different model on chromosome sequences or genomic scaffolds specified to belong to the X chromosome: on these sequences, the r parameter is taken to be $2/$ of its default, autosomal value: $r_x = (2/3)r$. Furthermore, the z parameter quantifies genetic drift and thus depends on the population size N, but the effective population size of the X chromosome is $3/4$ of the autosomal value, again owing to there being fewer copies of X chromosomes than autosomes in circulation in a population. We therefore recompute z using the same technique described earlier, but with $N_x = (3/4)N$, to give z_X, used in place of z on X chromosome sequences. Finally, users can specify which individuals are male, and for these individuals a haploid model is created instead of the default, diploid model on X chromosome sequences. In this model, there is no heterozygous ancestry state (AB) or heterozygous skip state (sAB), and transitioning from one ancestry state to the other only requires a single recombination event, followed by sampling the next base from the set of bases of the opposite type of ancestry:

$$p(BB|AA) = gr(1-p)$$

$$p(AA|BB) = grp$$

The transition probabilities from the start state are also different in the haploid X chromosome model, due to the absence of the heterozygous state:

$$p(AA|start) = p(1-s)$$

$$p(sAA|start) = ps$$

$$p(BB|start) = (1-p)(1-s)$$

$$p(sBB|start) = (1-p)s$$

If the organisms under study follow the ZW rather than the XY sex determination system, the same concept holds, except that the “X chromosome sequences” supplied to AD-LIBS should be the names of sequences or genomic scaffolds belonging to the Z chromosome, and the “males” specified to AD-LIBS should actually be females.

A.3 Emission probability distributions: expectation

For a description of emission "scores" used by AD-LIBS, see Materials and Methods. In this section, we describe the expected distributions of these scores used by AD-LIBS.

The expected distribution of IBS tracts between two haplotypes depends only on the parameter π, or average nucleotide diversity per site between those two haplotypes. We therefore, as a first step, compute the average genome-wide nucleotide diversity per site within population A, referred to as π_A, within population B, referred to as π_B, and between the two populations, referred to as π_{AB}. Given that π describes how often one expects to see a nucleotide difference, $1/\pi$ is the expected length of a haplotype before a difference is observed. IBS tract lengths tend to follow an exponential distribution with π as a parameter. Since our model considers samples of IBS tracts within genomic windows, however, we expect our mean IBS tract lengths to follow a normal distribution with a mean equal to the expected value and a standard deviation equal to the expected sample standard deviation, with the expected number of samples equal to the window size times π:

$$\mu = \frac{1}{w\pi}$$
$$\sigma = \frac{1}{\pi\sqrt{w\pi - 1}}$$

There are five such distributions to consider when computing scores. In genomic regions where a hybrid individual is homozygous for population A ancestry, sample mean IBS tracts with population A follow such a distribution with $\pi = \pi_A$ and sample mean IBS tracts with population B follow such a distribution with $\pi = \pi_{AB}$. Where there is homozygous population

B ancestry, mean IBS lengths with population A have $\pi = \pi_{AB}$ and IBS lengths with population B have $\pi = \pi_B$. In heterozygous regions, mean IBS lengths with population A have $\pi = (\pi_A + \pi_{AB})/2$ and mean IBS lengths with population B have $\pi = (\pi_B + \pi_{AB})/2$. For each of these five distributions, we calculate normal μ and σ as above and then transform the standard deviation into the equivalent for a log-normal distribution:

$$\sigma' = \sqrt{log(\frac{\sigma^2}{\mu^2} + 1)} = \sqrt{log(\frac{w^2}{2\pi - 1} + 1)}$$

We then use these values to compute the parameters for the three emission probability distributions. Since each is the ratio of two distributions, the variances of the emission probability distributions are the sum of the variances of the two mean IBS tract length distributions they compare:

$$\begin{aligned} \mu_{AA} &= log(\frac{1}{w\pi_A}) - log(\frac{1}{w\pi_{AB}}) \\ \mu_{BB} &= log(\frac{1}{w\pi_{AB}}) - log(\frac{1}{w\pi_B}) \\ \mu_{AB} &= log(\frac{1}{w(\pi_A + \pi_{AB})/2}) - log(\frac{1}{w(\pi_B + \pi_{AB})/2}) \end{aligned}$$

In AD-LIBS, all three emission probability distributions are modeled as normal distributions, due to successful performance on training data. We also reserve a specific value to be used as a “skip” score; distributions are set such that the skip score has zero probability under all three emission probability distributions. We also create an emission probability distribution for the three skip states that is only capable of emitting this value.

As with the transition probabilities, sequences belonging to the X chromosome present an edge case in which changes to the model are necessary. Since the effective population size of the X chromosome is $3/4$ that of the autosomes, we multiply π_A, π_B, and π_{AB} by $3/4$ before calculating the emission probability distribution parameters.

Whereas the need to keep transition probabilities below 1 sets an upper bound on window size, our expected emission probability distributions set a lower bound on window size. If π_A is the lowest value of π, then per the standard deviation calculations above, we require $w\pi_A > 1$ to avoid division by zero (standard deviation calculations can involve division by zero if the expected number of IBS tract observations in a given window is less than one). It is generally preferable to choose the smallest possible window size for which there is a reasonable lack of overlap among emission probability distributions (see Materials and Methods).

Appendix B

Supplementary methods for Chapter 3

B.1 Adding nodes to the ARG

Every node in the ARG must be "anchored" at one or more genomic positions. This is because each node's start and end coordinates depend on these positions, along with the propagation distance p. When a node's range is interrupted by a new node with which it fails the four haplotype test being created in the middle of its range, for example, the set of genomic sites the node "owns" are used to determine its new range. Additionally, to make it easier to look through the ARG, we store a mapping of sites to ARG nodes with those sites. Since it is set up this way, we do not allow any node to be created if the site at which it is originally anchored already is tied to an existing ARG node with which it fails the four haplotype test.

When a new node is to be created, we first check to see if it can be merged with an existing node. This is the case if there already exists a node with the same clade whose range overlaps the new node, and if no node that fails the four haplotype test with either of these nodes exists at any site in between them. If this is the case, then rather than create the new node, the existing node's range is expanded to take over the new node's range, and the new node's sites are added to the old node. A special case exists when the new node's range overlaps with two existing nodes (one upstream and one downstream). In this case, the two node's ranges are expanded to one another's outer limits, and then the two nodes are merged; sites belonging to the new node are then given to the merged existing node instead. It is also possible that a new node's range falls completely within the range of a single existing node; in this case, the new node's sites are given to the existing node.

If a new node does not merge with an existing node, then the first step is to determine its range by detecting all four haplotype test failures it will encounter. All nodes whose end coordinate plus p is greater than or equal to the new node's start coordinate minus p, or whose start coordinate minus p is less than or equal to the new node's end coordinate plus p, undergo the four haplotype test with the new node. Next, all relevant node's coordinates are adjusted according to these four haplotype test failures. If the new node is in the middle of an existing node's range and the new node fails the four haplotype test with that node, then that existing node is split into two nodes. Otherwise, start and end coordinates are adjusted so that if the new node is closer to an existing node than every other node with which the existing node fails the four haplotype test, then the end coordinates of both node's ranges are set to the furthest-reaching sites owned by those nodes. In other words, if p = 50 and node 1 with range [0,100] and a site at position 50 fails the four haplotype test with node 2, which has range [100, 200] and a site at position 150, then node 1 will now have range [0,50] and node 2 will now have range [150,200]. In another case, where node 3 has range [0,200] and sites 50 and 150 and fails the four haplotype test with node 4, which has range [50,150] and has site 100, node 3 will be split into one node with range [0,50] and site 50, and another node with range [150, 200] and site 150. Node 4 will have its range reduced to one site, [100,100] (Fig. B.1).

Once all node ranges have been adjusted according to four haplotype test failures, then all parent/child relationships are created. To do this, a depth-first search is performed across the ARG down from the root, across the entire range of the new node. If a node's clade is a superset of the new node's clade, then that node's sub-ARG is searched, and then the superset node is added to a set of parent nodes. Once all parent nodes are collected, they are sorted by increasing clade size and one by one added as parents of the new node. If the new node is a parent of an existing child of the parent node at a given range of sites, then that child is removed from the parent over that range and stored to later be added as a child of the new node (children are added in order of decreasing clade size). When adding parent/child edges, existing edges are sometimes split. All parent/child edges are bidirectional – the child node must store the same edge with the same coordinates to the parent as the parent stores to the child.

After all parent/child edges are added, recombination edges are added to the graph. Beginning with the set of four haplotype test-failing nodes stored from earlier (when start and end coordinates were adjusted and nodes were split as necessary), these nodes are divided into those upstream and those downstream of the new node. The upstream and downstream nodes are then both sorted by increasing distance from the new node. If there are two four haplotype test-failing nodes B and C downstream of node A, where B is closer to A than C, and B and C also fail the four haplotype test, then node C will not be connected to A with recombination edges. Otherwise, recombination edges are added. In this case, the ranges of the two nodes are expanded to the limits of the previously-solved recombination events, and any new clades that can be inferred are created as new nodes (see "Solving ancestral recombination events" section).

The process of adding recombination edges begins with checking to see if edges already exist between the two nodes (possibly from a solved or unsolvable recombination event). If so, nothing is done. If there are edges representing a solved recombination event connecting one of the two nodes to another node, and that recombination event's node matches one of the potential γ clades that could explain the new four haplotype test failure, then solved recombination edges are added between the two new nodes and the node for the γ clade from that solved recombination event.

If no previously-solved or unsolvable recombination event can explain the four haplotype test failure between any two nodes, then candidate recombination edges are added. In this case, each candidate moving clade is created as a node that is not inserted into the ARG via parent/child edges, but stored in a separate set of potential nodes. If a candidate γ clade already exists as a node in this set, then it is reused and its start and end coordinates are set to the narrowest interval between upstream and downstream nodes that fail the four haplotype test, with that γ clade connecting them. If there is more than one node in the set of candidate γ clade nodes whose range overlaps with the interval between the upstream and downstream nodes that fail the four haplotype test, then all such nodes are merged into one. Once an appropriate γ clade node is found, candidate recombination edges are added connecting it to both the upstream and downstream node that fail the four haplotype test.

Once start and end coordinates are adjusted, parent/child relationships are added, and recombination edges are added, the new node is “finished” and stored in the data structures that allow lookup by genomic position and clade.

A special case for adding nodes exists for clades where every haplotype shares the derived allele. These sites can only contribute to the branch length of the root node. Therefore, we store a single root node whose start and end coordinates span the entire chromosome. If a mutation is observed for which every haplotype shares the derived allele, it is added to the root node. Inferred (non-mutation) sites with this clade are ignored and not added to the root node.

Similarly, clades where every haplotype shares the ancestral allele are not informative for the ARG and are skipped altogether.

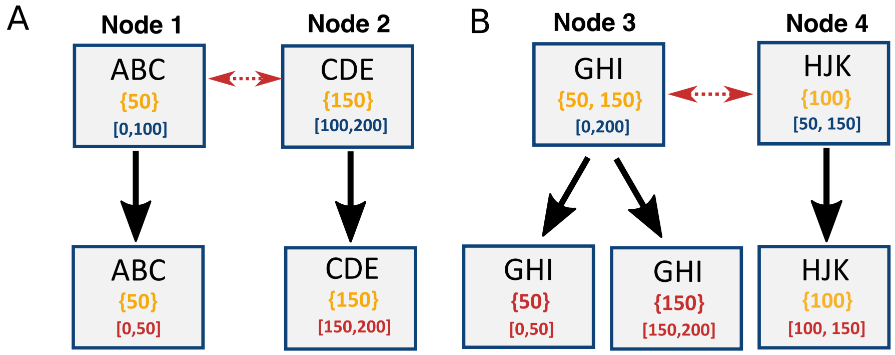


Figure B.1: How node indices are adjusted when four haplotype test failures are encountered. Black letters represent clades, yellow numbers in curly braces represent site indices, and blue numbers in brackets represent start and end coordinates (inclusive). Red text indicates an adjusted value. Red arrows show four haplotype test failures, and black arrows represent changes made to nodes. A: a simple case where the furthest donwnstream site owned by node 1 (50) is upstream of the furthest upstream site owned by node 2 (150). In this case, node 1’s end coordinate is set to its furthest downstream site, and node 2’s start coordinate is set to its furthest upstream site. B: Node 4 interrupts the range of node 3. Node 3 must be split into two nodes, and all three resulting nodes must have their ranges adjusted.

B.2 Solving ancestral recombination events

The process of “solving” ancestral recombination events consists of finding a node with unsolved recombination edges connecting it to one or more nodes downstream, finding a subgraph of the ARG containing other nodes involved in this or possibly other recombination events, filtering the subgraph so that it only describes a single recombination event, and then choosing the most likely node that could explain the recombination event (similar to the “two-trees” algorithm, Fig. 3.8). Finally, the chosen γ node is added to the ARG as a standard tree node, the start and end coordinates of all nodes involved are adjusted to account for the inferred recombination event, and any nodes that do not exist in the ARG but whose existence is implied by the recombination event are created (Fig. 3.2B). This process is the core of the ARG inference algorithm, as it allows for the creation of nodes not directly observed in the input SNP data.

One concept used by several stages of this algorithm is that of tree-compatibility between two nodes. Two nodes are tree-compatible if, according to their clades, genomic positions, and genomic positions of their four haplotype test-failing partner nodes, they can both exist in the same tree. At this stage in ARG building, start and end coordinates have not yet been finalized, so we cannot define compatibility based on coordinates alone. However, if two nodes already have overlapping start and end coordinates, then they must be compatible. In another simple case, two nodes that fail the four haplotype test cannot be compatible. Otherwise, we must rely, for upstream nodes, on the upstream-most start coordinate of all downstream tree nodes connected to the node via recombination edges. Likewise, for downstream nodes, we consider the downstream-most end coordinate of all upstream tree nodes connected to the node via recombination edges. We refer to this value, in both cases, as the "closest recombination partner" of the node; this determines how far, in theory, the nodes end coordinate (if upstream of a recombination event) or start coordinate (if downstream of a recombination event) could be extended in the ARG. Whether or not any two tree nodes are tree-compatible depends on the location of both nodes closest recombination partners. Imagining node A upstream of node B, node B must be upstream of node A's closest downstream recombination partner and node A must be downstream of node B's closest upstream recombination partner to be tree-compatible (Fig. B.2).

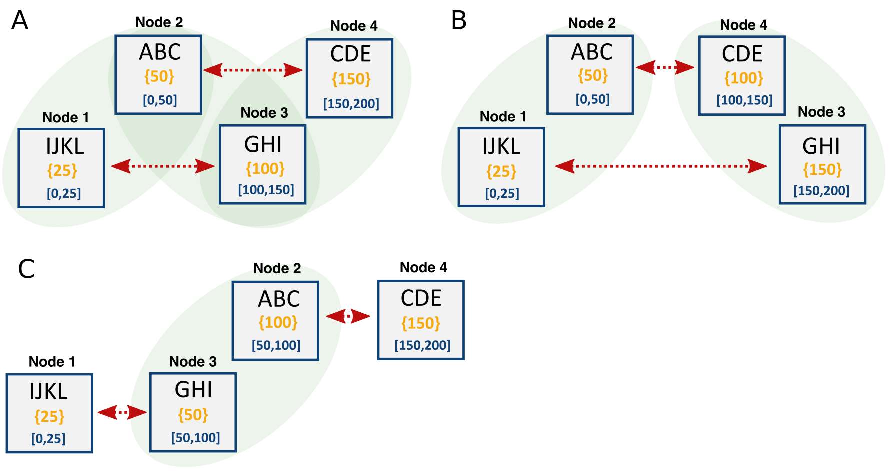


Figure B.2: Tree-compatibility in three different situations. Gray squares are nodes, black letters are clades, yellow numbers in curly braces are genomic positions, and blue numbers in brackets are start/end coordinates. Red arrows indicate four haplotype test failures, and green ovals denote tree compatibility. A: three pairs of nodes are compatible (can belong to the same trees as each other). B: only two pairs of nodes are tree-compatible. C: Only one pair of nodes is tree-compatible.

The central problem in solving ancestral recombination events is to find a subgraph of the ARG containing a set of tree-compatible upstream nodes U, a set of tree-compatible downstream nodes D, and a set of candidate γ nodes L that connect together nodes in U and D. Ideally, U, D, and L should correspond to a single ancestral recombination event; however, in practice, there are situations in which a single event is difficult or impossible to distinguish from multiple events (Fig. B.3). We will hereafter refer to this ARG subgraph used to infer ancestral recombination events as a "recombination graph."

Collecting a recombination graph begins with a "key" node k, which is a tree node in the ARG with unsolved recombination edges to downstream nodes. To begin, we visit each candidate γ node downstream of k and add it to L. Next, we visit all upstream tree nodes connected to every node in L and add them to U, if they are tree-compatible with k. We then follow all recombination edges from nodes in U, through candidate γ nodes, to tree nodes with start coordinates downstream of the end coordinate of k. These nodes are added to D, and all candidate γ nodes along their paths to nodes in U are added to L. We then revisit nodes in U; any that are not connected via recombination edges to nodes in D are removed from U.

Because ranges of candidate γ nodes are difficult to determine and subject to change, it is possible at this stage for L to contain multiple γ nodes that really represent the same candidate γ clade. To merge identical γ nodes, we compile all sets of nodes in L with the same clade, which also have start and end coordinates whose ranges fall within the range defined by the minimum end coordinate of nodes in U and the maximum start coordinate of nodes in D. If there is such a set of nodes M in L, and the nodes in M are connected to one or more upstream nodes that are tree-compatible with k, then these compatible upstream nodes are added to U, we follow their recombination edges through nodes in M to downstream tree nodes and add them to D, and then all nodes in M are merged together into a single node, which is added to L. We then undertake

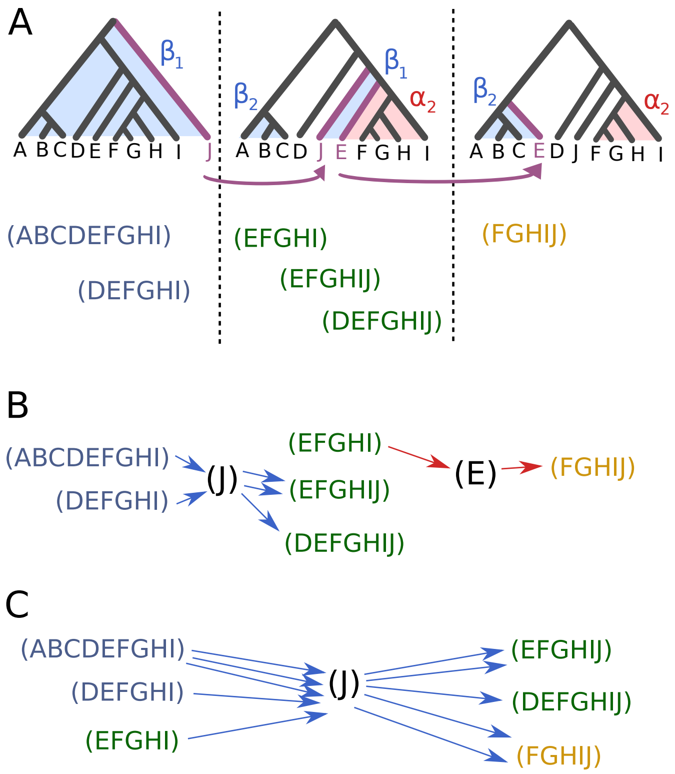


Figure B.3: An example case in which multiple ancestral recombination events may be considered as one. A: The true ARG across three adjacent genomic regions. Clades involved in recombination are marked α and β; subscripts denote the recombination event (first or second) to which they correspond. Clades observed in SNP data appear below each tree in the order in which they are observed; colors mark the true tree to which each clade belongs. Purple branches are true γ clades, and purple arrows show ancestral recombination events. B: The correct grouping of nodes path through them in a recombination graph. First, (J) moved downward from (ABCDEFGHIJ) to (EFGHIJ). Then, (E) moved from (EFGHI) to (ABCE). C: A likely incorrect inference made, if nodes are not grouped correctly into trees. It appears most parsimonious to say that (J) moved down from (ABCDEFGHIJ) in the first tree to (FGHIJ) in the third tree, skipping the middle tree altogether. If this choice is made, genomic positions for the ancestral recombination event will also be wrong, as it chooses the narrowest possible interval, which would place it between the first and second tree. Note that observing the clade (ABCE) in the third tree might help avoid this problem.

a simple filtering step designed to remove nodes involved in recombination events upstream or downstream of the main ancestral recombination event we are trying to solve. First, we sort all nodes in U by end coordinate and all nodes in D by start coordinate. If any node in U has a closest downstream recombination partner upstream of the furthest-upstream node in D, then the upstream node is removed from U. Likewise, if any node in D has a closest upstream recombination partner downstream of the furthest-downstream node in U, then the downstream node is removed from D. After this process, if any node in U lacks recombination edges to all nodes in D, or if any node in D lacks recombination edges to all nodes in U, then that node is removed from the set to which it belongs.

The next step is to filter U, D, and L to a set of nodes describing only a single recombination event. This is the most memory intensive part of the algorithm, as it must explore a large set of choices. At this stage, the recombination graph is likely to represent several different recombination events, which must be pared down to one before a branch movement can be inferred; the final recombination graph should have a set of fully tree-compatible U nodes and a set of fully tree-compatible D nodes that are all tree-incompatible with each other. In this step, we enumerate all possible inclusion/exclusion decisions and try to make the best one, or else defer decision making until later, when other ancestral recombination events will have been solved, reducing the number of recombination edges and thus making the recombination graph simpler the next time it is visited.

We store a collection of pairs of "decision" sets, each representing a choice to make either eliminate all nodes in the first set or eliminate all nodes in the second set. We call this collection of decision set pairs C, each consisting of (u,d), which is a pair of sets u is a set of nodes in U that can be eliminated, and d is a set of nodes in D that can be eliminated instead. For each node in U, the set of all nodes in D that are tree-compatible with it are gathered, and for each node in D, the set of all nodes in U that are tree-compatible with it are gathered. This forms a decision set pair; if starting with an upstream node, then u contains only that node and d contains all tree-compatible nodes in D. If starting with a downstream node, then d contains only that node and u contains all tree-compatible nodes in U. All sets are then combined into as few decision sets as possible: if in a preexisting decision set pair (u’, d’), u’ $\subseteq$ u and d’ $\subseteq$ d, then elements of u are added to u and elements of d are added to d (the same operation is carried out if u and d are flipped). Otherwise, if d = d’, then elements of u are added to u.

We also store a collection of pairs of “partner” sets P. These are the opposite of decision sets: they are pairs of nodes for which including one requires also including the other. The logic behind partner sets is that to include a node in U or D in the final recombination graph, we must also include its closest recombination partner. These sets of nodes are collapsed into as few as possible: for each set of an upstream node and its closest downstream partner, or a downstream node and its closest upstream partner (u,d), we visit each existing partner set (u’,d’); if u and u share members or d and d share members, then the node in u is added to u and the node in d is added to d. If this cannot be done, (u,d) is added to P.

Given all choices described by the node sets in C and P, we now build a set S, where each entry is a set of upstream and downstream nodes (U,D) that could describe a single recombination event. To populate S, we first enumerate all possible choices in C, then filter according to the constraints imposed by the pairs in P. If S is empty, we take a pair (u,d) in C, create two sets representing the two possible choices: ($\overline{u}$, d) and (u,$\overline{d}$), and add them to S. Otherwise, for each (u,d) in C, we visit all pairs (u’,d’) in S. If u and u share members and d and d share members, then we create the new pairs (u’$\setminus$ u, d’) and (u’,d’$\setminus$ d) and add them both to S, if both sets in the pair are non-empty. We then filter the node sets in S using the constraints in P. For each (u,d) in S, we visit each (u'd') in P. If u and u share members or d and d share members and it is not the case that u and d are both supersets or equal to their u and d counterparts, then members of u are removed from u and members of d are removed from d. If either u or d = ∅, then is removed from S.

We now have in S a set of choices of full recombination graphs. In the spirit of parsimony, we choose the set with the highest total node count (the recombination graph containing the most possible four haplotype test failures). If there is a tie, we choose the set of upstream and downstream nodes (u,d) with the shortest genomic span, defined by the furthest downstream end coordinate in D minus the furthest upstream start coordinate in U.

Before solving the recombination event, we check pairs of upstream and downstream nodes in the recombination graph and eliminate any pair whose four haplotype test failure could be explained by a previously-solved recombination event. Previously-solved recombination events are stored in a special type of recombination edge. By iterating through all previously-solved recombination events and transforming clades accordingly (removing γ clades from α clades or adding γ clades to β clades going downstream; doing the reverse going upstream), we can see if the two clades that fail the four haplotype test still fail the four haplotype test after being transformed according to previously-solved recombination events. If not, they are removed from the recombination graph.

We also check the recombination graph for evidence that it describes multiple recombination events, before attempting to solve it. There are many boundary cases in which it is impossible to determine if a given recombination graph describes one or more recombination events (Fig. B.3). Because of this, we keep track of how many times a given k node has been visited in the interest of solving recombination events. If we gather a recombination graph around k and see evidence that the graph might describe multiple recombination events, we stop trying to solve the event. Other recombination graphs with different k nodes might then be less ambiguous, and solving them can eliminate recombination edges that complicated solving the graph around k. The next time k is visited, then, the recombination event may be easier to solve. Regardless, the second time a given node is visited, its recombination events are "force solved," meaning that checks for evidence of multiple recombination events are skipped altogether.

The first sign that a recombination graph might describe multiple recombination events is if the downstream most node in U, U_d, does not fail the four haplotype test with the upstream-most node in D, D_u. If this is the case, then we could gather two alternative recombination graphs. One would exclude U_d and instead add to D other further-downstream recombination partners of other nodes in U; the other would exclude D_u in favor of other further-upstream recombination partners of other nodes in D. For both of these alternative recombination graphs, we calculate the genomic span of all nodes they contain; if either has a smaller span than the current chosen recombination graph, we take this as evidence of multiple recombination events and defer solving the recombination graph.

The last step before solving the recombination graph is to filter the nodes in L to a set of γ clades that are compatible with tree nodes already in the ARG. We therefore traverse nodes in the ARG that own SNP positions that fall between the downstream-most position owned by the downstream-most node in U and the upstream-most position owned by the upstream-most node in D. If any such nodes clade fails the four haplotype test with a candidate γ clade, the node with that clade is removed from L.

At this stage, we have a reasonably confident set of nodes U, D, and L, representing tree nodes upstream and downstream of a single ancestral recombination event, along with choices for a clade that changed positions between trees. The final step is to count the number of edges connecting each node in L to each pair of upstream and downstream nodes in U and D. If one node has the most edges, it is chosen as the most likely branch movement, analogous to the "two trees" algorithm (Fig. 3.2). If there is a tie, then we compute the maximum distance along the genome each candidate γ clade belonging to nodes in L could persist in the ARG (propagated in both directions) before encountering a four-haplotype test failure (or reaching the propagation distance parameter p). We also compute this number for each other clade implied by the γ clade: the difference of each upstream α clade and γ and the union of each upstream β clade and γ, propagated upstream, as well as the union of each downstream α clade and γ and the difference of each downstream β clade and γ, propagated downstream. The mean of all of these numbers gives a measure of how compatible a given γ clade is with the current ARG topology; therefore, the γ clade with the highest mean clade persistence distance across all clades it implies is chosen as the correct clade.

With a γ clade chosen, the remaining inferences to make are the genomic coordinates at which the recombination event happened, and which clade the γ belonged to immediately before and after the recombination event. To determine the parents of γ immediately before and after recombination, we follow all recombination edges between nodes in U and D that pass through the chosen γ node, noting the type (α or β) of each node implied by the recombination edge. We track the smallest α clade encountered in U, the smallest β clade encountered in D, and the union of all clades. If no downstream β was encountered, the branch movement is determined to be upward from the smallest upstream to the union of all clades. If no upstream α was encountered, the branch movement is determined to be downward from the union of all clades to the smallest downstream β. Otherwise, we infer a lateral branch movement from the smallest upstream α to the smallest downstream β.

Determining the coordinates of the recombination event is less straightforward. With the goal of eventually determining a site x immediately upstream of recombination and a site y immediately downstream of recombination, we define 3 possible starting points for x: x_1, x_2, and x_3. We also define two starting points for y: y_1 and y_2. We set x_1 to the downstream-most site in U and y_1 to the upstream-most site in D. We then set x_2 to the downstream-most upstream recombination partner of any node in D and y_2 to the upstream-most downstream recombination partner of any node in U. Finally, for each node in D, we gather an initial, unfiltered recombination graph (U’,D’) using that node as the key node k; x_3 is set to the downstream-most site in U across all such recombination graphs. Next, we set x to the maximum of x_1 and x_2 and y to the minimum of y_1 and y_2. At this point, if x_3 was set and is between x and y, then we set x to x_3. Finally, we locate a pair of adjacent sites in the ARG halfway between x and y and set x to the upstream site and y to the downstream site. If the choice is ambiguous (i.e. if the midpoint between x and y lands on a single site rather than between a pair of sites), we randomly either assign the middle site as x and the next site downstream as y, or assign the middle site as y and the next site upstream as x.

With all parameters of the ancestral recombination event inferred, the last steps are to adjust the ranges of all involved nodes, create new nodes implied by the recombination event, and add “solved” recombination edges between nodes marking this event. First, if any node in

U has an end coordinate upstream or downstream of x, it is set to x. Likewise, if any node in D has a start coordinate upstream or downstream of y, it is set to y. Next, the chosen γ clade is created as a node in the ARG, anchored to both sites x and y, with start coordinate x-p and end coordinate y+p. We create new versions of all upstream nodes, anchored at y and with the range [y,y+p], with clades adjusted by recombination (α clades have members of γ subtracted and β clades have members of γ added). We then create new versions of all downstream nodes, anchored at x and with the range [x-p,x], with clades adjusted by recombination (α clades have members of γ added and β clades have members of γ subtracted). If the recombination event was inferred to be upward or downward, we also create a clade containing the union of all clades in the recombination graph, with the range [y-p,y+p] and anchored at y (if upward) or [x-p,x+p] and anchored at x (if downward). If any of these nodes already existed in the ARG, this process will add sites to the node that do not affect its branch length and possibly extend its range. After this, we add "solved" recombination edges between each pair of upstream and downstream nodes that fail the four haplotype test.

B.3 Finalizing ARG node ranges

Because of the heuristic nature of our method, some ancestral recombination events go unsolved. Additionally, some may be unsolvable (for example, if all three candidate γ nodes for a four haplotype test failure fail the four haplotype test with other, existing ARG nodes at sites between the two four haplotype test failing nodes). When this is the case, we seek to expand the ranges of all nodes involved in recombination to their fullest extent. In other words, for every pair of nodes that fail the four haplotype test with each other, we want to ensure that the upstream nodes end coordinate and the downstream nodes start coordinate are set to sites approximately in the center of the genomic interval between the two nodes. If this is not done, there will be additional polytomies in the ARG. Therefore, when we are about to write a tree at site index s to disk, we seek to ensure that site index s + 1 will be covered either by a node in the tree covering s or by a downstream node that fails the four haplotype test with a node in the tree covering s.

To this end, we gather a set of upstream nodes U and a set of downstream nodes D, all of which are candidates to cover site s + 1. We visit each node n in the tree covering site s whose range ends upstream of s + 1. We then traverse all solved, unsolved, and unsolvable recombination edges from n to downstream nodes. If no such downstream nodes range covers s + 1 and s + 1 is within the eligible range of n (its downstream-most site + p), we add n to U. We then add each downstream recombination of partner n to D, if its start coordinate is downstream of s + 1. Finally, we add each upstream recombination partner of every node in D to U, if its end coordinate is upstream of or equal to s. For each site z between s and the upstream-most site owned by a node in D, we then gather a set U_z of upstream nodes eligible to cover z and a set D_z of downstream nodes eligible to cover z. To be eligible, an upstream node must have an end coordinate upstream of z, no downstream recombination partner upstream of z, and z must be within its eligible range (downstream-most site + p). Likewise, an upstream node must have a start coordinate downstream of z, no upstream recombination partner downstream of z, and z must be within its eligible range (upstream most site p). If $|U_z| = 0$, we expand the nodes in D_z upstream, and if $|D_z| = 0$, we expand the nodes in U_z downstream. Otherwise, we determine whether the downstream-most site owned by all nodes in U_z or the upstream-most site owned by all nodes in D_z is closer to z (if a tie, we choose randomly). Ranges of all nodes in the set found to be closer to z will be expanded to cover site z.

B.4 Collapsing to trees

To avoid making it necessary to hold the ARG over an entire chromosome in memory at once, or to load the entire ARG for all analyses, we represent the ARG on disk as a series of trees. At every site, the ARG collapses to a tree, so we write out each tree independently to disk, along with its chromosome and base position, in a custom serial binary format. We find that our files compress well with GZIP, and we provide utilities for indexing and retrieving specific genomic regions from files, and for converting our trees to Newick format.

Bibliography

[1] Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 1 2015.

[2] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 10 2015.

[3] R. Abbott, D. Albach, S. Ansell, J. W. Arntzen, S. J E Baird, N. Bierne, J. Boughman, a. Brelsford, C. a. Buerkle, R. Buggs, R. K. Butlin, U. Dieckmann, F. Eroukhmanoff, a. Grill, S. H. Cahan, J. S. Hermansen, G. Hewitt, a. G. Hudson, C. Jiggins, J. Jones, B. Keller, T. Marczewski, J. Mallet, P. Martinez-Rodriguez, M. Möst, S. Mullen, R. Nichols, a. W. Nolte, C. Parisod, K. Pfennig, a. M. Rice, M. G. Ritchie, B. Seifert, C. M. Smadja, R. Stelkens, J. M. Szymura, R. Väinölä, J. B W Wolf, and D. Zinner. Hybridization and speciation. *Journal of Evolutionary Biology*, 26(October 2011):229–246, 2013.

[4] Laurent Abi-Rached, Matthew J Jobin, Subhash Kulkarni, Alasdair McWhinnie, Klara Dalva, Loren Gragert, Farbod Babrzadeh, Baback Gharizadeh, Ma Luo, Francis A Plummer, Joshua Kimani, Mary Carrington, Derek Middleton, Raja Rajalingam, Meral Beksac, Steven G E Marsh, Martin Maiers, Lisbeth A Guethlein, Sofia Tavoularis, Ann-Margaret Little, Richard E Green, Paul J Norman, and Peter Parham. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science (New York, N.Y.)*, 334(6052):89–94, 10 2011.

[5] Viivi Karoliina Alaraudanjoki, Salla Koivisto, Paula Pesonen, Minna Männikkö, Jukka Leinonen, Leo Tjäderhane, Marja-Liisa Laitala, Adrian Lussi, and Vuokko Anna-Marketta Anttonen. Genome-Wide Association Study of Erosive Tooth Wear in a Finnish Cohort. *Caries research*, 53(1):49–59, 2019.

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

[7] Morten E. Allentoft, Martin Sikora, Karl-Gran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B. Damgaard, Hannes Schroeder, Torbjrn Ahlström, Lasse Vinner, Anna-Sapfo Malaspinas, Ashot Margaryan, Tom Higham, David Chivall,

Niels Lynnerup, Lise Harvig, Justyna Baron, Philippe Della Casa, Pawe Dąbrowski, Paul R. Duffy, Alexander V. Ebel, Andrey Epimakhov, Karin Frei, Mirosaw Furmanek, Tomasz Gralak, Andrey Gromov, Stanisaw Gronkiewicz, Gisela Grupe, Tams Hajdu, Radosaw Jarysz, Valeri Khartanovich, Alexandr Khokhlov, Viktria Kiss, Jan Kolář, Aivar Kriiska, Irena Lasak, Cristina Longhi, George McGlynn, Algimantas Merkevicius, Inga Merkyte, Mait Metspalu, Ruzan Mkrtchyan, Vyacheslav Moiseyev, Lszl Paja, Gyrgy Pálfi, Dalia Pokutta, ukasz Pospieszny, T. Douglas Price, Lehti Saag, Mikhail Sablin, Natalia Shishlina, Vclav Smrčka, Vasilii I. Soenov, Vajk Szeverényi, Gusztv Tóth, Synaru V. Trifanova, Liivi Varul, Magdolna Vicze, Levon Yepiskoposyan, Vladislav Zhitenev, Ludovic Orlando, Thomas Sicheritz-Pontén, Sren Brunak, Rasmus Nielsen, Kristian Kristiansen, and Eske Willerslev. Population genomics of Bronze Age Eurasia. *Nature*, 522:167–172, 2015.

[8] D Anderson, B J Holt, C E Pennell, P G Holt, P H Hart, and J M Blackwell. Genome-wide association study of vitamin D levels in children: replication in the Western Australian Pregnancy Cohort (Raine) study. *Genes and immunity*, 15(8):578–83, 12 2014.

[9] William J Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A Kostadima, John J Lambourne, Suthesh Sivapalaratnam, Kate Downes, Kousik Kundu, Lorenzo Bomba, Kim Berentsen, John R Bradley, Louise C Daugherty, Olivier Delaneau, Kathleen Freson, Stephen F Garner, Luigi Grassi, Jose Guerrero, Matthias Haimel, Eva M Janssen-Megens, Anita Kaan, Mihir Kamat, Bowon Kim, Amit Mandoli, Jonathan Marchini, Joost H A Martens, Stuart Meacham, Karyn Megy, Jared O'Connell, Romina Petersen, Nilofar Sharifi, Simon M Sheard, James R Staley, Salih Tuna, Martijn van der Ent, Klaudia Walter, Shuang-Yin Wang, Eleanor Wheeler, Steven P Wilder, Valentina Iotchkova, Carmel Moore, Jennifer Sambrook, Hendrik G Stunnenberg, Emanuele Di Angelantonio, Stephen Kaptoge, Taco W Kuijpers, Enrique Carrillo-de Santa-Pau, David Juan, Daniel Rico, Alfonso Valencia, Lu Chen, Bing Ge, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yang, Roderic Guigo, Stephan Beck, Dirk S Paul, Tomi Pastinen, David Bujold, Guillaume Bourque, Mattia Frontini, John Danesh, David J Roberts, Willem H Ouwehand, Adam S Butterworth, and Nicole Soranzo. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 167(5):1415–1429, 2016.

[10] Yael Baran, Bogdan Pasaniuc, Sriram Sankararaman, Dara G. Torgerson, Christopher Gignoux, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G. Ford, Pedro C. Avila, Jose Rodriguez-Santana, Esteban Gonzlez Burchard, and Eran Halperin. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10):1359–1367, 2012.

[11] Gary W Beecham, Kara Hamilton, Adam C Naj, Eden R Martin, Matt Huentelman, Amanda J Myers, Jason J Corneveaux, John Hardy, Jean-Paul Vonsattel, Steven G Younkin, David A Bennett, Philip L De Jager, Eric B Larson, Paul K Crane, M Ilyas

Kamboh, Julia K Kofler, Deborah C Mash, Linda Duque, John R Gilbert, Harry Gwirtsman, Joseph D Buxbaum, Patricia Kramer, Dennis W Dickson, Lindsay A Farrer, Matthew P Frosch, Bernardino Ghetti, Jonathan L Haines, Bradley T Hyman, Walter A Kukull, Richard P Mayeux, Margaret A Pericak-Vance, Julie A Schneider, John Q Trojanowski, Eric M Reiman, Alzheimer's Disease Genetics Consortium (ADGC), Gerard D Schellenberg, and Thomas J Montine. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. *PLoS genetics*, 10(9):e1004606, 9 2014.

[12] Heike Blockus and Alain Chédotal. Slit-Robo signaling. *Development*, 143(17):3037–3044, 9 2016.

[13] Sue C Bodine and Leslie M Baehr. Skeletal muscle atrophy and the E3 ubiquitin ligases MuRF1 and MAFbx/atrogin-1. *American journal of physiology. Endocrinology and metabolism*, 307(6):469–84, 9 2014.

[14] Adrian W Briggs, Udo Stenzel, Philip L F Johnson, Richard E Green, Janet Kelso, Kay Prüfer, Matthias Meyer, Johannes Krause, Michael T Ronan, Michael Lachmann, and Svante Pääbo. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, 104:14616–14621, 2007.

[15] Abra Brisbin, Katarzyna Bryc, Jake Byrnes, Fouad Zakharia, Larsson Omberg, Jeremiah Degenhardt, Andrew Reynolds, Harry Ostrer, Jason G Mezey, and Carlos D Bustamante. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology*, 84(4):343–64, 2012.

[16] P. Brotherton, P. Endicott, J. J. Sanchez, M. Beaumont, R. Barnett, J. Austin, and A. Cooper. Novel high-resolution characterization of ancient DNA reveals C ¿ U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Research*, 35(17):5717–5728, 2007.

[17] Sharon R. Browning and Brian L. Browning. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 11 2007.

[18] Sharon R Browning, Brian L Browning, Ying Zhou, Serena Tucci, and Joshua M Akey. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*, 173(1):53–61, 2018.

[19] James a Cahill, Richard E Green, Tara L Fulton, Mathias Stiller, Flora Jay, Nikita Ovsyanikov, Rauf Salamzade, John St John, Ian Stirling, Montgomery Slatkin, and Beth Shapiro. Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. *PLoS genetics*, 9(3):e1003345, 1 2013.

[20] James a Cahill, Ian Stirling, Logan Kistler, Rauf Salamzade, Erik Ersmark, Tara L. Fulton, Mathias Stiller, Richard E Green, and Beth Shapiro. Genomic evidence of geographically widespread effect of gene flow from polar bears into brown bears. *Molecular Ecology*, 24(6):1205–1217, 3 2015.

[21] Jonathan Casper, Ann S Zweig, Chris Villarreal, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, Christopher M Lee, Brian T Lee, Donna Karolchik, Angie S Hinrichs, Maximilian Haeussler, Luvina Guruvadoo, Jairo Navarro Gonzalez, David Gibson, Ian T Fiddes, Christopher Eisenhart, Mark Diekhans, Hiram Clawson, Galt P Barber, Joel Armstrong, David Haussler, Robert M Kuhn, and W James Kent. The UCSC Genome Browser database: 2018 update. *Nucleic acids research*, 46(D1):D762–D769, 1 2018.

[22] John C Chambers, Weihua Zhang, Graham M Lord, Pim van der Harst, Debbie A Lawlor, Joban S Sehmi, Daniel P Gale, Mark N Wass, Kourosh R Ahmadi, Stephan J L Bakker, Jacqui Beckmann, Henk J G Bilo, Murielle Bochud, Morris J Brown, Mark J Caulfield, John M C Connell, H Terence Cook, Ioana Cotlarciuc, George Davey Smith, Ranil de Silva, Guohong Deng, Olivier Devuyst, Lambert D Dikkeschei, Nada Dimkovic, Mark Dockrell, Anna Dominiczak, Shah Ebrahim, Thomas Eggermann, Martin Farrall, Luigi Ferrucci, Jurgen Floege, Nita G Forouhi, Ron T Gansevoort, Xijin Han, Bo Hedblad, Jaap J Homan van der Heide, Bouke G Hepkema, Maria Hernandez-Fuentes, Elina Hypponen, Toby Johnson, Paul E de Jong, Nanne Kleefstra, Vasiliki Lagou, Marta Lapsley, Yun Li, Ruth J F Loos, Jian'an Luan, Karin Luttropp, Cline Maréchal, Olle Melander, Patricia B Munroe, Louise Nordfors, Afshin Parsa, Leena Peltonen, Brenda W Penninx, Esperanza Perucha, Anneli Pouta, Inga Prokopenko, Paul J Roderick, Aimo Ruokonen, Nilesh J Samani, Serena Sanna, Martin Schalling, David Schlessinger, Georg Schlieper, Marc A J Seelen, Alan R Shuldiner, Marketa Sjögren, Johannes H Smit, Harold Snieder, Nicole Soranzo, Timothy D Spector, Peter Stenvinkel, Michael J E Sternberg, Ramasamyiyer Swaminathan, Toshiko Tanaka, Lielith J Ubink-Veltmaat, Manuela Uda, Peter Vollenweider, Chris Wallace, Dawn Waterworth, Klaus Zerres, Gerard Waeber, Nicholas J Wareham, Patrick H Maxwell, Mark I McCarthy, Marjo-Riitta Jarvelin, Vincent Mooser, Goncalo R Abecasis, Liz Lightstone, James Scott, Gerjan Navis, Paul Elliott, and Jaspal S Kooner. Genetic loci influencing kidney function and chronic kidney disease. *Nature genetics*, 42(5):373–5, 5 2010.

[23] Brian Charlesworth. Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205, 3 2009.

[24] W. V. Chen and T. Maniatis. Clustered protocadherins. *Development*, 140(16):3297–3302, 8 2013.

[25] Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 9 2005.

[26] Audrey Y Chu, Xuan Deng, Virginia A Fisher, Alexander Drong, Yang Zhang, Mary F Feitosa, Ching-Ti Liu, Olivia Weeks, Audrey C Choh, Qing Duan, Thomas D Dyer, John D Eicher, Xiuqing Guo, Nancy L Heard-Costa, Tim Kacprowski, Jack W Kent, Leslie A Lange, Xinggang Liu, Kurt Lohman, Lingyi Lu, Anubha Mahajan, Jeffrey R O'Connell, Ankita Parihar, Juan M Peralta, Albert V Smith, Yi Zhang, Georg Homuth, Ahmed H Kissebah, Joel Kullberg, Ren Laqua, Lenore J Launer, Matthias Nauck, Michael Olivier, Patricia A Peyser, James G Terry, Mary K Wojczynski, Jie Yao, Lawrence F Bielak, John Blangero, Ingrid B Borecki, Donald W Bowden, John Jeffrey Carr, Stefan A Czerwinski, Jingzhong Ding, Nele Friedrich, Vilmunder Gudnason, Tamara B Harris, Erik Ingelsson, Andrew D Johnson, Sharon L R Kardia, Carl D Langefeld, Lars Lind, Yongmei Liu, Braxton D Mitchell, Andrew P Morris, Thomas H Mosley, Jerome I Rotter, Alan R Shuldiner, Bradford Towne, Henry Völzke, Henri Wallaschofski, James G Wilson, Matthew Allison, Cecilia M Lindgren, Wolfram Goessling, L Adrienne Cupples, Matthew L Steinhauser, and Caroline S Fox. Multiethnic genome-wide meta-analysis of ectopic fat depots identifies loci associated with adipocyte development and differentiation. *Nature genetics*, 49(1):125–130, 2017.

[27] M. P. Cox, F. L. Mendez, T. M. Karafet, M. M. Pilkington, S. B. Kingan, G. Destro-Bisol, B. I. Strassmann, and M. F. Hammer. Testing for Archaic Hominin Admixture on the X Chromosome: Model Likelihoods for the Modern Human RRM2P4 Region From Summaries of Genealogical Topology Under the Structured Coalescent. *Genetics*, 178(January):427–437, 2008.

[28] Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet (London, England)*, 381(9875):1371–1379, 4 2013.

[29] Rute R. da Fonseca, Bruce D. Smith, Nathan Wales, Enrico Cappellini, Pontus Skoglund, Matteo Fumagalli, Jos Alfredo Samaniego, Christian Carøe, Mara C. Ávila-Arcos, David E. Hufnagel, Thorfinn Sand Korneliussen, Filipe Garrett Vieira, Mattias Jakobsson, Bernardo Arriaza, Eske Willerslev, Rasmus Nielsen, Matthew B. Hufford, Anders Albrechtsen, Jeffrey Ross-Ibarra, and M. Thomas P. Gilbert. The origin and evolution of maize in the Southwestern United States. *Nature Plants*, 1(January):14003, 2015.

[30] Gail Davies, Max Lam, Sarah E Harris, Joey W Trampush, Michelle Luciano, W David Hill, Saskia P Hagenaars, Stuart J Ritchie, Riccardo E Marioni, Chloe Fawns-Ritchie, David C M Liewald, Judith A Okely, Ari V Ahola-Olli, Catriona L K Barnes, Lars Bertram, Joshua C Bis, Katherine E Burdick, Andrea Christoforou, Pamela DeRosse, Srdjan Djurovic, Thomas Espeseth, Stella Giakoumaki, Sudheer Giddaluru, Daniel E Gustavson, Caroline Hayward, Edith Hofer, M Arfan Ikram, Robert Karlsson, Emma Knowles, Jari Lahti, Markus Leber, Shuo Li, Karen A Mather, Ingrid Melle, Derek Morris, Christopher Oldmeadow, Teemu Palviainen, Antony Payton, Raha Pazoki, Katja Petrovic, Chandra A Reynolds, Muralidharan Sargurupremraj, Markus Scholz, Jennifer A Smith, Albert V Smith, Natalie Terzikhan, Anbupalam Thalamuthu, Stella Trompet, Sven J van der Lee, Erin B Ware, B Gwen Windham, Margaret J Wright,

Jingyun Yang, Jin Yu, David Ames, Najaf Amin, Philippe Amouyel, Ole A Andreassen, Nicola J Armstrong, Amelia A Assareh, John R Attia, Deborah Attix, Dimitrios Avramopoulos, David A Bennett, Anne C Böhmer, Patricia A Boyle, Henry Brodaty, Harry Campbell, Tyrone D Cannon, Elizabeth T Cirulli, Eliza Congdon, Emily Drabant Conley, Janie Corley, Simon R Cox, Anders M Dale, Abbas Dehghan, Danielle Dick, Dwight Dickinson, Johan G Eriksson, Evangelos Evangelou, Jessica D Faul, Ian Ford, Nelson A Freimer, He Gao, Ina Giegling, Nathan A Gillespie, Scott D Gordon, Rebecca F Gottesman, Michael E Griswold, Vilmundur Gudnason, Tamara B Harris, Annette M Hartmann, Alex Hatzimanolis, Gerardo Heiss, Elizabeth G Holliday, Peter K Joshi, Mika Kähönen, Sharon L R Kardia, Ida Karlsson, Luca Kleineidam, David S Knopman, Nicole A Kochan, Bettina Konte, John B Kwok, Stephanie Le Hellard, Teresa Lee, Terho Lehtimäki, Shu-Chen Li, Tian Liu, Marisa Koini, Edythe London, Will T Longstreth, Oscar L Lopez, Anu Loukola, Tobias Luck, Astri J Lundervold, Anders Lundquist, Leo-Pekka Lyytikäinen, Nicholas G Martin, Grant W Montgomery, Alison D Murray, Anna C Need, Raymond Noordam, Lars Nyberg, William Ollier, Goran Papenberg, Alison Pattie, Ozren Polasek, Russell A Poldrack, Bruce M Psaty, Simone Reppermund, Steffi G Riedel-Heller, Richard J Rose, Jerome I Rotter, Panos Roussos, Suvi P Rovio, Yasaman Saba, Fred W Sabb, Perminder S Sachdev, Claudia L Satizabal, Matthias Schmid, Rodney J Scott, Matthew A Scult, Jeannette Simino, P Eline Slagboom, Nikolaos Smyrnis, Acha Soumaré, Nikos C Stefanis, David J Stott, Richard E Straub, Kjetil Sundet, Adele M Taylor, Kent D Taylor, Ioanna Tzoulaki, Christophe Tzourio, Andr Uitterlinden, Veronique Vitart, Aristotle N Voineskos, Jaakko Kaprio, Michael Wagner, Holger Wagner, Leonie Weinhold, K Hoyan Wen, Elisabeth Widen, Qiong Yang, Wei Zhao, Hieab H H Adams, Dan E Arking, Robert M Bilder, Panos Bitsios, Eric Boerwinkle, Ornit Chiba-Falek, Aiden Corvin, Philip L De Jager, Stphanie Debette, Gary Donohoe, Paul Elliott, Annette L Fitzpatrick, Michael Gill, David C Glahn, Sara Hägg, Narelle K Hansell, Ahmad R Hariri, M Kamran Ikram, J Wouter Jukema, Eero Vuoksimaa, Matthew C Keller, William S Kremen, Lenore Launer, Ulman Lindenberger, Aarno Palotie, Nancy L Pedersen, Neil Pendleton, David J Porteous, Katri Räikkönen, Olli T Raitakari, Alfredo Ramirez, Ivar Reinvang, Igor Rudan, Dan Rujescu, Reinhold Schmidt, Helena Schmidt, Peter W Schofield, Peter R Schofield, John M Starr, Vidar M Steen, Julian N Trollor, Steven T Turner, Cornelia M Van Duijn, Arno Villringer, Daniel R Weinberger, David R Weir, James F Wilson, Anil Malhotra, Andrew M McIntosh, Catharine R Gale, Sudha Seshadri, Thomas H Mosley, Jan Bressler, Todd Lencz, and Ian J Deary. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature communications*, 9(1):2098, 2018.

[31] Katrina M de Lange, Loukas Moutsianas, James C Lee, Christopher A Lamb, Yang Luo, Nicholas A Kennedy, Luke Jostins, Daniel L Rice, Javier Gutierrez-Achury, Sun-Gou Ji, Graham Heap, Elaine R Nimmo, Cathryn Edwards, Paul Henderson, Craig Mowat, Jeremy Sanderson, Jack Satsangi, Alison Simmons, David C Wilson, Mark Tremelling, Ailsa Hart, Christopher G Mathew, William G Newman, Miles Parkes, Charlie W Lees, Holm Uhlig, Chris Hawkey, Natalie J Prescott, Tariq Ahmad, John C Mansfield, Carl A

Anderson, and Jeffrey C Barrett. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, 49(2):256–261, 2 2017.

[32] S A J de With, S L Pulit, T Wang, W G Staal, W W van Solinge, P I W de Bakker, and R A Ophoff. Genome-wide association study of lymphoblast cell viability after clozapine exposure. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 168B(2):116–22, 3 2015.

[33] Olivier Delaneau, Jonathan Marchini, 1000 Genomes Project Consortium, and 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature communications*, 5:3934, 6 2014.

[34] Marcel den Hoed, Mark Eijgelsheim, Tnu Esko, Bianca J J M Brundel, David S Peal, David M Evans, Ilja M Nolte, Ayellet V Segrè, Hilma Holm, Robert E Handsaker, Harm-Jan Westra, Toby Johnson, Aaron Isaacs, Jian Yang, Alicia Lundby, Jing Hua Zhao, Young Jin Kim, Min Jin Go, Peter Almgren, Murielle Bochud, Gabrielle Boucher, Marilyn C Cornelis, Daniel Gudbjartsson, David Hadley, Pim van der Harst, Caroline Hayward, Martin den Heijer, Wilmar Igl, Anne U Jackson, Zoltn Kutalik, Jian'an Luan, John P Kemp, Kati Kristiansson, Claes Ladenvall, Mattias Lorentzon, May E Montasser, Omer T Njajou, Paul F O'Reilly, Sandosh Padmanabhan, Beate St Pourcain, Tuomo Rankinen, Perttu Salo, Toshiko Tanaka, Nicholas J Timpson, Veronique Vitart, Lindsay Waite, William Wheeler, Weihua Zhang, Harmen H M Draisma, Mary F Feitosa, Kathleen F Kerr, Penelope A Lind, Evelin Mihailov, N Charlotte Onland-Moret, Ci Song, Michael N Weedon, Weijia Xie, Loic Yengo, Devin Absher, Christine M Albert, Alvaro Alonso, Dan E Arking, Paul I W de Bakker, Beverley Balkau, Cristina Barlassina, Paola Benaglio, Joshua C Bis, Nabila Bouatia-Naji, Sren Brage, Stephen J Chanock, Peter S Chines, Mina Chung, Dawood Darbar, Christian Dina, Marcus Dörr, Paul Elliott, Stephan B Felix, Krista Fischer, Christian Fuchsberger, Eco J C de Geus, Philippe Goyette, Vilmundur Gudnason, Tamara B Harris, Anna-Liisa Hartikainen, Aki S Havulinna, Susan R Heckbert, Andrew A Hicks, Albert Hofman, Suzanne Holewijn, Femke Hoogstra-Berends, Jouke-Jan Hottenga, Majken K Jensen, Asa Johansson, Juhani Junttila, Stefan Kääb, Bart Kanon, Shamika Ketkar, Kay-Tee Khaw, Joshua W Knowles, Angrad S Kooner, Jan A Kors, Meena Kumari, Lili Milani, Pivi Laiho, Edward G Lakatta, Claudia Langenberg, Maarten Leusink, Yongmei Liu, Robert N Luben, Kathryn L Lunetta, Stacey N Lynch, Marcello R P Markus, Pedro Marques-Vidal, Irene Mateo Leach, Wendy L McArdle, Steven A McCarroll, Sarah E Medland, Kathryn A Miller, Grant W Montgomery, Alanna C Morrison, Martina Müller-Nurasyid, Pau Navarro, Mari Nelis, Jeffrey R O'Connell, Christopher J O'Donnell, Ken K Ong, Anne B Newman, Annette Peters, Ozren Polasek, Anneli Pouta, Peter P Pramstaller, Bruce M Psaty, Dabeeru C Rao, Susan M Ring, Elizabeth J Rossin, Diana Rudan, Serena Sanna, Robert A Scott, Jaban S Sehmi, Stephen Sharp, Jordan T Shin, Andrew B Singleton, Albert V Smith,

Nicole Soranzo, Tim D Spector, Chip Stewart, Heather M Stringham, Kirill V Tarasov, Andr G Uitterlinden, Liesbeth Vandenput, Shih-Jen Hwang, John B Whitfield, Cisca Wijmenga, Sarah H Wild, Gonneke Willemsen, James F Wilson, Jacqueline C M Witteman, Andrew Wong, Quenna Wong, Yalda Jamshidi, Paavo Zitting, Jolanda M A Boer, Dorret I Boomsma, Ingrid B Borecki, Cornelia M van Duijn, Ulf Ekelund, Nita G Forouhi, Philippe Froguel, Aroon Hingorani, Erik Ingelsson, Mika Kivimaki, Richard A Kronmal, Diana Kuh, Lars Lind, Nicholas G Martin, Ben A Oostra, Nancy L Pedersen, Thomas Quertermous, Jerome I Rotter, Yvonne T van der Schouw, W M Monique Verschuren, Mark Walker, Demetrius Albanes, David O Arnar, Themistocles L Assimes, Stefania Bandinelli, Michael Boehnke, Rudolf A de Boer, Claude Bouchard, W L Mark Caulfield, John C Chambers, Gary Curhan, Daniele Cusi, Johan Eriksson, Luigi Ferrucci, Wiek H van Gilst, Nicola Glorioso, Jacqueline de Graaf, Leif Groop, Ulf Gyllensten, Wen-Chi Hsueh, Frank B Hu, Heikki V Huikuri, David J Hunter, Carlos Iribarren, Bo Isomaa, Marjo-Riitta Jarvelin, Antti Jula, Mika Kähönen, Lambertus A Kiemeney, Melanie M van der Klauw, Jaspal S Kooner, Peter Kraft, Licia Iacoviello, Terho Lehtimäki, Marja-Liisa L Lokki, Braxton D Mitchell, Gerjan Navis, Markku S Nieminen, Claes Ohlsson, Neil R Poulter, Lu Qi, Olli T Raitakari, Eric B Rimm, John D Rioux, Federica Rizzi, Igor Rudan, Veikko Salomaa, Peter S Sever, Denis C Shields, Alan R Shuldiner, Juha Sinisalo, Alice V Stanton, Ronald P Stolk, David P Strachan, Jean-Claude Tardif, Unnur Thorsteinsdottir, Jaako Tuomilehto, Dirk J van Veldhuisen, Jarmo Virtamo, Jorma Viikari, Peter Vollenweider, Grard Waeber, Elisabeth Widen, Yoon Shin Cho, Jesper V Olsen, Peter M Visscher, Cristen Willer, Lude Franke, Global BPgen Consortium, CARDIoGRAM Consortium, Jeanette Erdmann, John R Thompson, PR GWAS Consortium, Arne Pfeufer, QRS GWAS Consortium, Nona Sotoodehnia, QT-IGC Consortium, Christopher Newton-Cheh, CHARGE-AF Consortium, Patrick T Ellinor, Bruno H Ch Stricker, Andres Metspalu, Markus Perola, Jacques S Beckmann, George Davey Smith, Kari Stefansson, Nicholas J Wareham, Patricia B Munroe, Ody C M Sibon, David J Milan, Harold Snieder, Nilesh J Samani, and Ruth J F Loos. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nature genetics*, 45(6):621–31, 6 2013.

[35] Eric Y. Durand, Nick Patterson, David Reich, and Montgomery Slatkin. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252, 2011.

[36] Richard M. Durbin, David L. Altshuler, Richard M. Durbin, Gonalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Francis S. Collins, De La Vega, M. Francisco, Peter Donnelly, Michael Egholm, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Bartha M. Knoppers, Eric S. Lander, Hans Lehrach, Elaine R. Mardis, Gil A. McVean, Debbie A. Nickerson, Leena Peltonen, Alan J. Schafer, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, David Deiros, Mike Metzker, Donna Muzny, Jeff Reid, David Wheeler, Jun Wang, Jingxiang Li, Min Jian, Guoqing Li, Ruiqiang Li, Huiqing Liang, Geng Tian, Bo Wang, Jian Wang, Wei Wang, Huanming Yang, Xiuqing Zhang, Huisong Zheng, Eric S. Lander, David L. Altshuler,

Lauren Ambrogio, Toby Bloom, Kristian Cibulskis, Tim J. Fennell, Stacey B. Gabriel, David B. Jaffe, Erica Shefler, Carrie L. Sougnez, David R. Bentley, Niall Gormley, Sean Humphray, Zoya Kingsbury, Paula Koko-Gonzales, Jennifer Stone, Kevin J. McKernan, Gina L. Costa, Jeffry K. Ichikawa, Clarence C. Lee, Ralf Sudbrak, Hans Lehrach, Tatiana A. Borodina, Andreas Dahl, Alexey N. Davydov, Peter Marquardt, Florian Mertes, Wilfiried Nietfeld, Philip Rosenstiel, Stefan Schreiber, Aleksey V. Soldatov, Bernd Timmermann, Marius Tolzmann, Michael Egholm, Jason Affourtit, Dana Ashworth, Said Attiya, Melissa Bachorski, Eli Buglione, Adam Burke, Amanda Caprio, Christopher Celone, Shauna Clark, David Conners, Brian Desany, Lisa Gu, Lorri Guccione, Kalvin Kao, Andrew Kebbel, Jennifer Knowlton, Matthew Labrecque, Louise McDade, Craig Mealmaker, Melissa Minderman, Anne Nawrocki, Faheem Niazi, Kristen Pareja, Ravi Ramenani, David Riches, Wanmin Song, Cynthia Turcotte, Shally Wang, Elaine R. Mardis, Richard K. Wilson, David Dooling, Lucinda Fulton, Robert Fulton, George Weinstock, Richard M. Durbin, John Burton, David M. Carter, Carol Churcher, Alison Coffey, Anthony Cox, Aarno Palotie, Michael Quail, Tom Skelly, James Stalker, Harold P. Swerdlow, Daniel Turner, Anniek De Witte, Shane Giles, Richard A. Gibbs, David Wheeler, Matthew Bainbridge, Danny Challis, Aniko Sabo, Fuli Yu, Jin Yu, Jun Wang, Xiaodong Fang, Xiaosen Guo, Ruiqiang Li, Yingrui Li, Ruibang Luo, Shuaishuai Tai, Honglong Wu, Hancheng Zheng, Xiaole Zheng, Yan Zhou, Guoqing Li, Jian Wang, Huanming Yang, Gabor T. Marth, Erik P. Garrison, Weichun Huang, Amit Indap, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Aaron R. Quinlan, Chip Stewart, Michael P. Stromberg, Alistair N. Ward, Jiantao Wu, Charles Lee, Ryan E. Mills, Xinghua Shi, Mark J. Daly, Mark A. DePristo, David L. Altshuler, Aaron D. Ball, Eric Banks, Toby Bloom, Brian L. Browning, Kristian Cibulskis, Tim J. Fennell, Kiran V. Garimella, Sharon R. Grossman, Robert E. Handsaker, Matt Hanna, Chris Hartl, David B. Jaffe, Andrew M. Kernytsky, Joshua M. Korn, Heng Li, Jared R. Maguire, Steven A. McCarroll, Aaron McKenna, James C. Nemesh, Anthony A. Philippakis, Ryan E. Poplin, Alkes Price, Manuel A. Rivas, Pardis C. Sabeti, Stephen F. Schaffner, Erica Shefler, Ilya A. Shlyakhter, David Neil Cooper, Edward Vincent Ball, Matthew Edwin Mort, Andrew David Phillips, Peter Daniel Stenson, Jonathan Sebat, Vladimir Makarov, Kenny Ye, Seungtai C. Yoon, Carlos D. Bustamante, Andrew G. Clark, Adam Boyko, Jeremiah Degenhardt, Simon Gravel, Ryan N. Gutenkunst, Mark Kaganovich, Alon Keinan, Phil Lacroute, Xin Ma, Andy Reynolds, Laura Clarke, Paul Flicek, Fiona Cunningham, Javier Herrero, Stephen Keenen, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Richard E. Smith, Vadim Zalunin, Xiangqun Zheng-Bradley, Jan O. Korbel, Adrian M. Stütz, Sean Humphray, Markus Bauer, R. Keira Cheetham, Tony Cox, Michael Eberle, Terena James, Scott Kahn, Lisa Murray, Aravinda Chakravarti, Kai Ye, De La Vega, M. Francisco, Yutao Fu, Fiona C. L. Hyland, Jonathan M. Manning, Stephen F. McLaughlin, Heather E. Peckham, Onur Sakarya, Yongming A. Sun, Eric F. Tsung, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Ralf Herwig, Dimitri V. Parkhomchuk, Stephen T. Sherry, Richa Agarwala, Hoda M. Khouri, Aleksandr O. Morgulis, Justin E. Paschall, Lon D. Phan, Kirill E. Rotmistrovsky, Robert D. Sanders, Martin F. Shumway, Chunlin Xiao, Gil A. McVean, Adam Auton, Zamin Iqbal, Gerton Lunter, Jonathan L. Marchini, Loukas Moutsianas, Simon Myers, Afidalina Tumian, Brian Desany, James Knight, Roger Winer, David W. Craig, Steve M. Beckstrom-Sternberg, Alexis Christoforides, Ahmet A. Kurdoglu, John V. Pearson, Shripad A. Sinari, Waibhav D. Tembe, David Haussler, Angie S. Hinrichs, Sol J. Katzman, Andrew Kern, Robert M. Kuhn, Molly Przeworski, Ryan D. Hernandez, Bryan Howie, Joanna L. Kelley, S. Cord Melton, Gonalo R. Abecasis, Yun Li, Paul Anderson, Tom Blackwell, Wei Chen, William O. Cookson, Jun Ding, Hyun Min Kang, Mark Lathrop, Liming Liang, Miriam F. Moffatt, Paul Scheet, Carlo Sidore, Matthew Snyder, Xiaowei Zhan, Sebastian Zöllner, Philip Awadalla, Ferran Casals, Youssef Idaghdour, John Keebler, Eric A. Stone, Martine Zilversmit, Lynn Jorde, Jinchuan Xing, Evan E. Eichler, Gozde Aksay, Can Alkan, Iman Hajirasouliha, Fereydoun Hormozdiari, Jeffrey M. Kidd, S. Cenk Sahinalp, Peter H. Sudmant, Elaine R. Mardis, Ken Chen, Asif Chinwalla, Li Ding, Daniel C. Koboldt, Mike D. McLellan, David Dooling, George Weinstock, John W. Wallis, Michael C. Wendl, Qunyuan Zhang, Richard M. Durbin, Cornelis A. Albers, Qasim Ayub, Senduran Balasubramaniam, Jeffrey C. Barrett, David M. Carter, Yuan Chen, Donald F. Conrad, Petr Danecek, Emmanouil T. Dermitzakis, Min Hu, Ni Huang, Matt E. Hurles, Hanjun Jin, Luke Jostins, Thomas M. Keane, Si Quang Le, Sarah Lindsay, Quan Long, Daniel G. MacArthur, Stephen B. Montgomery, Leopold Parts, James Stalker, Chris Tyler-Smith, Klaudia Walter, Yujun Zhang, Mark B. Gerstein, Michael Snyder, Alexej Abyzov, Suganthi Balasubramanian, Robert Bjornson, Jiang Du, Fabian Grubert, Lukas Habegger, Rajini Haraksingh, Justin Jee, Ekta Khurana, Hugo Y. K. Lam, Jing Leng, Xinmeng Jasmine Mu, Alexander E. Urban, Zhengdong Zhang, Yingrui Li, Ruibang Luo, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Aaron R. Quinlan, Chip Stewart, Michael P. Stromberg, Alistair N. Ward, Jiantao Wu, Charles Lee, Ryan E. Mills, Xinghua Shi, Steven A. McCarroll, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Chris Hartl, Joshua M. Korn, Heng Li, James C. Nemesh, Jonathan Sebat, Vladimir Makarov, Kenny Ye, Seungtai C. Yoon, Jeremiah Degenhardt, Mark Kaganovich, Laura Clarke, Richard E. Smith, Xiangqun Zheng-Bradley, Jan O. Korbel, Sean Humphray, R. Keira Cheetham, Michael Eberle, Scott Kahn, Lisa Murray, Kai Ye, De La Vega, M. Francisco, Yutao Fu, Heather E. Peckham, Yongming A. Sun, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Chunlin Xiao, Zamin Iqbal, Brian Desany, Tom Blackwell, Matthew Snyder, Jinchuan Xing, Evan E. Eichler, Gozde Aksay, Can Alkan, Iman Hajirasouliha, Fereydoun Hormozdiari, Jeffrey M. Kidd, Ken Chen, Asif Chinwalla, Li Ding, Mike D. McLellan, John W. Wallis, Matt E. Hurles, Donald F. Conrad, Klaudia Walter, Yujun Zhang, Mark B. Gerstein, Michael Snyder, Alexej Abyzov, Jiang Du, Fabian Grubert, Rajini Haraksingh, Justin Jee, Ekta Khurana, Hugo Y. K. Lam, Jing Leng, Xinmeng Jasmine Mu, Alexander E. Urban, Zhengdong Zhang, Richard A. Gibbs, Matthew Bainbridge, Danny Challis, Cristian Coafra, Huyen Dinh, Christie Kovar, Sandy Lee, Donna Muzny, Lynne Nazareth, Jeff Reid, Aniko Sabo, Fuli Yu, Jin Yu, Gabor T. Marth, Erik P. Garrison, Amit Indap, Wen Fung Leong, Aaron R. Quinlan, Chip Stewart, Alistair N. Ward, Jiantao Wu, Kristian Cibulskis, Tim J. Fennell, Stacey B. Gabriel, Kiran V. Garimella, Chris Hartl, Erica Shefler, Carrie L. Sougnez, Jane Wilkinson, Andrew G.

Clark, Simon Gravel, Fabian Grubert, Laura Clarke, Paul Flicek, Richard E. Smith, Xiangqun Zheng-Bradley, Stephen T. Sherry, Hoda M. Khouri, Justin E. Paschall, Martin F. Shumway, Chunlin Xiao, Gil A. McVean, Sol J. Katzman, Gonalo R. Abecasis, Tom Blackwell, Elaine R. Mardis, David Dooling, Lucinda Fulton, Robert Fulton, Daniel C. Koboldt, Richard M. Durbin, Senduran Balasubramaniam, Allison Coffey, Thomas M. Keane, Daniel G. MacArthur, Aarno Palotie, Carol Scott, James Stalker, Chris Tyler-Smith, Mark B. Gerstein, Suganthi Balasubramanian, Aravinda Chakravarti, Bartha M. Knoppers, Gonalo R. Abecasis, Carlos D. Bustamante, Neda Gharani, Richard A. Gibbs, Lynn Jorde, Jane S. Kaye, Alastair Kent, Taosha Li, Amy L. McGuire, Gil A. McVean, Pilar N. Ossorio, Charles N. Rotimi, Yeyang Su, Lorraine H. Toji, Chris Tyler-Smith, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Assya Abdallah, Christopher R. Juenger, Nicholas C. Clemm, Francis S. Collins, Audrey Duncanson, Eric D. Green, Mark S. Guyer, Jane L. Peterson, Alan J. Schafer, Gonalo R. Abecasis, David L. Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

[37] Arun Durvasula and Sriram Sankararaman. Recovering signals of ghost archaic admixture in the genomes of present-day Africans. *bioRxiv*, page 285734, 2018.

[38] D a R Eaton and R H Ree. Inferring phylogeny and introgression using RADseq data: an example from glowering plants (Pedicularis: Orobanchaceae). *Systematic Biology*, 62(5):689–706, 2013.

[39] Sean R Eddy. What is a hidden Markov model? *Nature biotechnology*, 22(10):1315–1316, 2004.

[40] Ceiridwen J. Edwards, Marc a. Suchard, Philippe Lemey, John J. Welch, Ian Barnes, Tara L. Fulton, Ross Barnett, Tamsin C. O'Connell, Peter Coxon, Nigel Monaghan, Cristina E. Valdiosera, Eline D. Lorenzen, Eske Willerslev, Gennady F. Baryshnikov, Andrew Rambaut, Mark G. Thomas, Daniel G. Bradley, and Beth Shapiro. Ancient hybridization and an irish origin for the modern polar bear matriline. *Current Biology*, 21(15):1251–1258, 2011.

[41] M Ehrlich and R Y Wang. 5-Methylcytosine in eukaryotic DNA. *Science (New York, N.Y.)*, 212(4501):1350–7, 6 1981.

[42] Valur Emilsson, Marjan Ilkov, John R Lamb, Nancy Finkel, Elias F Gudmundsson, Rebecca Pitts, Heather Hoover, Valborg Gudmundsdottir, Shane R Horman, Thor Aspelund, Le Shu, Vladimir Trifonov, Sigurdur Sigurdsson, Andrei Manolescu, Jun Zhu, rn Olafsson, Johanna Jakobsdottir, Scott A Lesley, Jeremy To, Jia Zhang, Tamara B Harris, Lenore J Launer, Bin Zhang, Gudny Eiriksdottir, Xia Yang, Anthony P Orth, Lori L Jennings, and Vilmundur Gudnason. Co-regulatory networks of human serum proteins link genetics to disease. *Science (New York, N.Y.)*, 361(6404):769–773, 2018.

[43] Jacob Enk, Alison Devault, Regis Debruyne, Christine E King, Todd Treangen, Dennis O'Rourke, Steven L Salzberg, Daniel Fisher, Ross MacPhee, and Hendrik Poinar. Complete Columbian mammoth mitogenome suggests interbreeding with woolly mammoths. *Genome Biology*, 12(5):R51, 2011.

[44] Patrick D Evans, Nitzan Mekel-Bobrov, Eric J Vallender, Richard R Hudson, and Bruce T Lahn. Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 103:18178–18183, 2006.

[45] Daniel Falush, Matthew Stephens, and Jonathan K. Pritchard. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.

[46] Alexander Favorov, Loris Mularoni, Leslie M Cope, Yulia Medvedeva, Andrey A Mironov, Vsevolod J Makeev, and Sarah J Wheelan. Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS computational biology*, 8(5):e1002529, 5 2012.

[47] Manuel A Ferreira, Judith M Vonk, Hansjrg Baurecht, Ingo Marenholz, Chao Tian, Joshua D Hoffman, Quinta Helmer, Annika Tillander, Vilhelmina Ullemar, Jenny van Dongen, Yi Lu, Franz Rüschendorf, Jorge Esparza-Gordillo, Chris W Medway, Edward Mountjoy, Kimberley Burrows, Oliver Hummel, Sarah Grosche, Ben M Brumpton, John S Witte, Jouke-Jan Hottenga, Gonneke Willemsen, Jie Zheng, Elke Rodríguez, Melanie Hotze, Andre Franke, Joana A Revez, Jonathan Beesley, Melanie C Matheson, Shyamali C Dharmage, Lisa M Bain, Lars G Fritsche, Maiken E Gabrielsen, Brunilda Balliu, 23andMe Research Team, AAGC collaborators, BIOS consortium, LifeLines Cohort Study, Jonas B Nielsen, Wei Zhou, Kristian Hveem, Arnulf Langhammer, Oddgeir L Holmen, Mari Løset, Gonalo R Abecasis, Cristen J Willer, Andreas Arnold, Georg Homuth, Carsten O Schmidt, Philip J Thompson, Nicholas G Martin, David L Duffy, Natalija Novak, Holger Schulz, Stefan Karrasch, Christian Gieger, Konstantin Strauch, Ronald B Melles, David A Hinds, Norbert Hübner, Stephan Weidinger, Patrik K E Magnusson, Rick Jansen, Eric Jorgenson, Young-Ae Lee, Dorret I Boomsma, Catarina Almqvist, Robert Karlsson, Gerard H Koppelman, and Lavinia Paternoster. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nature genetics*, 49(12):1752–1757, 12 2017.

[48] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, Doron Lancet, and Dana Cohen. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database : the journal of biological databases and curation*, 2017, 2017.

[49] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline

Chrast, Fiona Cunningham, Toms Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 1 2019.

[50] Qiaomei Fu, Mateja Hajdinjak, Oana Teodora Moldovan, Silviu Constantin, Swapan Mallick, Pontus Skoglund, Nick Patterson, Nadin Rohland, Iosif Lazaridis, Birgit Nickel, Bence Viola, Kay Prüfer, Matthias Meyer, Janet Kelso, David Reich, and Svante Pääbo. An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, 524, 2015.

[51] Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M. Slepchenko, Aleksei a. Bondarev, Philip L. F. Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare de Filippo, Matthias Meyer, Nicolas Zwyns, Domingo C. Salazar-García, Yaroslav V. Kuzmin, Susan G. Keates, Pavel a. Kosintsev, Dmitry I. Razhev, Michael P. Richards, Nikolai V. Peristov, Michael Lachmann, Katerina Douka, Thomas F. G. Higham, Montgomery Slatkin, Jean-Jacques Hublin, David Reich, Janet Kelso, T. Bence Viola, and Svante Pääbo. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514:8–13, 2014.

[52] Matteo Fumagalli. Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS ONE*, 8(11):14–17, 2013.

[53] 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, and A Manica. Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science (New York, N.Y.)*, 350(6262):820–2, 11 2015.

[54] Daniel Garrigan, Zahra Mobasher, Sarah B. Kingan, Jason a. Wilder, and Michael F. Hammer. Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. *Genetics*, 170(August):1849–1856, 2005.

[55] M. T. P. Gilbert, J. Binladen, W. Miller, C. Wiuf, E. Willerslev, H. Poinar, J. E. Carlson, J. H. Leebens-Mack, and S. C. Schuster. Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. *Nucleic Acids Research*, 35(1):1–10, 12 2006.

[56] Richard E Green, Johannes Krause, Adrian W Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F Hansen, Eric Y Durand, Anna-Sapfo Malaspinas, Jeffrey D Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernn a Burbano, Jeffrey M Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Zeljko Kucan, Ivan Gusic, Vladimir B Doronichev, Liubov V Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W Schmitz, Philip L F Johnson, Evan E Eichler, Daniel Falush, Ewan Birney, James C Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, and Svante Pääbo. A draft sequence of the Neandertal genome. *Science (New York, N.Y.)*, 328(5979):710–22, 5 2010.

[57] Richard E Green, Johannes Krause, Susan E Ptak, Adrian W Briggs, Michael T Ronan, Jan F Simons, Lei Du, Michael Egholm, Jonathan M Rothberg, Maja Paunovic, and Svante Pääbo. Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444(November):330–336, 2006.

[58] Richard E. Green, Anna Sapfo Malaspinas, Johannes Krause, Adrian W. Briggs, Philip L F Johnson, Caroline Uhler, Matthias Meyer, Jeffrey M. Good, Tomislav Maricic, Udo Stenzel, Kay Prüfer, Michael Siebauer, Hernn a. Burbano, Michael Ronan, Jonathan M. Rothberg, Michael Egholm, Pavao Rudan, Dejana Brajković, eljko Kućan, Ivan Gušić, Mrten Wikström, Liisa Laakkonen, Janet Kelso, Montgomery Slatkin, and Svante Pääbo. A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. *Cell*, 134:416–426, 2008.

[59] R C Griffiths and P Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of computational biology : a journal of computational molecular cell biology*, 3(4):479–502, 1996.

[60] Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, 5(10):e1000695, 10 2009.

[61] Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, Qiaomei Fu, Alissa Mittnik, Eszter Bánffy, Christos Economou, Michael Francken, Susanne Friederich, Rafael Garrido Pena, Fredrik Hallgren, Valery Khartanovich, Aleksandr Khokhlov, Michael Kunst, Pavel Kuznetsov, Harald Meller, Oleg Mochalov, Vayacheslav Moiseyev, Nicole Nicklisch, Sandra L. Pichler, Roberto Risch, Manuel a. Rojo Guerra, Christina Roth, Anna Szécsényi-Nagy, Joachim Wahl, Matthias Meyer, Johannes Krause, Dorcas Brown, David Anthony, Alan Cooper, Kurt Werner Alt, and David Reich. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522:207–11, 2015.

[62] M. F. Hammer, A. E. Woerner, F. L. Mendez, J. C. Watkins, and J. D. Wall. Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences*, 108(37):15123–15128, 2011.

[63] J Hardy, a Pittman, a Myers, K Gwinn-Hardy, H C Fung, R de Silva, M Hutton, and J Duckworth. Evidence suggesting that Homo neanderthalensis contributed the H2 MAPT haplotype to Homo sapiens. *Biochemical Society transactions*, 33:582–585, 2005.

[64] Kelley Harris and Rasmus Nielsen. Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, 9(6):e1003521, 6 2013.

[65] Kelley Harris and Rasmus Nielsen. The Genetic Cost of Neanderthal Introgression. *Genetics*, 203(2):881–91, 2016.

[66] Daniel L Hartl and Andrew G Clark. Random Genetic Drift. In *Principles of Population Genetics*, chapter Random Gen, pages 102–118. Sinauer Associates, Inc., Sunderland, Massachussetts, fourth edi edition, 2007.

[67] Masami Hasegawa, Hirohisa Kishino, and Taka Aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[68] John Hawks, Keith Hunley, S.-H. Lee, and M. Wolpoff. Population Bottlenecks and Pleistocene Human Evolution. *Molecular Biology and Evolution*, 17(1):2–22, 1 2000.

[69] S. Hayashi and M. Takeichi. Emerging roles of protocadherins: from self-avoidance to enhancement of motility. *Journal of Cell Science*, 128(8):1455–1464, 4 2015.

[70] R Higuchi, B Bowman, M Freiberger, O A Ryder, and A C Wilson. DNA sequences from the quagga, an extinct member of the horse family. *Nature*, 312:282–284.

[71] W D Hill, R E Marioni, O Maghzian, S J Ritchie, S P Hagenaars, A M McIntosh, C R Gale, G Davies, and I J Deary. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Molecular psychiatry*, 24(2):169–181, 2 2019.

[72] A S Hinrichs, D Karolchik, R Baertsch, G P Barber, G Bejerano, H Clawson, M Diekhans, T S Furey, R A Harte, F Hsu, J Hillman-Jackson, R M Kuhn, J S Pedersen, A Pohl, B J Raney, K R Rosenbloom, A Siepel, K E Smith, C W Sugnet, A Sultan-Qurraie, D J Thomas, H Trumbower, R J Weber, M Weirauch, A S Zweig, D Haussler, and W J Kent. The UCSC Genome Browser Database: update 2006. *Nucleic acids research*, 34(Database issue):590–8, 1 2006.

[73] C J Hoggart, M D Shriver, R a Kittles, D G Clayton, and P M McKeigue. Design and analysis of admixture mapping studies. *American journal of human genetics*, 74(McKeigue 1998):965–978, 2004.

[74] Clive J Hoggart, Eteban J Parra, Mark D Shriver, Carolina Bonilla, Rick a Kittles, David G Clayton, and Paul M McKeigue. Control of confounding of genetic associations in stratified populations. *American journal of human genetics*, 72:1492–1504, 2003.

[75] Kent E Holsinger and Bruce S Weir. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature reviews. Genetics*, 10(9):639–50, 9 2009.

[76] R R Hudson and N L Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–64, 9 1985.

[77] Richard R Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, 18(2):337–338, 2002.

[78] Adriana I Iglesias, Aniket Mishra, Veronique Vitart, Yelena Bykhovskaya, Ren Höhn, Henrit Springelkamp, Gabriel Cuellar-Partida, Puya Gharahkhani, Jessica N Cooke Bailey, Colin E Willoughby, Xiaohui Li, Seyhan Yazar, Abhishek Nag, Anthony P Khawaja, Ozren Polašek, David Siscovick, Paul Mitchell, Yih Chung Tham, Jonathan L Haines, Lisa S Kearns, Caroline Hayward, Yuan Shi, Elisabeth M van Leeuwen, Kent D Taylor, Blue Mountains Eye StudyGWAS group, Pieter Bonnemaijer, Jerome I Rotter, Nicholas G Martin, Tanja Zeller, Richard A Mills, Emmanuelle Souzeau, Sandra E Staffieri, Jost B Jonas, Irene Schmidtmann, Thibaud Boutin, Jae H Kang, Sionne E M Lucas, Tien Yin Wong, Manfred E Beutel, James F Wilson, NEIGHBORHOOD Consortium, Wellcome Trust Case Control Consortium 2 (WTCCC2), Andr G Uitterlinden, Eranga N Vithana, Paul J Foster, Pirro G Hysi, Alex W Hewitt, Chiea Chuen Khor, Louis R Pasquale, Grant W Montgomery, Caroline C W Klaver, Tin Aung, Norbert Pfeiffer, David A Mackey, Christopher J Hammond, Ching-Yu Cheng, Jamie E Craig, Yaron S Rabinowitz, Janey L Wiggs, Kathryn P Burdon, Cornelia M van Duijn, and Stuart MacGregor. Cross-ancestry genome-wide association analysis of corneal thickness strengthens link between complex and Mendelian eye diseases. *Nature communications*, 9(1):1864, 2018.

[79] Henry F Inman and Edwin L Bradley. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods*, 18(February 2016):3851–3874, 1989.

[80] Yuan Ji, Daniel J Schaid, Zeruesenay Desta, Michiaki Kubo, Anthony J Batzler, Karen Snyder, Taisei Mushiroda, Naoyuki Kamatani, Evan Ogburn, Daniel Hall-Flavin, David Flockhart, Yusuke Nakamura, David A Mrazek, and Richard M Weinshilboum. Citalopram and escitalopram plasma drug and metabolite concentrations: genome-wide associations. *British journal of clinical pharmacology*, 78(2):373–83, 8 2014.

[81] Sebastian Jünemann, Fritz Joachim Sedlazeck, Karola Prior, Andreas Albersmeier, Uwe John, Jrn Kalinowski, Alexander Mellmann, Alexander Goesmann, Arndt von Haeseler,

Jens Stoye, and Dag Harmsen. Updating benchtop sequencing performance comparison. *Nature Biotechnology*, 31:294–296, 2013.

[82] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

[83] Jerome Kelleher, Yan Wong, Patrick K Albers, W Anthony, and Gil Mcvean. Inferring the ancestry of everyone. *bioRxiv*, pages 1–42, 2018.

[84] Andreas Keller, Angela Graefen, Markus Ball, Mark Matzas, Valesca Boisguerin, Frank Maixner, Petra Leidinger, Christina Backes, Rabab Khairat, Michael Forster, Bjrn Stade, Andre Franke, Jens Mayer, Jessica Spangler, Stephen McLaughlin, Minita Shah, Clarence Lee, Timothy T. Harkins, Alexander Sartori, Andres Moreno-Estrada, Brenna Henn, Martin Sikora, Ornella Semino, Jacques Chiaroni, Siiri Rootsi, Natalie M. Myres, Vicente M. Cabrera, Peter a. Underhill, Carlos D. Bustamante, Eduard Egarter Vigl, Marco Samadelli, Giovanna Cipollini, Jan Haas, Hugo Katus, Brian D. O'Connor, Marc R.J. Carlson, Benjamin Meder, Nikolaus Blin, Eckart Meese, Carsten M. Pusch, and Albert Zink. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications*, 3:698, 2012.

[85] Michelle Kendall and Caroline Colijn. Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. *Molecular Biology and Evolution*, 33(10):2735–2743, 10 2016.

[86] Richard B Kennedy, Inna G Ovsyannikova, V Shane Pankratz, Iana H Haralambieva, Robert A Vierkant, and Gregory A Poland. Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. *Human genetics*, 131(9):1403–21, 9 2012.

[87] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. The human genome browser at UCSC. *Genome research*, 12(6):996–1006, 6 2002.

[88] BernardY. Kim and KirkE. Lohmueller. Selection and Reduced Population Size Cannot Explain Higher Amounts of Neandertal Ancestry in East Asian than in European Human Populations. *The American Journal of Human Genetics*, 96(3):454–461, 2015.

[89] Mi-Yeon Kim, Eun-Jung Ann, Jin-Young Kim, Jung-Soon Mo, Ji-Hye Park, Sun-Yee Kim, Mi-Sun Seo, and Hee-Sae Park. Tip60 histone acetyltransferase acts as a negative regulator of Notch1 signaling by means of acetylation. *Molecular and cellular biology*, 27(18):6506–19, 9 2007.

[90] M Kimura. Solution of a Process of Random Genetic Drift With a Continuous Model. *Proceedings of the National Academy of Sciences of the United States of America*, 41(2):144–150, 1955.

[91] Derek Klarin, Scott M Damrauer, Kelly Cho, Yan V Sun, Tanya M Teslovich, Jacqueline Honerlaw, David R Gagnon, Scott L DuVall, Jin Li, Gina M Peloso, Mark Chaffin, Aeron M Small, Jie Huang, Hua Tang, Julie A Lynch, Yuk-Lam Ho, Dajiang J Liu, Connor A Emdin, Alexander H Li, Jennifer E Huffman, Jennifer S Lee, Pradeep Natarajan, Rajiv Chowdhury, Danish Saleheen, Marijana Vujkovic, Aris Baras, Saiju Pyarajan, Emanuele Di Angelantonio, Benjamin M Neale, Aliya Naheed, Amit V Khera, John Danesh, Kyong-Mi Chang, Gonalo Abecasis, Cristen Willer, Frederick E Dewey, David J Carey, Global Lipids Genetics Consortium, Myocardial Infarction Genetics (MIGen) Consortium, Geisinger-Regeneron DiscovEHR Collaboration, VA Million Veteran Program, John Concato, J Michael Gaziano, Christopher J O'Donnell, Philip S Tsao, Sekar Kathiresan, Daniel J Rader, Peter W F Wilson, and Themistocles L Assimes. Genetics of blood lipids among ˜300,000 multi-ethnic participants of the Million Veteran Program. *Nature genetics*, 50(11):1514–1523, 11 2018.

[92] Johannes Krause, Qiaomei Fu, Jeffrey M Good, Bence Viola, Michael V Shunkov, Anatoli P Derevianko, and Svante Pääbo. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, 464(April):894–897, 2010.

[93] Hans R Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

[94] Verena E. Kutschera, Tobias Bidon, Frank Hailer, Julia L. Rodi, Steven R. Fain, and Axel Janke. Bears in a forest of gene trees: Phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Molecular Biology and Evolution*, 31(8):2004–2017, 2014.

[95] Joseph Lachance, Benjamin Vernot, Clara C Elbers, Bart Ferwerda, Alain Froment, Jean Marie Bodo, Godfrey Lema, Wenqing Fu, Thomas B Nyambo, Timothy R. Rebbeck, Kun Zhang, Joshua M Akey, and Sarah a Tishkoff. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell*, 150(3):457–469, 2012.

[96] John D. Lafferty, Andrew Mccallum, and Fernando C.N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning 2001*, pages 282–289, San Francisco, CA, 2001. Morgan Kaufmann Publishers Inc.

[97] Martina Lari, Ermanno Rizzi, Lucio Milani, Giorgio Corti, Carlotta Balsamo, Stefania Vai, Giulio Catalano, Elena Pilli, Laura Longo, Silvana Condemi, Paolo Giunti, Catherine Hänni, Gianluca de Bellis, Ludovic Orlando, Guido Barbujani, and David Caramelli. The microcephalin ancestral allele in a neanderthal individual. *PLoS ONE*, 5(5):6–11, 2010.

[98] Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, Mark Lipson, Bonnie Berger, Christos Economou, Ruth Bollongino, Qiaomei Fu, Kirsten I

Bos, Susanne Nordenfelt, Heng Li, Cesare de Filippo, Kay Prüfer, Susanna Sawyer, Cosimo Posth, Wolfgang Haak, Fredrik Hallgren, Elin Fornander, Nadin Rohland, Dominique Delsate, Michael Francken, Jean-Michel Guinet, Joachim Wahl, George Ayodo, Hamza a. Babiker, Graciela Bailliet, Elena Balanovska, Oleg Balanovsky, Ramiro Barrantes, Gabriel Bedoya, Haim Ben-Ami, Judit Bene, Fouad Berrada, Claudio M Bravi, Francesca Brisighelli, George B J Busby, Francesco Cali, Mikhail Churnosov, David E C Cole, Daniel Corach, Larissa Damba, George van Driem, Stanislav Dryomov, Jean-Michel Dugoujon, Sardana a. Fedorova, Irene Gallego Romero, Marina Gubina, Michael Hammer, Brenna M Henn, Tor Hervig, Ugur Hodoglugil, Aashish R. Jha, Sena Karachanak-Yankova, Rita Khusainova, Elza Khusnutdinova, Rick Kittles, Toomas Kivisild, William Klitz, Vaidutis Kučinskas, Alena Kushniarevich, Leila Laredj, Sergey Litvinov, Theologos Loukidis, Robert W. Mahley, Bla Melegh, Ene Metspalu, Julio Molina, Joanna Mountain, Klemetti Näkkäläjärvi, Desislava Nesheva, Thomas Nyambo, Ludmila Osipova, Jri Parik, Fedor Platonov, Olga Posukh, Valentino Romano, Francisco Rothhammer, Igor Rudan, Ruslan Ruizbakiev, Hovhannes Sahakyan, Antti Sajantila, Antonio Salas, Elena B. Starikovskaya, Ayele Tarekegn, Draga Toncheva, Shahlo Turdikulova, Ingrida Uktveryte, Olga Utevska, Ren Vasquez, Mercedes Villena, Mikhail Voevoda, Cheryl A Winkler, Levon Yepiskoposyan, Pierre Zalloua, Tatijana Zemunik, Alan Cooper, Cristian Capelli, Mark G. Thomas, Andres Ruiz-Linares, Sarah a. Tishkoff, Lalji Singh, Kumarasamy Thangaraj, Richard Villems, David Comas, Rem Sukernik, Mait Metspalu, Matthias Meyer, Evan E. Eichler, Joachim Burger, Montgomery Slatkin, Svante Pääbo, Janet Kelso, David Reich, and Johannes Krause. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–13, 2014.

[99] James J Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, Mark Alan Fontana, Tushar Kundu, Chanwook Lee, Hui Li, Ruoxi Li, Rebecca Royer, Pascal N Timshel, Raymond K Walters, Emily A Willoughby, Loc Yengo, 23andMe Research Team, COGENT (Cognitive Genomics Consortium), Social Science Genetic Association Consortium, Maris Alver, Yanchun Bao, David W Clark, Felix R Day, Nicholas A Furlotte, Peter K Joshi, Kathryn E Kemper, Aaron Kleinman, Claudia Langenberg, Reedik Mägi, Joey W Trampush, Shefali Setia Verma, Yang Wu, Max Lam, Jing Hua Zhao, Zhili Zheng, Jason D Boardman, Harry Campbell, Jeremy Freese, Kathleen Mullan Harris, Caroline Hayward, Pamela Herd, Meena Kumari, Todd Lencz, Jian'an Luan, Anil K Malhotra, Andres Metspalu, Lili Milani, Ken K Ong, John R B Perry, David J Porteous, Marylyn D Ritchie, Melissa C Smart, Blair H Smith, Joyce Y Tung, Nicholas J Wareham, James F Wilson, Jonathan P Beauchamp, Dalton C Conley, Tnu Esko, Steven F Lehrer, Patrik K E Magnusson, Sven Oskarsson, Tune H Pers, Matthew R Robinson, Kevin Thom, Chelsea Watson, Christopher F Chabris, Michelle N Meyer, David I Laibson, Jian Yang, Magnus Johannesson, Philipp D Koellinger, Patrick Turley, Peter M Visscher, Daniel J Benjamin, and David Cesarini. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in

1.1 million individuals. *Nature genetics*, 50(8):1112–1121, 8 2018.

[100] Christopher J Lessard, He Li, Indra Adrianto, John A Ice, Astrid Rasmussen, Kiely M Grundahl, Jennifer A Kelly, Mikhail G Dozmorov, Corinne Miceli-Richard, Simon Bowman, Sue Lester, Per Eriksson, Maija-Leena Eloranta, Johan G Brun, Lasse G Gøransson, Erna Harboe, Joel M Guthridge, Kenneth M Kaufman, Marika Kvarnström, Helmi Jazebi, Deborah S Cunninghame Graham, Martha E Grandits, Abu N M Nazmul-Hossain, Ketan Patel, Adam J Adler, Jacen S Maier-Moore, A Darise Farris, Michael T Brennan, James A Lessard, James Chodosh, Rajaram Gopalakrishnan, Kimberly S Hefner, Glen D Houston, Andrew J W Huang, Pamela J Hughes, David M Lewis, Lida Radfar, Michael D Rohrer, Donald U Stone, Jonathan D Wren, Timothy J Vyse, Patrick M Gaffney, Judith A James, Roald Omdal, Marie Wahren-Herlenius, Gabor G Illei, Torsten Witte, Roland Jonsson, Maureen Rischmueller, Lars Rönnblom, Gunnel Nordmark, Wan-Fai Ng, UK Primary Sjögren's Syndrome Registry, Xavier Mariette, Juan-Manuel Anaya, Nelson L Rhodus, Barbara M Segal, R Hal Scofield, Courtney G Montgomery, John B Harley, and Kathy L Sivils. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. *Nature genetics*, 45(11):1284–92, 11 2013.

[101] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, 27(21):2987–93, 11 2011.

[102] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 3 2013.

[103] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760, 2009.

[104] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079, 2009.

[105] Mark Lipson, Po Ru Loh, Alex Levin, David Reich, Nick Patterson, and Bonnie Berger. Efficient moment-based inference of admixture parameters and sources of gene flow. *Molecular Biology and Evolution*, 30(8):1788–1802, 2013.

[106] Jiaqi Liu, Yangzhong Zhou, Sen Liu, Xiaofei Song, Xin-Zhuang Yang, Yanhui Fan, Weisheng Chen, Zeynep Coban Akdemir, Zihui Yan, Yuzhi Zuo, Renqian Du, Zhenlei Liu, Bo Yuan, Sen Zhao, Gang Liu, Yixin Chen, Yanxue Zhao, Mao Lin, Qiankun Zhu, Yuchen Niu, Pengfei Liu, Shiro Ikegawa, You-Qiang Song, Jennifer E Posey, Guixing Qiu, DISCO (Deciphering disorders Involving Scoliosis and COmorbidities) Study, Feng Zhang, Zhihong Wu, James R Lupski, and Nan Wu. The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. *Human genetics*, 137(6-7):553–567, 7 2018.

[107] Jimmy Z Liu, Mohamed A Almarri, Daniel J Gaffney, George F Mells, Luke Jostins, Heather J Cordell, Samantha J Ducker, Darren B Day, Michael A Heneghan, James M Neuberger, Peter T Donaldson, Andrew J Bathgate, Andrew Burroughs, Mervyn H Davies, David E Jones, Graeme J Alexander, Jeffrey C Barrett, Richard N Sandford, Carl A Anderson, UK Primary Biliary Cirrhosis (PBC) Consortium, and Wellcome Trust Case Control Consortium 3. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nature genetics*, 44(10):1137–41, 10 2012.

[108] Shiping Liu, Eline D. Lorenzen, Matteo Fumagalli, Bo Li, Kelley Harris, Zijun Xiong, Long Zhou, Thorfinn Sand Korneliussen, Mehmet Somel, Courtney Babbitt, Greg Wray, Jianwen Li, Weiming He, Zhuo Wang, Wenjing Fu, Xueyan Xiang, Claire C. Morgan, Aoife Doherty, Mary J. O'Connell, James O. McInerney, Erik W. Born, Love Dalén, Rune Dietz, Ludovic Orlando, Christian Sonne, Guojie Zhang, Rasmus Nielsen, Eske Willerslev, and Jun Wang. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, 157:785–794, 2014.

[109] Y Liu, C A Fernandez, C Smith, W Yang, C Cheng, J C Panetta, N Kornegay, C Liu, L B Ramsey, S E Karol, L J Janke, E C Larsen, N Winick, W L Carroll, M L Loh, E A Raetz, S P Hunger, M Devidas, J J Yang, C G Mullighan, J Zhang, W E Evans, S Jeha, C-H Pui, and M V Relling. Genome-Wide Study Links PNPLA3 Variant With Elevated Hepatic Transaminase After Acute Lymphoblastic Leukemia Therapy. *Clinical pharmacology and therapeutics*, 102(1):131–140, 2017.

[110] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, Richard Durbin, and Alkes L Price. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics*, 48(11):1443–1448, 2016.

[111] Xi Lu, Xinde Hu, Lingzhen Song, Lei An, Minghui Duan, Shulin Chen, and Shanting Zhao. The SH2 domain is crucial for function of Fyn in neuronal migration and cortical lamination. *BMB reports*, 48(2):97–102, 2 2015.

[112] Kathryn L Lunetta, Ralph B D'Agostino, David Karasik, Emelia J Benjamin, Chao-Yu Guo, Raju Govindaraju, Douglas P Kiel, Margaret Kelly-Hayes, Joseph M Massaro, Michael J Pencina, Sudha Seshadri, and Joanne M Murabito. Genetic correlates of longevity and selected age-related phenotypes: a genome-wide association study in the Framingham Study. *BMC medical genetics*, 8 Suppl 1:S13, 9 2007.

[113] Sharon M Lutz, Michael H Cho, Kendra Young, Craig P Hersh, Peter J Castaldi, Merry-Lynn McDonald, Elizabeth Regan, Manuel Mattheisen, Dawn L DeMeo, Margaret Parker, Marilyn Foreman, Barry J Make, Robert L Jensen, Richard Casaburi, David A Lomas, Surya P Bhatt, Per Bakke, Amund Gulsvik, James D Crapo, Terri H Beaty, Nan M Laird, Christoph Lange, John E Hokanson, Edwin K Silverman, ECLIPSE Investigators, and COPDGene Investigators. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC genetics*, 16:138, 12 2015.

[114] VincentJ. Lynch, OscarC. Bedoya-Reina, Aakrosh Ratan, Michael Sulak, DanielaI. Drautz-Moses, GeorgeH. Perry, Webb Miller, and StephanC. Schuster. Elephantid Genomes Reveal the Molecular Bases of Woolly Mammoth Adaptations to the Arctic. *Cell Reports*, 12(2):217–228, 2015.

[115] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research*, 45(D1):D896–D901, 2017.

[116] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, Pontus Skoglund, Iosif Lazaridis, Sriram Sankararaman, Qiaomei Fu, Nadin Rohland, Gabriel Renaud, Yaniv Erlich, Thomas Willems, Carla Gallo, Jeffrey P. Spence, Yun S. Song, Giovanni Poletti, Francois Balloux, George van Driem, Peter de Knijff, Irene Gallego Romero, Aashish R. Jha, Doron M. Behar, Claudio M. Bravi, Cristian Capelli, Tor Hervig, Andres Moreno-Estrada, Olga L. Posukh, Elena Balanovska, Oleg Balanovsky, Sena Karachanak-Yankova, Hovhannes Sahakyan, Draga Toncheva, Levon Yepiskoposyan, Chris Tyler-Smith, Yali Xue, M. Syafiq Abdullah, Andres Ruiz-Linares, Cynthia M. Beall, Anna Di Rienzo, Choongwon Jeong, Elena B. Starikovskaya, Ene Metspalu, Jri Parik, Richard Villems, Brenna M. Henn, Ugur Hodoglugil, Robert Mahley, Antti Sajantila, George Stamatoyannopoulos, Joseph T. S. Wee, Rita Khusainova, Elza Khusnutdinova, Sergey Litvinov, George Ayodo, David Comas, Michael F. Hammer, Toomas Kivisild, William Klitz, Cheryl A. Winkler, Damian Labuda, Michael Bamshad, Lynn B. Jorde, Sarah A. Tishkoff, W. Scott Watkins, Mait Metspalu, Stanislav Dryomov, Rem Sukernik, Lalji Singh, Kumarasamy Thangaraj, Svante Pääbo, Janet Kelso, Nick Patterson, and David Reich. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 10 2016.

[117] Michael D. Martin, Filipe G. Vieira, Simon Y.W. Ho, Nathan Wales, Mikkel Schubert, Andaine Seguin-Orlando, Jean B. Ristaino, and M. Thomas P. Gilbert. Genomic Characterization of a South American Phytophthora Hybrid Mandates Reassessment of the Geographic Origins of Phytophthora infestans. *Molecular Biology and Evolution*, 33(2):478–491, 2 2016.

[118] Maureen D Mayes, Lara Bossini-Castillo, Olga Gorlova, Jos Ezequiel Martin, Xiaodong Zhou, Wei V Chen, Shervin Assassi, Jun Ying, Filemon K Tan, Frank C Arnett, John D Reveille, Sandra Guerra, Mara Teruel, Francisco David Carmona, Peter K Gregersen, Annette T Lee, Elena López-Isac, Eguzkine Ochoa, Patricia Carreira, Carmen Pilar Simeón, Ivn Castellví, Miguel ngel González-Gay, Spanish Scleroderma Group, Alexandra Zhernakova, Leonid Padyukov, Marta Alarcón-Riquelme, Cisca Wijmenga, Matthew Brown, Lorenzo Beretta, Gabriela Riemekasten, Torsten Witte, Nicolas Hunzelmann, Alexander Kreuter, Jrg H W Distler, Alexandre E Voskuyl, Annemie J Schuerwegh,

Roger Hesselstrand, Annika Nordin, Paolo Airó, Claudio Lunardi, Paul Shiels, Jacob M van Laar, Ariane Herrick, Jane Worthington, Christopher Denton, Fredrick M Wigley, Laura K Hummers, John Varga, Monique E Hinchcliff, Murray Baron, Marie Hudson, Janet E Pope, Daniel E Furst, Dinesh Khanna, Kristin Phillips, Elena Schiopu, Barbara M Segal, Jerry A Molitor, Richard M Silver, Virginia D Steen, Robert W Simms, Robert A Lafyatis, Barri J Fessler, Tracy M Frech, Firas Alkassab, Peter Docherty, Elzbieta Kaminska, Nader Khalidi, Henry Niall Jones, Janet Markland, David Robinson, Jasper Broen, Timothy R D J Radstake, Carmen Fonseca, Bobby P Koeleman, and Javier Martin. Immunochip analysis identifies multiple susceptibility loci for systemic sclerosis. *American journal of human genetics*, 94(1):47–61, 1 2014.

[119] Rajiv C. McCoy, Jon Wakefield, and Joshua M. Akey. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell*, 168(5):916–927, 2 2017.

[120] A. J. McKane and D. Waxman. Singular solutions of the diffusion equation of population genetics. *Journal of Theoretical Biology*, 247:849–858, 2007.

[121] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20:1297–1303, 2010.

[122] Fernando L Mendez, Joseph C Watkins, and Michael F Hammer. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *American journal of human genetics*, 91(2):265–74, 8 2012.

[123] Fernando L. Mendez, Joseph C. Watkins, and Michael F. Hammer. Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. *Molecular Biology and Evolution*, 29(6):1513–1520, 2012.

[124] P Menozzi, A Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies in Europeans. *Science (New York, N.Y.)*, 201(4358):786–92, 9 1978.

[125] Michael L Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, 2010.

[126] M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prufer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, a. Tandon, M. Siebauer, R. E. Green, K. Bryc, a. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, a. P. Derevianko, N. Patterson, a. M. Andres, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, and S. Paabo. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, 338(October):222–226, 2012.

[127] Matthias Meyer, Qiaomei Fu, Ayinuer Aximu-Petri, Isabelle Glocke, Birgit Nickel, Juan-Luis Arsuaga, Ignacio Martínez, Ana Gracia, Jos Mara Bermdez de Castro, Eudald Carbonell, and Svante Pääbo. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*, 505:403–6, 2014.

[128] Wolfgang Meyerhof, Claudia Batram, Christina Kuhn, Anne Brockhoff, Elke Chudoba, Bernd Bufe, Giovanni Appendino, and Maik Behrens. The molecular receptive ranges of human TAS2R bitter taste receptors. *Chemical senses*, 35(2):157–70, 2 2010.

[129] Webb Miller, Stephan C Schuster, Andreanna J Welch, Aakrosh Ratan, Oscar C Bedoya-Reina, Fangqing Zhao, Hie Lim Kim, Richard C Burhans, Daniela I Drautz, Nicola E Wittekindt, Lynn P Tomsho, Enrique Ibarra-Laclette, Luis Herrera-Estrella, Elizabeth Peacock, Sean Farley, George K Sage, Karyn Rode, Martyn Obbard, Rafael Montiel, Lutz Bachmann, Olafur Ingólfsson, Jon Aars, Thomas Mailund, Oystein Wiig, Sandra L Talbot, and Charlotte Lindqvist. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences of the United States of America*, 109(36):2382–90, 9 2012.

[130] Mark J Minichiello and Richard Durbin. Mapping trait loci by use of inferred ancestral recombination graphs. *American journal of human genetics*, 79(5):910–22, 11 2006.

[131] Sajad Mirzaei and Yufeng Wu. RENT+: An improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*, 33(7):1021–1030, 2017.

[132] Mayukh Mondal, Jaume Bertranpetit, and Oscar Lao. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, 10(1):246, 12 2019.

[133] Benjamin H Mullin, Jing Hua Zhao, Suzanne J Brown, John R B Perry, Jian'an Luan, Hou-Feng Zheng, Claudia Langenberg, Frank Dudbridge, Robert Scott, Nick J Wareham, Tim D Spector, J Brent Richards, John P Walsh, and Scott G Wilson. Genome-wide association study meta-analysis for quantitative ultrasound parameters of bone identifies five novel loci for broadband ultrasound attenuation. *Human molecular genetics*, 26(14):2791–2802, 2017.

[134] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, 12(6):443–51, 2011.

[135] James P Noonan, Graham Coop, Sridhar Kudaravalli, Doug Smith, Johannes Krause, Joe Alessi, Feng Chen, Darren Platt, Svante Pääbo, Jonathan K Pritchard, and Edward M Rubin. Sequencing and analysis of Neanderthal genomic DNA. *Science (New York, N.Y.)*, 314(November):1113–1118, 2006.

[136] M Nordborg. On the probability of Neanderthal ancestry. *American journal of human genetics*, 63(4):1237–40, 10 1998.

[137] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltn Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, Matthew Stephens, and Carlos D Bustamante. Genes mirror geography within Europe. *Nature*, 456(November):98–101, 2008.

[138] Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, Robert R Graham, Arun Manoharan, Ward Ortmann, Tushar Bhangale, Joshua C Denny, Robert J Carroll, Anne E Eyler, Jeffrey D Greenberg, Joel M Kremer, Dimitrios A Pappas, Lei Jiang, Jian Yin, Lingying Ye, Ding-Feng Su, Jian Yang, Gang Xie, Ed Keystone, Harm-Jan Westra, Tnu Esko, Andres Metspalu, Xuezhong Zhou, Namrata Gupta, Daniel Mirel, Eli A Stahl, Dorothe Diogo, Jing Cui, Katherine Liao, Michael H Guo, Keiko Myouzen, Takahisa Kawaguchi, Marieke J H Coenen, Piet L C M van Riel, Mart A F J van de Laar, Henk-Jan Guchelaar, Tom W J Huizinga, Philippe Dieudé, Xavier Mariette, S Louis Bridges, Alexandra Zhernakova, Rene E M Toes, Paul P Tak, Corinne Miceli-Richard, So-Young Bang, Hye-Soon Lee, Javier Martin, Miguel A Gonzalez-Gay, Luis Rodriguez-Rodriguez, Solbritt Rantapää-Dahlqvist, Lisbeth Arlestig, Hyon K Choi, Yoichiro Kamatani, Pilar Galan, Mark Lathrop, RACI consortium, GARNET consortium, Steve Eyre, John Bowes, Anne Barton, Niek de Vries, Larry W Moreland, Lindsey A Criswell, Elizabeth W Karlson, Atsuo Taniguchi, Ryo Yamada, Michiaki Kubo, Jun S Liu, Sang-Cheol Bae, Jane Worthington, Leonid Padyukov, Lars Klareskog, Peter K Gregersen, Soumya Raychaudhuri, Barbara E Stranger, Philip L De Jager, Lude Franke, Peter M Visscher, Matthew A Brown, Hisashi Yamanaka, Tsuneyo Mimori, Atsushi Takahashi, Huji Xu, Timothy W Behrens, Katherine A Siminovitch, Shigeki Momohara, Fumihiko Matsuda, Kazuhiko Yamamoto, and Robert M Plenge. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–81, 2 2014.

[139] Mariano Oppikofer, Tianyi Bai, Yutian Gan, Benjamin Haley, Peter Liu, Wendy Sandoval, Claudio Ciferri, and Andrea G Cochran. Expansion of the ISWI chromatin remodeler family with new active complexes. *EMBO reports*, 18(10):1697–1706, 2017.

[140] Svante Pääbo. Preservation of DNA in ancient Egyptian mummies. *Journal of Archaeological Science*, 12(6):411–417, 1985.

[141] Svante Pääbo. The diverse origins of the human gene pool. *Nature Reviews Genetics*, 16(6):313–314, 2015.

[142] Svante Pääbo, Russel G Higuchi, and Allan C Wilson. Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology. *The Journal of biological chemistry*, 264(17):9709–12, 6 1989.

[143] Eleftheria Palkopoulou, Swapan Mallick, Pontus Skoglund, Jacob Enk, Nadin Rohland, Heng Li, Aya Omrak, Sergey Vartanyan, Hendrik Poinar, Anders Götherström, David Reich, and Love Dalén. Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth. *Current Biology*, 25(10):1395–1400, 5 2015.

[144] Stephen D E Park, David A Magee, Paul A. McGettigan, Matthew D Teasdale, Ceiridwen J Edwards, Amanda J Lohan, Alison Murphy, Martin Braud, Mark T Donoghue, Yuan Liu, Andrew T Chamberlain, Kvin Rue-Albrecht, Steven Schroeder, Charles Spillane, Shuaishuai Tai, Daniel G Bradley, Tad S Sonstegard, Brendan J Loftus, and David E. MacHugh. Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. *Genome Biology*, 16:234, 2015.

[145] Murray Patterson, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W Klau, and Alexander Schönhuth. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Journal of computational biology : a journal of computational molecular cell biology*, 22(6):498–509, 6 2015.

[146] Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller, David a Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J O'Brien, David Altshuler, Mark J Daly, and David Reich. Methods for high-density admixture mapping of disease genes. *American journal of human genetics*, 74:979–1000, 2004.

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

[148] Nick Patterson, Alkes L. Price, and David Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):2074–2093, 2006.

[149] Joshua S. Paul and Yun S. Song. A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. *Genetics*, 186(1776):321–338, 2010.

[150] John Rb Perry, Felix Day, Cathy E Elks, Patrick Sulem, Deborah J Thompson, Teresa Ferreira, Chunyan He, Daniel I Chasman, Tnu Esko, Gudmar Thorleifsson, Eva Albrecht, Wei Q Ang, Tanguy Corre, Diana L Cousminer, Bjarke Feenstra, Nora Franceschini, Andrea Ganna, Andrew D Johnson, Sanela Kjellqvist, Kathryn L Lunetta, George McMahon, Ilja M Nolte, Lavinia Paternoster, Eleonora Porcu, Albert V Smith, Lisette Stolk, Alexander Teumer, Natalia Tšernikova, Emmi Tikkanen, Sheila Ulivi, Erin K Wagner, Najaf Amin, Laura J Bierut, Enda M Byrne, Jouke-Jan Hottenga, Daniel L Koller, Massimo Mangino, Tune H Pers, Laura M Yerges-Armstrong, Jing Hua Zhao, Irene L Andrulis, Hoda Anton-Culver, Femke Atsma, Stefania Bandinelli, Matthias W Beckmann, Javier Benitez, Carl Blomqvist, Stig E Bojesen, Manjeet K Bolla, Bernardo Bonanni, Hiltrud Brauch, Hermann Brenner, Julie E Buring, Jenny Chang-Claude, Stephen Chanock, Jinhui Chen, Georgia Chenevix-Trench, J Margriet Collée, Fergus J Couch, David Couper, Andrea D Coveillo, Angela Cox, Kamila Czene, Adamo Pio D'adamo, George Davey Smith, Immaculata De Vivo, Ellen W Demerath, Joe Dennis, Peter Devilee, Aida K Dieffenbach, Alison M Dunning, Gudny Eiriksdottir, Johan G Eriksson, Peter A Fasching, Luigi Ferrucci, Dieter Flesch-Janys, Henrik Flyger, Tatiana Foroud,

Lude Franke, Melissa E Garcia, Montserrat García-Closas, Frank Geller, Eco Ej de Geus, Graham G Giles, Daniel F Gudbjartsson, Vilmundur Gudnason, Pascal Guénel, Suiqun Guo, Per Hall, Ute Hamann, Robin Haring, Catharina A Hartman, Andrew C Heath, Albert Hofman, Maartje J Hooning, John L Hopper, Frank B Hu, David J Hunter, David Karasik, Douglas P Kiel, Julia A Knight, Veli-Matti Kosma, Zoltan Kutalik, Sandra Lai, Diether Lambrechts, Annika Lindblom, Reedik Mägi, Patrik K Magnusson, Arto Mannermaa, Nicholas G Martin, Gisli Masson, Patrick F McArdle, Wendy L McArdle, Mads Melbye, Kyriaki Michailidou, Evelin Mihailov, Lili Milani, Roger L Milne, Heli Nevanlinna, Patrick Neven, Ellen A Nohr, Albertine J Oldehinkel, Ben A Oostra, Aarno Palotie, Munro Peacock, Nancy L Pedersen, Paolo Peterlongo, Julian Peto, Paul Dp Pharoah, Dirkje S Postma, Anneli Pouta, Katri Pylkäs, Paolo Radice, Susan Ring, Fernando Rivadeneira, Antonietta Robino, Lynda M Rose, Anja Rudolph, Veikko Salomaa, Serena Sanna, David Schlessinger, Marjanka K Schmidt, Mellissa C Southey, Ulla Sovio, Meir J Stampfer, Doris Stöckl, Anna M Storniolo, Nicholas J Timpson, Jonathan Tyrer, Jenny A Visser, Peter Vollenweider, Henry Völzke, Gerard Waeber, Melanie Waldenberger, Henri Wallaschofski, Qin Wang, Gonneke Willemsen, Robert Winqvist, Bruce Hr Wolffenbuttel, Margaret J Wright, Australian Ovarian Cancer Study, GENICA Network, kConFab, LifeLines Cohort Study, InterAct Consortium, Early Growth Genetics (EGG) Consortium, Dorret I Boomsma, Michael J Econs, Kay-Tee Khaw, Ruth Jf Loos, Mark I McCarthy, Grant W Montgomery, John P Rice, Elizabeth A Streeten, Unnur Thorsteinsdottir, Cornelia M van Duijn, Behrooz Z Alizadeh, Sven Bergmann, Eric Boerwinkle, Heather A Boyd, Laura Crisponi, Paolo Gasparini, Christian Gieger, Tamara B Harris, Erik Ingelsson, Marjo-Riitta Järvelin, Peter Kraft, Debbie Lawlor, Andres Metspalu, Craig E Pennell, Paul M Ridker, Harold Snieder, Thorkild Ia Sørensen, Tim D Spector, David P Strachan, Andr G Uitterlinden, Nicholas J Wareham, Elisabeth Widen, Marek Zygmunt, Anna Murray, Douglas F Easton, Kari Stefansson, Joanne M Murabito, and Ken K Ong. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, 514(7520):92–97, 10 2014.

[151] Nitin Phadnis and H. Allen Orr. A single gene causes both male sterility and segregation distortion in Drosophila hybrids. *Science (New York, N.Y.)*, 323(5912):376–9, 1 2009.

[152] J. K. Pickrell, N. Patterson, P.-R. Loh, M. Lipson, B. Berger, M. Stoneking, B. Pakendorf, and D. Reich. Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences*, 111(7):2632–2637, 2 2014.

[153] Joseph K. Pickrell and Jonathan K. Pritchard. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, 8(11):e1002967, 11 2012.

[154] Vincent Plagnol and Jeffrey D. Wall. Possible ancestral structure in human populations. *PLoS genetics*, 2(7), 2006.

[155] Hendrik N Poinar, Carsten Schwarz, Ji Qi, Beth Shapiro, Ross D E Macphee, Bernard Buigues, Alexei Tikhonov, Daniel H Huson, Lynn P Tomsho, Alexander Auch, Markus

Rampp, Webb Miller, and Stephan C Schuster. Metagenomics to Paleogenomics. *Science*, 311(January):392–394, 2006.

[156] Renato Polimanti, Hongyu Zhao, Lindsay A Farrer, Henry R Kranzler, and Joel Gelernter. Ancestry-specific and sex-specific risk alleles identified in a genome-wide gene-by-alcohol dependence interaction study of risky sexual behaviors. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 174(8):846–853, 12 2017.

[157] J. E. Pool and R. Nielsen. Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. *Genetics*, 181:711–719, 2008.

[158] Iain M Porter, Katharina Schleicher, Michael Porter, and Jason R Swedlow. Bod1 regulates protein phosphatase 2A at mitotic kinetochores. *Nature communications*, 4:2677, 2013.

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

[160] Alkes L. Price, Arti Tandon, Nick Patterson, Kathleen C. Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H. Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genetics*, 5(6):e1000519, 6 2009.

[161] J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[162] Kay Prüfer. snpAD: An ancient DNA genotype caller. *Bioinformatics (Oxford, England)*, 6 2018.

[163] Kay Prüfer, Cesare de Filippo, Steffi Grote, Fabrizio Mafessoni, Petra Korlević, Mateja Hajdinjak, Benjamin Vernot, Laurits Skov, Pinghsun Hsieh, Stphane Peyrégne, David Reher, Charlotte Hopfe, Sarah Nagel, Tomislav Maricic, Qiaomei Fu, Christoph Theunert, Rebekah Rogers, Pontus Skoglund, Manjusha Chintalapati, Michael Dannemann, Bradley J Nelson, Felix M Key, Pavao Rudan, eljko Kućan, Ivan Gušić, Liubov V Golovanova, Vladimir B Doronichev, Nick Patterson, David Reich, Evan E Eichler, Montgomery Slatkin, Mikkel H Schierup, Aida M Andrés, Janet Kelso, Matthias Meyer, and Svante Pääbo. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science (New York, N.Y.)*, 358(6363):655–658, 2017.

[164] Kay Prüfer, Bjoern Muetzel, Hong-Hai Do, Gunter Weiss, Philipp Khaitovich, Erhard Rahm, Svante Pääbo, Michael Lachmann, and Wolfgang Enard. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC bioinformatics*, 8:41, 2007.

[165] Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C Mullikin, Samuel H Vohr, Richard E Green, Ines Hellmann, Philip L F Johnson, Hlne Blanche, Howard Cann, Jacob O Kitzman, Jay Shendure, Evan E Eichler, Ed S Lein, Trygve E Bakken, Liubov V Golovanova, Vladimir B Doronichev, Michael V Shunkov, Anatoli P Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505:43–9, 2014.

[166] Pengfei Qin and Mark Stoneking. Denisovan Ancestry in East Eurasian and Native American Populations. *Molecular Biology and Evolution*, 32(10):2665–2674, 10 2015.

[167] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842, 2010.

[168] R Core Team. R: A Language and Environment for Statistical Computing, 2015.

[169] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[170] Fernando Racimo, Davide Marnetto, and Emilia Huerta-Sánchez. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Molecular biology and evolution*, 34(2):296–317, 2017.

[171] Fernando Racimo, Sriram Sankararaman, Rasmus Nielsen, and Emilia Huerta-Sánchez. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371, 2015.

[172] Maanasa Raghavan, Michael DeGiorgio, Anders Albrechtsen, Ida Moltke, Pontus Skoglund, Thorfinn S Korneliussen, Bjarne Grønnow, Martin Appelt, Hans Christian Gulløv, T Max Friesen, William Fitzhugh, Helena Malmström, Simon Rasmussen, Jesper Olsen, Linea Melchior, Benjamin T Fuller, Simon M Fahrni, Thomas Stafford, Vaughan Grimes, M A Priscilla Renouf, Jerome Cybulski, Niels Lynnerup, Marta Mirazon Lahr, Kate Britton, Rick Knecht, Jette Arneborg, Mait Metspalu, Omar E Cornejo, Anna-Sapfo Malaspinas, Yong Wang, Morten Rasmussen, Vibha Raghavan, Thomas V O Hansen, Elza Khusnutdinova, Tracey Pierre, Kirill Dneprovsky, Claus Andreasen, Hans Lange, M Geoffrey Hayes, Joan Coltrain, Victor A Spitsyn, Anders Götherström, Ludovic Orlando, Toomas Kivisild, Richard Villems, Michael H Crawford, Finn C Nielsen, Jrgen Dissing, Jan Heinemeier, Morten Meldgaard, Carlos Bustamante, Dennis H O'Rourke, Mattias Jakobsson, M Thomas P Gilbert, Rasmus Nielsen, and Eske Willerslev. The genetic prehistory of the New World Arctic. *Science (New York, N.Y.)*, 345:1255832, 2014.

[173] Maanasa Raghavan, Pontus Skoglund, Kelly E Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, Thomas W Stafford, Ludovic Orlando, Ene Metspalu, Monika Karmin, Kristiina Tambets, Siiri Rootsi, Reedik Mägi, Paula F Campos, Elena Balanovska, Oleg Balanovsky, Elza Khusnutdinova, Sergey Litvinov, Ludmila P Osipova, Sardana a Fedorova, Mikhail I Voevoda, Michael DeGiorgio, Thomas Sicheritz-Ponten, Sren Brunak, Svetlana Demeshchenko, Toomas Kivisild, Richard Villems, Rasmus Nielsen, Mattias Jakobsson, and Eske Willerslev. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, 505:87–91, 2014.

[174] Maanasa Raghavan, M Steinrucken, Kelley Harris, Stephan Schiffels, S. Rasmussen, M DeGiorgio, Anders Albrechtsen, Cristina Valdiosera, M C Avila-Arcos, A-S Malaspinas, Anders Eriksson, Ida Moltke, M. Metspalu, Julian R Homburger, Jeff Wall, Omar E Cornejo, J V Moreno-Mayar, Thorfinn S Korneliussen, Tracey Pierre, Morten Rasmussen, Paula F Campos, P D B Damgaard, Morten E Allentoft, John Lindo, Ene Metspalu, R Rodriguez-Varela, Josefina Mansilla, Celeste Henrickson, A Seguin-Orlando, H Malmstrom, T Stafford, S S Shringarpure, A Moreno-Estrada, Monika Karmin, Kristiina Tambets, A Bergstrom, Yali Xue, Vera Warmuth, Andrew D Friend, Joy Singarayer, Paul Valdes, Francois Balloux, Iln Leboreiro, J L Vera, H Rangel-Villalobos, Davide Pettener, Donata Luiselli, Loren G Davis, Evelyne Heyer, Christoph P E Zollikofer, M S Ponce de Leon, C. I. Smith, Vaughan Grimes, K-a Pike, Michael Deal, Benjamin T Fuller, B Arriaza, V Standen, M F Luz, F Ricaut, N Guidon, L Osipova, M I Voevoda, O L Posukh, O Balanovsky, M Lavryashina, Y Bogunov, Elza Khusnutdinova, M Gubina, Elena Balanovska, S. Fedorova, Sergey Litvinov, B. Malyarchuk, M. Derenko, M J Mosher, D Archer, J Cybulski, B Petzelt, J. Mitchell, R Worl, P J Norman, Peter Parham, B M Kemp, Toomas Kivisild, Chris Tyler-Smith, M S Sandhu, M Crawford, Richard Villems, D G Smith, M R Waters, T Goebel, J R Johnson, R S Malhi, Mattias Jakobsson, D J Meltzer, Andrea Manica, Richard Durbin, C D Bustamante, Y S Song, Rasmus Nielsen, and Eske Willerslev. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*, 349(6250):aab3884–aab3884, 8 2015.

[175] Andrew Rambaut and Nicholas C. Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 1997.

[176] Matthew D. Rasmussen, Melissa J. Hubisz, Ilan Gronau, and Adam Siepel. Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics*, 10(5):e1004342, 5 2014.

[177] Morten Rasmussen, Sarah L Anzick, Michael R Waters, Pontus Skoglund, Michael DeGiorgio, Thomas W Stafford, Simon Rasmussen, Ida Moltke, Anders Albrechtsen, Shane M Doyle, G David Poznik, Valborg Gudmundsdottir, Rachita Yadav, Anna-Sapfo Malaspinas, Samuel Stockton White, Morten E Allentoft, Omar E Cornejo, Kristiina Tambets, Anders Eriksson, Peter D Heintzman, Monika Karmin, Thorfinn Sand Korneliussen, David J Meltzer, Tracey L Pierre, Jesper Stenderup, Lauri Saag, Vera M War-

muth, Margarida C Lopes, Ripan S Malhi, Sren Brunak, Thomas Sicheritz-Ponten, Ian Barnes, Matthew Collins, Ludovic Orlando, Francois Balloux, Andrea Manica, Ramneek Gupta, Mait Metspalu, Carlos D Bustamante, Mattias Jakobsson, Rasmus Nielsen, and Eske Willerslev. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, 506:225–9, 2014.

[178] Morten Rasmussen, Martin Sikora, Anders Albrechtsen, Thorfinn Sand Korneliussen, J. Vctor Moreno-Mayar, G. David Poznik, Christoph P. E. Zollikofer, Marcia S. Ponce de León, Morten E. Allentoft, Ida Moltke, Hkon Jónsson, Cristina Valdiosera, Ripan S. Malhi, Ludovic Orlando, Carlos D. Bustamante, Thomas W. Stafford, David J. Meltzer, Rasmus Nielsen, and Eske Willerslev. The ancestry and affiliations of Kennewick Man. *Nature*, 523:455–458, 6 2015.

[179] David Reich, Richard E Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y Durand, Bence Viola, Adrian W Briggs, Udo Stenzel, Philip L F Johnson, Tomislav Maricic, Jeffrey M Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V Shunkov, Anatoli P Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin, and Svante Pääbo. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468:1053–1060, 2010.

[180] David Reich, Nick Patterson, Desmond Campbell, Arti Tandon, Stphane Mazieres, Nicolas Ray, Maria V. Parra, Winston Rojas, Constanza Duque, Natalia Mesa, Luis F. García, Omar Triana, Silvia Blair, Amanda Maestre, Juan C. Dib, Claudio M. Bravi, Graciela Bailliet, Daniel Corach, Tbita Hünemeier, Maria Ctira Bortolini, Francisco M. Salzano, Mara Luiza Petzl-Erler, Victor Acuña-Alonzo, Carlos Aguilar-Salinas, Samuel Canizales-Quinteros, Teresa Tusié-Luna, Laura Riba, Maricela Rodríguez-Cruz, Mardia Lopez-Alarcón, Ramn Coral-Vazquez, Thelma Canto-Cetina, Irma Silva-Zolezzi, Juan Carlos Fernandez-Lopez, Alejandra V. Contreras, Gerardo Jimenez-Sanchez, Maria Jos Gómez-Vázquez, Julio Molina, ngel Carracedo, Antonio Salas, Carla Gallo, Giovanni Poletti, David B. Witonsky, Gorka Alkorta-Aranburu, Rem I. Sukernik, Ludmila Osipova, Sardana a. Fedorova, Ren Vasquez, Mercedes Villena, Claudia Moreau, Ramiro Barrantes, David Pauls, Laurent Excoffier, Gabriel Bedoya, Francisco Rothhammer, Jean-Michel Dugoujon, Georges Larrouy, William Klitz, Damian Labuda, Judith Kidd, Kenneth Kidd, Anna Di Rienzo, Nelson B. Freimer, Alkes L. Price, and Andrs Ruiz-Linares. Reconstructing Native American population history. *Nature*, 488:370–374, 2012.

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

[182] David E. Reich and Eric S. Lander. On the allelic spectrum of human disease. *Trends in Genetics*, 17(9):502–510, 2001.

[183] E Rhéaume, Y Lachance, H F Zhao, N Breton, M Dumont, Y de Launoit, C Trudel, V Luu-The, J Simard, and F Labrie. Structure and expression of a new complementary

DNA encoding the almost exclusive 3 beta-hydroxysteroid dehydrogenase/delta 5-delta 4-isomerase in human adrenals and gonads. *Molecular endocrinology (Baltimore, Md.)*, 5(8):1147–57, 8 1991.

[184] Jos Rino, Joana M P Desterro, Teresa R Pacheco, Theodorus W J Gadella, and Maria Carmo-Fonseca. Splicing factors SF1 and U2AF associate in extraspliceosomal complexes. *Molecular and cellular biology*, 28(9):3045–57, 5 2008.

[185] Katherine S Ruth, Purdey J Campbell, Shelby Chew, Ee Mun Lim, Narelle Hadlow, Bronwyn G A Stuckey, Suzanne J Brown, Bjarke Feenstra, John Joseph, Gabriela L Surdulescu, Hou Feng Zheng, J Brent Richards, Anna Murray, Tim D Spector, Scott G Wilson, and John R B Perry. Genome-wide association study with 1000 genomes imputation identifies signals for nine sex hormone-related phenotypes. *European journal of human genetics : EJHG*, 24(2):284–90, 2 2016.

[186] Sriram Sankararaman, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David Reich. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507:354–7, 2014.

[187] Sriram Sankararaman, Swapan Mallick, Nick Patterson, and David Reich. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology*, 26(9):1241–1247, 5 2016.

[188] Sriram Sankararaman, Nick Patterson, Heng Li, Svante Pääbo, and David Reich. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS Genetics*, 8(10):e1002947, 10 2012.

[189] Nathan K Schaefer, Beth Shapiro, and Richard E Green. Detecting hybridization using ancient DNA. *Molecular ecology*, 25(11):2398–2412, 1 2016.

[190] Stephen F Schaffner. The X chromosome in population genetics. *Nat Rev Genet*, 5(January):43–51, 2004.

[191] Carina M Schlebusch, Helena Malmström, Torsten Günther, Per Sjödin, Alexandra Coutinho, Hanna Edlund, Arielle R Munters, Mrio Vicente, Maryna Steyn, Himla Soodyall, Marlize Lombard, and Mattias Jakobsson. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science (New York, N.Y.)*, 358(6363):652–655, 2017.

[192] Andaine Seguin-Orlando, Thorfinn S Korneliussen, Martin Sikora, A.-S. Malaspinas, Andrea Manica, Ida Moltke, Anders Albrechtsen, Amy Ko, Ashot Margaryan, Vyacheslav Moiseyev, Ted Goebel, Michael Westaway, David Lambert, Valeri Khartanovich, Jeffrey D. Wall, Philip R. Nigst, Robert A. Foley, Marta Mirazon Lahr, Rasmus Nielsen, Ludovic Orlando, and Eske Willerslev. Genomic structure in Europeans dating back at least 36,200 years. *Science*, 346(6213):1113–1118, 11 2014.

[193] David Serre, Andr Langaney, Mario Chech, Maria Teschler-Nicola, Maja Paunovic, Philippe Mennecier, Michael Hofreiter, Gran Possnert, and Svante Pääbo. No evidence of neandertal mtDNA contribution to early modern humans. *PLoS biology*, 2(3):313–317, 2004.

[194] B Shapiro and M Hofreiter. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science (New York, N.Y.)*, 343(January):1236573, 2014.

[195] Jin Na Shin, Injune Kim, Jung Sup Lee, Gou Young Koh, Zang Hee Lee, and Hong-Hee Kim. A novel zinc finger protein that inhibits osteoclastogenesis and the function of tumor necrosis factor receptor-associated factor 6. *The Journal of biological chemistry*, 277(10):8346–53, 3 2002.

[196] Adam Siepel. Phylogenomics of primates and their ancestral populations. *Genome Research*, 19:1929–1941, 2009.

[197] C. N. Simonti, B. Vernot, L. Bastarache, E. Bottinger, D. S. Carrell, R. L. Chisholm, D. R. Crosslin, S. J. Hebbring, G. P. Jarvik, I. J. Kullo, R. Li, J. Pathak, M. D. Ritchie, D. M. Roden, S. S. Verma, G. Tromp, J. D. Prato, W. S. Bush, J. M. Akey, J. C. Denny, and J. A. Capra. The phenotypic legacy of admixture between modern humans and Neandertals. *Science*, 351(6274):737–741, 2 2016.

[198] Abanish Singh, Michael A Babyak, Daniel K Nolan, Beverly H Brummett, Rong Jiang, Ilene C Siegler, William E Kraus, Svati H Shah, Redford B Williams, and Elizabeth R Hauser. Gene by stress genome-wide interaction analysis and path analysis identify EBF1 as a cardiovascular and metabolic risk gene. *European journal of human genetics : EJHG*, 23(6):854–62, 6 2015.

[199] Pontus Skoglund, Swapan Mallick, Maria Ctira Bortolini, Niru Chennagiri, Tbita Hünemeier, Maria Luiza Petzl-Erler, Francisco Mauro Salzano, Nick Patterson, and David Reich. Genetic evidence for two founding populations of the Americas. *Nature*, 525:104–108, 7 2015.

[200] Yun S Song and Jotun Hein. Constructing minimal ancestral recombination graphs. *Journal of computational biology*, 12(2):147–169, 2005.

[201] Paul R Staab, Sha Zhu, Dirk Metzler, and Gerton Lunter. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics (Oxford, England)*, 31(10):1680–2, 5 2015.

[202] Matthias Steinrücken, Joshua S. Paul, and Yun S. Song. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theoretical Population Biology*, 87:51–61, 2013.

[203] Ravi P Subramanian, Julia H Wildschutte, Crystal Russo, and John M Coffin. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology*, 8:90, 11 2011.

[204] Sayaka Sukegawa, Eri Miyagi, Fadila Bouamr, Helena Farkašová, and Klaus Strebel. Mannose Receptor 1 Restricts HIV Particle Release from Infected Macrophages. *Cell reports*, 22(3):786–795, 2018.

[205] Benjamin B Sun, Joseph C Maranville, James E Peters, David Stacey, James R Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, Clare Oliver-Williams, Mihir A Kamat, Bram P Prins, Sheri K Wilcox, Erik S Zimmerman, An Chi, Narinder Bansal, Sarah L Spain, Angela M Wood, Nicholas W Morrell, John R Bradley, Nebojsa Janjic, David J Roberts, Willem H Ouwehand, John A Todd, Nicole Soranzo, Karsten Suhre, Dirk S Paul, Caroline S Fox, Robert M Plenge, John Danesh, Heiko Runz, and Adam S Butterworth. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 2018.

[206] Andreas Sundquist, Eugene Fratkin, Chuong B. Do, and Serafim Batzoglou. Effects of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, 18(4):676–682, 2008.

[207] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research*, 45(D1):D362–D368, 2017.

[208] Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing Genetic Ancestry Blocks in Admixed Individuals. *The American Journal of Human Genetics*, 79(1):1–12, 7 2006.

[209] Fasil Tekola Ayele, Ayo Doumatey, Hanxia Huang, Jie Zhou, Bashira Charles, Michael Erdos, Jokotade Adeleye, Williams Balogun, Olufemi Fasanmade, Thomas Johnson, Johnnie Oli, Godfrey Okafor, Albert Amoah, Benjamin A Eghan, Kofi Agyenim-Boateng, Joseph Acheampong, Clement A Adebamowo, Alan Herbert, Norman Gerry, Michael Christman, Guanjie Chen, Daniel Shriner, Adebowale Adeyemo, and Charles N Rotimi. Genome-wide associated loci influencing interleukin (IL)-10, IL-1Ra, and IL-6 levels in African Americans. *Immunogenetics*, 64(5):351–9, 5 2012.

[210] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, Christopher T Johansen, Sigrid W Fouchier, Aaron Isaacs, Gina M Peloso, Maja Barbalic, Sally L Ricketts, Joshua C Bis, Yurii S Aulchenko, Gudmar Thorleifsson, Mary F Feitosa, John Chambers, Marju Orho-Melander, Olle Melander, Toby Johnson, Xiaohui Li, Xiuqing Guo, Mingyao Li, Yoon Shin Cho, Min Jin Go, Young Jin Kim, Jong-Young Lee, Taesung Park, Kyunga Kim, Xueling Sim, Rick Twee-Hee Ong, Damien C Croteau-Chonka, Leslie A Lange, Joshua D Smith, Kijoung Song, Jing Hua Zhao, Xin Yuan, Jian'an Luan, Claudia Lamina, Andreas Ziegler, Weihua Zhang, Robert Y L Zee, Alan F Wright, Jacqueline C M Witteman, James F Wilson, Gonneke Willemsen, H-Erich Wichmann, John B Whitfield, Dawn M Waterworth,

Nicholas J Wareham, Grard Waeber, Peter Vollenweider, Benjamin F Voight, Veronique Vitart, Andre G Uitterlinden, Manuela Uda, Jaakko Tuomilehto, John R Thompson, Toshiko Tanaka, Ida Surakka, Heather M Stringham, Tim D Spector, Nicole Soranzo, Johannes H Smit, Juha Sinisalo, Kaisa Silander, Eric J G Sijbrands, Angelo Scuteri, James Scott, David Schlessinger, Serena Sanna, Veikko Salomaa, Juha Saharinen, Chiara Sabatti, Aimo Ruokonen, Igor Rudan, Lynda M Rose, Robert Roberts, Mark Rieder, Bruce M Psaty, Peter P Pramstaller, Irene Pichler, Markus Perola, Brenda W J H Penninx, Nancy L Pedersen, Cristian Pattaro, Alex N Parker, Guillaume Pare, Ben A Oostra, Christopher J O'Donnell, Markku S Nieminen, Deborah A Nickerson, Grant W Montgomery, Thomas Meitinger, Ruth McPherson, Mark I McCarthy, Wendy McArdle, David Masson, Nicholas G Martin, Fabio Marroni, Massimo Mangino, Patrik K E Magnusson, Gavin Lucas, Robert Luben, Ruth J F Loos, Marja-Liisa Lokki, Guillaume Lettre, Claudia Langenberg, Lenore J Launer, Edward G Lakatta, Reijo Laaksonen, Kirsten O Kyvik, Florian Kronenberg, Inke R König, Kay-Tee Khaw, Jaakko Kaprio, Lee M Kaplan, Asa Johansson, Marjo-Riitta Jarvelin, A Cecile J W Janssens, Erik Ingelsson, Wilmar Igl, G Kees Hovingh, Jouke-Jan Hottenga, Albert Hofman, Andrew A Hicks, Christian Hengstenberg, Iris M Heid, Caroline Hayward, Aki S Havulinna, Nicholas D Hastie, Tamara B Harris, Talin Haritunians, Alistair S Hall, Ulf Gyllensten, Candace Guiducci, Leif C Groop, Elena Gonzalez, Christian Gieger, Nelson B Freimer, Luigi Ferrucci, Jeanette Erdmann, Paul Elliott, Kenechi G Ejebe, Angela Döring, Anna F Dominiczak, Serkalem Demissie, Panagiotis Deloukas, Eco J C de Geus, Ulf de Faire, Gabriel Crawford, Francis S Collins, Yii-der I Chen, Mark J Caulfield, Harry Campbell, Noel P Burtt, Lori L Bonnycastle, Dorret I Boomsma, S Matthijs Boekholdt, Richard N Bergman, Inłs Barroso, Stefania Bandinelli, Christie M Ballantyne, Themistocles L Assimes, Thomas Quertermous, David Altshuler, Mark Seielstad, Tien Y Wong, E-Shyong Tai, Alan B Feranil, Christopher W Kuzawa, Linda S Adair, Herman A Taylor, Ingrid B Borecki, Stacey B Gabriel, James G Wilson, Hilma Holm, Unnur Thorsteinsdottir, Vilmundur Gudnason, Ronald M Krauss, Karen L Mohlke, Jose M Ordovas, Patricia B Munroe, Jaspal S Kooner, Alan R Tall, Robert A Hegele, John J P Kastelein, Eric E Schadt, Jerome I Rotter, Eric Boerwinkle, David P Strachan, Vincent Mooser, Kari Stefansson, Muredach P Reilly, Nilesh J Samani, Heribert Schunkert, L Adrienne Cupples, Manjinder S Sandhu, Paul M Ridker, Daniel J Rader, Cornelia M van Duijn, Leena Peltonen, Gonalo R Abecasis, Michael Boehnke, and Sekar Kathiresan. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–13, 8 2010.

[211] Serena Tucci, Samuel H. Vohr, Rajiv C. McCoy, Benjamin Vernot, Matthew R. Robinson, Chiara Barbieri, Brad J. Nelson, Wenqing Fu, Gludhug A. Purnomo, Herawati Sudoyo, Evan E. Eichler, Guido Barbujani, Peter M. Visscher, Joshua M. Akey, and Richard E. Green. Evolutionary history and adaptation of a human pygmy population of Flores Island, Indonesia. *Science*, 361(6401):511–516, 2018.

[212] Pim van der Harst and Niek Verweij. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation research*, 122(3):433–443, 2 2018.

[213] B. Vernot, S. Tucci, J. Kelso, J. G. Schraiber, A. B. Wolf, R. M. Gittelman, M. Dannemann, S. Grote, R. C. McCoy, H. Norton, L. B. Scheinfeldt, D. A. Merriwether, G. Koki, J. S. Friedlaender, J. Wakefield, S. Paabo, and J. M. Akey. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science*, 352(6282):235–239, 4 2016.

[214] Benjamin Vernot and Joshua M Akey. Resurrecting surviving Neandertal lineages from modern human genomes. *Science (New York, N.Y.)*, 343:1017–21, 2014.

[215] Benjamin Vernot and Joshua M. Akey. Complex history of admixture between modern humans and Neandertals. *American journal of human genetics*, 96(3):448–53, 3 2015.

[216] Jeffrey D. Wall. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, 154:1271–1279, 2000.

[217] Jeffrey D. Wall, Kirk E. Lohmueller, and Vincent Plagnol. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular Biology and Evolution*, 26:1823–1827, 2009.

[218] Jeffrey D. Wall, Melinda a. Yang, Flora Jay, Sung K. Kim, Eric Y. Durand, Laurie S. Stevison, Christopher Gignoux, August Woerner, Michael F. Hammer, and Montgomery Slatkin. Higher levels of Neanderthal ancestry in east Asians than in Europeans. *Genetics*, 194(May):199–209, 2013.

[219] L Wang, H Oota, N Saitou, F Jin, T Matsushita, and S Ueda. Genetic structure of a 2,500-year-old human population in China and its spatiotemporal changes. *Molecular biology and evolution*, 17:1396–1400, 2000.

[220] Cavin K Ward-Caviness, Lucas M Neas, Colette Blach, Carol S Haynes, Karen LaRocque-Abramson, Elizabeth Grass, Elaine Dowdy, Robert B Devlin, David Diaz-Sanchez, Wayne E Cascio, Marie Lynn Miranda, Simon G Gregory, Svati H Shah, William E Kraus, and Elizabeth R Hauser. Genetic Variants in the Bone Morphogenic Protein Gene Family Modify the Association between Residential Exposure to Traffic and Peripheral Arterial Disease. *PloS one*, 11(4):e0152670, 2016.

[221] Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, Jin Chen, Martin L Buchkovich, Samia Mora, Jacques S Beckmann, Jennifer L Bragg-Gresham, Hsing-Yi Chang, Aye Demirkan, Heleen M Den Hertog, Ron Do, Louise A Donnelly, Georg B Ehret, Tnu Esko, Mary F Feitosa, Teresa Ferreira, Krista Fischer, Pierre Fontanillas, Ross M Fraser, Daniel F Freitag, Deepti Gurdasani, Kauko Heikkilä, Elina Hyppönen, Aaron Isaacs, Anne U Jackson, sa Johansson, Toby Johnson, Marika Kaakinen, Johannes Kettunen, Marcus E Kleber, Xiaohui Li, Jian'an Luan, Leo-Pekka Lyytikäinen, Patrik K E Magnusson, Massimo Mangino, Evelin Mihailov, May E Montasser, Martina Müller-Nurasyid, Ilja M Nolte, Jeffrey R O'Connell, Cameron D Palmer, Markus Perola, Ann-Kristin Petersen, Serena Sanna, Richa Saxena, Susan K Service, Sonia Shah, Dmitry Shungin, Carlo Sidore,

Ci Song, Rona J Strawbridge, Ida Surakka, Toshiko Tanaka, Tanya M Teslovich, Gudmar Thorleifsson, Evita G Van den Herik, Benjamin F Voight, Kelly A Volcik, Lindsay L Waite, Andrew Wong, Ying Wu, Weihua Zhang, Devin Absher, Gershim Asiki, Inłs Barroso, Latonya F Been, Jennifer L Bolton, Lori L Bonnycastle, Paolo Brambilla, Mary S Burnett, Giancarlo Cesana, Maria Dimitriou, Alex S F Doney, Angela Döring, Paul Elliott, Stephen E Epstein, Gudmundur Ingi Eyjolfsson, Bruna Gigante, Mark O Goodarzi, Harald Grallert, Martha L Gravito, Christopher J Groves, Gran Hallmans, Anna-Liisa Hartikainen, Caroline Hayward, Dena Hernandez, Andrew A Hicks, Hilma Holm, Yi-Jen Hung, Thomas Illig, Michelle R Jones, Pontiano Kaleebu, John J P Kastelein, Kay-Tee Khaw, Eric Kim, Norman Klopp, Pirjo Komulainen, Meena Kumari, Claudia Langenberg, Terho Lehtimäki, Shih-Yi Lin, Jaana Lindström, Ruth J F Loos, Franois Mach, Wendy L McArdle, Christa Meisinger, Braxton D Mitchell, Gabrielle Müller, Ramaiah Nagaraja, Narisu Narisu, Tuomo V M Nieminen, Rebecca N Nsubuga, Isleifur Olafsson, Ken K Ong, Aarno Palotie, Theodore Papamarkou, Cristina Pomilla, Anneli Pouta, Daniel J Rader, Muredach P Reilly, Paul M Ridker, Fernando Rivadeneira, Igor Rudan, Aimo Ruokonen, Nilesh Samani, Hubert Scharnagl, Janet Seeley, Kaisa Silander, Alena Stančáková, Kathleen Stirrups, Amy J Swift, Laurence Tiret, Andre G Uitterlinden, L Joost van Pelt, Sailaja Vedantam, Nicholas Wainwright, Cisca Wijmenga, Sarah H Wild, Gonneke Willemsen, Tom Wilsgaard, James F Wilson, Elizabeth H Young, Jing Hua Zhao, Linda S Adair, Dominique Arveiler, Themistocles L Assimes, Stefania Bandinelli, Franklyn Bennett, Murielle Bochud, Bernhard O Boehm, Dorret I Boomsma, Ingrid B Borecki, Stefan R Bornstein, Pascal Bovet, Michel Burnier, Harry Campbell, Aravinda Chakravarti, John C Chambers, Yii-Der Ida Chen, Francis S Collins, Richard S Cooper, John Danesh, George Dedoussis, Ulf de Faire, Alan B Feranil, Jean Ferrières, Luigi Ferrucci, Nelson B Freimer, Christian Gieger, Leif C Groop, Vilmundur Gudnason, Ulf Gyllensten, Anders Hamsten, Tamara B Harris, Aroon Hingorani, Joel N Hirschhorn, Albert Hofman, G Kees Hovingh, Chao Agnes Hsiung, Steve E Humphries, Steven C Hunt, Kristian Hveem, Carlos Iribarren, Marjo-Riitta Järvelin, Antti Jula, Mika Kähönen, Jaakko Kaprio, Antero Kesäniemi, Mika Kivimaki, Jaspal S Kooner, Peter J Koudstaal, Ronald M Krauss, Diana Kuh, Johanna Kuusisto, Kirsten O Kyvik, Markku Laakso, Timo A Lakka, Lars Lind, Cecilia M Lindgren, Nicholas G Martin, Winfried März, Mark I McCarthy, Colin A McKenzie, Pierre Meneton, Andres Metspalu, Leena Moilanen, Andrew D Morris, Patricia B Munroe, Inger Njølstad, Nancy L Pedersen, Chris Power, Peter P Pramstaller, Jackie F Price, Bruce M Psaty, Thomas Quertermous, Rainer Rauramaa, Danish Saleheen, Veikko Salomaa, Dharambir K Sanghera, Jouko Saramies, Peter E H Schwarz, Wayne H-H Sheu, Alan R Shuldiner, Agneta Siegbahn, Tim D Spector, Kari Stefansson, David P Strachan, Bamidele O Tayo, Elena Tremoli, Jaakko Tuomilehto, Matti Uusitupa, Cornelia M van Duijn, Peter Vollenweider, Lars Wallentin, Nicholas J Wareham, John B Whitfield, Bruce H R Wolffenbuttel, Jose M Ordovas, Eric Boerwinkle, Colin N A Palmer, Unnur Thorsteinsdottir, Daniel I Chasman, Jerome I Rotter, Paul W Franks, Samuli Ripatti, L Adrienne Cupples, Manjinder S Sandhu, Stephen S Rich, Michael Boehnke, Panos Deloukas, Sekar Kathiresan, Karen L Mohlke, Erik Ingelsson, Gonalo R Abecasis, and Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274–1283, 11 2013.

[222] Amy L Williams, Suzanne B R Jacobs, Hortensia Moreno-Macías, Alicia Huerta-Chagoya, Claire Churchhouse, Carla Márquez-Luna, Humberto García-Ortíz, Mara Jos Gómez-Vázquez, Nol P Burtt, Carlos a Aguilar-Salinas, Clicerio González-Villalpando, Jose C Florez, Lorena Orozco, Christopher a Haiman, Teresa Tusié-Luna, and David Altshuler. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*, 506(7486):97–101, 2014.

[223] Aaron B. Wolf and Joshua M. Akey. Outstanding questions in the study of archaic hominin admixture. *PLOS Genetics*, 14(5):e1007349, 5 2018.

[224] Chung I. Wu. The genic view of the process of speciation. *Journal of Evolutionary Biology*, 14:851–865, 2001.

[225] Chung-I Wu and Chau-Ti Ting. Genes and speciation. *Nature reviews. Genetics*, 5(February):114–122, 2004.

[226] Melinda a. Yang, Anna Sapfo Malaspinas, Eric Y. Durand, and Montgomery Slatkin. Ancient structure in Africa unlikely to explain neanderthal and non-african genetic similarity. *Molecular Biology and Evolution*, 29(10):2987–2995, 2012.

[227] Wen-Yun Yang, John Novembre, Eleazar Eskin, and Eran Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics*, 44(6):725–731, 2012.

[228] K W Yau and D A Baylor. Cyclic GMP-activated conductance of retinal photoreceptor cells. *Annual review of neuroscience*, 12:289–327, 1989.

[229] Berran Yucesoy, Kenneth M Kaufman, Zana L Lummus, Matthew T Weirauch, Ge Zhang, Andr Cartier, Louis-Philippe Boulet, Joaquin Sastre, Santiago Quirce, Susan M Tarlo, Maria-Jesus Cruz, Xavier Munoz, John B Harley, and David I Bernstein. Genome-Wide Association Study Identifies Novel Loci Associated With Diisocyanate-Induced Occupational Asthma. *Toxicological sciences : an official journal of the Society of Toxicology*, 146(1):192–201, 7 2015.

[230] Guangju Zhai, Alexander Teumer, Lisette Stolk, John R B Perry, Liesbeth Vandenput, Andrea D Coviello, Annemarie Koster, Jordana T Bell, Shalender Bhasin, Joel Eriksson, Anna Eriksson, Florian Ernst, Luigi Ferrucci, Timothy M Frayling, Daniel Glass, Elin Grundberg, Robin Haring, Asa K Hedman, Albert Hofman, Douglas P Kiel, Heyo K Kroemer, Yongmei Liu, Kathryn L Lunetta, Marcello Maggio, Mattias Lorentzon, Massimo Mangino, David Melzer, Iva Miljkovic, MuTHER Consortium, Alexandra Nica, Brenda W J H Penninx, Ramachandran S Vasan, Fernando Rivadeneira, Kerrin S Small, Nicole Soranzo, Andr G Uitterlinden, Henry Völzke, Scott G Wilson, Li Xi, Wei Vivian Zhuang, Tamara B Harris, Joanne M Murabito, Claes Ohlsson, Anna Murray, Frank H

de Jong, Tim D Spector, and Henri Wallaschofski. Eight common genetic variants associated with serum DHEAS levels suggest a key role in ageing mechanisms. *PLoS genetics*, 7(4):e1002025, 4 2011.

[231] Xiaofeng Zhu, Richard S Cooper, and Robert C Elston. Linkage analysis of a complex disease through use of admixed populations. *American journal of human genetics*, 74:1136–1153, 2004.