THE UNIVERSITY OF CHICAGO

# STATISTICAL METHODS FOR GENETIC DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON EVOLUTIONARY BIOLOGY

BY
HUSSEIN AL-ASADI

CHICAGO, ILLINOIS
AUGUST 2018

Copyright © 2018 by Hussein Al-Asadi

All Rights Reserved

Dedicated to my wife, Nour.

# TABLE OF CONTENTS

LIST OF FIGURES . . . . . vi

ACKNOWLEDGMENTS . . . . . xv

ABSTRACT . . . . . xvi

1 INTRODUCTION . . . . . 1

2 ESTIMATING RECENT MIGRATION AND POPULATION SIZE SURFACES . 9
- 2.1 Introduction . . . . . 9
- 2.2 Results . . . . . 11
  - 2.2.1 Outline of the MAPS method . . . . . 11
  - 2.2.2 Differences from EEMS . . . . . 14
  - 2.2.3 Evaluation of performance under a stepping-stone coalescent model . 14
  - 2.2.4 Applying MAPS to the POPRES data . . . . . 18
- 2.3 Discussion . . . . . 23
- 2.4 Methods . . . . . 27
  - 2.4.1 MAPS configuration . . . . . 27
  - 2.4.2 Inferring PSC segments from the data . . . . . 27
  - 2.4.3 Model . . . . . 27
- 2.5 Supplementary Notes . . . . . 33
  - 2.5.1 The model . . . . . 33
  - 2.5.2 Transformation of migration rates to dispersal rates . . . . . 38
  - 2.5.3 Diversity rates versus coalescent rates . . . . . 40
  - 2.5.4 The prior . . . . . 41
  - 2.5.5 MCMC . . . . . 43

3 SUBSTITUTION RATE HETEROGENEITY CAUSES FALSE POSITIVE INFERENCES IN PHYLOGENETIC BASED-TESTS OF POSITIVE SELECTION . . 46
- 3.1 Introduction . . . . . 46
- 3.2 Results . . . . . 50
  - 3.2.1 Simulating codons with among-site rate heterogeneity. . . . . 50
  - 3.2.2 Moderate substitution rate heterogeneity causes false inferences of positive selection. . . . . 51
  - 3.2.3 Systematic bias is not alleviated by better sequence sampling. . . . . 55
  - 3.2.4 Analysis of empirical sequence data. . . . . 57
  - 3.2.5 Bias vs. loss of power. . . . . 59
  - 3.2.6 Realistic substitution rate heterogeneity is sufficient to produce false inferences of positive selection. . . . . 61
  - 3.2.7 Causes of bias. . . . . 63
- 3.3 Discussion . . . . . 67
- 3.4 Methods . . . . . 71

3.4.1 Controlled simulations . . . . . 71
3.4.2 Analysis . . . . . 72
3.4.3 Sites test and sites test-RH . . . . . 73
3.4.4 Options used for other tests of positive selection . . . . . 74
3.4.5 Analysis of empirical mammalian dataset . . . . . 75
3.4.6 Causes of bias . . . . . 76

4 INFERENCE AND VISUALIZATION OF DNA DAMAGE PATTERNS USING A GRADE OF MEMBERSHIP MODEL. . . . . 78
4.1 Introduction . . . . . 78
4.2 Methods . . . . . 80
4.3 Results . . . . . 82
4.3.1 `aRchaic` clustering of modern and ancient individuals . . . . . 84
4.3.2 The effects of contamination on inferred grades of membership . . . . 85
4.3.3 `aRchaic` can identify both DNA damage and technical artifacts . . . 87
4.4 Discussion . . . . . 89

REFERENCES . . . . . 93

APPENDIX: SUPPLEMENTARY FIGURES . . . . . 105

# LIST OF FIGURES

2.1 **Schematic overview of MAPS.** (a) Coalescent times between a pair of haplotypes (A and B) will vary across the genome in discrete segments bordered by recombination breakpoints. On average, longer segments represent shorter pairwise coalescent times ($T_{AB}$) (b) Flow diagram of MAPS. i) We start with a matrix of called genotypes; ii) lPSC segments between all pairs of chromosomes across the genome are identified from the data using external methods (such as BEAGLE, Browning and Browning [2011]); iii) lPSC segments between pairs of individuals are aggregated at the levels of pairs of populations; iv) A grid is constructed and individuals are assigned to the most nearby node; v) The probability of the PSC sharing matrix can be computed under a stepping-stone model where each node represents a population and each edge represents symmetric migration; vi) We use an MCMC scheme to sample from the posterior distribution of migration rates and population sizes. The final MAPS output is the mean over these posterior samples, and the averaged rates can be transformed to units of dispersal rate and population density. The diagram does not show a bootstrapping step used to estimate likelihood weights to account for correlations between lPSC segments, see Equation (2.6) in Methods. . . . . 12

2.2 **Simulations comparing migration rates inferred with MAPS against effective migration rates inferred with EEMS.** (a) We simulated data under uniform migration rates equal to 0.01 and applied EEMS and MAPS using PSC segments in the range 2-6cM and ≥6cM. Like EEMS, MAPS correctly infers a uniform migration surface. Additionally, MAPS provides accurate estimates of the migration rates for both PSC segments 2-6cM (mean 0.01) and PSC segments ≥6cM (mean 0.0086). (b) We simulated a recent sudden migration barrier formation 10 generations ago. Here, EEMS is unable to infer a barrier, while MAPS correctly infers the historical uniform surface (2-6cM) and a barrier in the more recent time scale (≥6cM). (c) We simulated a long-standing migration barrier that recently dissipated 20 generations ago. EEMS infers a barrier, while MAPS correctly infers both the historical migration barrier (2-6cM) and the uniform migration surface in the more recent time scale (≥6cM). In all cases shown here, we simulated a 20 deme stepping stone model such that the population sizes all equal to 10,000, and 10 diploid individuals were sampled at each deme. . . . . 16

2.3 **Simulations comparing population sizes inferred with MAPS and "diversity-rates" inferred with EEMS.** We simulated uniform migration rates of 0.01 and a trough of low population sizes in the center of the habitat such that population sizes equal to 1,000 at the center and 10,000 otherwise. Under these simulations, EEMS infers a barrier in effective migration and infers uniform diversity rates. However, MAPS correctly infers a uniform migration surface (mean 0.01) and provides accurate estimates of deme sizes (mean 985 at the center and 9100 at the edges) . . . . . 17

2.4 **Inferred Dispersal Surfaces and Population Density Surfaces over time for Europe.** We apply MAPS to a European subset of POPRES Nelson et al. [2008] with 2,234 individuals and plot the inferred dispersal $\sigma(\vec{x})$ and population density $D_e(\vec{x})$ surfaces for PSC length bins (a) > 1cM (b) 5-10cM and (c) >10cM. We transform estimates of $\vec{N}$ and $\mathbf{M}$ to estimates of $\sigma(\vec{x})$ and $D_e(\vec{x})$ by scaling the migration rates and population sizes by the grid step-size and area (see Equations (2.17) and (2.18)). Generally, we observe the patterns of dispersal to be relatively constant over time periods, however, we see a sharp increase in population density in the most recent time scale (>10cM). Note the wider plotting limits in inferred densities in the most recent time scale. . . . . 20

3.1 **Rate heterogeneity causes false inferences of positive selection.** a) Simulation scheme for generating sequences with rate heterogeneity. A proportion ($p$) of a total of N codons were generated using a codon model with non-synonymous/ synonymous rate A and a tree on which all branches have lengths $b_A$; the remaining $(1-p)$ sites were generated using the same model but with independent parameters B and $b_B$. All data were generated with both A and B $\leq 1$. b) Rate heterogeneity causes false positive inferences. For each generating condition, the proportion of 50 replicate alignments for which the sites test found a signature of positive selection (p < 0.05) is shown (circles). Also shown are the proportion of positive inferences using the branch-sites test (triangles) and an a priori partitioned version of the sites test that incorporates rate heterogeneity (crosses). Dashed line, acceptable false positive rate of 5%. Solid line, best-fit logistic function to sites test results. c) Effect of the prevalence of rate heterogeneity on false positive inference rates by the sites test. Sequences were simulated under the conditions shown, and the rate of positive inferences by the sites test was plotted against the number of sites with the slow evolutionary rate ($p$ in the generating model). d) Rate heterogeneity causes false inferences of selection even under purifying selection. The sites test false positive rate is shown as a function of in the generating model, under the conditions shown. . . . . 53

3.2 **Systematic bias towards false positive inferences caused by rate heterogeneity.** a) The sites test false positive inference rate is shown as a function of the number of codons in the alignment (N) when sequences are generated under the conditions shown. Dashed line, acceptable false positive rate (5%). b) The sites test false positive rate is shown as a function of the number of taxa in the tree; the total depth of the tree from root to tips was kept constant (0.2 and 0.4 in the two partitions) to mimic improved taxon sampling. . . . . 56

3.3 **Relationship of rate heterogeneity to positive selection signatures in empirical data.** a) The signal of positive selection in empirical data is weakened when rate heterogeneity is incorporated. 1596 mammalian genes previously found to have a signature of positive selection [Kosiol et al., 2008] were reanalyzed using the standard implementation of the sites test and an alternative version that incorporates synonymous rate variation (sites test-RH). The left portion of the graph shows the number of genes with a signature of positive selection ($p < 0.05$) in each test. The right portion shows the number of genes retaining this signature after adjusting for multiple testing (1596 tests) at a 5% false discovery rate is shown. b) Genes inferred to be under positive selection contain substantial rate variation. For each of the 397 mammalian genes inferred to be under positive selection by the sites test, we inferred maximum likelihood estimates of the degree of rate heterogeneity using the null model of sites test-RH. The distribution of the coefficient of rate variation (CVRV) across these genes is shown. Dotted line indicates CVRV=0.206, the degree of rate heterogeneity required under the conditions in Figure 3.1a to yield unacceptably high false positive rates in the sites test. c) Selection signatures are correlated with rate heterogeneity. For each gene in the set of 1596, the degree of rate heterogeneity was estimated as the coefficient of variation of the branch length rate multipliers in the mixture model of sites test-RH. Genes were binned by this metric, and the proportion of genes in each bin under positive selection according to the sites test ($p \leq 0.05$) was calculated. Spearmans correlation coefficient=0.97. D) Incorporating rate heterogeneity eliminates selection signals rather than reducing statistical power. For genes with a significant signature of selection in the sites test (p-value $< 0.05$), we plotted the sites-test p-value against the sites test-RH p-value. Pearson and Spearman correlation coefficients=0.21 and 0.26, respectively. . . . . 58

3.4 **Empirically-derived rate heterogeneity causes false positive inferences.** For each mammalian gene with a signature of selection and complete coverage, the parameters of the sites-test-RHs null model were inferred, and 25 replicate sequence alignments were then simulated under these conditions (without positive selection). The false positive rate for each gene is the fraction of replicate alignments that yielded a significant signature of selection ($p < 0.05$). The distribution of false positive inference rates across mammalian genes is shown. Dotted line, acceptable false positive inference rate. . . . . 62

3.5 **Causes of spurious support for M2a by rate heterogeneity.** a) Support for the positive selection model vs. null model by category of codon state patterns. Data were generated on a two-taxon tree with no selection ($\omega$=1) in two partitions of equal size with branch lengths 0.2 and 1.6. Codon sites in the alignment were categorized by the number of nucleotide differences between taxa; for sites with 1 difference, codons were further as having a nonsynonymous (N) or synonymous (S) difference. The height of each bar represents the difference in log-likelihood between the positive selection model M2a and the null model M1a (using the ML parameter values for each), summed over all instances of sites in each category in the dataset. Parentheses indicate the number of codon sites of each category in an alignment of 1000 codons. b) Actual and predicted frequency of codon state patterns under various models. Data were generated under the conditions described for panel a. The first column shows the actual proportion of codons in each category in the concatenated dataset; parentheses show the percentage of codons in each category generated in the partition with the slower rate. The next three columns show the predicted proportion given the maximum-likelihood optimized parameters of models M0, M1a, and M2a. c) Model M2a mimics rate heterogeneity by having three very different effective branch lengths for its three submodels. Scaling factors and effective branch lengths for codon mixture models are shown, using ML parameter estimates given data generated as in panel a. For each submodel, the bar length shows the effective branch length the expected number of substitutions given the scaling factor for the model ($S$, shown x1000), the ML estimate of the branch length for the model (b), and the total rate of substitution for the submodel ($t_k$). When $S_{t_k}$=1, the effective and given branch lengths are the same. Note the broken scale. . . . . 65

4.1 **Illustration of the `aRchaic` grades of membership and mismatch profiles.** (a) The features of a mismatch modeled by `aRchaic` (b) A depiction of an ancient DNA sample that has 80% of its reads assigned to cluster 1 and 20% of its reads assigned to cluster 2. Each cluster is defined by a *mismatch profile* showing the enrichment of the mismatch type, bases flanking the mismatch, the distance of the mismatch from the nearest end of the read, and the base immediately 5' to the strand-break. To produce a mismatch profile for a cluster, mismatch features are aggregated across reads assigned to the cluster, and their frequencies are represented by an *EDLogo* plot [Dey et al., 2017b]. In the *EDLogo* plot, the frequencies are scaled against a background frequency computed from Consortium et al. [2012]. . . . . 83

4.2 `aRchaic` **clearly distinguishes between modern, ancient (UDG), and ancient (non-UDG) samples.** `aRchaic` is applied with $K = 3$ to a collection of ancient individuals from four studies Skoglund et al. [2014a], Gamba et al. [2014], Lazaridis et al. [2016], Lipson et al. [2017] along with modern individuals randomly sampled from the 1000 Genomes Project and 10 individuals from the Human Genome Diversity Panel [Consortium et al., 2012, Cann et al., 2002, Meyer et al., 2012]. Modern samples have high membership in the red cluster. The *EDLogo* representation of this cluster does not show strong enrichment against a modern background. The ancient (non-UDG) samples are representative of the blue cluster. The *EDLogo* plot for the blue cluster shows a strong enrichment in C-to-T mismatches at the end of reads, a depletion of guanine in the right flanking base, and a depletion of cytosine at the 5' strand-break. The ancient (UDG) samples have partial membership both in the red cluster and in the gold cluster. The *EDLogo* plot for the gold cluster is enriched in C-to-T mismatches at the terminal ends of the reads, shows an enrichment of guanine at the right flanking base, and a depletion of thymine one base 5' upstream of the strand break. 86

4.3 **Estimated grades of membership reflect levels of contamination.** (a) Reads from one ancient individual (KO1 from Gamba et al. [2014]) were split into 10 equally sized groups. Reads were added from a distinct individual in the 1000 Genomes Project [Consortium et al., 2012] to each group (S1-S10) at varying levels of percentages (indicating levels of contamination). (b) We applied `aRchaic` with $K = 2$ on a combined dataset comprised of these 10 contaminated groups of reads (S1-S10) along with 40 other modern individuals from 1000 Genomes (c) The grades of membership in cluster 1 ("modern" cluster) were plotted as a function of the percentage of contamination before (red curve) and after (green curve) applying the correction factor discussed in the last paragraph of Section (3.2). (d) We repeated the same experiment as described in panel a, except we discarded all reads with no mismatches or greater than one mismatch. The grades of membership in cluster 1 were plotted as a function of the "mismatch contamination rate" which is defined as the proportion of mismatches that originate from a contaminated read (e) Each group (S1-S10) was further sub-sampled to 10,000 reads, and `aRchaic` was applied with $K = 2$ to the new subsampled groups and the same 40 modern individuals as in panel b. . . . . 88

4.4 **DNA damage and library preparation techniques drive grades of membership.** (a) We applied `aRchaic` with $K = 2$ to 25 modern samples from Lindo et al. [2016]. The samples prepared with the TruSeq kit show nearly full membership in the pink cluster. Samples prepared with the Nextera kit show partial membership in the pink cluster and the tan cluster. The tan cluster shows a blip at the 12th position from the end of the read (b) We applied `aRchaic` with $K = 2$ to 25 ancient samples from Lindo et al. [2016]. The two clusters show an enrichment of C-to-T mismatches at the ends of reads and an enrichment of purines at the 5' strand-break. . . . . 90

S1 **The performance of MAPS on a recent barrier scenario under different PSC length bins.** Here, we investigate the ability of MAPS to detect a recent barrier (< 10 generations) for various PSC length bins (a) Simulation scenario. Population sizes were set to 10,000 per deme and 10 diploids were sampled per deme, replicating the conditions in Figure 2.2b. (b) Results for different PSC length bins. Length bins that encompass shorter segments (2-4cM 2-6cM 2-8cM) recover the higher uniform migration surface; while length bins with longer segments (>4, >6, >8) recover the recent ancestral barrier. For the last length scale (> 8cM), the signature of low migration extends across the habitat. The variation in migration rates is missed presumably because of the small number of shared segments at this length scale. . . . . 106

S2 **The performance of MAPS on a past barrier scenario under different PSC length bins.** a) Simulation scenario. Population sizes were set to 10000 per deme and 10 diploids were sampled per deme, replicating the conditions in Figure 2.2c. (b) Results for different PSC length bins. Length bins that encompass shorter segments (2-4cM, 2-6cM, 2-8cM) recover the ancestral barrier; while length bins with longer segments (>4, >6, >8) recover the recent constant migration surface. . . . . 107

S3 **The performance of MAPS under a jointly heterogeneous migration rate and population size surface.** a) Simulation Scenario. Heterogeneous population-sizes and migration rates (as shown) were simulated, and 10 diploid individuals were sampled per deme. (b) Results for PSC segments greater than 2cM are shown. . . . . 108

S4 **Visualizing normalized sharing of PSC segments that are 1-5cM.** The color scheme is the same as used in Ralph and Coop [2013] where the colors give categories based on the regional groupings: W Western Europe, S Southern Europe, and E Eastern Europe (a) The average sharing within each sample locale is transformed to population sizes using the simple single deme estimator by Palamara et al. [2012]. This transformation can be roughly summarized as to say that $N_\alpha \propto \frac{1}{\bar{x}_{\alpha,\alpha}}$ where $N_\alpha$ is the effective population size in deme $\alpha$ and $\bar{x}_{\alpha,\alpha}$ is the average pairwise PSC sharing between individuals in deme $\alpha$. (b) Similar to Ralph and Coop [2013], for each focal population (marked with an x), we plot the normalized average pairwise sharing between that population and all others (normalized by the average sharing within the focal population), i.e. if $\alpha$ is the focal population, we show $\frac{\bar{x}_{\alpha,\beta}}{\bar{x}_{\alpha,\alpha}}$ for each other country $\beta$. . . . . 109

S5 **The correlation between census size and inverse average PSC sharing as a function of minimum PSC length considered.** We use census size compiled from the The World Bank [2016] and National Records of Scotland [2011]. The smooth black curve denotes the loess fit. Longer PSC segments correlate more strongly with census size than shorter PSC segments . . . . . 110

S6 **Census size versus MAPS estimated population sizes.** Using the MAPS output, we estimate a total size per population by summing the estimated deme-level sizes across the area of each respective country (whether's a deme's location falls within a country was determined by querying The GeoNames Geographical Database). Finally, we plot the results on a log10 scale for different length scales (a) 1-5cM, (b) 5-10cM, and (c) >10cM. The red curve denotes the linear fit on the absolute scale. . . . . 111

S7 **Plots of estimated average log10 differences in demographic parameters between adjacent time scales**. (a) We plot estimates of $E[\log10(\frac{\sigma'}{\sigma})]$ and $E[\log10(\frac{D_e{}'}{D_e})]$ across the spatial habitat where $\sigma'$ ($D_e'$) denotes the dispersal rates (population densities) in the 5-10cM length bin and $\sigma$ ($D_e$) denotes the dispersal rates (population densities) in the 1-5cM length bin. (b) The results here are similarly plotted as above, however, the adjacent length scales are given by: 5-10cM and >10cM. The log10 differences are estimated in such a way so that the mean log10 difference is shrunk to zero. For example, for estimating dispersal in 5-10cM, we assume $\log10(\sigma') = E[\log10(\sigma)] + \epsilon$ where $E[\log10(\sigma)]$ is estimated using PSC segments 1-5cM and $\epsilon \sim N(0, \omega^2)$ is estimated from PSC segments 5-10cM. Consequently, the log ratio between dispersal rates from the two lengths bins is constructed to have mean zero *apriori* (i.e. $E[\log10(\frac{\sigma'}{\sigma})] = 0$). . . . . 112

S8 **EEMS applied to the POPRES dataset.** We apply EEMS to the same set of individuals as used in Figure 2.4 (see Methods). (a) The effective migration rates (b) The effective diversity rates. Here, we ran EEMS with 200 demes (as in Figure 2.4) with default parameters and averaged over 10 independent replicate chains. Each chain ran with 50e6 MCMC iterations, 25e6 set as burn-in, and we thinned every 5000 iterations. . . . . 113

S9 **Genetic distance vs PSC sharing** (a) The averaged genetic distance (as used in EEMS) is plotted against the average number of PSC segments (> 1cM) for each pair of populations. Each point denotes a pair, the symbols represent groupings from Ralph and Coop [2013] (W Western Europe, S Southern Europe, and E Eastern Europe), and the colors represent the pair of regions. We see a negative correlation between the two summary statistics (Pearson's $\rho$ = -0.38, p-value = 7e-11), with the largest deviations occurring in comparisons between Eastern European populations. (b) EEMS results on PSC data transformed to a distance matrix. First, we encoded the PSC sharing statistics into a similarity matrix $S$ such that $S_{i,j}$ is the number of shared PSC segments between samples i and j and $S_{i,i}$ is the maximum number of shared segments in the dataset (which we denote as $c$) to ensure $S$ is a similarity matrix. Next, we transformed $S$ to a genetic distance matrix $D$ such that $D = c11^T - S + E$ where $E \approx 0$ is a random genetic distance matrix of normal vectors with mean 0 and standard deviation of 0.01 added to ensure $D$ is full rank. Finally, we applied EEMS to the distance matrix $D$. Though this procedure is heuristic, we see shared features between this surface and the MAPS dispersal surface shown in Figure 2.4. . . . . 114

S10 **The Sites Test is biased towards false positives under different implementations and variations.** (a) Rate heterogeneity causes false positive inferences by the sites test as implemented in Hyphy. Sequences were simulated under the evolutionary conditions shown and then analyzed using the Nielsen-Yang method as implemented in the Hyphy software package. The false positive rate out of 25 repetitions per conditions is plotted as a function of $b_B/b_A$. (b) Rate heterogeneity causes high false positives in the sites test when more complex models are used. Sequences were simulated under the conditions shown and analyzed using the sites test and models M7 vs. M8 as implemented in PAML software. Each false positive rate is calculated out of 40 repetitions. . . . . 115

S11 **Rate heterogeneity with longer branches increases the sites tests false positive inference rate.** The rate of false inferences of positive selection is shown with constant rate heterogeneity $b_B/b_A$ and increasing branch lengths in both partitions $b_B$ under the conditions shown. The false positive rate is calculated as the number of positive inferences out of 50 replicate alignments analyzed for each conditions . . . . . 115

S12 **Site-specific likelihoods of submodels comprised by M1a and M2a. For each category of state pattern, the likelihoods of each mixture models submodels are shown, summed over all instances in the alignment.** Each column represents submodel 0, 1, or 2 in models M1a or M2a. Values of $\omega$ and the mixing proportion $p$ for each submodel are shown. The generating conditions are the rate-heterogeneous conditions with $\omega = 1$ as specified in Figure 3.5a. A dash indicates likelihood $< 0.001$. . . . . 116

S13 **The non-parametric counting method of Nei & Gojobori, 1986 is robust to model violation.** 50 replicate alignments at each condition were generated by evolutionary simulation with $p$=0.5, $\omega = 1$, $N$=996, and number of taxa=4. We implemented and applied the dN/dS inference method described by Nei & Gojobori, 1986. Confidence intervals were estimated by bootstrapping. . . . . 116

S14 **Rate heterogeneity causes false inferences of positive selection in the BUSTED test of positive selection.** 10 replicate alignments at each condition were generated by evolutionary simulation with $p$=0.5, $\omega$=1, $N$=996, $b_A$=0.1, and number of taxa=4. We applied MEME to test for gene-wide positive selection. The false positive increases as the degree of rate hetereogeneity ($b_B/b_A$) increases. 117

S15 **Rate heterogeneity does not cause an elevated false positive rate in the MEME test of positive selection. 10 replicate alignments at each condition were generated by evolutionary simulation with $p$=0.5, $\omega$=1, $N$=996, $b_A$=0.1, and number of taxa=4.** We applied MEME to test for positive selection at each site. Here, we plot the average (over 10 replicates) number of sites inferred to be under positive selection at a significant level of 0.05. Under the null hypothesis, we expect $996 * 0.05 \approx 50$ sites to have a p-value $< 0.05$ (red line). The number of sites at that significance level increases as the degree of rate heterogeneity ($b_B/b_A$) increases but is far below the 50. . . . . 117

S16 **Four examples of how the reference strand for a mismatch is designated.** The dark yellow line denotes the mapped read, and the blue and teal line represent the reference genome at two different strand orientations. The reference strand is always designated as the strand that contains the C mismatch or T mismatch (latter not shown). . . . . 118
S17 `aRchaic` **grades of membership for the example in Fig 4.1 corresponding to 3 different values of** $K$ **(**$K = 4, 5, 6$**).** Higher values of $K$ distinguish among the ancient studies, reflecting lab and study specific biases. . . . . 119
S18 `aRchaic` **plot for** $K = 2$ **on the combined data of 25 moderns and 25 ancients from Lindo et al. [2016].** `aRchaic` clearly distinguishes the moderns from the ancients. The ancients are primarily presented by the blue cluster. This cluster shows an enrichment of C-to-T mismatches and depletion of T-to-C mismatches with respect to modern background, as well as enrichment of G and depletion of T at the 5' strand break. The red cluster shows a blip at 12th position from the end of the read, the explanation for which is provided in Figure S20. . . . . 120
S19 **We apply** `aRchaic` **with** $K = 6$ **on the data from Fig 4.4**. In addition to separating out the ancients from the moderns, `aRchaic` now distinguishes between moderns individuals based on library kit (Nextera vs Tru-seq.) . . . . . 121
S20 **The frequency of all mismatch types plotted against the position of the read (from the 5' end) for each of the 25 moderns samples in Lindo et al. [2016].** Each sample was prepared by one of two library kits: Nextera and TruSeq. Most of the samples prepared with the Nextera kit show a spike in frequency at the 12th position from the 5' end of the read . . . . . 122

# ACKNOWLEDGMENTS

First, I would like to thank my wonderful PhD advisors Matthew Stephens and John Novembre. In addition to being supportive and understanding, they have taught me a great deal about statistical genetics and how to break-down scientific problems. I am very grateful for their mentorship, and working with them over the years was both scientifically engaging and a lot of fun. I especially would like to thank Matthew for helping me see the simplicity in seemingly complicated things, and John for helping me realize the power of good visualizations. I would like to thank my committee members Dick Hudson and Joe Thornton. I am grateful to Dick for his constant support during my PhD and having taught me coalescent theory and many fundamental ideas in population genetics. I would like to thank Joe for teaching me phylogenetics, a productive rotation project, and his support and sound advice during committee meetings.

I would like to thank my collaborators, Kushal Dey and Joe Marcus. Over the years, I've come to realize that I learn the best when working with others. The debates and brainstorming sessions in both projects were fundamental in my development as a scientist.

I would like to thank Peter Carbonetto and John Blishak for their scientific advice and computational support. I would also like to thank Carolyn Johnson and Mike Coates for their continued administrative support during my PhD.

And, I would like to thank my friends and family. First, I owe everything to my mother who raised me single-handedly with my two brothers after we came to the United States as political refugees from Iraq. I would like to thank my wife and kids for their love, understanding and support. I would like to thank my friends: Joe Marcus, Kushal Dey, Joel Smith, Evan Koch, Daniel Rice, John Lindo, Aarti Venkat, Dallas Krentzel, John Park, Mark Bitter, and Desi Petkova; and I thank my good fishing friends (who have a special place in my life): Tetsuya Nakamura, Matt Bonakdarpour, Shigeki Nakogome, and John Loudis.

# ABSTRACT

Statistical methods have proven to be fundamental in the analyses of genetic data and have been used as a means to arrive at many novel biological discoveries; for example, these methods have helped researchers better understand human history, long-term evolutionary processes between species, and increased the power of studies with linkage-based imputation [Stephens et al., 2001, Rosenberg et al., 2002, Felsenstein, 2004, Li and Durbin, 2011]. Many methods in statistical genetics model sequence data along the genome and assumes that the data is generated from a tree; while some methods do not. Here, I talk about three PhD projects that encompass many classes of methods in statistical genetics. Two projects are connected in that they model nucleotide variation along the genome and assumes the data is generated from a tree, while the third project models data at the levels of reads and utilizes a popular class of methods based on an "Admixture" model, which assumes no underlying tree.

Trees connect all biological organisms on earth. At the most basic level, individuals within a population are related through a genealogical tree. These trees are influenced by population level parameters and can be characterized using coalescent theory [Hein et al., 2004]. In my first project, the underlying tree structure is utilized to infer dispersal and population density in Europe across space and different time periods. At a higher level, individuals between species are related through a phylogenetic tree, which can be studied using methods from phylogenetics [Felsenstein, 2004]. In my second PhD project, phylogenetic-based methods for inferring positive selection are studied in both simulated and empirical datasets.

However, many methods in statistical genetics do not assume an underlying tree, such as methods based on "Admixture" models. In my third project, an Admixture model is used to model mismatches from sequencing reads in ancient DNA. Unlike tree models, Admixture models assume that each sample is unrelated and has an estimated grade of membership in each of $K$ unrelated clusters that are estimated from the data [Pritchard et al., 2000]. Taken

together, my three PHD projects span a wide class of models in statistical genetics.

# CHAPTER 1

# INTRODUCTION

Life can only be understood going backwards, but it must be lived going forwards.

Soren Kierkegaard (1813-1855)

Individuals in a population undergo evolution forward in time. However, evolution can sometimes be more easily modeled backwards in time because individuals are sampled in the present time and shaped by past processes. A popular approach for modeling genetic data is to posit latent genealogies, which themselves are modeled as random variables shaped by population-level parameters [Hein et al., 2004]. These genealogy-based approaches are commonly referred to as "coalescent models", and were first independently formulated by Sir John Kingman, Richard Hudson, and Fumio Tajima [Kingman, 1982, Hudson, 1983, Tajima, 1983].

The coalescent model is based on the simple idea that individuals are related through an (unobserved) genealogy, and given the genealogy, data can be easily generated from it [Hudson et al., 1990]. The data is assumed to be in the form of sequence data along the genome, and furthermore, is often further reduced into bi-alleleic SNP data i.e. the data can be summarized with a genotype matrix $G$ such that $g_{i,l} \in \{0, 1\}$ denotes the allele at locus $l$ for individual $i$ [Hudson, 1983]. For concreteness, I provide a brief description of the coalescent model for a sample size of two in a one population model. Here, the underlying length of the tree $T_{i,j}$ (in units of generations) between a pair of samples $i$ and $j$ is given by

$$T_{i,j} \sim \exp(\frac{1}{2N_e}), \tag{1.1}$$

where $N_e$ is the effective population size [Hudson et al., 1990]. Intuitively, equation (1.1) captures the idea that individuals are more related to each other (i.e. shorter trees) in smaller populations. After generating the genealogy, data can be generated by sprinkling mutations

uniformly on the tree according to a Poisson process. The number of differences between individuals (or number of "Segregating sites") $S_{i,j}$ becomes

$$S_{i,j}|T_{i,j} \sim \text{Poisson}(2T_{i,j}\mu), \tag{1.2}$$

where $\mu$ is the mutation rate, and under the very good approximation that there is at most one mutation at any one site (this approximation is often to referred to as the "infinite sites" assumption because every mutation occurs on a new site in the genome) [Watterson, 1975]. This basic one-deme coalescent model has been used to reveal many interesting biological insights. For example, a seminal paper by Li and Durbin [2011] infers population-sizes across time under a single deme model, and has been widely used to learn about demographic history of many species (e.g. plants, birds, reptiles, mamamals) [Skoglund et al., 2015, Zhao et al., 2013, Green et al., 2014, Nadachowska-Brzyska et al., 2016].

The coalescent was extended to capture the spatial structure of populations, which is done by modeling many populations (or "demes") arranged on a lattice [Notohara, 1990], here we refer to this extension as the "multi-deme coalescent". In the multi-deme coalescent, the lattice is constructed so to only allow migration between adjacent demes, also known as a "stepping stone" model [Kimura and Weiss, 1964]. The stepping stone assumption is important because it captures the decrease of genetic correlation with distance observed in genetic data (this pattern is also known as "isolation-by-distance") [Kimura and Weiss, 1964, Weiss and Kimura, 1965, Sawyer, 1976]. The multi-deme coalescent can be mathematically described with a Continuous Time Markov chain (CTMC) [Bahlo and Griffiths, 2001]. Let $i$ and $j$ denote the two sampled lineages, $d$ denote the number of demes (or populations), and $(\alpha, \beta) \in \{1, \cdots, d\} \times \{1, ..., d\}$ denote the locations of the two samples, and let $c$ denote the coalescent state. The state of this CTMC is given by $\{1, \cdots, d\} \times \{1, ..., d\} \bigcup c$. The

infinitesimal rate matrix is given by

$$
\begin{aligned}
R_{(\alpha,\beta),(\gamma,\beta)} &= m_{\alpha,\gamma} \;\; \beta = 1, ..., d, \; \gamma \neq \alpha \\
R_{(\alpha,\beta),(\alpha,\gamma)} &= m_{\beta,\gamma} \;\; \alpha = 1, ..., d, \; \gamma \neq \beta \\
R_{(\alpha,\alpha),(c)} &= \frac{1}{2N_\alpha} \\
R_{(\alpha,\beta),(\alpha,\beta)} &= -(m_{\alpha_+} + m_{\beta_+}) - \delta_{\alpha\beta} q_\alpha \\
R_{(c),(c)} &= 0 \\
R_{(\alpha,\beta),(\gamma,\kappa)} &= 0 \;\; \gamma, \kappa = 1, ..., d, \; \gamma \neq \alpha, \; \kappa \neq \beta,
\end{aligned}
\tag{1.3}
$$

where $M = \langle m_{\alpha,\beta} \rangle$ denotes the migration rate matrix, and $m_{\alpha,\beta}$ is the migration rate between demes $\alpha, \beta$ and $N_\alpha$ the population size at deme $\alpha$. Let $T_{i,j}$ denote the (random) coalescent time between the pair of sampled lineages. Under this framework, lineages make transitions from state to state, independent of the past according to a discrete-time Markov chain defined by the ratio of rates. For example, a lone lineage found in deme $\alpha$ moves to a neighboring deme $\beta$ with probability $\frac{m_{\alpha,\beta}}{\sum_\gamma m_{\alpha,\gamma}}$. But once a lineage enters a state, it remains in the state, independent of the past, for an exponentially distributed amount of time (determined by the rates) before changing state again. The coalescent time between lineage $i$ and $j$ ($T_{i,j}$) is simply twice the total time before lineages $i$ and $j$ reach the coalescent state $c$. Like before, conditional on knowing $T_{i,j}$, mutations can be added according to a Poisson process.

Many studies have used the multi-deme coalescent to infer migration rates and population sizes [Beerli and Felsenstein, 1999, 2001, Wilson and Rannala, 2003, Hey and Nielsen, 2004, Broquet et al., 2009, Barton et al., 2002]. Historically, these methods accommodated only a handful of demes because computation under the coalescent model can be unfeasible for many demes. Under the coalescent, the state space grows quadratic with the number of demes. Furthermore, the number of parameters grow with the number of models and therefore some

regularization is required, which is a non-trivial problem.

Recently, a significant improvement in this area was made with the development of the "EEMS" method (Estimating Effective Migration Surfaces). EEMS infers effective migration rates under the coalescent model, and can accommodate a large number of demes [Petkova et al., 2015]. Furthermore, the EEMS method can both flexibly share information across adjacent demes and accommodate abrupt changes in migration with a prior parameterized by a Voronoi tessellation. The underlying machinery of EEMS is based on modeling the genetic distance between all pairs of individuals, and assuming that the expected distances are proportional to their expected coalescent times. However, the EEMS method cannot infer very recent population structure because it models the data through the expected coalescent time, which can occur a relatively long time ago in the past. More formally, the expected coalescent time is approximately on the order of the effective population size $\mathcal{O}(N_e)$, which for example, is tens of thousands in human populations. Furthermore, EEMS infers the effective migration rate which is a compound parameter of the migration rates and the population sizes, and thereby make it hard for users to disentangle the effects of the two on the data.

In order to address these problems, in my first project, I develop a method (MAPS) to infer population-sizes and migration rates separately under the coalescent model. Importantly, I use long pairwise shared coalescent (lPSC) segments as a summary statistic, which are a better summary statistic for inferring recent population structure. For instance, several methods focus on analyzing these long segments to infer recent demographic history because these segments correspond to recent coalescent times, and as a result, are far more sensitive to recent population structure than the average genetic distance as used in EEMS [Schiffels and Durbin, 2014, Sheehan et al., 2013, Terhorst et al., 2016, Palamara et al., 2012, Han et al., 2017]. Using extensive coalescent simulations, I find the method is capable of revealing time-varying migration rates and population sizes, including changes that are not detectable

with data summaries that ignore haplotypic structure. I apply the method to a European subset of the the POPRES dataset (consisting of 2224 European individuals), and provide insights on recent population structure in Europe [Nelson et al., 2008]

Far back in time, the coalescent genealogies eventually segregate into species trees (or "phylogenies") when comparing individuals between different species, and a vast array of methods exist to estimate these trees [Maddison, 1997]. At the species level, coalescent models become less useful and phylogentic methods are preferred. The earliest of phylogenetic methods were developed to model nucleotide data on a phylogeny [Jukes et al., 1969, Kimura, 1980, Felsenstein, 1981]. Many of these methods work by modeling evolution with a CTMC. For example, F81 is a simple of model nucleotide substitution where the transition matrix between the four states (A, C, G, T) are given by

$$Q_{i,j} = \pi_j, \text{ for } j \neq i \tag{1.4}$$

where $\pi_A, \pi_G, \pi_C, \pi_T$ represent the equilibrium frequencies of the four different nucleotides.

This simple CTMC framework was further extended to model codon evolution a tree [Goldman and Yang, 1994]. Codons are composed of three nucloetides and specify the DNA code from the nucleotide level to protein (i.e. a three-nucleotide codon specifies a single amino acid). Similar to nucleotides, the evolution of codons can be described by a CTMC such that the states constitute the different types of codons [Goldman and Yang, 1994]. Codon models were further extended to model positive selection. These family of methods detect positive selection by assessing the ratio of the rate of non-synonymous substitution (dN) to that of synonymous substitution (dS). Synonymous changes do not alter the protein sequence, so they are assumed to evolve under the influence of mutation and drift alone; thus, dN/dS $= 1$ is expected in the absence of selection, dN/dS $< 1$ is taken as evidence of purifying selection, and positive selection is inferred when dN/dS $> 1$ at some or all sites.

In these models of positive selection, the null hypothesis of neutral evolution and the

alternative hypothesis of positive selection are each represented by a CTMC. The likelihood of each model defined as the probability that all the extant sequence data would have evolved given the phylogeny, the model, and its parameters is calculated, and these are compared to determine whether a statistically significant improvement results when positive selection is incorporated into the model. In the simplest version of these tests, the models contain a single selection-related parameter $\omega$, which represents the ratio of the instantaneous rates of non-synonymous and synonymous substitution; in the null model, $\omega$ is constrained to values $\leq 1$, and the positive selection model allows any value, including $\omega > 1$. The rate matrix is given by

$$Q_{i,j} = \begin{cases} 0, & \text{if two codons differ at more than one position,} \\ \pi_j & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases} \tag{1.5}$$

where $\pi_j$ is the equilbirum frequency of codon $j$, $\kappa$ is the transition/transversion ratio, and $\omega$ = dN/dS. Many published inferences of positive selection have been based with these codon models.

In my second PhD project, I show that the phylogenetic-based tests of positive selection yield frequent, strongly supported false inferences of positive selection under rate heterogeneity. Using simulations, I show that even a moderate degree of this kind of heterogeneity causes the Sites-Test and its popular variant the Branch-Sites test to yield frequent, strongly supported false inferences of positive selection. In a whole-genome empirical dataset from mammals, the Sites-Test finds positive selection in hundreds of genes, but rate heterogeneity is widespread, and a method that incorporates it drastically reduces the number of inferred positively selected genes. These findings indicate that the bias is strong and empirically

relevant, and many reported cases of positive selection using the Sites-Test and related tests could be artifacts of unincorporated rate heterogeneity.

Thus far, phylogenies (trees between species) and genealogies (trees within species) were discussed, however many methods in statistical genetics are not tree-based. For example, a widely used class of methods in statistical genetics are based on "Admixture" models (also known "grade of membership" or "topic" models). Instead of assuming that individuals are related under a tree, admixture models postulate $K$ unrelated populations, and that individuals are an mixture of these $K$ populations. In addition to inferring structure in admixed populations [Pritchard et al., 2000], admixture models are widely used to infer structure in document collections [Blei et al., 2003], RNA-seq data Dey et al. [2017a] and somatic mutation data [Shiraishi et al., 2015] for example.

Admixture models have a parameter $K$ (the number of "source populations") that the user must postulate. The statistical problem now is equivalent to estimating allele frequencies of the $K$ populations for each locus $l$, and the admixture proportions of each individual in each of the $K$ populations. The classical admixture model can be summarized as

$$E[g_{i,l}] = \sum_{k=1}^{K} \omega_{i,k} 2 f_{k,l} \tag{1.6}$$

where $g_{i,l}$ is the genotype of individual $i$ at locus $l$, $\omega_{i,k}$ is the admixture proportion of individual $i$ in population $k$, and $f_{k,l}$ is the allele frequency in population $k$ at locus $l$.

Using an Admixture model, in my third project, I develop a method to estimate DNA damage. These methods start with a Binary Alignment Map (BAM) file, obtained by aligning each read to a reference genome. The BAM file includes information on the mismatches that occur in each read (vs the reference). We characterize each mismatch by several relevant features, including its type (e.g. C-to-G, etc), flanking bases, and distance from the end of the read. We then use these features to cluster the mismatches into groups, which we call *mismatch profiles*. Intuitively, a mismatch profile associated with post-mortem damage

is expected to show high levels of C-to-T mismatches at the ends of reads. On the other hand, a mismatch profile that is typical of modern DNA polymorphism will show a different pattern, such as a transition to transversion ratio of 2:1 [Goldman and Yang, 1994]. These mismatch profiles are analogous to estimating the allele frequencies of the $K$ populations in the classical admixture model. Finally I estimate the relative frequency of each mismatch profile in each sample, which I refer to as the "Grade of membership" [Erosheva, 2006] of that sample in that mismatch profile. These grades of membership should reflect which processes generated mismatches in each sample. For example ancient samples should have a high grade of membership in mismatch profiles characteristic of post-mortem DNA damage.

# CHAPTER 2

# ESTIMATING RECENT MIGRATION AND POPULATION SIZE SURFACES

*(with Desislava Petkova, Matthew Stephens, and John Novembre)*

## 2.1 Introduction

Populations exist on a physical landscape and often have limited dispersal. As a result, most genetic data exhibit a pattern of isolation by distance [Wright, 1943], which is simply to say that populations closer to each other geographically are more similar genetically. Furthermore, the degree of isolation by distance can vary across space and time [Manel et al., 2003]. For instance, in a mountainous area of a terrestrial species' range, a pair of individuals may be more divergent from each other than a pair of individuals separated by the same distance in a flat and open area of the habitat. Additionally, the degree of isolation by distance can change over time – for example, if dispersal patterns are changing over time. Such spatial and temporal heterogeneity is an important aspect of population biology, and understanding it is crucial to solving problems in ecology [Turner et al., 2001], conservation genetics [Segelbacher et al., 2010], evolution [Rousset, 2004], and human genetics [Rosenberg et al., 2005].

Several methods have been developed to reveal spatial heterogeneity in patterns of isolation by distance [Womble, 1951, Barbujani et al., 1989, Guillot et al., 2005, 2009, Caye et al., 2016, Petkova et al., 2016, Bradburd et al., 2016, 2017]. Some methods are based on explicitly modeling the spatial structure in the data [Guillot et al., 2005, 2009, Petkova et al., 2016, Bradburd et al., 2016, 2017]; others take non-parametric approaches [e.g. Womble, 1951, Barbujani et al., 1989]; while other methods ignore the spatial configuration of the

samples and rely on researchers to make a *post hoc* geographic interpretation of the results [e.g. Pritchard et al., 2000, Patterson et al., 2006]. However, none of these methods can be flexibly applied to address temporal heterogeneity in isolation by distance patterns, and new methods are needed.

One source of information for inferring changes in demography across time is the density of mutations observed in pairwise sequence comparisons [Li and Durbin, 2011, Schraiber and Akey, 2015]. For example, when individuals are similar along a long segment of their chromosomes, it suggests that these segments share a recent common ancestor [Palamara et al., 2012]. These segments are often called "identity-by-descent" tracts, although here we prefer the term "long pairwise shared coalescence" (lPSC) segments (as identity by descent traditionally required a definition of a founder generation, which is not clear in most data applications). A key feature of these segments is that filtering them by length provides a means to interrogate different periods of population history. The longest segments reflect the most recent population history, whereas shorter segments reflect longer periods of time. Recent analyses using lPSC segments suggest that they can reveal fine-scale spatial and temporal patterns of population structure that are not evident with genotype-based methods such as principal components analysis [Ralph and Coop, 2013, Lawson et al., 2012, Leslie et al., 2015].

Here we develop a new method to infer spatial and temporal heterogeneity in population sizes and migration rates. The method takes as input geographic coordinates for a set of individuals sampled across a spatial landscape, and a matrix of their genetic similarities as measured by sharing of lPSC segments. It then infers two maps, one representing dispersal rates across the landscape, and another representing population density. Crucially, building these maps using different lengths of lPSC segments can help reveal changes in dispersal rates and population sizes over time.

Our method is based on a stepping-stone model where randomly-mating subpopulations

are connected to neighboring subpopulations in a grid. Such models are parameterized by a vector of population sizes ($\vec{N}$) and a sparse migration rate matrix ($\mathbf{M}$). Stepping-stone models with a large number of demes can approximate spatially continuous population models [Barton et al., 2002, Baharian et al., 2016], and this can be exploited to produce maps of approximate dispersal rates and population density across continuous space.

Our method builds upon a method developed for estimating effective migration surfaces (EEMS) [Petkova et al., 2016]. While EEMS infers local rates of effective migration relative to a global average, here we can explicitly infer absolute parameter values by leveraging lPSC segments and modeling the recombination process [$\vec{N}$ and $\mathbf{M}$ values in the stepping-stone model, and effective spatial density function $D_e(\vec{x})$ and dispersal rate function $\sigma(\vec{x})$ in the continuous limit]. We call this method MAPS, for inferring Migration And Population-size Surfaces.

We test MAPS on coalescent simulations and apply it to a European subset of 2,224 individuals from the POPRES data [Nelson et al., 2008]. In simulations, we show that MAPS can infer both time-resolved migration barriers and population sizes across the habitat. In empirical data, we infer dispersal rates $\sigma(\vec{x})$ and population densities $D_e(\vec{x})$ across different time periods in Europe.

## 2.2 Results

### *2.2.1 Outline of the MAPS method*

MAPS estimates demography using the number of Pairwise Shared Coalescence (PSC) segments of different lengths shared between individuals. We define a PSC segment between (haploid) individuals to be a genomic segment with a single coalescent time across its length (Figure 2.1A). Long PSC (lPSC) segments tend to have a recent coalescent time, and so manifest themselves in genotype data as unusually long regions of high pairwise similarity, which

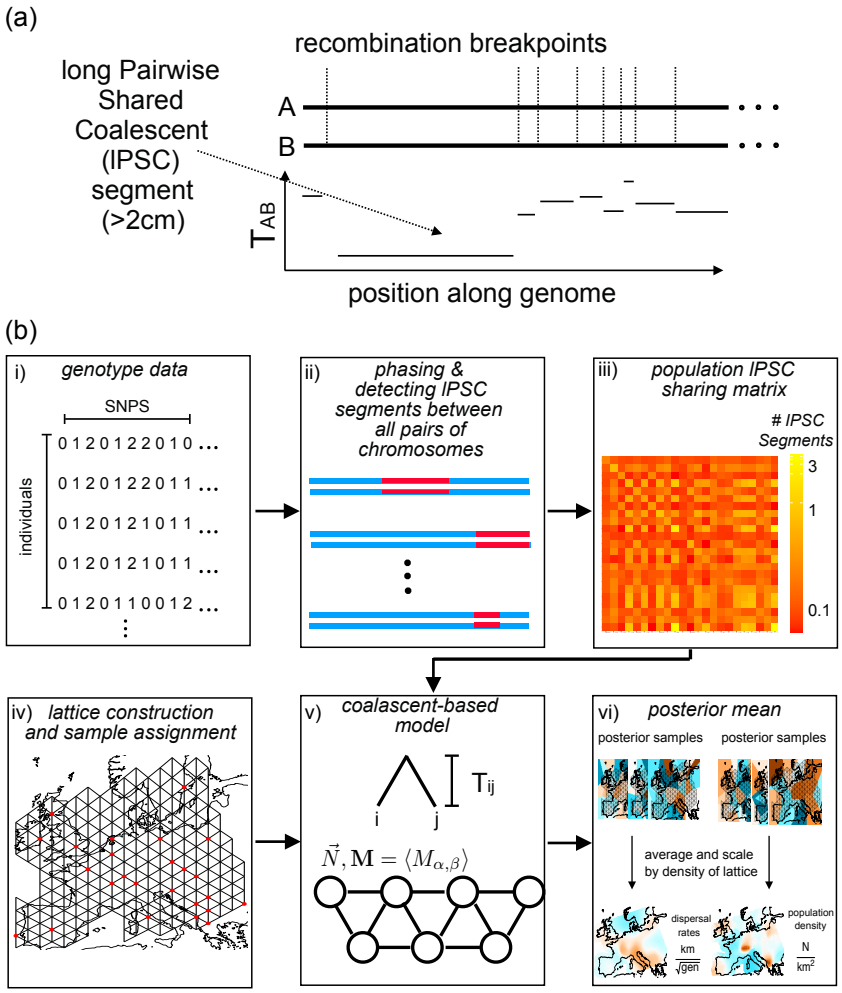



Figure 2.1: **Schematic overview of MAPS.** (a) Coalescent times between a pair of hapolotypes (A and B) will vary across the genome in discrete segments bordered by recombination breakpoints. On average, longer segments represent shorter pairwise coalescent times ($T_{AB}$) (b) Flow diagram of MAPS. i) We start with a matrix of called genotypes; ii) lPSC segments between all pairs of chromosomes across the genome are identified from the data using external methods (such as BEAGLE, Browning and Browning [2011]); iii) lPSC segments between pairs of individuals are aggregated at the levels of pairs of populations; iv) A grid is constructed and individuals are assigned to the most nearby node; v) The probability of the PSC sharing matrix can be computed under a stepping-stone model where each node represents a population and each edge represents symmetric migration; vi) We use an MCMC scheme to sample from the posterior distribution of migration rates and population sizes. The final MAPS output is the mean over these posterior samples, and the averaged rates can be transformed to units of dispersal rate and population density. The diagram does not show a bootstrapping step used to estimate likelihood weights to account for correlations between lPSC segments, see Equation (2.6) in Methods.

can be detected by various software packages [Gusev et al., 2009, Browning and Browning, 2011, 2013, Chiang et al., 2016]. Because lPSC segments typically reflect recent coalescent events, counts of lPSC segments are especially informative for recent population structure [Ringbauer et al., 2017, Palamara et al., 2012, Baharian et al., 2016]. And partitioning lPSC segments into different lengths bins (e.g. 2-8cM, $\geq$8cM) can help focus inference on different (recent) temporal scales.

The MAPS model involves two components: i) a likelihood function, which relates the observed data (genetic similarities, as measured by sharing of lPSC segments) to the underlying demographic parameters (migration rates and population sizes); and ii) a prior distribution on the demographic parameters, which captures the idea that nearby locations will often have similar demographic parameters. The likelihood function comes from a coalescent-based "stepping-stone" model in which discrete populations (demes) arranged on a spatial grid exchange migrants with their neighbors (Figure 2.1b). The parameters of this model are the migration rates between neighboring demes ($M_{\alpha,\beta}$) and the population sizes within each deme ($N_\alpha$). The prior distribution is similar to that from Petkova et al. [2016], and is based on partitioning the habitat into cells using Voronoi tesselations (one for migration and one for population size), and assuming that migration rates (or population sizes) are constant in each cell. We use an MCMC scheme to sample from the posterior distribution on the model parameters (migration rates, population sizes, and Voronoi cell configurations). We can summarize these results by surfaces showing the posterior means of demographic parameters across the habitat.

The inferred migration rates and population sizes will depend on the density of the grid used. However, using ideas from Barton et al. [2002] and Baharian et al. [2016] we convert them to corresponding parameters in continuous space, whose interpretation is independent of the grid for suitably dense grids. Specifically, we convert the migration rates to a spatial diffusion parameter $\sigma(\vec{x})$, often referred to as the "root mean square dispersal distance",

which can be interpreted roughly as the expected distance an individual disperses in one generation; and we convert the population sizes ($\vec{N}$) to an "effective population density" $D_e(\vec{x})$ which can be interpreted as the number of individuals per square kilometer. Similar to the original grid-based demographic parameters, we can summarize MAPS results by surfaces showing the posterior means of $\sigma(\vec{x})$ and $D_e(\vec{x})$ across the habitat.

### *2.2.2 Differences from EEMS*

Our MAPS approach is closely related to the EEMS method from Petkova et al. [2016], but there are some important differences. First, the MAPS likelihood is based on lPSC sharing, rather than a simple average genetic distance across markers. This was primarily motivated by the fact that, by considering lPSC segments in different length bins, MAPS can interrogate demographic parameters across different recent time periods. However, this change also allows MAPS, in principle, to estimate absolute values for the parameters $\mathbf{M}$ and $\vec{N}$, whereas EEMS can estimate only "effective" parameters which represent the combined effects of $\mathbf{M}$ and $\vec{N}$. This ability of MAPS to estimate absolute values stems from its use of a known recombination map, which acts as an independent clock to calibrate the decay of PSC segments. Finally, MAPS uses a coalescent model, whereas Petkova et al. [2016] uses a resistance distance approximation [McRae and Nürnberger, 2006].

### *2.2.3 Evaluation of performance under a stepping-stone coalescent model*

We assess the performance of MAPS with several simulations, and compare and contrast the results with EEMS. We used the program MACS [Chen et al., 2009] to simulate data under a coalescent stepping stone model and refinedIBD [Browning and Browning, 2011, 2013] to identify lPSC segments. All simulations involved twenty demes, each containing 10,000 diploid individuals, and each exchanging migrants with their neighbors. We analyzed each simulated data set using PSC segments of length 2-6cM and $\geq$6cM, which correspond to time-

scales of approximately 50 generations and 12.5 generations respectively (see Supplementary Notes). Results for other length bins are qualitatively similar (Supplementary Figure S1 & S2).

## Migration Rate Inference

First, we simulated under a uniform (constant) migration surface with migration rate 0.01 (Figure 2.2a), assumed to have stayed constant over time. In this case both EEMS and MAPS correctly infer uniform migration (Figure 2.2a), and MAPS provides accurate estimates of the migration rate (posterior mean 0.010 when using segments 2-6cM and 0.0086 using segments $\geq$6cM). As noted earlier, EEMS does not estimate the absolute migration rate; it estimates only the *relative* (effective) migration rates.

Next, we considered a scenario where the migration surface changed across time. Specifically the migration surface matches the constant migration scenario (above) until 10 generations ago, when a complete barrier to gene flow instantaneously arose (a "vicariance event", Figure 2.2b). In this setting EEMS again infers a uniform migration surface. This is because EEMS is based on pairwise genetic distances, which are negligibly influenced by the recent barrier. In contrast, by applying MAPS with different PSC segment lengths, we can see both the historically uniform migration surface (for segments 2-6cM) and the recent barrier (segments $\geq$6cM).

Next we consider a complementary time-varying scenario: an ancestral barrier disappeared 20 generations ago to allow uniform migration (Figure 2.2c). Here the EEMS results again reflect the longer-term processes, and a barrier is evident. And again, by applying MAPS with different PSC segment lengths, we can see different migration surfaces corresponding to different time scales, which are here reversed compared with the previous scenario: the historical barrier (for segments 2-6cM) and the recent uniform migration (segments $\geq$6cM).

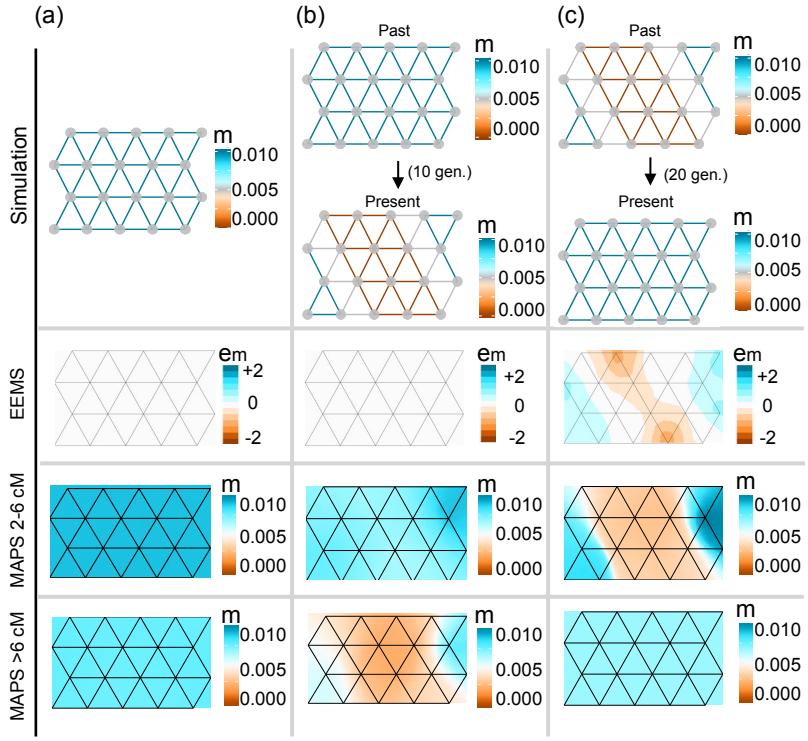



Figure 2.2: **Simulations comparing migration rates inferred with MAPS against effective migration rates inferred with EEMS.** (a) We simulated data under uniform migration rates equal to 0.01 and applied EEMS and MAPS using PSC segments in the range 2-6cM and ≥6cM. Like EEMS, MAPS correctly infers a uniform migration surface. Additionally, MAPS provides accurate estimates of the migration rates for both PSC segments 2-6cM (mean 0.01) and PSC segments ≥6cM (mean 0.0086). (b) We simulated a recent sudden migration barrier formation 10 generations ago. Here, EEMS is unable to infer a barrier, while MAPS correctly infers the historical uniform surface (2-6cM) and a barrier in the more recent time scale (≥6cM). (c) We simulated a long-standing migration barrier that recently dissipated 20 generations ago. EEMS infers a barrier, while MAPS correctly infers both the historical migration barrier (2-6cM) and the uniform migration surface in the more recent time scale (≥6cM). In all cases shown here, we simulated a 20 deme stepping stone model such that the population sizes all equal to 10,000, and 10 diploid individuals were sampled at each deme.

## Population Size Inference

As noted above, and discussed in [Petkova et al., 2016], EEMS estimates an "effective" migration surface that reflects the combined effects of population sizes $\vec{N}$ and migration rates $\mathbf{M}$; consequently it cannot distinguish between variation in $\mathbf{M}$ and variation in $\vec{N}$. In contrast, MAPS has the potential to distinguish these two types of variation.

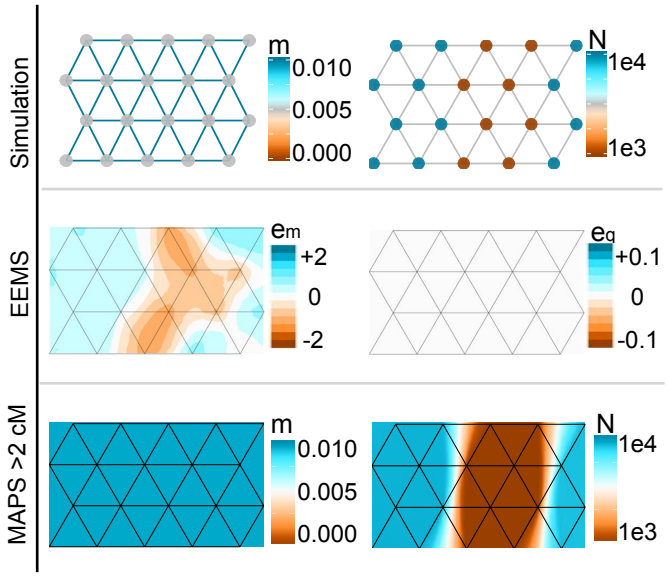



Figure 2.3: **Simulations comparing population sizes inferred with MAPS and "diversity-rates" inferred with EEMS.** We simulated uniform migration rates of 0.01 and a trough of low population sizes in the center of the habitat such that population sizes equal to 1,000 at the center and 10,000 otherwise. Under these simulations, EEMS infers a barrier in effective migration and infers uniform diversity rates. However, MAPS correctly infers a uniform migration surface (mean 0.01) and provides accurate estimates of deme sizes (mean 985 at the center and 9100 at the edges)

To illustrate this difference we simulate data with a constant migration surface, and a population size surface that has a 10-fold "dip" in the middle of the habitat (deme size 1,000 vs 10,000; Figure 2.3). Petkova et al. [2016] performed a similar simulation, and showed that EEMS estimated an effective migration surface with an "effective barrier" in the middle, caused by the dip in population size. As expected, we obtain a similar result for EEMS here. Further, the EEMS inferred diversity surface is also approximately constant, because the

diversity surface reflects changes in within-deme heterozygosity, and these vary little in this simulation. In contrast, MAPS is able to separate the influence of migration and population sizes: the estimated migration surface is approximately constant (with mean migration rate equal to the true value 0.01) and the estimated population size surface shows a dip in the middle, with accurate estimates of deme sizes (mean 985 at the center and 9100 at the edges). Additional simulations with non-uniform migration rates reinforce these results; see Supplementary Figure S3.

### *2.2.4 Applying MAPS to the POPRES data*

To illustrate MAPS on real data, we analyze a genome-wide SNP dataset on individuals of European ancestry [the "POPRES" study Nelson et al., 2008]. Previous analyses of these data have shown the strong influence of geography on patterns of genetic similarity [Novembre et al., 2008, Lao et al., 2008, Ralph and Coop, 2013]. In particular Ralph and Coop [2013] analyzed spatial patterns in the sharing of PSC segments across Europe. To facilitate comparison with their results, we use their PSC segment calls, focusing on a subset of 2224 individuals after filtering (see Methods).

We applied MAPS to these data using three different PSC segment length bins: $1-5$cM, $5-10$cM, and $>10$cM. The longer bins correspond to more recent demography because as PSC lengths increase, the average coalescent times decrease. Indeed, the average coalescent times for each of these three length bins is inferred to be 90, 23 and 7.5 generations respectively, which correspond to 2700 years, 675 years and 225 years if we assume 30 years per generation.

We note that the accuracy of called PSC segments will vary across these bins: based on simulations in Ralph and Coop [2013] PSC segment calls in the smallest bin (1-5cM) will likely suffer from both false positives and false negatives, whereas for the longer bins PSC calls should be very reliable. Nonetheless, even in the smallest bin, closely-related individuals will

still tend to show higher PSC sharing, and so the estimated MAPS surfaces should provide a useful qualitative summary of spatial patterns of variation even if quantitative estimates may be less reliable.

## Inferring dispersal and population density surfaces

The inferred MAPS dispersal rates (migration rates scaled by grid step size) and population densities (population sizes scaled by grid area size) for each PSC length bin are shown in Figure 2.4.

Largely speaking, the spatial variation in inferred dispersal rates and population densities is remarkably consistent across the different time scales (Figure 2.4). In the MAPS dispersal surfaces, several regions with consistently low estimated dispersal rates coincide with geographic features that would be expected to reduce gene flow, including the English Channel, Adriatic Sea and the Alps. In addition we see consistently high dispersal across the region between the UK and Norway, which may reflect the known genetic effects of the Viking expansion [e.g Leslie et al., 2015]. The MAPS population density surfaces consistently show lowest density in Ireland, Switzerland, Iberia, and the southwest region of the Balkans. This is consistent with samples within each of these areas having among the highest PSC segment sharing (Supplementary Figure S4a). The MAPS inferred country population sizes are also highly correlated with estimated current census population sizes from The World Bank [2016] and National Records of Scotland [2011] (Supplementary Figure S6).

The most notable variation among the estimated surfaces from different time scales is a dramatic increase in the mean estimated population density in the most recent time scale (Figure 2.4 and Supplementary Figure S7). Indeed, the estimated mean for the last time scale – 1.4 individuals per square km – is 6-9 fold higher than those for the earlier time scales (0.16 and 0.22 respectively). This increase is consistent with the recent exponential growth of human population sizes [Cohen, 1995]. The estimates themselves are lower than

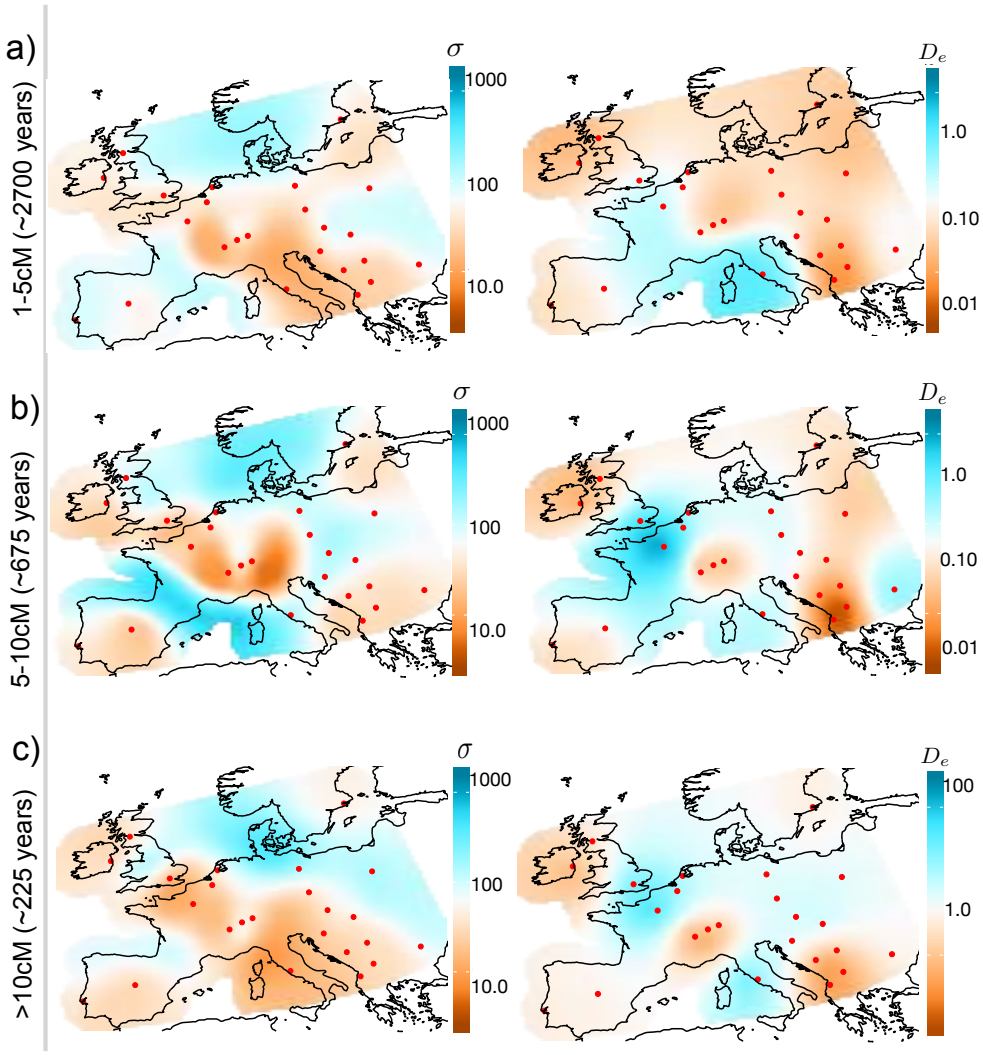



Figure 2.4: **Inferred Dispersal Surfaces and Population Density Surfaces over time for Europe.** We apply MAPS to a European subset of POPRES Nelson et al. [2008] with 2,234 individuals and plot the inferred dispersal $\sigma(\vec{x})$ and population density $D_e(\vec{x})$ surfaces for PSC length bins (a) > 1cM (b) 5-10cM and (c) >10cM. We transform estimates of $\vec{N}$ and $\mathbf{M}$ to estimates of $\sigma(\vec{x})$ and $D_e(\vec{x})$ by scaling the migration rates and population sizes by the grid step-size and area (see Equations (2.17) and (2.18)). Generally, we observe the patterns of dispersal to be relatively constant over time periods, however, we see a sharp increase in population density in the most recent time scale (>10cM). Note the wider plotting limits in inferred densities in the most recent time scale.

historical estimates of ≈1-30 individuals per square km based on archaeological data [e.g. Zimmermann et al., 2009].

The dispersal surfaces show more minor changes between time periods (Figure 2.4 and Supplementary Figure S7). In particular, the estimated mean dispersal rates are relatively constant across time, being 73, 103 and 72 respectively (in units of km in a single generation). These mean estimates are consistent with empirical estimates of 10-100 km in a single generation compiled by Kaplanis et al. [2018] using pedigrees of individuals living between 1650 and 1950 AD. We do note the lower estimated dispersal rates between Portugal and Spain in the analyses of longer PSC segments (5-10 and > 10cM), and the higher estimated dispersal rates through the Baltic Sea (> 10cM segments), possibly reflecting changing gene flow in these regions in recent history.

## Comparison to Ringbauer et al. [2017]

Ringbauer et al. [2017] also estimate a mean dispersal rate and population density from the Eastern European subset of the data analyzed here. Their estimates are based on PSC segments > 4cM, which is most comparable with our analysis of 5-10cM. Unlike our analysis, their estimates are based on a spatially homogeneous model. To compare with their estimates we computed the mean of the estimated dispersal rate and population densities in Eastern Europe (but based on an analysis of the full data). For the dispersal rate this yields an estimate of 88 km in a single generation, which is consistent with the range of 50-100 given by [Ringbauer et al., 2017]. For the population density, it yields an estimate of 0.10 individuals per square km, which is somewhat higher than the estimate of 0.05 obtained under a comparable (time-homogeneous) population model in [Ringbauer et al., 2017]. Possibly our higher estimate partly reflects the influence of our spatial modeling approach, which will tend to shift the estimate for Eastern Europe toward the estimated mean across all of Europe (which is 0.22). In addition, the difference in length thresholds

(> 4cM versus 5-10cM) may also be contributing; if segments in the Ringbauer et al. [2017] analysis are on average shorter and hence older, one would expect lower density estimates, based on our results that suggest lower densities in the past (Figure 2.4).

## Comparison with EEMS

The EEMS results for these data (Figure S8) show non-trivial differences with the MAPS results (Figure 2.4a). Two potential causes are: i) differences in the summary data used (PSC segment sharing vs genetic distances) and hence sensitivity to different timescales; and ii) differences in the underlying models (e.g. composite Poisson likelihood vs Wishart likelihood, and different parameterizations/approximations to the coalescent model; see Discussion). To evaluate the impact of i) we compared the PSC segment sharing and genetic distances, and found their correlation to be only modest (Pearson's $\rho$ = -0.38), with the most notable deviation for comparisons between countries in Eastern Europe (Figure S9a). Furthermore, most of this correlation is due to geographic distance: after controlling for geographic distance the correlation is only -0.18, which may be a more relevant metric because inferred spatial heterogeneity in gene flow (barriers and corridors) is driven by departures from simple isolation by distance.

To better assess the impact of ii) we applied EEMS on a distance matrix constructed to have the same similarity patterns as the PSC segment sharing matrix input to MAPS ($1-5$cM length bin). The resulting EEMS surface is more similar to the corresponding MAPS dispersal surface (Supplementary Figure S9b vs Figure 2.4a), but there remain substantial differences. This investigation confirms what we expected *a priori* — the two surfaces should be different because the underlying models and inferred parameters of MAPS and EEMS are different. As noted before, EEMS infers the "effective migration rate" which reflects the effects of both the migration rates and population sizes, while MAPS infers them separately.

## 2.3 Discussion

We developed a method (MAPS) for inferring migration rates and population sizes across space and time periods from geo-referenced samples. Our method builds upon a previous method developed for estimating effective migration surfaces (EEMS) [Petkova et al., 2016]. However there are several differences between MAPS and EEMS. Most fundamentally, MAPS draws inferences from observed levels of PSC sharing between samples, whereas EEMS draws inferences from the genetic distance. These two data summaries capture different information about the coalescent distributions: in essence, PSC sharing captures the frequency of recent coalescent events, whereas genetic distance captures the mean coalescent time. Consequently MAPS inferences largely reflect the recent past ($\lessapprox$ 1000 years for human recombination rates and generation times with PSC segments $>$ 2cM), whereas EEMS inferences reflect demographic history on a longer timescale across which pairwise coalescence occurs (99% of events $>$ 6000 years old, assuming diploid $N_e$ of 10,000 for humans, exponential coalescent time distribution).

Another consequence of modelling PSC sharing, rather than genetic distance, is that MAPS can separately estimate demographic parameters related to migration rates ($\mathbf{M}$) and population sizes ($\vec{N}$), as in Figure 2.3 for example. In essence MAPS does this by using the known recombination map as an additional piece of information to help calibrate inferences. In contrast EEMS, which makes no use of recombination maps, cannot separate $\mathbf{M}$ and $\vec{N}$. Instead EEMS infers a compound parameter referred to as the "effective migration rate", which is influenced by changes in both $\mathbf{M}$ and $\vec{N}$; see Figure 2.3. In principle, if applied to sequence data instead of genotype data at ascertained SNPs, the genetic distances used by EEMS could perhaps also separately estimate $\mathbf{M}$ and $\vec{N}$ by exploiting known mutation rates to calibrate inferences. However, this would require non-trivial additional changes to the current EEMS likelihood, which was designed to be applicable to ascertained SNPs and does not explicitly model variation in population sizes. (The EEMS likelihood instead

uses a "diversity rate" $e_q$, which reflects within-deme heterozygosity but is not explicitly a population size parameter.)

An additional useful feature of PSC segments is that, by varying the lengths analyzed, one can infer parameter values across different time scales. For example, our simulations show how by contrasting shorter and longer PSC segments, the method can reveal different gene flow patterns in scenarios with recent changes (see Figures 2.2 and 2.3). Further support comes from our empirical analysis of the POPRES data-set, where we found population sizes inferred from longer PSC segments to be more correlated with census sizes The World Bank [2015 census 2016] and National Records of Scotland [2011, 2011 census] than sizes inferred from shorter segments (e.g. Spearman's $\rho = 0.71$ for $1 - 5$cM and $\rho = 0.84$ for $> 10$cM; see Supplementary Figures S5 and S6). Also, not surprisingly, PSC segments greatly outperform using heterozygosity as an indicator of census population size (the Spearman's correlation coefficient between heterozygosity and census size was insignificant, p-value = 0.25).

Our estimates of dispersal distances and population density from the POPRES data are among the first such estimates using a spatial model for Europe (though see [Ringbauer et al., 2017]). The features observed in the dispersal and population density surfaces are in principle discernible by careful inspection of the numbers of shared PSC segments between pairs of countries (e.g. using average pairwise numbers of shared segments, Supplementary Figure S4b, as in Ralph and Coop [2013]). For example, high connectivity across the North Sea is reflected in the raw PSC calls: samples from the British Isles share a relatively high number of PSC segments with those from Sweden (Supplementary Figure S4b). Also the low estimated dispersal between Switzerland and Italy is consistent with Swiss samples sharing relatively few PSC segments with Italians given their close proximity (Supplementary Figure S4b). However, identifying interesting patterns directly from the PSC segment sharing data is not straightforward, and one goal of MAPS (and EEMS) is to produce visualizations that point to patterns in the data that suggest deviations from simple isolation by distance.

Our results suggest that several features of dispersal in Europe have been relatively stable over the last ∼3000 years, whereas the population sizes have been increasing. The relative stability of the gene flow patterns is perhaps surprising given ancient DNA results suggest a continually dynamic history of population movements. One possibility is that much of European population structure may have been established by the end of the Bronze Age ( 4,000 years ago), with relatively more stable patterns in the intervening period that is reflected in lPSC segments. Nonetheless, the dispersal is not completely stable– our results suggest changes in Iberia, the Baltic, and to minor degrees in other areas.

The inferred population size surfaces for the POPRES data show a general increase in sizes through time, with small fluctuations across geography; for instance, Polish samples have a relatively larger population size in inferred values from the largest length scale (> 10cM). In our results, the smallest inferred population sizes are in the Balkans and Eastern Europe more generally. This is in agreement with the signal seen by Ralph and Coop [2013]; however, taken at face value, our results suggest that high PSC sharing in these regions may be due more to consistently low population densities than to historical expansions (such as the Slavic or Hunnic expansions).

Although consistent with previous results, our estimates of dispersal and population sizes do not exactly agree with empirical estimates. For example, our estimates of population sizes are consistently lower than the census sizes (Supplementary Figure S6). This is to be expected for several reasons. First, census sizes include non-breeding individuals (juvenile and post-reproductive age) that do not impact the formation of PSC segments. Second, MAPS is fitting a single population size per location, and in a growing population the best fit population size will be an under-estimate of contemporary size. Third, in a wide class of population genetic models, the effective size, even among reproductive age individuals, is lower than the census size because of factors that inflate the variance in offspring number. Fourth, some discrepancy is expected simply because the stepping-stone population genetic

model used here is only a coarse approximation to the complex spatial dynamics of human populations. Finally, recombination rate mis-specification can bias the inferred parameters. Furthermore, we caution that our results must be interpreted in the light of the fact that we have limited spatial sampling across Europe, and only very coarse geographical origin data (country of origin).

Here, as in Petkova et al. [2016] we use a discrete stepping-stone model to approximate a process that might be more naturally modelled as continuously varying in space. Recent work [Ringbauer et al., 2017, Baharian et al., 2016] exploits continuous models to estimate dispersal and population density parameters from sharing of lPSC segments. However, these methods assume that dispersal and population density are constant across space: extending them to allow these parameters to vary across space could be an interesting avenue for future work.

A major achievement in method development in population genetics would be to jointly infer migration rates and population sizes across both space and time. MAPS is a step towards this goal. However, we do not infer demography explicitly as a function of time and instead infer surfaces in time blocks defined by PSC length bins. In principle, our method allows for inference of demography across time by treating PSC segments as independent across length bins, see Equation (2.27) in Supplementary Notes. However, this requires fitting multiple migration/population surfaces and is computationally unfeasible with our current MCMC routine. Other computational techniques (e.g. Variational Bayes or fast optimization of the likelihood) might make this goal possible.

## 2.4 Methods

### *2.4.1 MAPS configuration*

For the empirical data analysis, we ran MAPS with 200 demes. The MAPS output was obtained by averaging over 20 independent replicates (the number of MCMC iterations in each replicate was to set 5e6, number of burn-in iterations set to 2e6, and we thinned every 2000 iterations). We provide a software tool here: www.github.com/halasadi/MAPS

### *2.4.2 Inferring PSC segments from the data*

When calling PSC segments, we follow the recommendations of Browning and Browning [2011, 2013] and Ralph and Coop [2013] by running BEAGLE multiple times and merging shorter segments.

For the empirical data analysis, we use the PSC segments ("IBD") calls from Ralph and Coop [2013]. We further applied a filter to retain countries with at least 5 sampled individuals, and removed Russian and Greek individuals to restrict the habitat to a smaller spatial scale

### *2.4.3 Model*

MAPS assumes a population genetic model consisting of triangular grid of $d$ demes (or populations) with symmetric migration. The density of the grid is pre-specified by the user with the consideration that the computational complexity is $O(d^3)$. We use Bayesian inference to estimate the MAPS parameters: the migration rates and coalescent rates $M$ and $\underline{q}$ respectively. Its key components are the likelihood, which measures how well the parameters explain the observed data, and the prior, which captures the expectation that $M$ and $\underline{q}$ have some spatial structure (in particular, the idea that nearby edges will tend to have similar migration rates and nearby demes have similar coalescent rates).

MAPS estimate the posterior distribution of $\Theta = M, \underline{q}$ given the data. The data used for MAPS consists of a similarity matrix $X^R = \{X^R_{i,j}\}$ which denotes the number of PSC segments in a range $R = [u, v]$ base-pairs shared between pairs of haploid individuals $(i, j) \in \{1, \cdots, n\} \times \{1, \cdots, n\}$ where $n$ is the number of (haploid) individuals. Furthermore, a recombination rate map is required as input for MAPS. The likelihood is a function of the expected value of $X^R_{i,j}$ ($E[X^R_{i,j}]$). Below we describe the computation of $E[X^R_{i,j}]$ and the other key components of the likelihood. Finally, we briefly describe the prior used and an MCMC scheme to sample from the posterior distribution of $\Theta$.

## The likelihood function

Let $\alpha, \beta$ denote the demes that (haploid) individuals $i$ and $j$ are sampled in, we define,

$$\lambda^{\Theta}_{\alpha,\beta} = E[X^R_{i,j}|\Theta]. \tag{2.1}$$

For the marginal distribution, we assume

$$X^R_{i,j}|\Theta \sim \text{Pois}(\lambda^{\Theta}_{\alpha,\beta}|\Theta), \tag{2.2}$$

and one option for computing the joint distribution of the data is to assume independence between pairs of individuals $(i, j)$ as done previously [Palamara et al., 2012, Palamara and Peer, 2013, Ralph and Coop, 2013, Ringbauer et al., 2017]. This assumption leads to the log-likelihood,

$$log\mathcal{L}(\Theta; \bar{X}) = \sum_{\alpha \leq \beta} n_{\alpha,\beta} \left( \bar{X}_{\alpha,\beta} \log(\lambda^{\Theta}_{\alpha,\beta}) - \lambda^{\Theta}_{\alpha,\beta} \right), \tag{2.3}$$

where $\bar{X} = \{\bar{X}_{\alpha,\beta}\}$ such that $(\alpha, \beta) \in \{1, \cdots, d\} \times \{1, \cdots, d\}$ and $d$ is the number of demes. Furthermore

$$\bar{X}_{\alpha,\beta} = \begin{cases} \frac{1}{n_\alpha n_\beta} \sum_{i \in d_\alpha, j \in d_\beta} X^R_{ij} & \text{if } \alpha \neq \beta \\ \frac{1}{\binom{n_\alpha}{2}} \sum_{i \in d_\alpha, i<j} X^R_{ij} & \text{if } \alpha = \beta \end{cases}, \tag{2.4}$$

where $n_\alpha$ is the number of sampled individuals in deme $\alpha$, $d_\alpha$ is the set of all individuals in deme $\alpha$, and

$$n_{\alpha,\beta} = \begin{cases} n_\alpha n_\beta & \text{if } \alpha \neq \beta \\ \binom{n_\alpha}{2} & \text{if } \alpha = \beta \end{cases}. \tag{2.5}$$

However, we found that there were significant correlations in lPSC segments between individuals. To deal with this, we down-weighted the likelihood function to reflect the "effective" number of samples ($e_{\alpha,\beta}$) instead of the number of pairs ($n_{\alpha,\beta}$). The effective number of samples between demes $\alpha$,$\beta$ is given by,

$$e_{\alpha,\beta} = \frac{\bar{X}_{\alpha,\beta}}{\text{Var}[\bar{X}_{\alpha,\beta}]}. \tag{2.6}$$

In the case of independence, $\text{Var}[\bar{X}_{\alpha,\beta}] \approx \frac{\bar{X}_{\alpha,\beta}}{n_{\alpha,\beta}}$. However, because of correlations in the data, the actual variance is significantly larger than the variance computed under an independence model. Here, we estimate $\text{Var}[\bar{X}_{\alpha,\beta}]$ by bootstrapping individuals with replacement. This way, we model the correlations between pairs of individuals for within and between-deme comparisons. The loglikelihood adjusted for correlations is given by,

$$log\mathcal{L}(\Theta; \bar{X}) = \sum_{\alpha \leq \beta} e_{\alpha,\beta} \left( \bar{X}_{\alpha,\beta} \log(\lambda^{\Theta}_{\alpha,\beta}) - \lambda^{\Theta}_{\alpha,\beta} \right). \tag{2.7}$$

## Computing the expectation of $X^R_{i,j}|\Theta$

Next, we derive expressions to compute the expectation of the number of PSC segments of length greater than $u$ ($X^{R=[\mu,\infty)}_{i,j}$) conditional on the demography $\Theta$. From results in

Palamara et al. [2012] it is easy to show that

$$E[X_{i,j}^{R=[\mu,\infty)}|\Theta] \approx G \int_u^\infty f_L(l|\Theta)/l \, dl, \tag{2.8}$$

where $G$ denotes the length of the genome (in base-pairs), $L$ denotes the random length (in base-pairs) of the PSC segment between $i$ and $j$ containing a pre-specified position in the genome (base $b$ say), and $f_L$ is its probability density. Intuitively, $Gf_L(l|\Theta)$ is the expected number of base-pairs that lie in PSC segments of length $l$, making $\frac{Gf_L(l|\Theta)}{l}$ the expected number of PSC segments of length $l$. Integrating the latter quantity from $\mu$ to $\infty$ gives the desired result.

To help compute (2.8) we introduce $T_{ij}$ to denote the (random) coalescent time in generations between $i$ and $j$ at base $b$, with density $f_{T_{ij}}(t|\Theta)$. Then (2.8) can be written as an integral over $T_{ij}$:

$$E[X_{i,j}^{R=[\mu,\infty)}|\Theta] \approx G \int_\mu^\infty f_L(l|\Theta)/l \, dl \tag{2.9}$$

$$= G \int_\mu^\infty \int_0^\infty f_{L,T_{i,j}}(l,t|\Theta)/l \, dt \, dl \tag{2.10}$$

$$= G \int_0^\infty f_{T_{i,j}}(t|\Theta) \int_\mu^\infty f_L(l|t)/l \, dl \, dt, \tag{2.11}$$

using the relation that $f_{L,T_{i,j}}(l,t|\Theta) = f_L(l|t,\Theta) f_{T_{i,j}}(t|\Theta) = f_L(l|t) f_{T_{i,j}}(t|\Theta)$. A key simplification here comes from the fact that, given $T_{ij}$, $L$ is conditionally independent of $\Theta$.

It can be shown that the conditional distribution of $L$ given $T_{ij}$ is an erlang-2 distribution [Palamara et al., 2012, Palamara and Peer, 2013, Hein et al., 2004] with density

$$f_L(l|t) = 4r^2t^2le^{-2trl}, \tag{2.12}$$

where $r$ is the recombination rate per base-pair. Substituting this into the inner integral of

(2.11) and integrating analytically yields

$$\int_u^\infty f_L(l|t)/l\, dt = 2rte^{-2tru}, \tag{2.13}$$

leading to

$$E[X_{i,j}^{R=[\mu,\infty)}|\Theta] \approx G \int_0^\infty f_{T_{i,j}}(t|\Theta) 2rte^{-2tru} dt. \tag{2.14}$$

Here, we assume the probability density of $T_{i,j}$ is given by,

$$f_{T_{i,j}}(t|\Theta) \approx \sum_\kappa q_k (e^{-Mt})_{\alpha,\kappa} (e^{-Mt})_{\beta,\kappa}, \tag{2.15}$$

where demes $\alpha, \beta$ denote the deme where lineages $i$ and $j$ are sampled from, $q_\kappa = \frac{1}{2N_\kappa}$ is the coalescent rate in deme $\kappa$, and $M = \langle m_{\alpha,\beta} \rangle$ is the migration rate matrix between all $d$ demes such that $(\alpha, \beta) \in \{1, \cdots, d\} \times \{1, ..., d\}$. We compute the matrix exponential by first diagonalizing the matrix $M = PDP^T$ and compute $e^{-Mt} = Pe^{-Dt}P^T$.

Having computed all individual components of $\int_0^\infty f_{T_{i,j}}(t|\Theta) 2rte^{-2tru} dt$, we are left to evaluate a one-dimensional integral which we do by Gaussian quadrature (with 50 weights).

To compute the expected number of PSC segments in a range $R = (\mu, \nu)$

$$E[X_{i,j}^{R=[\mu,\nu]}] = E[X_{i,j}^{R=[\mu,\infty)}] - E[X_{i,j}^{R=[\nu,\infty)}]. \tag{2.16}$$

As mentioned previously, the units of $\mu, \nu$ are in base-pairs. However, we can transform to units of centiMorgans (cM) by : $\mu_{cM} = 100\mu r$.

## The Prior

MAPS uses a hierarchical prior parameterized by Voronoi tessellation (similar to EEMS). The Voronoi tessellation partitions the habitat into $C$ cells. Given a Voronoi tessellation of the habitat, each cell $c \in \{1, \cdots, C\}$ is associated with a migration rate ($\mathcal{M}_c$) and population

size ($\mathcal{N}_c$). Demes ($\alpha$) that fall into cell $c$ will have population size $N_\alpha = \mathcal{N}_c$, and similarly, migration rates between demes $\alpha, \beta$ equal $m_{\alpha,\beta} = \frac{\mathcal{M}_{c_1}+\mathcal{M}_{c_2}}{2}$ if demes $\alpha, \beta$ fall into cells $c_1$ and $c_2$. We use an MCMC to integrate over the distribution on partitions of Voronoi cells.

## The MCMC

We break up the MCMC updates into updating a series of conditionally independent distributions. Provided the conditional posterior distributions for each part give support to all the parameter space, this will define an irreducible Markov chain with the correct joint posterior distribution Stephens [2000]. We use Metropolis-Hastings to update all parameters, and random-walk proposals for most updates, with exception of birth-death updates for updating the number of Voronoi cells.

## Transformation of parameters to continuous space

Given an inferred population size at a particular deme $\alpha$ and a grid with uniform spacing, the transformation from population size to population density is given by

$$D_e(x) = \frac{N_\alpha}{\Delta A}, \tag{2.17}$$

where $\Delta A = \frac{\mathcal{A}_H}{d}$ is the area covered per deme such that $\mathcal{A}_H$ is the area of the habitat (in km$^2$), $d$ is the number of demes, and $x$ corresponds to the spatial position of deme $\alpha$. Intuitively, (2.17) implies that the density multiplied by the area equals population size, i.e. $D_e(x)\Delta A \approx N_\alpha$. Equation (2.17) can is analogous to equation 7 in [Baharian et al., 2016].

Given a migration rate ($m$), the transformation to dispersal distances is given by,

$$\sigma = \sqrt{m}\Delta x, \tag{2.18}$$

where $\Delta x$ is the step size of the grid (km). The dispersal distance represents the distance

traveled by an individual after one generation, and sometimes is called the "root mean square distance" or "dispersal rate" [Barton et al., 2002]. Please see derivation of (2.18) below in Supplementary Notes.

## 2.5 Supplementary Notes

### *2.5.1 The model*

The coalescent process for two samples under a multi-deme model can be described by a continuous time Markov chain (CTMC) [Bahlo and Griffiths, 2001]. Let $i, j$ represent sampled lineages and $\alpha, \beta$ their locations, respectively, $d$ is the number of demes (or populations) and $(\alpha, \beta) \in \{1, \cdots, d\} \times \{1, ..., d\}$. Let $c$ denote the coalescent state. The infinitesimal rate matrix $R$ of this CTMC is

$$\begin{aligned}
R_{(\alpha,\beta),(\gamma,\beta)} &= m_{\alpha,\gamma} \;\; \beta = 1, ..., d, \; \gamma \neq \alpha \\
R_{(\alpha,\beta),(\alpha,\gamma)} &= m_{\beta,\gamma} \;\; \alpha = 1, ..., d, \; \gamma \neq \beta \\
R_{(\alpha,\alpha),(c)} &= q_\alpha \\
R_{(\alpha,\beta),(\alpha,\beta)} &= -(m_{\alpha_+} + m_{\beta_+}) - \delta_{\alpha\beta} q_\alpha \\
R_{(c),(c)} &= 0 \\
R_{(\alpha,\beta),(\gamma,\kappa)} &= 0 \;\; \gamma, \kappa = 1, ..., d, \; \gamma \neq \alpha, \; \kappa \neq \beta,
\end{aligned} \tag{2.19}$$

where $M = \langle m_{\alpha,\beta} \rangle$ denotes the migration rate matrix, and $m_{\alpha,\beta}$ is the migration rate between demes $\alpha, \beta$ and $q_\alpha = \frac{1}{2N_\alpha}$ is the coalescent rate of deme $\alpha$ which is proportional to the inverse of the population size at deme $\alpha$ ($N_k$). Let $T_{i,j}$ denote the (random) coalescent time between the pair of sampled lineages, and $f_{T_{i,j}}(t)$ denote the probability density of a coalescent event at time $t$. Here, we derive $f_{T_{i,j}}(t)$ by conditioning on the position of the two lineages.

**Lemma 2.5.1.** *Let* $\big(X_i(t), X_j(t)\big) \in \{1, \cdots, d\} \times \{1, ..., d\}$ *denote the position of lineage* $i$ *and lineage* $j$ *at time* $t$ *respectively. The probability density* $f_{T_{i,j}}(t)$ *that lineage* $i$ *and* $j$ *coalesce at time* $t$ *is given by* $\sum_{\kappa=1}^{d} q_\kappa P(X_i(t) = \kappa, X_j(t) = \kappa)$.

For $\Delta t \approx 0$,

$$P(T_{i,j} \in [t, t + \Delta t]) \tag{2.20}$$

$$\approx \sum_{\kappa=1}^{d} P(T_{i,j} \in [t, t + \Delta t] | X_i(t) = \kappa, X_j(t) = \kappa) P(X_i(t) = \kappa, X_j(t) = \kappa) \tag{2.21}$$

$$\approx \sum_{\kappa=1}^{d} q_\kappa \Delta t P(X_i(t) = \kappa, X_j(t) = \kappa). \tag{2.22}$$

Taking the limit $\Delta t \to 0$, we arrive at the density

$$f_{T_{i,j}}(t) = \lim_{\Delta t \to 0} \; P(T_{i,j} \in [t, t + \Delta t])/\Delta t = \sum_{\kappa=1}^{d} q_\kappa P(X_i(t) = \kappa, X_j(t) = \kappa). \tag{2.23}$$

## The random walk approximation to the coalescent

Here, we introduce an approximation,

$$P(X_i(t) = \kappa, X_j(t) = \kappa) \approx P(X_i(t) = \kappa) P(X_j(t) = \kappa). \tag{2.24}$$

The intuition is that probability that lineage $i$ and $j$ coalesce before time $t$ is extremely small such that the two lineages approximately behave like two independently moving particles. Each lineage can be modeled by a random walk with transition matrix $M$. These assumptions were also made in the context of continuous spatial diffusion models for haplotype sharing Baharian et al. [2016], Ringbauer et al. [2017] , and even further back, as a general

approximation to the two-dimensional continuous-space coalescent process [Barton et al., 2002, Wilkins, 2004, Blum et al., 2004, Novembre and Slatkin, 2009, Robledo-Arnuncio and Rousset, 2010].

This approximation implies that

$$f_{T_{i,j}}(t) \approx \sum_{\kappa} q_k (e^{-Mt})_{\alpha,\kappa} (e^{-Mt})_{\beta,\kappa}, \tag{2.25}$$

where lineages $i, j$ are initially sampled in deme $\alpha, \beta$. Or equivalently in matrix form,

$$f_{T_{i,j}}(t) \approx \left( e^{-Mt} Q e^{-Mt} \right)_{i,j}, \tag{2.26}$$

where $Q = diag(q_1, ..., q_d)$.

## Varying migration rates and population sizes across time

**Corollary 2.5.1.1.** *Let time slice $k$ be defined by the interval $t_{k-1} < t < t_k$, $M_k$ denote the migration rate matrix in time slice $k$, and $Q_k = diag(q_1^k, ..., q_d^k)$ where $q_\alpha^k$ denotes the coalescent rate in deme $\alpha$ at time slice $k$. Let $T_{i,j}$ denote the coalescent time between lineage $i, j$ sampled in demes $\alpha, \beta$, then under the independence assumption, for $t \in (t_{K-1}, t_K)$,*

$$f_{T_{i,j}}(t) \approx \left( G_K(t) Q_K G_K(t) \right)_{\alpha,\beta}, \tag{2.27}$$

*where* $G_K(t) = \exp\left( - \sum_{k=1}^{K-1} (t_k - t_{k-1}) M_k - (t - t_K) M_K \right)$.

## Expected number of lPSC segments given the demography $\Theta$

**Lemma 2.5.2.** *Let $X_{i,j}^{\mu}$ denote the number of PSC segment greater than $\mu$ basepairs shared between haploid individuals $i, j$, $\Theta$ denote the demographic model, $G$ the size of the genome, $L$ denotes the random length (in base-pairs) of the PSC segment between $i$ and $j$ containing*

*a pre-specified position in the genome, then* $E[X^{\mu}_{i,j}|\theta] \approx G \int_u^{\infty} f_L(l|\Theta)/l\, dl$.

Let $E[\mathcal{F}^{\mu}|\Theta]$ denote the expected fraction of the genome between $i, j$ that lies in PSC segments greater than $\mu$, and $E[s^{\mu}|\Theta]$ the expected size of a PSC segment conditional on it being at least length $\mu$. According to equations 9-14 from [Palamara et al., 2012],

$$E[X^{\mu}_{i,j}|\theta] \approx \frac{G\, E[\mathcal{F}^{\mu}|\theta]}{E[s^{\mu}|\Theta]}, \tag{2.28}$$

$$E[\mathcal{F}^{\mu}|\Theta] = \int_{\mu}^{\infty} f_L(l|\Theta)dl, \tag{2.29}$$

$$E[s^{\mu}|\Theta] = \frac{\int_{\mu}^{\infty} f_L(l|\Theta)dl}{\int_{\mu}^{\infty} f_L(l|\Theta)/l\, dl}. \tag{2.30}$$

We obtain the desired result by substituting (2.29) and (2.30) into (2.28) and canceling like-terms.

## Expected age of a segment

We choose PSC segment lengths based on their expected age which is derived below.

**Lemma 2.5.3.** *The expected coalescent time (t, in generations) of an PSC segment between between length* $L_1$ *centiMorgans and* $L_2$ *centiMorgans is approximately* $\frac{300}{4}(\frac{1}{L_1} + \frac{1}{L_2})$ *if the effective population size (N) is sufficiently large.*

We choose to work in units of basepairs, and will convert back to units of morgans at the end. We convert $L_1$ into units of base-pairs with the transformation: $\mu = \frac{L_1}{100r}$ and similarly $\nu = \frac{L_2}{100r}$.

Let us denote $T|l, N$ as the random coalescent time of a PSC segment that is at least length $l$ under a single-deme demography model with population size $N$. The expected

coalescent time of an PSC segment longer than $\mu$ base-pairs can be expressed as

$$E[T|l \geq \mu, N] = \int_0^\infty t f_T(t|l \geq \mu, N)dt = \int_0^\infty t \frac{f_L(l \geq \mu|t) f_T(t|N)}{f_L(l \geq \mu|N)} dt$$
$$= \frac{\int_0^\infty t f_L(l \geq \mu|t) f_T(t|N)dt}{\int_0^\infty f_L(l \geq \mu|t) f_T(t|N)dt}, \tag{2.31}$$

where $f_L(l|t) = 4r^2t^2le^{-2trl}$ denotes the probability density that a PSC segment is of length $l$ given it has a common ancestor event at time $t$, $f_T(t|N)$ denotes the probability density that a coalescent event occurs at time $t$ under the demography model with population size $N$.

Next, we expand a key term in equation (2.31)

$$f_L(l \geq \mu|t) = \int_\mu^\infty f_L(l|t)dl = (2rt\mu + 1)\exp\Big( - 2rt\mu + 1\Big) \tag{2.32}$$

and assume,

$$f_T(t|N) = \frac{e^{-t/N}}{N}. \tag{2.33}$$

Putting everything together,

$$E[T|l \geq \mu, N] = \frac{N(1 + 6Nr\mu)/(1 + 2Nr\mu)^3}{(1 + 4Nr\mu)/(1 + 2Nr\mu)^2} = \frac{N(1 + 6Nr\mu)}{1 + 6Nr\mu + 8N^2(r\mu)^2}. \tag{2.34}$$

We can remove the dependence of $N$ by taking $\lim_{N\to\infty}$ as done similarly in Baharian et al. [2016],

$$\lim_{N\to\infty} E[T|l \geq \mu, N] = \frac{3}{4r\mu} \tag{2.35}$$

Now that we have derived the expected age of PSC segment longer than $\mu$, it is quite simple

to expand the equation for PSC segments between $\mu$ and $\nu$ base-pairs,

$$E[T|\mu \leq l \leq \nu] = \frac{\int_0^\infty t f_L(\mu \leq l \leq \nu|t) f_{T|N}(t)dt}{\int_0^\infty f_L(\mu \leq l \leq \nu|t) f_{T|N}(t)dt} = \frac{\int_0^\infty t\Big(f_L(l \geq \nu|t) - f_L(l \geq \mu|t)\Big) f_{T|N}(t)dt}{\int_0^\infty \Big(f_L(l \geq \nu|t) - f_L(l \geq \mu|t)\Big) f_{T|N}(t)dt}$$
$$= \frac{3}{4}\left(\frac{1}{r\mu} + \frac{1}{r\nu}\right) \tag{2.36}$$

We transform back to units of centimorgans: let $L_1 = 100r\mu$ and $L_2 = 100r\nu$ be in units of centiMograns, and taking the limit, we get the desired result

$$\lim_{N\to\infty} E[t|\mu \leq l \leq \nu] = \frac{300}{4}\left(\frac{1}{L_1} + \frac{1}{L_2}\right). \tag{2.37}$$

### *2.5.2 Transformation of migration rates to dispersal rates*

Migration rates inferred under a discrete model can be transformed to dispersal distances representing parameters in continuous space. Here, we derive the transformation.

**Lemma 2.5.4.** *Consider a random walk on a 2D grid, where steps are taken according to a Poisson process with rate $m$, and let $\underline{X}(t)$ be a vector denoting the coordinates of the particle at time $t$. The distribution of $\underline{X}(t)$ approximately only depends on the compound parameter $m(\Delta x)^2$ (or equivalently $\sqrt{m}\Delta x$).*

$$\underline{X}(t) = \sum_{i=1}^{N(t)} \underline{Z}_i, \tag{2.38}$$

where $N(t)$ is the number of steps taken by time $t$, and $\underline{Z}_i$ is a random variable representing the direction and magnitude taken at step $i$. Since $X(t)$ is a sum of iid variables, a form of the central limit theorem applies here and $X(t)$ converges to the normal distribution [Rényi, 1960].

In a random walk on a triangular grid, a particle can move in one of the 6 directions (upper-right, right, lower-right, left, upper-left, and lower-left):

$$
\begin{aligned}
&\underline{Z}_i \\
&\quad = (1/2, \Delta x\sqrt{3}/2)^T \text{ with } p = 1/6 \\
&\quad = (\Delta x, 0)^T \text{ with } p = 1/6 \\
&\quad = (\Delta x/2, -\Delta x\sqrt{3}/2)^T \text{ with } p = 1/6 \\
&\quad = (-\Delta x, 0)^T \text{ with } p = 1/6 \\
&\quad = (-\Delta x/2, \Delta x\sqrt{3}/2)^T \text{ with } p = 1/6 \\
&\quad = (-\Delta x/2, -\Delta x\sqrt{3}/2)^T \text{ with } p = 1/6
\end{aligned}
$$

where $\Delta x$ represents the step size in the grid (i.e. edge length). The mean and variance are given by,

$$E[\underline{X}(t)] = 0 \tag{2.39}$$

and,

$$Var[\underline{X}(t)] = \frac{mt(\Delta x)^2}{2} I_2. \tag{2.40}$$

where $I_2$ is the identity matrix. Under normality, the mean and variance are sufficient statistics. Note that (2.39) and (2.40) also hold for square grids.

## Interpretation of the migration diffusion parameter $m(\Delta x)^2$

In addition, we provide a physical interpretation to $(\Delta x)^2$ in terms of the squared distance from the origin per generation. Let the distance $d = \|\underline{X}(t)\| = \sqrt{x_1^2 + x_2^2}$, then

$$E[d^2]/t = E[x_1^2 + x_2^2]/t = E[x_1^2]/t + E[x_2^2]/t = \frac{m(\Delta x)^2}{2} + \frac{m(\Delta x)^2}{2} = m(\Delta x)^2. \quad (2.41)$$

$\sqrt{\frac{E[d^2]}{t}} = \sqrt{m}\Delta x$ can be interpreted as the distance traveled by an individual after one generation, and sometimes is called the "dispersal" distance or the "root mean square distance".

### *2.5.3 Diversity rates versus coalescent rates*

For computational efficiency, the EEMS software uses a combination of the resistance distance model and within-deme "diversity rates" to approximate expected pairwise coalescent times, in which,

$$E[\hat{T}_{\alpha,\beta}] = \begin{cases} \frac{R_{\alpha,\beta}}{4} + \frac{e_{q_\alpha}+e_{q_\beta}}{2} & \text{if } \alpha \neq \beta \\ e_{q_\alpha} & \text{if } \alpha = \beta \end{cases}. \quad (2.42)$$

where $E[\hat{T}_{\alpha,\beta}]$ is the resistance distance approximation to the expected coalescent time between deme $\alpha$ and deme $\beta$, $e_{q_\alpha}$ is the "diversity rate" in deme $\alpha$, and $R_{\alpha,\beta}$ is the resistance distance between demes $\alpha, \beta$ [Petkova et al., 2016]. The diversity rates have no simple expression in terms of population-genetic parameters under the multi-deme coalescent model. As an alternative, diversity rates can be interpreted as reflecting average within deme heterozygosity since $e_q = E[\hat{T}_w] \propto H_\alpha$ where the heterozygosity for deme $\alpha$ ($H_\alpha$) is defined as,

$$H_\alpha = \frac{1}{\binom{n_\alpha}{2}} \sum_{i<j, i\in\alpha, j\in\alpha} D_{i,j}, \quad (2.43)$$

where $D_{i,j}$ is the average number of differences between (haploid) individuals $i$ and $j$.

### Migration and population sizes are identifiable in MAPS

MAPS models the recombination process using rates estimated from a recombination rate map. In this model, population sizes and migration rates can be inferred separately rather than as a joint parameter. Intuitively, the recombination rate serves an independent clock to calibrate estimates.

More formally, a statement of identifiability is a statement regarding the likelihood. MAPS models the expected number of lPSC segments shared between pairs of (haploid) individuals, and can be computed with an integral (2.14). The integral can be broken up into a product of two functions: a function describing the decay of PSC segments as a function of time ("recombination rate clock"), and the coalescent time probability density $f_{T_{i,j}}(t)$ (2.15). The migration rates and population sizes only appear in $f_{T_{i,j}}(t)$, and cannot be cannot be factored into parameters involving combinations of the migration rates and population sizes.

## *2.5.4 The prior*

The structure of the prior closely resembles the prior in the EEMS method Petkova et al. [2016]. The tessellation for the migration rates ($T_m$) is encoded by a list $(\underline{l^m}, \underline{m}, c_m, \mu_m)$ where $\underline{l^m}$ are the locations of each cell, $\underline{m}$ the rates of each cell, and are vectors of length $c_m$ (i.e. number of Voronoi cells), and $\mu_m$ is the overall mean migration rate. The Voronoi tessellation for the coalescent rates is $T_q = (\underline{l^q}, \underline{q}, c_q, \mu_q)$.

The location of each (unordered) Voronoi cell is distributed uniformly across the habitat,

$$l_c^m \overset{iid}{\sim} U(H), \tag{2.44}$$

and the number of cells (a-priori) are drawn from a negative binomial distribution,

$$c_m \sim \text{NegBi}(r_m, p_m). \tag{2.45}$$

The effects of each Voronoi cell is normally distributed with variance $\omega^2$.

$$\log10(m_i) \overset{iid}{\sim} N(\mu_m, \omega_m^2) \tag{2.46}$$

$$\log10(q_i) \overset{iid}{\sim} N(\mu_q, \omega_q^2) \tag{2.47}$$

The probability of a particular (unordered) cell configuration is,

$$p(\underline{m}|c_m) = c_m! \prod_{i=1}^{c_m} N(m_i|\mu_m, \omega_m^2) \tag{2.48}$$

We assume,

$$\log10(\omega_m) \sim U(-3, \log10(1.5)) \tag{2.49}$$

$$\log10(\omega_q) \sim U(-3, log10(1)) \tag{2.50}$$

We set log10(2) as the upper bound for $\log10(\omega_m)$ so the $m$ so the probability that it is within 3 orders of magnitude from the mean is 0.95 a priori, and we set log10(1) as the upper bound for $\log10(\omega_q)$ to restrict the population sizes so to be within 2 orders of magnitude from the mean with probability 0.95 a priori.

We assume,

$$\mu_m \sim U(-10, 4) \tag{2.51}$$

$$\mu_q \sim U(-10, 4). \tag{2.52}$$

We place a uniform prior on the log of the mean rates to reflect that we are uncertain about the order of magnitude. Here, the data is highly informative of the mean, as a result, we can allow the support of the prior to vary by many orders of magnitude.

### *2.5.5 MCMC*

#### Re-parameterization

We re-parameterize the model to improve mixing of the MCMC. We decouple the migration (or coalescent) rates from the mean rate ($\mu$), and variance ($\omega$) by introducing a new variable $e_i$,

$$e_i \overset{iid}{\sim} N(0, 1) \tag{2.53}$$

and the cell specific migration (or coalescent) rates are computed as,

$$\log10(m_i) = e_i\omega + \mu, \tag{2.54}$$

which allows us to update the magnitude of the parameters ($\mu$) and the variance scale ($\omega$) separately.

We add MH joint random-walk updates to $\mu$ and $e_i$ to ensure that $\bar{e} = \frac{\sum_i e_i}{c} \approx 0$. To do this, we jointly update $\mu$ and $e_i$ by,

$$\mu' = \mu + \epsilon \tag{2.55}$$

$$e_i' = e_i - \frac{\epsilon}{\omega} \tag{2.56}$$

where $\epsilon \sim N(0,1)$. We do this for both the migration rates and population sizes.

## Updating the number of cells

The number of cells change the dimension of the likelihood, and a result, we must use a Reversible Jump MCMC step so that the ratio of densities in the Metropolis-Hastings acceptance ratio is well-defined [Green, 1995]. We choose to update the number of cells with a birth-death update [Stephens, 2000]. Fortunately, in such a case, the updates reduce to standard Metropolis-Hastings because the dimension matching constant (i.e. the "Jacobian") equals one [Petkova et al., 2016, Stephens, 2000]. See equations S31 and S32 in Petkova et al. [2016] for formulas regarding the birth-death update. Here, we use nearly identical updates (with a slight modification).

When increasing the number of cells from $c$ to $c+1$ (i.e. a birth-update), we randomly choose a location uniformly across the habitat, and the new migration is proposed from a standard normal because our cell effects are standardized. In contrast, EEMS proposes cell effects migration to be normally distributed around a cell effect at a randomly chosen point in the habitat. Here we set, p(birth) = p(death) = 0.5 if the number cells $\geq 1$, otherwise p(birth) = 1.

The acceptance ratio for a birth update (going from $c$ cells to $c+1$ cells) is

$$\alpha(x, x') = min(1, \frac{p(\text{death})}{p(\text{birth})} \frac{l(x')p(x')\frac{1}{c+1}}{l(x)p(x)N(e_{c+1}|0,1)}), \tag{2.57}$$

where $x$ denotes the current state of the MCMC, $x'$ the proposed state, $e_{c+1}$ is the proposed cell effect drawn from a standard normal. Conversely, in a death-update, we randomly choose one cell uniformly to kill. In this case, the acceptance ratio for a death proposal (going from

$c+1$ cells to $c$ cells) is

$$\alpha(x, x') = min(1, \frac{p(\text{birth})}{p(\text{death})} \frac{l(x')p(x')N(e_c|0,1)}{l(x)p(x)\frac{1}{c+1}}). \tag{2.58}$$

# CHAPTER 3

# SUBSTITUTION RATE HETEROGENEITY CAUSES FALSE POSITIVE INFERENCES IN PHYLOGENETIC BASED-TESTS OF POSITIVE SELECTION

*(with Joseph Thornton)*

## 3.1 Introduction

Positive selection is the evolutionary process by which beneficial alleles rapidly increase in frequency in a population. A central program in modern studies of molecular evolution is to identify genes involved in adaptive evolution based on statistical signatures of positive selection in gene sequence data [Yang and Bielawski, 2000, Nielsen et al., 2005, Przeworski and Bustamante, 2004]. One major family of methods detects positive selection by assessing the ratio of the rate of non-synonymous substitution (dN) to that of synonymous substitution (dS). Synonymous changes do not alter the protein sequence, so they are assumed to evolve under the influence of mutation and drift alone; thus, $dN/dS = 1$ is expected in the absence of selection, $dN/dS < 1$ is taken as evidence of purifying selection, and positive selection is inferred when $dN/dS > 1$ at some or all sites. The earliest positive selection tests simply counted the number of synonymous and non-synonymous substitutions between sequences [Nei and Gojobori, 1986, Suzuki and Gojobori, 1999]. These methods have been largely supplanted by a family of statistical tests that use maximum likelihood (ML) phylogeny-based approaches and probabilistic codon models of evolution [Zhai et al., 2012, Yang and Bielawski, 2000]; these ML tests have now been used to support thousands of published inferences of genes that have evolved under historical positive selection.

In the ML tests, the null hypothesis of neutral evolution and the alternative hypothesis of

positive selection are each represented by a probabilistic Markov model of sequence change. The likelihood of each model  defined as the probability that all the extant sequence data would have evolved given the phylogeny, the model, and its parameters  is calculated, and these are compared to determine whether a statistically significant improvement results when positive selection is incorporated into the model [Nielsen and Yang, 1998, Yang, 1998, Yang and Bielawski, 2000]. In the simplest version of these tests, the models contain a single selection-related parameter $\omega$, which represents the ratio of the instantaneous rates of non-synonymous and synonymous substitution; in the null model, $\omega$ is constrained to values $\leq 1$, and the positive selection model allows any value, including $\omega > 1$. The sites test elaborate on this approach to incorporate the possibility that positive selection might affect only a subset of codons in a gene, while others drift neutrally or are constrained by purifying selection [Nielsen and Yang, 1998, Yang and Bielawski, 2000] . This scenario is represented by a mixture model that sums likelihoods over submodels that have different values of $\omega$. The null model sums the likelihoods of two submodels  one in which sites are subject to purifying selection ($\omega < 1$) and another that evolves by drift ($\omega = 1$). The positive selection model adds a third submodel in which $\omega$ may exceed 1, representing sites that repeatedly undergo rapid substitution at a rate greater than drift. The null hypothesis is a constrained version of the positive selection model, so the latter will always have a greater or equal likelihood; a likelihood ratio test using a $\tilde{\chi}^2$ distribution is therefore used to determine whether the observed improvement in likelihood is greater than would be expected due to fitting stochastic variation in the data if the null hypothesis were true. A further elaboration, the branch-sites test, uses a similar approach to test for positive selection in a subset of sites on specified lineages of the tree.

The sites and branch-sites tests have been applied to many thousands of genes and gene families and have provided the foundation for a great number of inferences of positive selection. In recent years, new methods and models have been introduced, but the sites and

branch-sites test remain in widespread use. Simulation studies have shown that the sites test is powerful and accurate when the underlying model used for the analysis is the same one used to generate the data. Specifically, when evolution is simulated *in silico* under the sites tests null model and the resulting sequences are then analyzed using the sites test, the rate of false positive inferences is lower than the acceptable error rate [Anisimova et al., 2001, Wong et al., 2004]. In real phylogenetic analyses, however, the true model is never known, and it is unlikely that it is ever available, because real protein coding sequences evolve under complex dynamics [Thornton and Kolaczkowski, 2005]. The models used in the standard implementation of the sites test are relatively simple: along with three categories of dN/dS ratio, they include a parameter for different rates of nucleotide transitions and transversions and independent equilibrium frequency parameters for each of the four nucleotides at the three codon positions. They do not incorporate many other ways in which evolutionary dynamics may vary among sequence sites and lineages. For example, they assume that the same rate of synonymous substitution applies to all codons in the sequence, that the rate of nonsynonymous substitutions does not depend on the amino acids being exchanged, and that the equilibrium frequency of codons is identical among sites and does not depend on the amino acid the codon produces. If these assumptions are violated, then both the null and positive selection models would be incorrect; in principle, the likelihood ratio test might reject the null model even when positive selection is absent, because the extra parameters in the positive selection model might allow a significantly better fit to the data even if those parameters are different from those that generated the data [Zhai et al., 2012, Wong et al., 2004].

Heterogeneity in evolutionary rates is a particularly pervasive form of evolutionary complexity, but little is known about the effects of unincorporated rate heterogeneity on the sites test. One form of rate heterogeneity  variation due to differences in the dN/dS ratio among sites – is incorporated into the sites tests mixture models. There are, however, other impor-

tant forms of rate heterogeneity that have not been thoroughly assessed for their effects. For example, codons may differ in their substitution rates for reasons unrelated to variability in dN/dS if, for example, mutation rates or selective constraints on synonymous substitution vary among codons, or the recombination rate varies along a sequence. In such cases, the rate of synonymous substitution, or the overall rate of both synonymous and nonsynonymous substitution, would differ across codons. These kinds of heterogeneity – which we refer to simply as substitution rate variation, to distinguish them from variation in dN/dS per se are pervasive in real datasets [Yang, 1996, Pond and Muse, 2005a, Dimitrieva and Anisimova, 2014, Shriner et al., 2003, Begun et al., 2007].

Previous work suggests that substitution rate heterogeneity might bias the sites test [Pond and Muse, 2005b]. In particular, among-codon variation in the synonymous substitution rate has been shown to inflate estimates of the dN/dS parameter $\omega$ and to affect the post-hoc classification of individual sites as positively selected or not [Pond and Muse, 2005b, Scheffler et al., 2006, Spielman and Wilke, 2015]. There has not, however, been a systematic analysis of the effects of substitution rate heterogeneity on the sites tests primary use as a statistical test of the hypothesis that positive selection was involved in the evolution of some gene. It is unknown whether substitution rate heterogeneity causes elevated rates of false inferences of positive selection or if it does – how much heterogeneity is necessary to lead to high error rates. Further, if a bias is produced, whether it strongly affects conclusions under realistic conditions is unknown.

We therefore used several strategies to assess the impact on the sites test of heterogeneity in the substitution rate. First, we simulated data under simplified, controlled evolutionary conditions in order to directly test how heterogeneity in substitution rate affects the sites tests propensity to produce false positive inferences. Second, to understand how substitution rate hetereogeneity observed in real coding sequences affects the sites test, we analyzed a genome-wide empirical dataset using the sites test and a modified version of the test that in-

corporates such rate heterogeneity. Third, we simulated sequence evolution under conditions derived from this large empirical dataset to determine whether the degree of substitution rate heterogeneity present in real sequences is sufficient to produce frequent false inferences of adaptive evolution. Fourth, we analyzed the specific patterns in sequence data that can be produced by substitution rate heterogeneity that may lead to false positive inferences. Finally, we examined more recently introduced positive selection tests to determine the extent to which they can remedy any bias observed in the classical tests.

## 3.2 Results

### *3.2.1 Simulating codons with among-site rate heterogeneity.*

To understand the effects of substitution rate heterogeneity on the sites test, we first applied it to sequence data generated by simulation under simple evolutionary conditions with this kind of heterogeneity. We varied specific parameters of the generating model while holding the others constant, an approach that allowed us to make direct causal inferences about factors that affect the rate of false positive inference. We simulated codon evolution using a neutral model of evolution corresponding to the sites tests null hypothesis of drift and constraint, except that the sequence data were generated in two partitions with different overall substitution rates and then concatenated into a single alignment. In this partitioned model, a proportion of sites ($p$) were generated in partition A, with the remainder ($1-p$) in partition B. The underlying tree topology was identical for both partitions, but the branch lengths differed, yielding different overall substitution rates for the two sets of sites: in partition A, all branches have length $b_A$, while in partition B, every branch has length $b_B$ (see Figure 3.1a). The flexible dN/dS $\omega_0$ parameter of the null model was allowed to differ between partitions, but no sites were subject to positive selection ($\omega_A$ and $\omega_B$ were always assigned values $\leq 1$). In most but not all of our simulations, we set $\omega_A = \omega_B$, causing codons

in the two partitions to have different overall substitution rates but identical dN/dS ratios, as might occur if mutation rates differ between the partitions (affecting synonymous and nonsynonymous substitution rates equally), or if synonymous and nonsynymous substitutions are both subject to selective constraints that differ among codons. The other parameters of the model (the transition/transversion ratio and the equilibrium frequencies) were identical between partitions. Other than the heterogeneity in substitution rate between the two partitions, this generating model represents a specific instance of the sites tests null model. Although it is highly simplified, the model makes it possible to specifically test the effects of overall substitution rate heterogeneity *per se* on the sites tests accuracy under controlled conditions.

We analyzed the simulated sequences using the sites test as implemented in CODEML and HYPHY software, comparing the models M1a and M2a. As is customary, we used the likelihood ratio test and a $\tilde{\chi}^2$ distribution (df=2) to evaluate the statistical significance of inferences of positive selection, applying a significance cutoff of p-value $< 0.05$. Because the sequences were simulated with $\omega \leq 1$, every inference of positive selection by the sites test represents a false positive error. If the method is unbiased, this approach should yield positive inferences for no more than 5% of replicate datasets simulated under any set of evolutionary conditions.

### *3.2.2 Moderate substitution rate heterogeneity causes false inferences of positive selection.*

We began by simulating data on a simple four-taxon tree, with all codons evolving neutrally ($\omega_A = \omega_B = 1$), equally sized partitions ($p = 0.5$), and relatively short branch lengths in partition A ($b_A = 0.1$). We imposed substitution rate heterogeneity by varying the ratio $b_B/b_A$.

When heterogeneity was absent, false positive inferences were rare. However, even a

small degree of rate heterogeneity between partitions ($b_B = 0.15$, corresponding to a 1.5-fold difference in substitution rate between partitions) is sufficient to produce an unacceptably high rate of false positive errors (Figure 3.1b). False positive error rates increased monotonically as the degree of heterogeneity between partitions increased. A three-fold difference in rates between partitions yielded a false positive error rate of 60%, and a four-fold difference produced a 95% error rate. When substitution rate heterogeneity was incorporated into the model used for analysis using a partitioned version of the sites test which allows the branch lengths of the two partitions to be optimized independently, false positive errors occurred rarely and at an acceptable rate, and their frequency did not increase with greater heterogeneity (Figure 3.1b). This result indicates that the false positives we observed were specifically caused by heterogeneity in the overall substitution rate among sites and the inability of the sites test to incorporate it.

This bias is not unique to a specific implementation of the sites test. We observed nearly identical results when using either PAML or Hyphy software (Supplementary Figure S10a) and when using a different model comparison that models among-site variation in dN/dS ratio using a flexible continuous distribution (M7 vs. M8, Supplementary Figure S10b).

We also assessed the effects of substitution rate heterogeneity on three other methods for detecting positive selection by analyzing the simulated data described above. First, the branch-sites test uses the same underlying model as the sites test but seeks evidence of selection acting on a subset of codons on a subset of phylogeny: we found a high rate of false inferences of positive selection in the face of substitution rate heterogeneity, and this rate increased with the degree of heterogeneity between the two partitions. (Figure 3.1b). Second, the more recent method BUSTED tests for site-and-lineage-specific positive selection by allowing selection to vary stochastically over branches and sites. We found that BUSTED was subject to a false-positive bias induced by increasing substitution rate heterogeneity (Supplementary Figure S14). Finally, we assessed MEME, which uses a fixed-

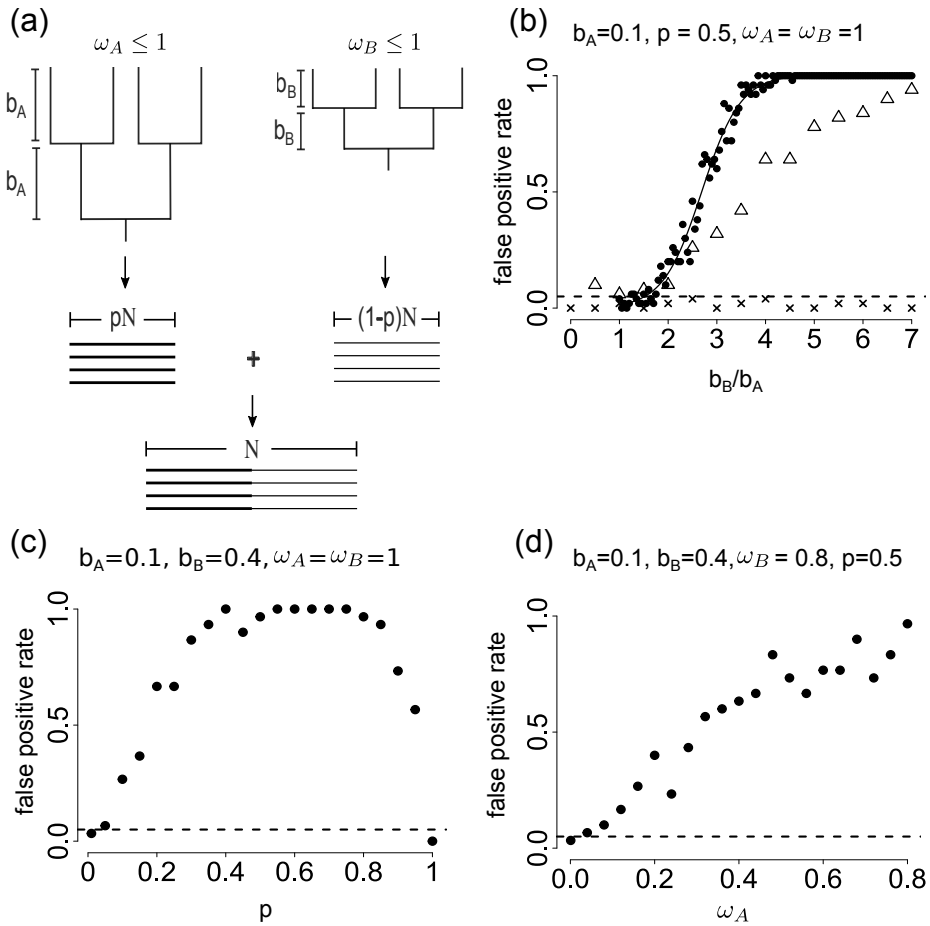



Figure 3.1: **Rate heterogeneity causes false inferences of positive selection.** a) Simulation scheme for generating sequences with rate heterogeneity. A proportion ($p$) of a total of N codons were generated using a codon model with non-synonymous/ synonymous rate A and a tree on which all branches have lengths $b_A$; the remaining $(1-p)$ sites were generated using the same model but with independent parameters B and $b_B$. All data were generated with both A and B $\leq 1$. b) Rate heterogeneity causes false positive inferences. For each generating condition, the proportion of 50 replicate alignments for which the sites test found a signature of positive selection (p $<$ 0.05) is shown (circles). Also shown are the proportion of positive inferences using the branch-sites test (triangles) and an a priori partitioned version of the sites test that incorporates rate heterogeneity (crosses). Dashed line, acceptable false positive rate of 5%. Solid line, best-fit logistic function to sites test results. c) Effect of the prevalence of rate heterogeneity on false positive inference rates by the sites test. Sequences were simulated under the conditions shown, and the rate of positive inferences by the sites test was plotted against the number of sites with the slow evolutionary rate ($p$ in the generating model). d) Rate heterogeneity causes false inferences of selection even under purifying selection. The sites test false positive rate is shown as a function of in the generating model, under the conditions shown.

effect likelihood mode over sites that allows the non-synonymous and synonymous rate is estimated separately for every site: this test identifies specific sites with evidence of positive selection rather than testing on a gene-wide basis. We found that, after a correction for multiple testing (FDR $< 0.05$), MEME displayed no evidence of bias, with no sites falsely inferred to be under positive selection, even as heterogeneity increased. Thus, whereas methods that hold dS invariant across sites are biased by substitution rate variation, a method that relaxed this assumption and allowed dS to independently vary did not display this bias.

To determine whether the effect of substitution rate heterogeneity on the false positive error rate depends on the branch lengths of the underlying tree, we imposed a constant degree of heterogeneity ($b_B/b_A = 2$) but this time varied the branch lengths of the two partitions. We found that longer branch lengths monotonically increase the rate of false positive errors (Supplementary Figure S11). Even at this moderate level of heterogeneity, for example, relatively short branch lengths ($b_A = 0.2$, $b_B = 0.4$) yielded a false positive error rate of approximately 40%.

We next asked how the prevalence of substitution rate heterogeneity  defined as the portion of sites in each partition – affects the accuracy of the sites test. We found that heterogeneity that affects only a relatively small number of sites can still produce a high rate of false positive inference. Specifically, when the two partitions differ in rate by a factor of 4, an elevated false positive rate is obtained when only 10% of sites are in the slower evolving partition (or in the faster evolving partition) (Supplementary Figure S10c), and the error rate is at a maximum of 1.0 across a broad range of intermediate values of $p$.

Neutral evolution with $\omega_0 = 1.0$ represents a borderline condition in which false inferences of positive selection might be particularly likely to occur. Real coding sequences typically evolve with many sites under purifying selection, with $\omega$ values much less than 1; such conditions might buffer the test against false positive inferences. We therefore investigated

how the sites test performs in the face of moderate rate heterogeneity ($b_A = 0.1$, $b_B = 0.2$, $p = 0.5$) when one partition evolves under strong purifying selection ($\omega_A$ variable but $< 1$) and the other evolves under weak purifying selection ($\omega_B = 0.8$). Even under these conditions, we observed an elevated false positive rate. For example, when $\omega_A = 0.2$, the false positive rate is twice the acceptable rate (Figure S10d). When $\omega_A = 0.4$, the error rate is close to 50%.

### *3.2.3 Systematic bias is not alleviated by better sequence sampling.*

The high rate of false positive inferences that we observed might be due to a systematic bias, or it might be limited to cases in which sequence data are sparsely sampled. We therefore investigated whether the susceptibility of the sites test to high error rates could be alleviated by increasing the amount of data analyzed.

First, we asked whether increasing the length of simulated sequences analyzing more sites and thereby reducing sampling error improved the false positive error rate. To the contrary, we found that increasing the sequence length monotonically increased the false positive rate (Figure S11a). For conditions that yield a ≈10% false positive rate when 1000 codons are analyzed, the false positive rate rises to almost 50% when 5000 codons are analyzed and approaches 100% as the dataset continues to grow larger. This finding indicates that the sites test suffers from a systematic bias towards false positive inferences, which grows worse as the sampling of sequence sites becomes more complete.

Second, we asked whether using denser taxon sampling to reduce branch lengths and improve phylogenetic signal could reduce the false positive inference rate. Previous studies have shown that the accuracy of phylogenetic inference *per se* the probability of inferring the correct tree topology – can be dramatically improved by sampling additional sequences to break up long branches on the tree [Hillis, 1996, Zwickl and Hillis, 2002]. We simulated sequences with weak heterogeneity and no positive selection as above, but increased the

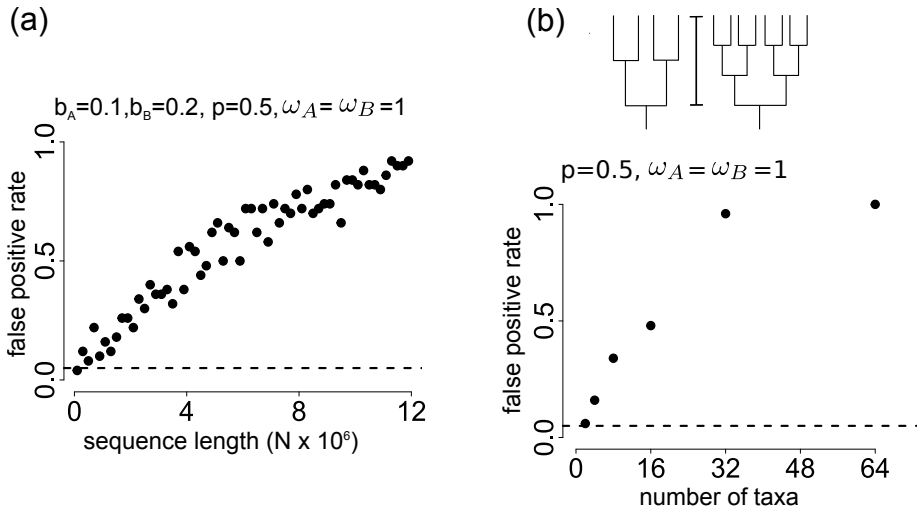



Figure 3.2: **Systematic bias towards false positive inferences caused by rate heterogeneity.** a) The sites test false positive inference rate is shown as a function of the number of codons in the alignment (N) when sequences are generated under the conditions shown. Dashed line, acceptable false positive rate (5%). b) The sites test false positive rate is shown as a function of the number of taxa in the tree; the total depth of the tree from root to tips was kept constant (0.2 and 0.4 in the two partitions) to mimic improved taxon sampling.

number of taxa sampled while keeping the total tree height constant (Figure 3.2b). Surprisingly, we observed that increasing the number of taxa monotonically increases the rate of false positive inferences (Figure 3.2b). For example, the false positive rate increases from about 10 percent when four sequences are sampled to 75 percent when 32 are sampled, and it approaches 100 percent when 64 sequences are analyzed.

Taken together, these experiments indicate that the sites test (and its variant the branch-sites test) is sensitive to substitution rate heterogeneity caused by dynamics other than positive selection or variation in the dN/dS ratio. Even weak or moderate heterogeneity is sufficient to produce unacceptably high rates of false positive inference. This bias is systematic and cannot be alleviated using the popular strategies of sampling more taxa or sequence sites.

### *3.2.4 Analysis of empirical sequence data.*

The simulation conditions we evaluated in the experiments described above are simple and unrealistic. We therefore investigated whether substitution rate heterogeneity affects inferences of positive selection in real data. We focused on a classic genome-wide dataset previously used to analyze the extent of positive selection in mammals a set of ≈16,500 aligned orthologous genes from human, chimp, macaque, mouse, rat, and dog [Kosiol et al., 2008]. The previously published analysis of these data used a slightly modified version of the sites test and found that 1596 genes had a statistically significant signature of positive selection (p-value $< 0.05$); of these, 400 remained significant after correcting for multiple testing (false discovery rate FDR $< 0.05$). These results led to a conclusion of widespread positive selection in mammalian genomes [Kosiol et al., 2008].

We sought to understand whether some inferences of positive selection in this large dataset might be due to the biasing effects of substitution rate heterogeneity. We therefore focused on the subset of 1596 genes previously found to contain signatures of positive selection. We first reanalyzed these data using the standard implementation of the sites test in CODEML and found that 397 genes yielded a signature of positive selection at p-value $<$ 0.05. (This smaller set reflects differences between the two implementations of the sites test in the specific ways that model parameters are optimized and p-values calculated.)

To determine the role of unincorporated substitution rate heterogeneity in producing signatures of positive selection in this dataset, we analyzed the same 1596 genes using a variant of the sites test that incorporates rate heterogeneity in both nonsynonymous and synonymous substitutions rates [Scheffler et al., 2006, Pond and Muse, 2005b]. This test (which we call sites test-RH) has the same general form as the traditional sites test, but the models include an additional flexible rate multiplier parameter $s$, which affects both synonymous and nonsynonymous rates and therefore serves as a scalar on branch lengths that can differ among submodels. As implemented, $s$ can take on three independent values,

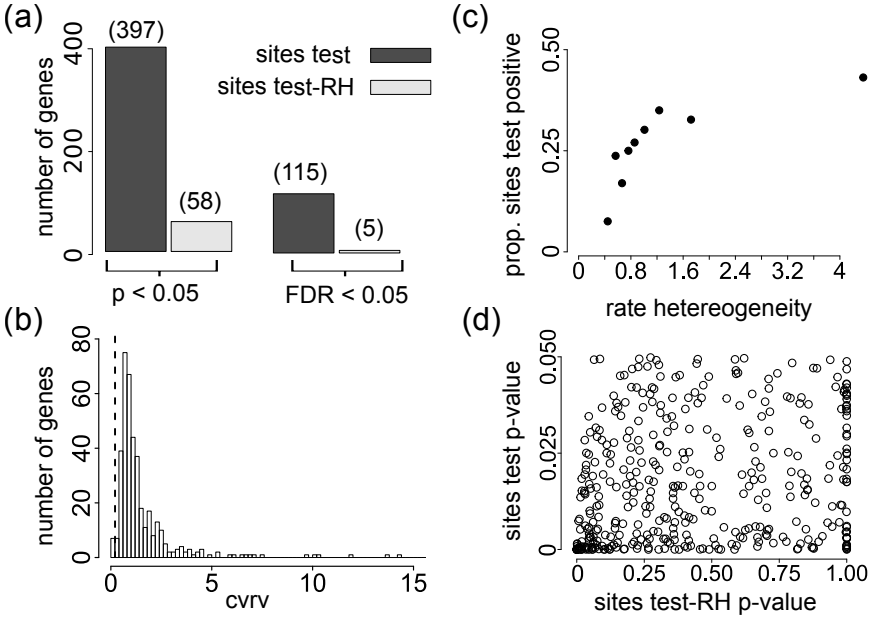



Figure 3.3: **Relationship of rate heterogeneity to positive selection signatures in empirical data.** a) The signal of positive selection in empirical data is weakened when rate heterogeneity is incorporated. 1596 mammalian genes previously found to have a signature of positive selection [Kosiol et al., 2008] were reanalyzed using the standard implementation of the sites test and an alternative version that incorporates synonymous rate variation (sites test-RH). The left portion of the graph shows the number of genes with a signature of positive selection ($p < 0.05$) in each test. The right portion shows the number of genes retaining this signature after adjusting for multiple testing (1596 tests) at a 5% false discovery rate is shown. b) Genes inferred to be under positive selection contain substantial rate variation. For each of the 397 mammalian genes inferred to be under positive selection by the sites test, we inferred maximum likelihood estimates of the degree of rate heterogeneity using the null model of sites test-RH. The distribution of the coefficient of rate variation (CVRV) across these genes is shown. Dotted line indicates CVRV=0.206, the degree of rate heterogeneity required under the conditions in Figure 3.1a to yield unacceptably high false positive rates in the sites test. c) Selection signatures are correlated with rate heterogeneity. For each gene in the set of 1596, the degree of rate heterogeneity was estimated as the coefficient of variation of the branch length rate multipliers in the mixture model of sites test-RH. Genes were binned by this metric, and the proportion of genes in each bin under positive selection according to the sites test ($p \leq 0.05$) was calculated. Spearmans correlation coefficient=0.97. D) Incorporating rate heterogeneity eliminates selection signals rather than reducing statistical power. For genes with a significant signature of selection in the sites test (p-value $< 0\ .05$), we plotted the sites-test p-value against the sites test-RH p-value. Pearson and Spearman correlation coefficients=0.21 and 0.26, respectively.

so the overall null model is a mixture over six submodels (three values of $s$ times two values of $\omega$, with $\omega$ always $< 1$) and the positive selection model is a mixture over nine submodels (three values of $s$ times three values of $\omega$, one of which may exceed 1). This test incorporates the type of rate heterogeneity we produced in our simplified simulations, but it does so via a mixture rather than partitioned model.

We found that taking into account substitution rate heterogeneity dramatically curtailed inferences of positive selection. Using sites test-RH, only 58 of the 1596 genes tested retained a signature of positive selection at p-value $< 0.05$. Thus, the set of genes inferred to be under positive selection by the sites test shrank from 397 to just 58 when rate heterogeneity was incorporated (Fig 3.3a), an 85% reduction. When a large number of genes are independently analyzed, some statistically significant results are expected by chance due to multiple testing. We therefore adjusted our analysis of the 1596 genes for multiple testing [Benjamini and Hochberg, 1995] at a false discovery rate $< 0.05$. When this correction was applied to the results of the sites test, 115 genes retained a significant signature of positive selection. When the results of the sites test-RH were corrected in the same way, only 5 genes retained statistically significant support for the positive selection model  a 96 percent reduction caused by incorporating rate heterogeneity (Figure 3.3a). Thus, out of hundreds of genes in mammalian genomes originally found to carry signatures of positive selection, just five retain a signal of positive selection when rate heterogeneity and multiple testing artifacts are taken into account.

### *3.2.5 Bias vs. loss of power.*

The near-complete disappearance of the signal of positive selection when the sites test-RH is used could have either of two explanations. First, consistent with our simulation studies, unincorporated substitution rate heterogeneity in the data might cause artifactual positive results in the sites test. Alternatively, the more highly parameterized models used in the

sites test-RH might result in a loss of statistical power, so that genes that in fact evolved under positive selection might fall below the threshold of statistical significance, leading to false negative inferences by the sites test-RH.

These two explanations each make specific testable predictions. If the inferences of positive selection are caused by heterogeneity-induced bias, then rate heterogeneity should be rampant. Further, statistical support for the positive selection model by each gene should be related to the degree of rate heterogeneity present in that gene. To test these predictions, we estimated the coefficient of rate variation (CVRV) for each of the 1596 genes, which expresses the normalized deviation of the three estimated weighted rate multipliers s from their mean.

Both predictions were corroborated. First, substitution rate heterogeneity was pervasive. The mean CVRV was 1.6, indicating very substantial variation among codons within each gene in their underlying substitution rates that cannot be captured by differences in the dN/dS ratio. For comparison, a far lower degree of heterogeneity CVRV=0.206 was sufficient to cause the sites test to produce an unacceptable false positive error rate under the simulation conditions we reported in Figure 3.1b ($b_B/b_A$=1.5). In fact, of the 397 mammalian genes with positive results in the sites test, 98 percent have CVRV $>$ 0.206, as do 94% of all 1596 genes analyzed (Figure 3.3b).

Second, rate heterogeneity was a good predictor that a gene would have a sites-test signature of positive selection. We binned genes based on their CVRV and, for each bin, calculated the proportion of genes inferred to be under selection by the sites test at p-value $<$ 0.05. We observed a nearly perfect monotonic relationship between the degree of rate heterogeneity and the propensity of genes to carry a signature of positive selection in the sites test (Spearmans correlation = 0.97 with p-value=0.000165, Figure 3.3c).

The second hypothesis – that the signal of positive selection in most genes disappears because the sites test-RH test has reduced statistical power predicts that there should be a relationship between the strength of support for positive selection in the sites test and

that in the sites test-RH, and genes with significant evidence of positive selection in the sites test should be marginally significant in the sites test-RH. We tested this prediction by plotting for each gene its p-value in the sites test-RH against its p-value in the sites test. As shown in Figure 3.3d, the sites test p-value does not accurately predict the sites test-RH p-value (Pearson correlation $r^2$=0.213). Further, of the genes that have a significant p-value in the sites test, the majority (77%) have p-values in the sites test-RH that are far from the threshold of significance.

Taken together, these results indicate that by incorporating rate heterogeneity, the sites test-RH eliminates an artifactual signature of positive selection in the vast majority of genes previously thought to be under positive selection. However, reduced power may also play an additional role.

### *3.2.6 Realistic substitution rate heterogeneity is sufficient to produce false inferences of positive selection.*

The analyses above do not directly assess whether the degree of rate heterogeneity in real data is sufficient to produce false positive inferences in the sites test. We therefore simulated sequences under conditions derived from real data to address this question.

For each of the 397 mammalian genes inferred to be under positive selection by the sites test, we used the sites test-RH to estimate (by maximum likelihood) all other parameters of the neutral null model, including the rate multipliers that represent rate heterogeneity, the branch lengths, and the values of $\omega$. We then simulated sequences in multiple replicates under the conditions derived from these genes and then analyzed these sequences using the traditional sites test. Only 84 of the 397 genes had complete coverage for all six taxa in the dataset, and branch lengths cannot be estimated from missing sequences, so we restricted our simulations to these genes. All sites in the simulations evolved without positive selection, so any positive inference is false. For the evolutionary conditions associated with each gene,

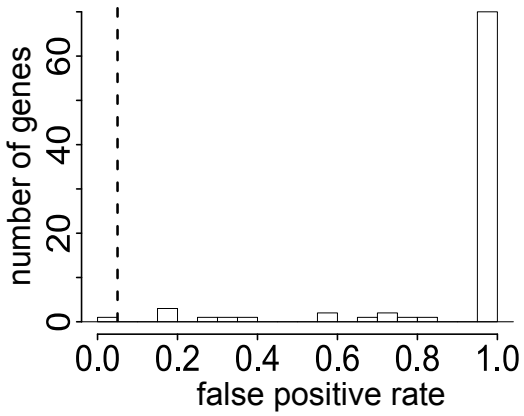



Figure 3.4: **Empirically-derived rate heterogeneity causes false positive inferences.** For each mammalian gene with a signature of selection and complete coverage, the parameters of the sites-test-RHs null model were inferred, and 25 replicate sequence alignments were then simulated under these conditions (without positive selection). The false positive rate for each gene is the fraction of replicate alignments that yielded a significant signature of selection ($p < 0.05$). The distribution of false positive inference rates across mammalian genes is shown. Dotted line, acceptable false positive inference rate.

we determined the frequency of false positive inferences as the fraction of replicate datasets simulated under those conditions that yielded a signature of positive selection ($p < 0.05$).

We found that the inferred evolutionary conditions for 83 of the 84 genes produced false positive inferences at an unacceptable rate (p-value $> 0.05$, Figure 3.4). More than 80% of the genes yielded an error rate of 1.0, and more than 90% of genes had an error rate $> 0.5$. The one gene that did not yield an unacceptably high error rate had an inferred coefficient of rate variation of zero, yielding sequences simulated with no heterogeneity.

This result establishes that for virtually every mammalian gene inferred by the sites test to have been under positive selection, the degree of substitution rate heterogeneity present in that gene is sufficient to have produced a false positive inference. Some of these genes could also have evolved under authentic positive selection, but, because of this bias, the sites test does not provide reliable evidence of this possibility. Taken together with our finding that virtually none of the genes with signatures of positive selection in the sites test maintain

this signature when rate heterogeneity is taken into account, this result indicates that a large proportion of genes inferred to be under positive selection using the sites test could be artifacts of the biasing effect of unincorporated rate heterogeneity.

### *3.2.7 Causes of bias.*

Why does substitution rate heterogeneity produce a false signal of positive selection in the sites test? The likelihood ratio test evaluates whether the difference in likelihood between the simpler and more complex models is greater than would be expected by fitting stochastic variation in the data if the simpler hypothesis were true. If neither model is correct, the null model will be rejected if additional flexible parameters in the positive selection model yield a significantly improved fit to patterns in the sequence data, even if those parameters are different from the ones that generated the data. We therefore hypothesized that the additional category of evolutionary rates in the positive selection mixture model allows it to artifactually improve the fit to data simulated under neutrality with substitution rate heterogeneity.

To evaluate this hypothesis, we first determined what kinds of codon character patterns and sequence substitutions provide false support for the positive selection model when rate heterogeneity is present. We also analyzed which submodels contribute the most to the difference in likelihood between the positive selection model (called M2a in CODEML) and the null model (called M1a). To facilitate this analysis, we generated sequences under very simple heterogeneous conditions that produce false inferences of positive selection; we used the simplest possible tree  a two-taxon tree, which, because it has only one branch, allows sequence differences to be unambiguously mapped onto the tree and categorized as synonymous or nonsynonymous. We calculated the site-specific log-likelihoods of the positive selection and null models and then categorized codon sites into classes according to the number and types of sequence differences between the two taxa on the tree: constant sites (the

same codon state in both taxa), one nonsynonymous difference, one synonymous difference, etc. For any category of sites, the support provided for the positive selection model is defined as the difference between the total log-likelihoods of M2a and M1a given the data at sites in that class.

One might expect that M2a would be supported primarily by codon patterns with multiple nonsynonymous changes. Surprisingly, however, we found that the majority of the support for the positive selection model comes from constant sites, in which the same codon is conserved in both sequences. Constant sites favor M2a by an average of 3.57 log-likelihood units per codon, a statistically significant difference in likelihood at each codon. Such sites make up more than half of the entire alignment, so together they yield overwhelming support for the positive selection model (Figure 3.5a). Codons with three differences between the two sequences also support M2a; these sites provide stronger support for M2a per site, but they are very rare, so their overall contribution to support for the positive selection model is smaller.

Why would constant sites, which manifest neither nonsynonymous nor synonymous substitutions, provide significantly greater support for a model that includes an elevated dN/dS parameter? Further, why would codons with multiple substitutions support this model so strongly, if those codons were produced by a process with an equal rate of nonsynoynmous and synonymous substitution? The answer is that substitution rate heterogeneity produces an excess of constant and multiple-hit codons, and the more flexible model M2a can better accommodate these patterns. Under the strictly neutral conditions we used for our simulations, the data partition with a slow substitution rate produces a very large number of constant codons, and the partition with a fast rate produces an excess of codons with three substitutions: when combined into a single dataset, the frequency of these two categories of pattern is greater than would be expected under a homogeneous model with one underlying rate, such as that used in the sites test (Figure 3.5b). When these data are analyzed

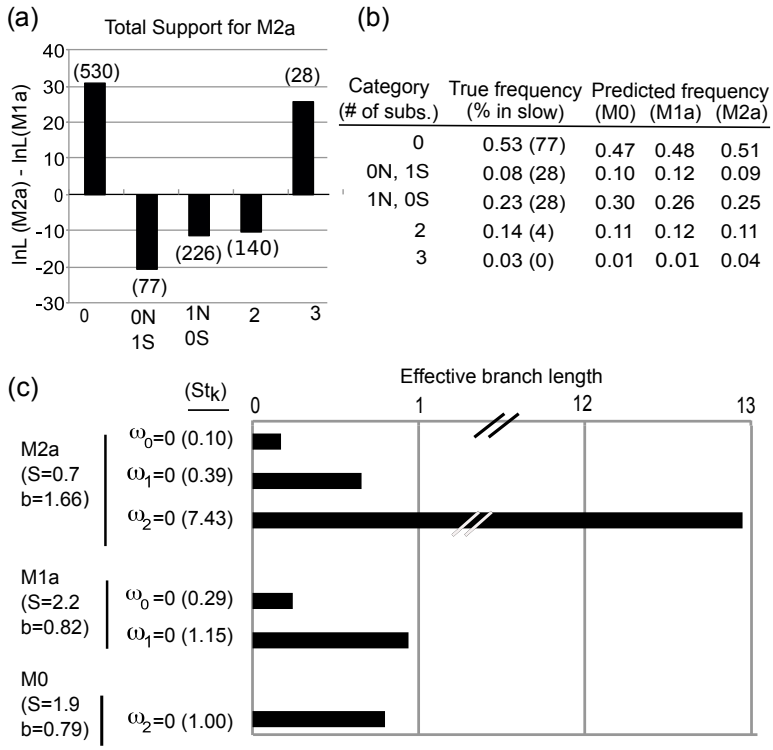



Figure 3.5: **Causes of spurious support for M2a by rate heterogeneity.** a) Support for the positive selection model vs. null model by category of codon state patterns. Data were generated on a two-taxon tree with no selection ($\omega$=1) in two partitions of equal size with branch lengths 0.2 and 1.6. Codon sites in the alignment were categorized by the number of nucleotide differences between taxa; for sites with 1 difference, codons were further as having a nonsynonymous (N) or synonymous (S) difference. The height of each bar represents the difference in log-likelihood between the positive selection model M2a and the null model M1a (using the ML parameter values for each), summed over all instances of sites in each category in the dataset. Parentheses indicate the number of codon sites of each category in an alignment of 1000 codons. b) Actual and predicted frequency of codon state patterns under various models. Data were generated under the conditions described for panel a. The first column shows the actual proportion of codons in each category in the concatenated dataset; parentheses show the percentage of codons in each category generated in the partition with the slower rate. The next three columns show the predicted proportion given the maximum-likelihood optimized parameters of models M0, M1a, and M2a. c) Model M2a mimics rate heterogeneity by having three very different effective branch lengths for its three submodels. Scaling factors and effective branch lengths for codon mixture models are shown, using ML parameter estimates given data generated as in panel a. For each submodel, the bar length shows the effective branch length  the expected number of substitutions given the scaling factor for the model ($S$, shown x1000), the ML estimate of the branch length for the model (b), and the total rate of substitution for the submodel ($t_k$). When $S_{t_k}$=1, the effective and given branch lengths are the same. Note the broken scale.

using the sites test, model M2a which allows three different nonsynonymous substitution rates can better fit the excess of these patterns than can the two-class model M1a or the one-category model M0.

When we examined the contribution of each of the submodels to the likelihood for each type of codon pattern, we noticed a surprising pattern (Supplementary Figure S12): for constant codons, the likelihoods of submodel 0 ($\omega < 1$) and of submodel 1 ($\omega = 1$) are much higher in M2a than they are in model M1a, despite the fact that the ML parameters of these two submodels are identical ($\omega = 0$ and $\omega = 1$, respectively). We hypothesized that this apparent paradox occurs because the presence of a third submodel in M2a affects the scaling factor applied to the rate matrices and branch lengths of submodels 0 and 1, changing the overall rate of evolution implied by those submodels. Instantaneous rate matrices are scaled to insure that the total number of substitutions predicted along a branch by a model equals the branch length used as an input, which is accomplished by multiplying each element of the matrix by a scaling factor $s$ and the branch length b, where $s$ is the reciprocal of the sum of all the instantaneous substitution rates among different codons within the matrix (weighted by the starting codons frequency) before scaling. Mixture models use a single scaling factor, which is calculated from the weighted average of the scaling factors that would have applied to each of the submodel matrices. This "average" scaling factor is then applied to all submodels rate matrices, which causes the total number of expected substitutions across the mixture model to be the same as the branch length, and the synonymous substitution rate remains the same for all submodels [Yang, 2014].

This procedure has two relevant consequences. First, because M2a contains a third submodel with an elevated $\omega_2$ , its overall scaling factor ends up being smaller than that of M1a; in turn, the "effective branch length" after scaling the expected number of substitutions given submodels 0 and 1 in M2a will be much smaller than in M1a, even if the corresponding submodels contain the exact same parameters. For example, under the conditions we

examined, the scaling factor for M2a is only 0.3 times that of M1a (Figure 3.5c), and the scaled effective branch lengths for the submodels 0 and 1 are much shorter. The probability of a codon remaining identical is higher when branches are short than when they are long, so the likelihood of the positive selection model is higher given a constant codon, even given the same values of $\omega_0$ and $\omega_1$.

The second consequence is that, because the submodels are scaled by the same "average" factor, they all imply different total substitution rates; M2as three submodels have more divergent substitution rates and thus can fit much more heterogeneity than M1a can. Under the conditions we examined, the estimated effective branch lengths for M2as class 2 and class 0 differ by a factor of $> 70$; thus, submodel 2 has a high likelihood given codons with multiple differences, and submodel 0 has a high likelihood given constant codons. Model M1a incorporates much less heterogeneity, with a $< 4$-fold difference in effective branch lengths between its submodels (Figure 3.5c). Because M2as mixture model incorporates submodels with far more divergent effective branch lengths than M1as does, it can better fit the large number of codons with no substitutions or many substitutions that are produced by rate heterogeneity without positive selection.

## 3.3 Discussion

The analyses we report here indicate that substitution rate hetereogeneity causes a strong, systematic bias in the sites test towards false inferences of positive selection. The initial simulations we conducted involved very simple conditions; this approach allowed us to vary specific parameters and demonstrate a causal relationship between substitution rate heterogeneity and elevated false positive error rates. These experiments establish that even a moderate amount of heterogeneity causes very high rates of false positive error, and that this bias cannot be alleviated with better sampling of taxa or codons. Our results are consistent with previous findings that unincorporated substitution rate heterogeneity, such as

that produced by variability in synonymous rate among codons, can bias estimates of the dN/dS parameter $\omega$ [Spielman and Wilke, 2015] and increase support for the a posteriori classification of codons as under positive selection [Pond and Muse, 2005b]. These prior studies, however, did not address the effects of rate heterogeneity on the performance of the likelihood ratio test of positive selection itself.

We examined in a preliminary but not systematic fashion whether a similar bias affects other ML-based positive selection tests. Our limited analysis suggests that two methods that do not allow for substitution rate heterogeneity are also biased towards false positive inferences as heterogeneity increases. MEME, in contrast, which does allow dS to vary over sites, did not display this bias. Further, our analyses of empirical data using the sites+RH test ameliorates some or all of the heterogeneity-induced bias. These data suggest that methods that allow dS to vary, in either a fixed or random effects likelihood framework, represent a promising approach to incorporating substitution-rate heterogeneity and reducing the rate of false positive inferences.

Our simulations under very simplified conditions say little about the propensity of real-world conditions to bias the sites test, but our analysis of a genome-wide empirical dataset does. We analyzed a very large set of aligned genes that previously was used to support the conclusion that positive selection is widespread in mammalian genomes. Our analyses of these data establish that bias in the sites test caused by substitution rate heterogeneity is a matter of concern in real-world settings. The sites test finds that hundreds of mammalian genes were subject to positive selection, but incorporating heterogeneity eliminates this signal for virtually all of these genes. And a loss of power by the more complex model does not appear to be the major cause, suggesting that a meaningful proportion of positive selection inferences in the mammalian genome made using the sites test may be artifacts of unincorporated substitution rate heterogeneity.

Using simulations under empirically derived conditions but without any positive selection,

we were able to determine how often the sites test produces false positive inferences when confronted with realistic levels of substitution rate heterogeneity. We found that the degree of heterogeneity present in virtually every mammalian gene with a significant signature of selection in the sites test is sufficient to yield a very high rate of positive inferences. It is therefore plausible that a large number of the sites tests positive inferences in this dataset were artifacts produced by use of an overly simple model.

Are our findings from the mammalian dataset generalizable? It is possible that mammalian genes harbor more substitution rate heterogeneity than genes in other taxa, but we know of no reason to expect this to be the case. A recent study, for example, found widespread variation in synonymous substitution rates among codons in over 7,000 groups of homologous genes from a wide variety of taxa [Dimitrieva and Anisimova, 2014]. The mammalian dataset we analyzed comprises a small number of taxa, but our experiments show that more complete taxon sampling tends to increase false inferences of positive selection due to substitution rate heterogeneity, not ameliorate it. We must therefore consider the possibility that many – or perhaps even most  published reports of positive selection based on the sites test could be artifacts of unincorporated rate heterogeneity.

We do not claim that positive selection is unimportant  it obviously is a crucial evolutionary process  or that every gene for which the sites test has obtained a signature of positive selection evolved without positive selection. Rather, our work merely indicates that the sites and branch-sites tests provide no reliable evidence of an adaptive history – not even suggestive or *prima facie* evidence. There is no way at present to distinguish between authentic and artifactual positive results in the sites test, so inferences of positive selection by the sites test may be correct in some cases. For example, there are cases of ongoing adaptive evolution involving host-parasite and intracellular genetic conflicts that produce sequence signatures of positive selection that are very likely to be authentic [Barber and Elde, 2014, Daugherty and Malik, 2012]. Indeed, viral genes likely to be subject to ongoing Red Queen

positive selection by host immune systems maintain a signature of positive selection when rate heterogeneity is incorporated into the sites test [Pond and Muse, 2005b]. The persuasive evidence for adaptive evolution in these cases, however, comes from sources other than the sites and related tests.

If the sites test is unreliable in the face of widespread evolutionary heterogeneity, what should researchers do? We see several options, all of which have advantages and disadvantages. First, sites test-RH, MEME, and similar approaches [Kosakovsky Pond and Frost, 2005] can be used to accommodate substitution rate heterogeneity into likelihood ratio tests of positive selection. This strategy would address the specific form of heterogeneity on which we focus in this study and seem likely to reduce the rate of false positive inferences. However, there are numerous other forms of evolutionary complexity that are not incorporated into these tests, such as heterotachy (different substitutions rates among both sites and lineages [Lopez et al., 2002]), among-site differences in amino acid preferences [Bloom, 2014], differences in the exchange rate among pairs of amino acids [Miyazawa and Jernigan, 1993], biases in the mutational processes that cause mutations within a codon [Schrider et al., 2011], and variation in the dN/dS in forms other than that incorporated into the sites tests models. Failure to incorporate these or other forms of complexity might, in principle, continue to bias the sites test-RH even after substitution rate heterogeneity among sites is taken into account. Future studies are therefore necessary to assess the robustness of tests of positive selection to these and other forms of model violation and, if necessary, to develop improved techniques.

A second alternative is to use non-parametric "counting" methods to estimate dN/dS [Nei and Gojobori, 1986, Suzuki and Gojobori, 1999]. These methods are not biased by among-site rate heterogeneity (Supplementary Figure S13), but they do not provide a framework for statistical hypothesis testing, and they have dramatically reduced power to detect positive selection [Yang and Bielawski, 2000, Wong et al., 2004]. For instance, the test of Nei &

Gojobori requires dN to exceed dS across the entire gene, rather than at only a subset of sites, so it is likely to suffer from a high rate of false negative inferences. Likelihood-based counting methods have also been developed to identify specific sites with an excess of nonsynonymous substitutions, and some of these methods allow hypothesis testing at specific sites [Kosakovsky Pond and Frost, 2005].

Finally, researchers can use functional experiments to explicitly test hypotheses of adaptive evolution. For example, they might experimentally determine whether historical changes in protein sequence actually produce the predicted changes in protein function or phenotype, and they might measure effects on fitness in evolutionarily relevant environments (e.g., [Barber and Elde, 2014, Yokoyama et al., 2008, Storz et al., 2009, Chen et al., 2009]. This approach requires time-consuming and challenging laboratory and field work [Barrett and Hoekstra, 2011] and cannot be implemented on a genome-wide scale. Ideally, future work will develop more robust models to detect positive selection, and experimental analyses of biological cause and effect can be used to test and enrich the inferences made using these approaches.

## 3.4 Methods

### *3.4.1 Controlled simulations*

We simulated codon sequence evolution using the software INDELible [Fletcher and Yang, 2009] and the basic GY codon substitution model specified by [Nielsen and Yang, 1998], a continuous time Markov chain whose state space consists of all 61 non-stop codons and is

defined by the rate matrix

$$Q_{i,j} = \begin{cases} 0, & \text{if two codons differ at more than one position,} \\ \pi_j & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases} \tag{3.1}$$

where $Q_{i,j}$ specifies the relative instantaneous substitution rates from codon state $i$ to state $j$, $\pi_j$ is the equilibrium frequency of codon $j$, $\kappa$ is the transition-transversion ratio, and $\omega$ is the ratio of the nonsynonymous to synonymous substitution rates. We fixed $\kappa = 1$ and the frequencies of each codon state to 1/61. Sequences of length $N$ codons were simulated in two partitions, with partition A containing $pN$ codons and partition B containing $(1-p)N$ codons. Both partitions evolved on a symmetric tree in which all branches have length $b$, but $b_A$ and $b_B$ are independent. For most conditions, $\omega_0$ for both partitions was set at 1 to represent neutral evolution at all sites. Simulated sequences from the two partitions were then concatenated and analyzed as a single dataset. For each set of conditions, 50 replicate alignments were generated.

### *3.4.2 Analysis*

All simulations were conducted under conditions with $\omega \leq 1$, so any inference of positive selection is false. We imposed a statistical significance cutoff of p-value $\leq 0.05$; an unbiased test should produce false positive inferences at a frequency of $\leq 0.05$.

To determine the degree of rate heterogeneity ($b_B/b_A$) at which the false positive inference rates exceeds the acceptable false positive rate of 0.05, we fitted a logistic function to the data in Figure 3.1b using the nls R package, and analytically solved for the value of the indepen-

dent variable $b_B$ at which the function equals 0.05. The coefficient of rate variation (CVRV) for this degree of heterogeneity is 0.206, calculated as CVRV=$\sqrt{p_a(b_a-\mu)^2+p_b(b_B-\mu)^2}/\mu$, where $\mu$ equals the mean of the two-class discrete branch length distribution ($\omega$=1, $b_B$=0.15191, $b_A$=0.1, $p_A$=$p_B$=0.5).

### *3.4.3 Sites test and sites test-RH*

The standard sites test compares the likelihoods of the null model M1a with the alternate model M2a using a likelihood ratio test [Yang and Bielawski, 2000]. M1a is a mixture model with two submodels, which represent different evolutionary dynamics. Submodel 0 represents purifying selection with $\omega \leq 1$; submodel 1 represents neutral evolution with $\omega_1 = 1$. M2a is a model of positive selection, which extends M1a by incorporating an additional submodel with $\omega > 1$. All free parameters, including the weight applied to each submodel, are estimated by maximum likelihood, and the total likelihood of a model is the sum of the weighted likelihoods of its submodels. The likelihood ratio test compares the likelihood-ratio statistic (LRS twice the log of the likelihood ratio) to a $\tilde{\chi}^2$ distribution with two degrees of freedom, which describes the expected distribution of LRS when the null model is true. We used the implementations of the Sites Test in the CODEML/PAML software [Yang, 2007] and Hyphy software packages [Pond and Muse, 2005b]. Analyses were conducted using CODEML unless noted otherwise. For each set of simulation conditions, the rate of false positive inference is the fraction of replicates yielding an inference of selection at p-value $\leq 0.05$.

For the partitioned version of the sites test, we separately analyzed each partition of sites (with accurate partitioning known a priori), optimized the parameters of model M1a and M2a for that partition, summed the log-likelihoods across the two partitions for each model, and then compared the total log-likelihoods of M2a and M1a as described above (df=4). Sites test-RH is described in Scheffler et al. [2006]. Briefly, the test is constructed as is the traditional sites test, but the submodels of both M1a and M2a incorporate an additional

variable parameter $s$, which scales all branch lengths of the phylogeny to represent variation in substitution rate (equivalent to varying the underlying synonymous substitution rate). Each submodel is represent by a rate matrix with the following form

$$Q_{i,j} = \begin{cases} 0, & \text{if two codons differ at more than one position,} \\ \pi_j s & \text{for synonymous transversion,} \\ \kappa\pi_j s, & \text{for synonymous transition,} \\ \omega\pi_j s, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j s, & \text{for nonsynonymous transition.} \end{cases} \tag{3.2}$$

The value of $s$ is drawn from a discrete distribution with three categories, scaled so the weighted mean $s = 1$. Sites test-RHs mixture model M1a consists of six submodels (three values of s times two values of $\omega$), and its M2a consists of nine submodels (three values of $s$ times three of $\omega$ ). Sites test-RH is implemented in the Hyphy batch language (PARRIS.bf), packaged with Hyphy [Pond and Muse, 2005b]. In our analysis, we used the following options: F3x4 estimated, HKY85, Dual Variable Rates Model, non-synonymous rate omega is multiplied by synonymous rate, 1 synonymous rate per codon, and PARRIS model comparison.

Sites test-RH as implemented in Hyphy calculates the coefficient of rate variation for a gene as CVRV=$\sqrt{\sum_k p_k(s_k - \bar{s})^2}/s$, where $s_k$ is the synonymous rate multiplier for submodel $k$, $p_k$ is the probability or weight on submodel $k$, and $\bar{s}$ is the weighted mean values of $s$ across submodels.

### *3.4.4 Options used for other tests of positive selection*

We simulated the data in the same way as Figure 3.1b with varying degrees of rate heterogeneity. We tested for positive selection using the branch-sites test, BUSTED, and MEME.

For the branch-sites, we arbitrarily chose one of the four branches to test for positive selection (because of symmetry the choice of branch does not matter). For BUSTED, we used the option to test for selection on the entire phylogeny and default options. Both the branch-sites test and BUSTED is a gene-level test of positive selection. However, MEME tests individual sites for positive selection. We computed two measures of false positive rate for MEME: (1) a site-level false positive rate by counting the number of sites with p-value $< 0.05$ which should total to less than $0.05*996 \approx 50$ sites if the test is well-behaved (2) a gene-wide false positive rate by designating a gene to be under positive selection if a site is found to have qvalue $< 0.05$. For MEME, we used default options and estimated dn/ds w/o branch correction.

### *3.4.5 Analysis of empirical mammalian dataset*

For analysis of empirical data, we used aligned coding sequences of orthologous genes from 6 mammalian genomes [Kosiol et al., 2008]. We extracted the 1596 genes previously found to be under positive selection using a version of the sites in which the models are the same as the traditional sites test, but the methods used to optimize parameters and compare likelihoods are slightly different [Kosiol et al., 2008]. We analyzed the alignment for each gene using the sites test and the sites-test RH as described above. To reduce convergence artifacts in parameter estimation, we inferred model parameters for each gene under the null model of sites test-RH in 25 independent runs and chose the set of conditions with the highest likelihood. The reported coefficient of rate variation under the null model was recorded, as was the p-value for rejecting the null model using the sites test-RH. In analyzing the correlation of p-value with the coefficient of rate variation, we excluded 68 genes (4% of total genes) with no inferred rate variation (cvrv=0), because we found that such estimates occur as a convergence artifact, as reported also in other studies [Dimitrieva and Anisimova, 2014].

For simulations under empirically-derived conditions, we used sites test-RH as described above to identify the maximum likelihood estimates of all parameters under the null model for each of the 397 mammalian genes with statistically significant support for positive selection in the sites test. Of these, 84 contained complete sequence for all six taxa. We used Hyphy to simulate sequence evolution in 25 replicates under the maximum likelihood parameters of the sites test-RHs null model for each of these genes. We then analyzed these sequences for a signature of positive selection using the traditional sites test (p-value $< 0.05$). Sequences were generated without positive selection, so all positive inferences are false. The rate of false positive inference for each gene is the fraction of replicates generated under the conditions inferred from that gene that yield a significant inference of selection.

### *3.4.6 Causes of bias*

To analyze the causes of support for the positive selection model, we used Indelible to simulate sequences of length 106 codons using a partitioned model consisting of two taxa, a single $\omega = 1$, $\kappa = 1$, equal codon frequencies, and branch lengths of 0.2 and 1.6 in the two partitions [Fletcher and Yang, 2009]. We then analyzed these data using Hyphy to identify the ML parameter estimates for M2a, M1a, and the non-mixture model M0. We then calculated the exact log-likelihoods given these parameters for every possible codon state pattern, which we define as every possible pair of codons in the two taxa (of which there are 61x61); this was accomplished by preparing a pseudoalignment consisting of one instance each codon state patterns and analyzing this pseudoalignment in Hyphy with parameter values fixed to the ML values as calculated above. Support for M2a vs. M1a by any single instance of a codon state pattern was calculated as the difference in site-specific log-likelihood ($LLD_i$) for the two models given the data at one site containing that pattern, or

$$LLD_i = lnP(D_i|\text{M2a}, \theta_{\text{M2a}}) - lnP(D_i|\text{M1a}, \theta_{\text{M1a}}), \tag{3.3}$$

where $\theta_M$ is the set of maximum likelihood estimates of the model parameters and branch lengths for model $M$, given the entire alignment. For a class of sites in an alignment for example, the set of constant sites, which have the same codon state in the two taxa the total support for M2a is the sum over all sites in that class of the site-specific log-likelihood differences, or

$$LLD = \sum_i lnP(D_i|\text{M2a}, \theta_{\text{M2a}}) - lnP(D_i|\text{M1a}, \theta_{\text{M1a}}). \tag{3.4}$$

The expected number of sites in a group was calculated as the sum of the expected frequency of each codon state pattern in the group given the generating conditions, times the length of the sequence. The expected frequency of each codon state pattern was calculated using Hyphy by fixing parameter values to the true generating conditions, calculating the site-specific likelihoods of each codon state pattern, and summing those likelihoods over partitions.

The scale factor $s$ for any simple model is defined as $s = 1/t$, where $t$ is the total unscaled instantaneous substitution rate, calculated as the sum of all the off-diagonal elements of the rate matrix, each weighted by the frequency of the starting codon, or $t = \frac{1}{\sum_i \pi_i \sum_{j \neq i} q_{ij}}$ For equal codon frequencies and $\kappa = 1$, $t = \frac{60}{61(392+134)}$. For a mixture model, the overall scaling factor $S$ is computed by determining $t$ for each submodel $k$ and then taking the inverse of the weighted sum over submodels: $S = \frac{1}{\sum_k p_k t_k}$ Likelihoods are then calculated by multiplying each element of the rate matrix by the scaling factor times the branch length $b$ and then exponentiating the resulting matrix. The effective branch length for any submodel is defined as the expected number of substitutions given the rate matrix, the given branch length $b$, and the scaling factor applied, and are calculated as $S_{t_k} b$. For a simple model and the scaling factor derived from it, $st = 1$, so the effective branch length is $b$. For a submodel of a mixture with a scaling factor calculated as described above, $S_{t_k} \neq 1$, so the effective branch length $\neq b$.

# CHAPTER 4

# INFERENCE AND VISUALIZATION OF DNA DAMAGE PATTERNS USING A GRADE OF MEMBERSHIP MODEL.

*(with Kushal Dey, John Novembre, and Matthew Stephens)*

## 4.1 Introduction

Ancient DNA (aDNA) research has seen rapid growth with the recent advancements in recovery of short DNA fragments, increased throughput, and lower per-base cost in sequencing [Shapiro and Hofreiter, 2014]. Both the number and size of aDNA datasets have grown rapidly over the last five years, and several recent studies sequenced hundreds of ancient individuals [Mathieson et al., 2015, Allentoft et al., 2015, Mathieson et al., 2017, Olalde et al., 2017, Lazaridis et al., 2016, Lipson et al., 2017].

This rapid recent growth in aDNA research has provided novel insights into human history. However, working with aDNA remains challenging. For example, ancient samples often contain very little endogenous DNA because in many environments DNA degrades rapidly post-mortem [Sawyer et al., 2012]. Furthermore, ancient samples are often contaminated by microbes and exogenous human DNA [Malmström et al., 2007]. Both these factors mean that many sequence reads generated by an aDNA study may not actually come from the ancient sample.

Because of these challenges, aDNA researchers pay careful attention to quality control (QC), including checking sequencing reads from each sample for signatures of endogenous aDNA. These signatures include: short fragment length, an enrichment of purines before strand breaks, and a high frequency of cytosine to thymine substitutions (C-to-T) at the ends of fragments [Sawyer et al., 2012, Ginolhac et al., 2011, Jónsson et al., 2013, Briggs

et al., 2007, Skoglund et al., 2014b]. One common QC procedure is to plot, for each sample, the C-to-T mismatch rate as a function of position from the end of the read, and to look for an elevated rate near the ends of reads as an indication of the presence of endogenous aDNA. Another common procedure is to look for elevated rates of purines (A & G) before the 5' strand-breaks. Both these procedures are implemented in the software *mapdamage* [Ginolhac et al., 2011, Jónsson et al., 2013], for example.

These commonly-used QC checks, though simple and useful, have several limitations. For example, they produce a plot for each sample, which can be inconvenient to work with and difficult to compare across many samples. This issue becomes increasingly important with the growing size of aDNA datasets. These plots can also be difficult to interpret, in part because they do not contrast observed patterns with expected patterns in modern samples. Finally, these approaches can detect only pre-defined damage signatures, and may fail to capture other key features or artifacts in the data.

Here we introduce methods to help address these problems. These methods start with a Binary Alignment Map (BAM) file, obtained by aligning each read to a reference genome. The BAM file includes information on the mismatches that occur in each read (vs the reference). We characterize each mismatch by several relevant features, including its type (e.g. C-to-G, etc), flanking bases, and distance from the end of the read. We then use these features to cluster the mismatches into groups, which we call *mismatch profiles*. Intuitively, a mismatch profile associated with post-mortem damage is expected to show high levels of C-to-T mismatches at the ends of reads. On the other hand, a mismatch profile that is typical of modern DNA polymorphism will show a different pattern, such as a transition to transversion ratio of 2:1 [Goldman and Yang, 1994]. Finally we estimate the relative frequency of each mismatch profile in each sample, which we refer to as the "Grade of membership" [Erosheva, 2006] of that sample in that mismatch profile. These grades of membership should reflect which processes generated mismatches in each sample. For example ancient samples

should have a high grade of membership in mismatch profiles characteristic of post-mortem DNA damage. Grade of membership models are widely used to infer structure in admixed populations [Pritchard et al., 2000], document collections [Blei et al., 2003], RNA-seq data Dey et al. [2017a] and somatic mutation data [Shiraishi et al., 2015] for example.

We have implemented methods to fit this model, and visualize the results in a software package, `aRchaic`. For example, the grades of membership for all samples are succinctly displayed in a single STRUCTURE plot [Rosenberg et al., 2002], and each mismatch profile is displayed using simple intuitive plots [Dey et al., 2017b]. Together these plots provide a concise visual summary of DNA damage patterns, as well as other processes generating mismatches in the data.

## 4.2 Methods

For each sample $i = 1, \ldots, I$ we first obtain a BAM file. From this BAM file we extract information on the mismatches (vs a reference) that occur in reads from the sample. First we filter out low-quality reads (mapping $\leq 30$), low-quality mismatches (base quality $\leq 20$), and mismatches that occur more than 20bp from the end of a read (since these are unlikely to reflect damage patterns [Briggs et al., 2007]). When a read carries more than one mismatch we treat these as independent (an assumption we verified by checking that the probability of a mismatch conditional on the occurrence of another mismatch on the same read is not significantly different from the marginal probability, p-value = 0.43).

Let $J_i$ denote the total number of remaining mismatches. For each mismatch $j = 1, \cdots, J_i$, we first identify the strand in the reference genome that carries a C or T allele; we call this strand the "reference strand" and denote it by $S_j$. Let $x_{i,j} = (x_{i,j,1}, x_{i,j,2}, x_{i,j,3}, x_{i,j,4}, x_{i,j,5})$ denote the following features of the mismatch (see Supplementary Figure S16 for illustration):

1. $x_{i,j,1} \in \{\text{T-to-A}, \text{T-to-C}, \text{T-to-G}, \text{C-to-A}, \text{C-to-G}, \text{C-to-T}\}$ denotes the mismatch (with respect to the reference strand $S_j$).

2. $x_{i,j,2} \in \{\text{A}, \text{C}, \text{G}, \text{T}\}$ denotes the nucleotide immediately 5' to the mismatch on $S_j$.

3. $x_{i,j,3} \in \{\text{A}, \text{C}, \text{G}, \text{T}\}$ denotes the nucleotide immediately 3' to the mismatch on $S_j$.

4. $x_{i,j,4} \in \{0, ...20\}$ denotes the distance (in base-pairs) from the mismatch to the nearest end of the read.

5. $x_{i,j,5} \in \{\text{A}, \text{C}, \text{G}, \text{T}\}$ denotes the nucleotide that occurs one base upstream from the 5' end of the reference strand that is closest to the location of mismatch. Since mismatches resulting from damage primarily occur at 5' ends of the reads, this features captures the enrichment of purines immediately one base 5' upstream of the strand-break for samples with sufficient DNA damage [Sawyer et al., 2012, Briggs et al., 2007].

These features are designed to reflect the major modes of DNA damage [Briggs et al., 2007, Sawyer et al., 2012, Prüfer et al., 2014, Sawyer et al., 2012]). For each feature $l \in \{1, \ldots, 5\}$ we let $M_l$ denote the number of possible values of $x_{i,j,l}$, and for notational convenience we relabel the possible outcomes such that $x_{i,j,l}$ takes on integer values in $\{1, \ldots, M_l\}$. For example we represent $x_{i,j,1} = \text{T-to-A}$ by $x_{i,j,1} = 1$.

Our model assumes that each mismatch in each individual arose from one of $K$ mismatch profiles ("clusters"). We introduce latent variables $z_{i,j} \in \{1, .., K\}$ to denote the profile that gave rise to mismatch $j$ in individual $i$. We assume

$$\Pr(z_{i,j} = k) = q_{i,k}, \tag{4.1}$$

where $q_{i,k}$ represents the membership proportion of individual $i$ in mismatch profile (cluster) $k \in 1, .., K$.

We further assume that, given $z_{i,j} = k$,

$$\Pr(x_{i,j,l} = m | z_{i,j} = k) = f_{k,l}(m), \tag{4.2}$$

where $m \in \{1, \ldots, M_l\}$, and $f_{k,l}(m)$ denotes the relative frequency of $m$ at feature $l$ in cluster $k$. We follow [Shiraishi et al., 2015] in assuming independence among features within each cluster.

Putting this all together, and assuming independence of observations yields the likelihood:

$$L(q, f; x) = \prod_{i,j,l} \sum_{k} f_{k,l}(x_{i,j,l}) q_{i,k}. \tag{4.3}$$

We fit this model, and estimate the individual parameters ($q$) and cluster parameters ($f$) by maximum likelihood using an accelerated EM algorithm. We use the same EM updates as in equations 2-4 in [Shiraishi et al., 2015], and we add first-order quasi-Newton acceleration to improve convergence [Lange, 1995, Alexander et al., 2009, Taddy, 2012].

For each cluster $k$, we visualize the cluster parameters $f_k$ using an *EDLogo* plot [Dey et al., 2017b], see Figure 4.1. The *EDLogo* plot allows one to visualize both enrichment and depletion of mismatch features scaled against a reference frequency. In our application, the reference frequency was computed by averaging the proportion of the five `aRchaic` features across individuals in the 1000 Genomes Project. This allows us to to compute an enrichment score which effectively compares mismatch profiles in our samples against that of modern individuals from Consortium et al. [2012]. We use a STRUCTURE plot [Rosenberg et al., 2002] to visualize the estimates of $q_{i,k}$ for each sample.

## 4.3 Results

We demonstrate the utility of `aRchaic` using three case-studies.

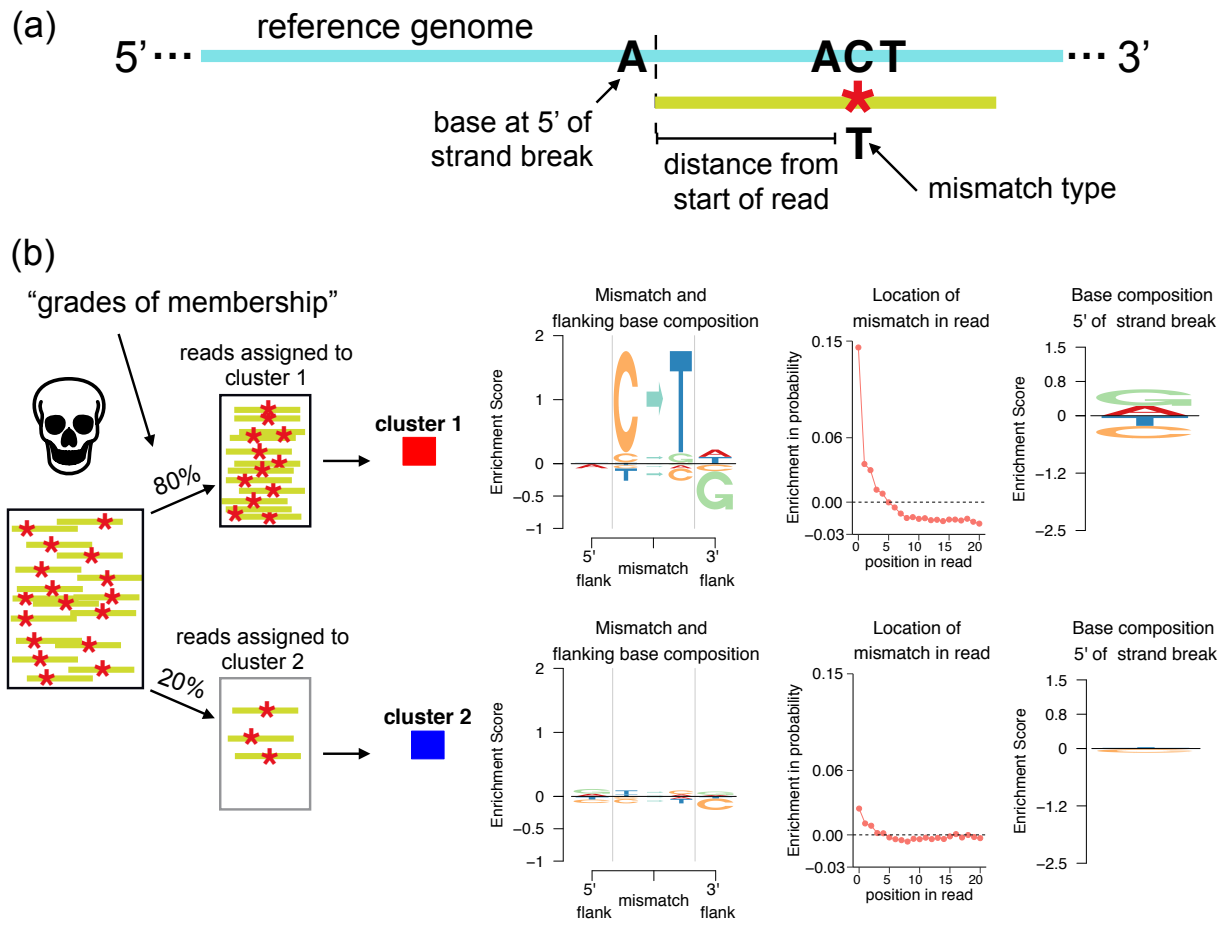



Figure 4.1: **Illustration of the `aRchaic` grades of membership and mismatch profiles.** (a) The features of a mismatch modeled by `aRchaic` (b) A depiction of an ancient DNA sample that has 80% of its reads assigned to cluster 1 and 20% of its reads assigned to cluster 2. Each cluster is defined by a *mismatch profile* showing the enrichment of the mismatch type, bases flanking the mismatch, the distance of the mismatch from the nearest end of the read, and the base immediately 5' to the strand-break. To produce a mismatch profile for a cluster, mismatch features are aggregated across reads assigned to the cluster, and their frequencies are represented by an *EDLogo* plot [Dey et al., 2017b]. In the *EDLogo* plot, the frequencies are scaled against a background frequency computed from Consortium et al. [2012].

### 4.3.1 `aRchaic` clustering of modern and ancient individuals

We applied `aRchaic` to a combined dataset of 52 ancient samples from four recent studies Skoglund et al. [2014a], Gamba et al. [2014], Lazaridis et al. [2016], Lipson et al. [2017] and 60 modern samples from Consortium et al. [2012] (n=50) and 10 individuals from HGDP [Cann et al., 2002] individuals sequenced by Meyer et al. [2012]. Two of the aDNA studies used partial-UDG treated libraries, which removes most – but not all – of the C-to-T deamination [Rohland et al., 2015].

Figure 4.2 shows results from `aRchaic` with $K = 3$ (see Supplementary Figure S17 for $K = 4, 5, 6$). To give a sense of computational requirements, these results took approximately 23 minutes to generate on a single modern compute node. The results clearly highlight differences between modern, ancient (UDG), and ancient (non-UDG) samples. The modern samples show very strong membership in a single cluster (red). As expected, this "modern" cluster shows only modest enrichment in its mismatch type, flanking base composition, and mismatch location, relative to the modern background.

Ancient (non-UDG) samples show high membership in a second cluster (blue). This cluster is characterized by a very strong enrichment of C-to-T mismatches at the ends of the reads, which is a typical sign of DNA damage [Rohland et al., 2015], and this enrichment is accompanied by a depletion of guanine just 3' to the mismatch. We see this depletion in guanine because the blue cluster is driven by mismatches at cytosine sites which seldom precedes a guanine because CpG sites occur less frequently than expected [Shen et al., 1994].

The ancient UDG-treated individuals show high membership in the third (orange) cluster, and partial membership in the first (red) cluster. The membership in the red cluster presumably reflects the fact that the UDG-treatment repairs much of the damage in these samples, making them look more "modern" in their mismatch profiles. The orange cluster is characterized by enrichment of C-to-T mismatches very close to the ends of reads, with a strong enrichment of guanine at the right flanking base. That is, an enrichment

of CpG-to-TpG mismatches at the ends of reads. This may be explained by the fact that when a methylated cytosine undergoes deamination it becomes thymine (in contrast to unmethylated cytosines, which deaminate to uracil) and these thymines are not repaired by the UDG-treatment [Duncan and Miller, 1980]. Furthermore, we see a depletion of thymine 5' upstream of the strand-break, which consequently is manifested as a depletion of thymine at the left flanking base.

### *4.3.2 The effects of contamination on inferred grades of membership*

We next sought to examine the effects of exogenous modern contamination on inferred grades of memberships in ancient samples.

We performed an *in-silico* experiment to artificially contaminate ancient samples with modern data from the 1000 Genomes Project [Consortium et al., 2012]. We selected one BAM file from an ancient sample (K01 from [Gamba et al., 2014]), and split its reads into 10 equal subsets. We then contaminated each subset with reads from a distinct modern individual from the 1000 Genomes Project, varying the contamination level from 0% to 100% (Figure 4.3a). This results in 10 samples (S1-S10) representing 10 contaminated ancient samples with known levels of modern contamination.

We applied `aRchaic` with $K = 2$ on the contaminated samples (S1-S10) plus 40 other modern individuals (randomly sampled from the 1000 Genomes Project; Figure 4.3b). Modern individuals showed high grades of membership in one cluster (red). Fully contaminated individuals nearly showed full membership in the red cluster, while uncontaminated ancient samples showed essentially no membership in this cluster. For samples in between, membership in the red cluster increased approximately monotonically with the level of contamination (Figure 4.3c). We obtained similar results even with only 10000 randomly-sampled reads for each sample (Figure 4.3e) implying that these results are robust to low sequencing depth.

We find that the grades of membership in the red cluster (representing moderns) are

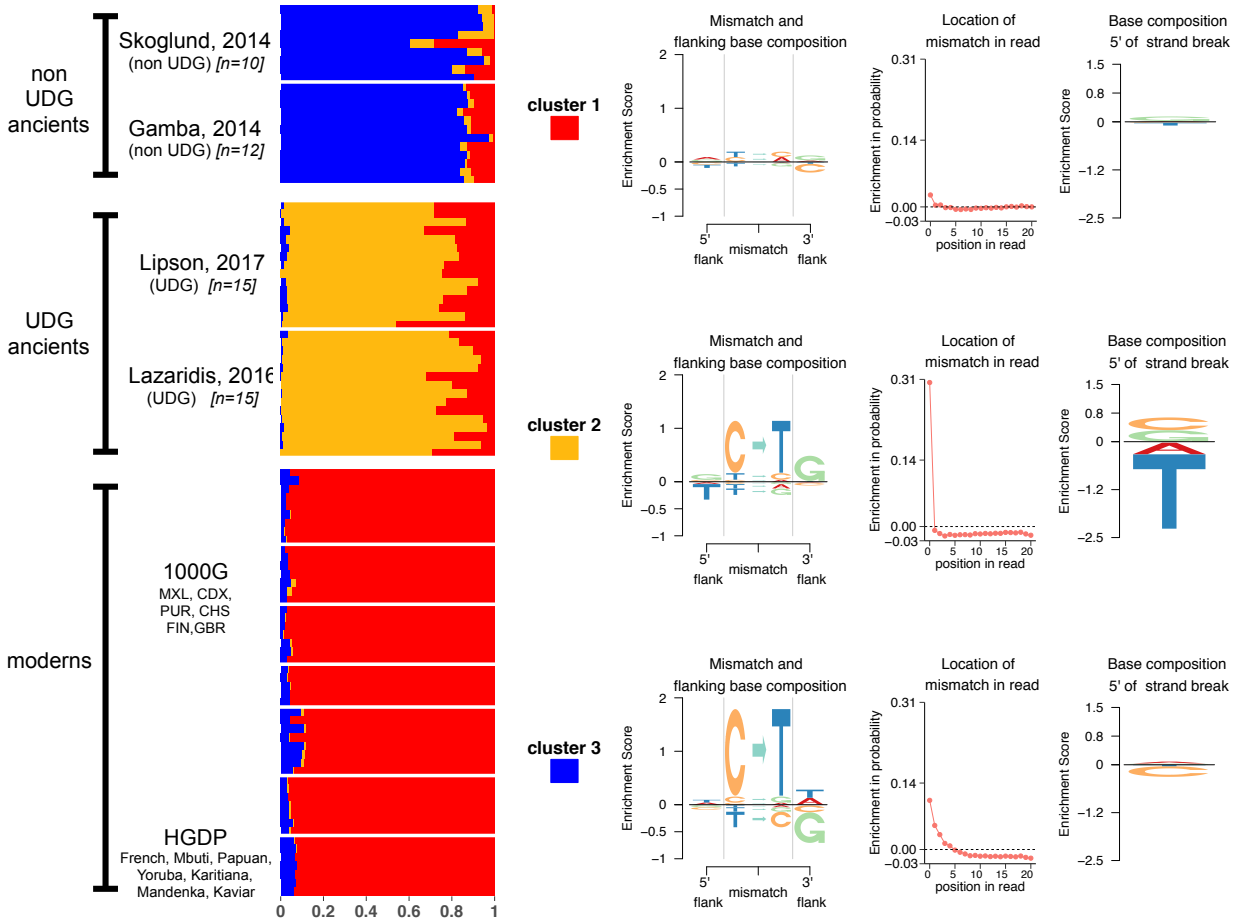



Figure 4.2: `aRchaic` **clearly distinguishes between modern, ancient (UDG), and ancient (non-UDG) samples.** `aRchaic` is applied with $K = 3$ to a collection of ancient individuals from four studies Skoglund et al. [2014a], Gamba et al. [2014], Lazaridis et al. [2016], Lipson et al. [2017] along with modern individuals randomly sampled from the 1000 Genomes Project and 10 individuals from the Human Genome Diversity Panel [Consortium et al., 2012, Cann et al., 2002, Meyer et al., 2012]. Modern samples have high membership in the red cluster. The *EDLogo* representation of this cluster does not show strong enrichment against a modern background. The ancient (non-UDG) samples are representative of the blue cluster. The *EDLogo* plot for the blue cluster shows a strong enrichment in C-to-T mismatches at the end of reads, a depletion of guanine in the right flanking base, and a depletion of cytosine at the 5' strand-break. The ancient (UDG) samples have partial membership both in the red cluster and in the gold cluster. The *EDLogo* plot for the gold cluster is enriched in C-to-T mismatches at the terminal ends of the reads, shows an enrichment of guanine at the right flanking base, and a depletion of thymine one base 5' upstream of the strand break.

consistently less than the proportion of contaminated reads (Figure 4.3c red curve). This pattern can be partly explained by the fact that the vast majority of reads from a DNA library contain no damage. To elaborate, only a fraction of contaminated DNA from a modern source have mismatches, and this fraction will typically be less than for ancient DNA. Thus, the estimated proportion of mismatches arising from a "modern DNA" cluster should generally be expected to be less than the contamination fraction of the library. This implies `aRchaic` will typically provide a lower bound on contamination rates.

In this experiment, if we define the "mismatch contamination rate" as proportion of mismatches that originate from a contaminating read, we see that the `aRchaic` grades of membership are approximately linearly proportional to the mismatch contamination rate, and only slightly underestimate the true proportion of contaminated reads (Figure 4.3d). This encourages the use of a possible correction factor to infer a true read-level contamination rate. The fraction of a DNA library with mismatches can be measured ($\hat{f}$, e.g. $\hat{f} = 0.3$), and the fraction of modern DNA reads with mismatches to reference ($\hat{d}$, e.g. $\hat{d} = 0.09$) can be approximated using modern samples. With those values a simple moment estimator of the read level contamination rate would be: $\hat{c} = f(d/q + f - d)^{-1}$, where $q$ is the fraction of membership in the "modern" cluster. Applying this estimator to the grade of membership in the the modern cluster in Figure 4.3c (red curve) provides values that are closer to the true contamination rates (Figure 4.3c green curve). However, a complication is that high quality aDNA may generate mismatches that cluster with "modern" mismatch profiles even without contamination. This effect is not accounted for by this simple estimator, and is difficult to predict.

### *4.3.3* `aRchaic` *can identify both DNA damage and technical artifacts*

As a final case study, we compiled data from 25 modern and 25 ancient Native Americans from the Northwest Coast of North America [Lindo et al., 2016]. This dataset offers us a

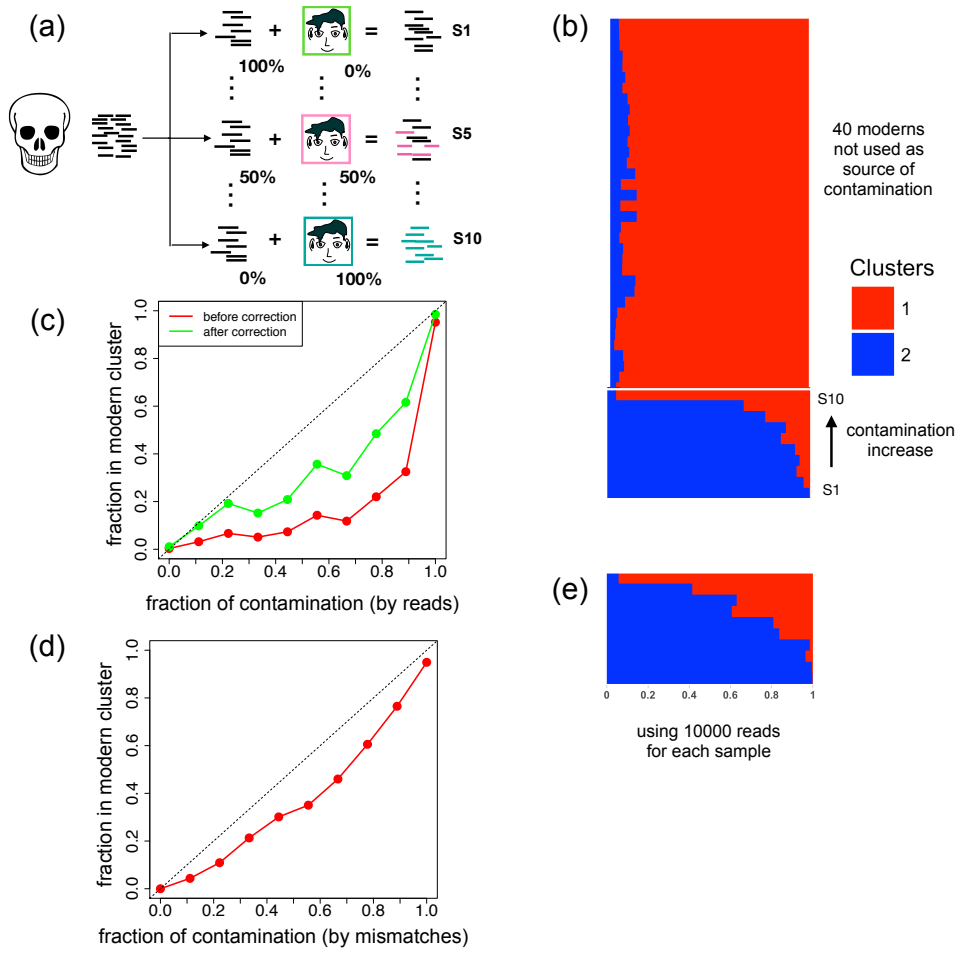



Figure 4.3: **Estimated grades of membership reflect levels of contamination.** (a) Reads from one ancient individual (KO1 from Gamba et al. [2014]) were split into 10 equally sized groups. Reads were added from a distinct individual in the 1000 Genomes Project [Consortium et al., 2012] to each group (S1-S10) at varying levels of percentages (indicating levels of contamination). (b) We applied `aRchaic` with $K = 2$ on a combined dataset comprised of these 10 contaminated groups of reads (S1-S10) along with 40 other modern individuals from 1000 Genomes (c) The grades of membership in cluster 1 ("modern" cluster) were plotted as a function of the percentage of contamination before (red curve) and after (green curve) applying the correction factor discussed in the last paragraph of Section (3.2). (d) We repeated the same experiment as described in panel a, except we discarded all reads with no mismatches or greater than one mismatch. The grades of membership in cluster 1 were plotted as a function of the "mismatch contamination rate" which is defined as the proportion of mismatches that originate from a contaminated read (e) Each group (S1-S10) was further sub-sampled to 10,000 reads, and `aRchaic` was applied with $K = 2$ to the new subsampled groups and the same 40 modern individuals as in panel b.

opportunity to apply `aRchaic` on modern and ancient DNA samples collected from the same population and sequenced in the same laboratory. In these data, the first two positions from the 5' end of each read had been removed by the original authors in an attempt to mitigate effects of DNA damage. Despite this, `aRchaic`, when applied with $K = 2$ to all 50 samples, clearly distinguishes between modern and ancient individuals (Supplementary Figure S18).

When we applied `aRchaic` to just modern samples we were surprised to find two clear clusters (Figure 4.4a). These clusters turned out to reflect the fact that the modern samples had been processed using two different library preparation kits, Nextera & TruSeq [Lindo et al., 2016]. Samples prepared with the TruSeq kit showed nearly full membership in one cluster (beige), and those prepared with the Nextera kit showed nearly full membership in a second cluster (pink). These clusters show only small differences in mismatch patterns, but the pink cluster is characterized by a strong excess of mismatches at position 12, apparently an artifact introduced by the Nextera preparation (Supplementary Figure S20).

We also applied `aRchaic` with $K = 2$ to just the ancient samples (see Figure 4.4b). Unlike the moderns, this yielded a continuous gradient of memberships in the two clusters (cyan and gold). One cluster (gold) is dominated by the strong enrichment of C-to-T mismatches relative to the modern background that is typical of ancient samples. The other cluster (cyan) is enriched primarily in C-to-A mismatches, possibly representing another type of damage, or other artifacts, in the ancient DNA. Interestingly, the individual with highest membership in this cyan cluster was much older than all the others ($\approx$ 6000 years BP; all other samples are $\approx$ 2000 years BP).

## 4.4 Discussion

We developed a method (`aRchaic`) for clustering and visualization of samples based on DNA mismatch patterns. Our method is based on a Grade-of-Membership (GoM) model, which generalizes the concept of clustering to allow samples to have membership in multiple

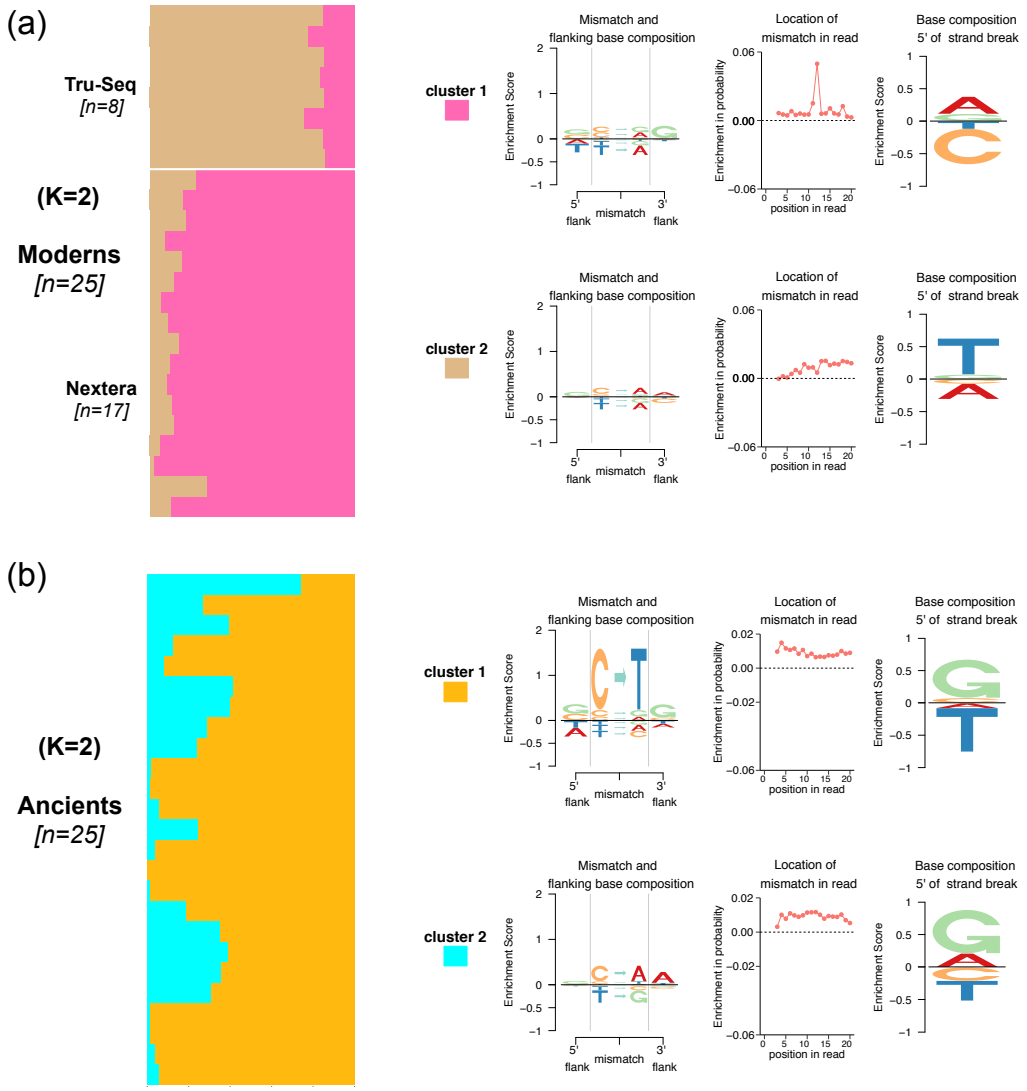



Figure 4.4: **DNA damage and library preparation techniques drive grades of membership.** (a) We applied `aRchaic` with $K = 2$ to 25 modern samples from Lindo et al. [2016]. The samples prepared with the TruSeq kit show nearly full membership in the pink cluster. Samples prepared with the Nextera kit show partial membership in the pink cluster and the tan cluster. The tan cluster shows a blip at the 12th position from the end of the read (b) We applied `aRchaic` with $K = 2$ to 25 ancient samples from Lindo et al. [2016]. The two clusters show an enrichment of C-to-T mismatches at the ends of reads and an enrichment of purines at the 5' strand-break.

clusters. We provide a visual representation of the grades of membership using a "Structure plot" [Rosenberg et al., 2002] and visualization of the mismatch profiles (or clusters) with an *EDLogo* plot [Dey et al., 2017b].

In GoM models, the choice of the number of clusters $K$ (or mismatch profiles) is a contentious issue. In our analyses we selected values of $K$ that highlight interpretable structure in the data. We emphasize that there will typically be no single "true" $K$, and that examining results with different $K$ can often provide additional insights [Dey et al., 2017a]. For example, Figure 4.2 shows results for $K = 3$, but higher values of $K$ reveal additional structure within each ancient subgroup (Supplementary Figure S17). Similarly, Supplementary Figure S18 shows results for $K = 2$ where the model fails to distinguish between Nextera and TruSeq modern samples, but analysis with $K = 6$ does pick up this difference (Supplementary Figure S19). More generally, `aRchaic` is useful for detecting batch effects as Figure 4.4 and Supplementary Figure S19 suggest.

A key challenge in analyzing ancient DNA is that data are often contaminated with exogenous modern DNA. Several approaches have been suggested to estimate the amount of contamination. One approach is to compute the rate of polymorphism across the X chromosome in males [Korneliussen et al., 2014, Rasmussen et al., 2011], where the presence of polymorphism would suggest contamination because males have only one X chromosome. Another approach is to quantify the contribution of a panel of modern mitochondrial haplotypes to the ancient DNA [Renaud et al., 2015, Fu et al., 2014]. Neither of these approaches leverages autosomal DNA. Although `aRchaic` does not provide explicit estimates of contamination levels, in some settings (e.g. Figure 4.3) the inferred grades of membership can reflect relative levels of contamination even with low sequencing depth, and may be a useful complement to these other methods. Some caution is necessary though. Under conditions where an ancient sample has undergone a noticeable amount of DNA damage, we underestimate the proportion of contamination (Figure 4.3). On the other hand, when ancient

mismatches are indistinguishable from modern, the proportion in modern cluster of a sample might be greater than the proportion of contamination (as in the case of an ancient sample with little to no damage). We encourage users to keep these caveats in mind, and suggest that in practice `aRchaic` will be most useful for flagging potentially contaminated samples as those that have an above average clustering with modern samples.

Here we have chosen to model features at the level of mismatches, which have been shown to be informative of DNA damage in previous studies [Ginolhac et al., 2011, Jónsson et al., 2013, Briggs et al., 2007]. Alternatively, one could formulate a model at the level of reads. For example, the method *PMDtools* computes a score for every read representing the probability that the read is damaged [Skoglund et al., 2014b]. This method models mismatches along the read; additionally, one can incorporate indels and fragment length along with mismatches. One reason we chose not to model these extra features was to reduce the feature and computational complexity of our method. Furthermore, these extra features may not actually be driven by DNA damage. For example, we explored fragment length profiles in several aDNA data-sets and found their distributions to be primarily driven by lab-specific effects rather than DNA damage. Another limitation of fragment length is that it can be used only in studies using paired-end sequencing.

# REFERENCES


David H Alexander, John Novembre, and Kenneth Lange. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Research*, 19(9):1655–1664, 2009.

Morten E Allentoft, Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B Damgaard, Hannes Schroeder, Torbjörn Ahlström, Lasse Vinner, et al. Population Genomics of Bronze Age Eurasia. *Nature*, 522(7555): 167–172, 2015.

Maria Anisimova, Joseph P Bielawski, and Ziheng Yang. Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. *Molecular Biology and Evolution*, 18(8):1585–1592, 2001.

Soheil Baharian, Maxime Barakatt, Christopher R Gignoux, Suyash Shringarpure, Jacob Errington, William J Blot, Carlos D Bustamante, Eimear E Kenny, Scott M Williams, Melinda C Aldrich, et al. The Great Migration and African-American Genomic Diversity. *PLoS genetics*, 12(5):e1006059, 2016.

Melanie Bahlo and Robert C Griffiths. Coalescence Time for Two Genes from a Subdivided Population. *Journal of Mathematical Biology*, 43(5):397–410, 2001.

Matthew F Barber and Nels C Elde. Escape from Bacterial Iron Piracy through Rapid Evolution of Transferrin. *Science*, 346(6215):1362–1366, 2014.

Guido Barbujani, Neal L Oden, and Robert R Sokal. Detecting regions of abrupt change in maps of biological variables. *Systematic Zoology*, 38(4):376–389, 1989.

Rowan DH Barrett and Hopi E Hoekstra. Molecular Spandrels: Tests of Adaptation at the Genetic Level. *Nature Reviews Genetics*, 12(11):767, 2011.

Nick H Barton, Frantz Depaulis, and Alison M Etheridge. Neutral Evolution in Spatially Continuous Populations. *Theoretical population biology*, 61(1):31–48, 2002.

Peter Beerli and Joseph Felsenstein. Maximum-Likelihood Estimation of Migration Rates and Effective Population Numbers in Two Populations using a Coalescent Approach. *Genetics*, 152(2):763–773, 1999.

Peter Beerli and Joseph Felsenstein. Maximum Likelihood Estimation of a Migration Matrix and Effective Population Sizes in n Subpopulations by using a Coalescent Approach. *Proceedings of the National Academy of Sciences*, 98(8):4563–4568, 2001.

David J Begun, Alisha K Holloway, Kristian Stevens, LaDeana W Hillier, Yu-Ping Poh, Matthew W Hahn, Phillip M Nista, Corbin D Jones, Andrew D Kern, Colin N Dewey, et al. Population Genomics: Whole-Genome analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biol*, 5(11):e310, 2007.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Jesse D Bloom. An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit. *Molecular Biology and Evolution*, 31(8):1956–1978, 2014.

Michael GB Blum, Christophe Damerval, Stephanie Manel, and Olivier François. Brownian Models and Coalescent Structures. *Theoretical Population Biology*, 65(3):249–261, 2004.

Gideon Bradburd, Graham Coop, and Peter Ralph. Inferring Continuous and Discrete Population Genetic Structure across Space. *bioRxiv*, page 189688, 2017.

Gideon S Bradburd, Peter L Ralph, and Graham M Coop. A Spatial Framework for Understanding Population Structure and Admixture. *PLoS Genet*, 12(1):e1005703, 2016.

Adrian W Briggs, Udo Stenzel, Philip LF Johnson, Richard E Green, Janet Kelso, Kay Prüfer, Matthias Meyer, Johannes Krause, Michael T Ronan, Michael Lachmann, et al. Patterns of Damage in Genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, 104(37):14616–14621, 2007.

Thomas Broquet, Jonathan Yearsley, Alexandre H Hirzel, Jerome Goudet, and Nicolas Perrin. Inferring Recent Migration Rates from Individual Genotypes. *Molecular Ecology*, 18 (6):1048–1060, 2009.

Brian L Browning and Sharon R Browning. A Fast, Powerful Method for Detecting Identity-by-Descent. *The American Journal of Human Genetics*, 88(2):173–182, 2011.

Brian L Browning and Sharon R Browning. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics*, 194(2):459–471, 2013.

Howard. Cann et al. A Human Genome Diversity Cell Line Panel. *Science*, 296(5566): 261–262, 2002.

Kevin Caye, Flora Jay, Olivier Michel, and Olivier Francois. Fast Inference of Individual Admixture Coefficients Using Geographic Data. *bioRxiv*, page 080291, 2016.

Gary K Chen, Paul Marjoram, and Jeffrey D Wall. Fast and flexible simulation of DNA sequence data. *Genome Research*, 19(1):136–142, 2009.

Charleston WK Chiang, Joseph H Marcus, Carlo Sidore, Hussein Al-Asadi, Magdalena Zoledziewska, Maristella Pitzalis, Fabio Busonero, Andrea Maschio, Giorgio Pistis, Maristella Steri, et al. Population History of the Sardinian People Inferred from Whole-Genome Sequencing. *BioRxiv*, page 092148, 2016.

Joel E Cohen. Population Growth and Earth's Human Carrying Capacity. *Science*, 269 (5222):341–346, 1995.

1000 Genomes Project Consortium et al. An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature*, 491(7422):56–65, 2012.

Matthew D Daugherty and Harmit S Malik. Rules of Engagement: Molecular Insights from Host-virus Arms Races. *Annual Review of Genetics*, 46:677–700, 2012.

Kushal K Dey, Chiaowen Joyce Hsiao, and Matthew Stephens. Visualizing the Structure of RNA-seq Expression Data using Grade of Membership Models. *PLoS Genetics*, 13(3): e1006599, 2017a.

Kushal K Dey, Dongyue Xie, and Matthew Stephens. A New Sequence Logo Plot to Highlight Enrichment and Depletion. *BioRxiv*, 2017b. doi: 10.1101/226597. URL `https://www.biorxiv.org/content/early/2017/11/29/226597`.

Slavica Dimitrieva and Maria Anisimova. Unraveling Patterns of Site-toSsite Synonymous Rates Variation and Associated Gene Properties of Protein Domains and Families. *PLoS One*, 9(6):e95034, 2014.

Bruce K Duncan and Jeffrey H Miller. Mutagenic Deamination of Cytosine Residues in DNA. *Nature*, 287(5782):560, 1980.

Elena A Erosheva. Latent Class Representation of the Grade of Membership Model. *Seatle: University of Washington*, 2006.

Joseph Felsenstein. Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

Joseph Felsenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

William Fletcher and Ziheng Yang. INDELible: a Flexible Simulator of Biological Sequence Evolution. *Molecular Biology and eEolution*, 26(8):1879–1888, 2009.

Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M Slepchenko, Aleksei A Bondarev, Philip LF Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare de Filippo, et al. Genome Sequence of a 45,000-year-old Modern Muman from Western Siberia. *Nature*, 514(7523): 445–449, 2014.

Cristina Gamba, Eppie R Jones, Matthew D Teasdale, Russell L McLaughlin, Gloria Gonzalez-Fortes, Valeria Mattiangeli, László Domboróczki, Ivett Kővári, Ildikó Pap, Alexandra Anders, et al. Genome Flux and Stasis in a Five Millennium Transect of European Prehistory. *Nature communications*, 5:5257, 2014.

Aurelien Ginolhac, Morten Rasmussen, M Thomas P Gilbert, Eske Willerslev, and Ludovic Orlando. mapDamage: Testing for Damage Patterns in Ancient DNA Sequences. *Bioinformatics*, 27(15):2153–2155, 2011.

Nick Goldman and Ziheng Yang. A Codon-Based Model of Nucleotide Substitution for Protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736, 1994.

Peter J Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732, 1995.

Richard E Green, Edward L Braun, Joel Armstrong, Dent Earl, Ngan Nguyen, Glenn Hickey, Michael W Vandewege, John A St John, Salvador Capella-Gutiérrez, Todd A Castoe, et al. Three Crocodilian Genomes Reveal Ancestral Patterns of Evolution among Archosaurs. *Science*, 346(6215):1254449, 2014.

Gilles Guillot, Frédéric Mortier, and Arnaud Estoup. GENELAND: A Computer Package for Landscape Genetics. *Molecular Ecology Resources*, 5(3):712–715, 2005.

Gilles Guillot, Raphael Leblois, Aurelie Coulon, and Alain C Frantz. Statistical Methods in Spatial Genetics. *Molecular Ecology*, 18(23):4734–4756, 2009.

Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole Population, Genome-Wide Mapping of Hidden Relatedness. *Genome Research*, 19(2):318–326, 2009.

Eunjung Han, Peter Carbonetto, Ross E Curtis, Yong Wang, Julie M Granka, Jake Byrnes, Keith Noto, Amir R Kermany, Natalie M Myres, Mathew J Barber, et al. Clustering of 770,000 Genomes Reveals Post-Colonial Population Structure of North America. *Nature Communications*, 8:14238, 2017.

Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, USA, 2004.

Jody Hey and Rasmus Nielsen. Multilocus Methods for Estimating Population Sizes, Migration Rates and Divergence Time, with Applications to the Divergence of Drosophila Pseudoobscura and D. Persimilis. *Genetics*, 167(2):747–760, 2004.

David M Hillis. Inferring Complex Phylogenies. *Nature*, 383(6596):130, 1996.

Richard R Hudson. Properties of a Neutral Allele Model with Intragenic Recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

Richard R Hudson et al. Gene Genealogies and the Coalescent Process. *Oxford Surveys in Evolutionary Biology*, 7(1):44, 1990.

Hákon Jónsson, Aurélien Ginolhac, Mikkel Schubert, Philip LF Johnson, and Ludovic Orlando. mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters. *Bioinformatics*, 29(13):1682–1684, 2013.

Thomas H Jukes, Charles R Cantor, et al. Evolution of Protein Molecules. *Mammalian protein metabolism*, 3(21):132, 1969.

Joanna Kaplanis, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, Barak Markus, Mona Sheikh, Melissa Gymrek, et al. Quantitative analysis of population-scale family trees with millions of relatives. *Science*, (early online): eaam9309, 2018.

Motoo Kimura. A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

Motoo Kimura and George H Weiss. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics*, 49(4):561–576, 1964.

John Frank Charles Kingman. The Coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, 15(1):356, 2014.

Sergei L Kosakovsky Pond and Simon DW Frost. Not so Different after all: a Comparison of Methods for Detecting Amino Acid Sites under Selection. *Molecular Biology and Eolution*, 22(5):1208–1222, 2005.

Carolin Kosiol, Tomáš Vinař, Rute R da Fonseca, Melissa J Hubisz, Carlos D Bustamante, Rasmus Nielsen, and Adam Siepel. Patterns of Positive Selection in Six Mammalian Genomes. *PLoS Genetics*, 4(8):e1000144, 2008.

Kenneth Lange. A Quasi-Newton Acceleration of the EM Algorithm. *Statistica Sinica*, pages 1–18, 1995.

Oscar Lao, Timothy T Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, Jaume Bertranpetit, Laurence A Bindoff, David Comas, et al. Correlation between Genetic and Geographic Structure in Europe. *Current Biology*, 18(16):1241–1248, 2008.

Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, 8(1):e1002453, 2012.

Iosif Lazaridis, Dani Nadel, Gary Rollefson, Deborah C Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes, Mario Novak, Beatriz Gamarra, Kendra Sirak, et al. Genomic Insights into the Origin of Farming in the Ancient Near East. *Nature*, 536(7617):419, 2016.

Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C Royrvik, Barry Cunliffe, Daniel J Lawson, et al. The Fine-Scale Genetic Structure of the British Population. *Nature*, 519(7543):309–314, 2015.

Heng Li and Richard Durbin. Inference of Human Population History from Individual Whole-Genome Sequences. *Nature*, 475(7357):493–496, 2011.

John Lindo, Emilia Huerta-Sánchez, Shigeki Nakagome, Morten Rasmussen, Barbara Petzelt, Joycelynn Mitchell, Jerome S Cybulski, Eske Willerslev, Michael DeGiorgio, and Ripan S Malhi. A Time Transect of Exomes from a Native American Population before and after European Contact. *Nature Communications*, 7:13175, 2016.

Mark Lipson, Anna Szécsényi-Nagy, Swapan Mallick, Annamária Pósa, Balázs Stégmár, Victoria Keerl, Nadin Rohland, Kristin Stewardson, Matthew Ferry, Megan Michel, et al. Parallel Palaeogenomic Transects Reveal Complex Genetic History of Early European Farmers. *Nature*, 551(7680):368, 2017.

Philippe Lopez, Didier Casane, and Hervé Philippe. Heterotachy, an Important Process of Protein Evolution. *Molecular Biology and Evolution*, 19(1):1–7, 2002.

Wayne P Maddison. Gene Trees in Species Trees. *Systematic Biology*, 46(3):523–536, 1997.

Helena Malmström, Emma M Svensson, M Thomas P Gilbert, Eske Willerslev, Anders Götherström, and Gunilla Holmlund. More on Contamination: the Use of Asymmetric Molecular Behavior to Identify Authentic Ancient Human DNA. *Molecular Biology and Evolution*, 24(4):998–1004, 2007.

Stéphanie Manel, Michael K Schwartz, Gordon Luikart, and Pierre Taberlet. Landscape Genetics: Combining Landscape Ecology and Population Genetics. *Trends in Ecology & Evolution*, 18(4):189–197, 2003.

Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak, et al. Genome-wide Patterns of Selection in 230 Ancient Eurasians. *Nature*, 528(7583): 499–503, 2015.

Iain Mathieson et al. The Genomic History Of Southeastern Europe. *BioRxiv*, 2017. doi: 10.1101/135616. URL `https://www.biorxiv.org/content/early/2017/09/19/135616`.

Brad H McRae and B Nürnberger. Isolation by Resistance. *Evolution*, 60(8):1551–1561, 2006.

Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G Schraiber, Flora Jay, Kay Prüfer, Cesare De Filippo, et al. A High-coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, 338 (6104):222–226, 2012.

Sanzo Miyazawa and Robert L Jernigan. A New Substitution Matrix for Protein Sequence Searches based on Contact Frequencies in Protein Structures. *Protein Engineering, Design and Selection*, 6(3):267–278, 1993.

Krystyna Nadachowska-Brzyska, Reto Burri, Linnéa Smeds, and Hans Ellegren. PSMC Analysis of Effective Population Sizes in Molecular Ecology and its Application to Black-and-White Ficedula Flycatchers. *Molecular Ecology*, 25(5):1058–1072, 2016.

National Records of Scotland. Scotland's 2011 census, 2011. census size retrieved from, `http://www.scotlandscensus.gov.uk/statistical-bulletins`.

Masatoshi Nei and Takashi Gojobori. Simple methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions. *Molecular Biology and Evolution*, 3(5):418–426, 1986.

Matthew R Nelson, Katarzyna Bryc, Karen S King, Amit Indap, Adam R Boyko, John Novembre, Linda P Briley, Yuka Maruyama, Dawn M Waterworth, Gérard Waeber, et al. The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. *The American Journal of Human Genetics*, 83(3): 347–358, 2008.

Rasmus Nielsen and Ziheng Yang. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics*, 148(3):929–936, 1998.

Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and Carlos Bustamante. Genomic Scans for Selective Sweeps using SNP Data. *Genome Research*, 15(11):1566–1575, 2005.

M Notohara. The Coalescent and the Genealogical Process in Geographically Structured Population. *Journal of Mathematical Biology*, 29(1):59–75, 1990.

John Novembre and Montgomery Slatkin. Likelihood-Based Inference in Isolation-by-Distance Models Using the Spatial Distribution of Low-Frequency Alleles. *Evolution*, 63(11):2914–2925, 2009.

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes Mirror Geography within Europe. *Nature*, 456(7218):98–101, 2008.

Iñigo Olalde et al. The Beaker Phenomenon And The Genomic Transformation Of Northwest Europe. *BioRxiv*, 2017. doi: 10.1101/135962. URL `https://www.biorxiv.org/content/early/2017/05/09/135962`.

Pier Francesco Palamara and Itsik Peer. Inference of historical migration rates via haplotype sharing. *Bioinformatics*, 29(13):i180–i188, 2013.

Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Peer. Length Distributions of Identity by Descent Reveal Fine-scale Demographic History. *The American Journal of Human Genetics*, 91(5):809–822, 2012.

Nick Patterson, Alkes L Price, and David Reich. Population Structure and Eigenanalysis. *PLoS Genet*, 2(12):e190, 2006.

Desislava Petkova, John Novembre, and Matthew Stephens. Visualizing Spatial Population Structure with Estimated Effective Migration Surfaces. Technical report, Nature Publishing Group, 2015.

Desislava Petkova, John Novembre, and Matthew Stephens. Visualizing Spatial Population Structure with Estimated Effective Migration Surfaces. *Nat Genet*, 48(1):94–100, Jan 2016. doi: 10.1038/ng.3464.

Sergei Kosakovsky Pond and Spencer V Muse. Site-to-site Variation of Synonymous Substitution Rates. *Molecular Biology and Evolution*, 22(12):2375–2385, 2005a.

Sergei L Kosakovsky Pond and Spencer V Muse. HyPhy: Hypothesis Testing using Phylogenies. In *Statistical Methods in Molecular Evolution*, pages 125–181. Springer, 2005b.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of Population Structure using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 2000.

Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H Sudmant, Cesare De Filippo, et al. The Complete Genome Sequence of a Neandertal from the Altai Mountains. *Nature*, 505(7481): 43, 2014.

Molly Przeworski and Carlos D Bustamante. Genetic Signatures of Natural Selection. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, 2004.

Peter Ralph and Graham Coop. The Geography of Recent Genetic Ancestry across Europe. *PLoS Biol*, 11(5):e1001555, 2013.

Morten Rasmussen, Xiaosen Guo, Yong Wang, Kirk E Lohmueller, Simon Rasmussen, Anders Albrechtsen, Line Skotte, Stinus Lindgreen, Mait Metspalu, Thibaut Jombart, et al. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. *Science*, 334(6052):94–98, 2011.

Gabriel Renaud, Viviane Slon, Ana T Duggan, and Janet Kelso. Schmutzi: Estimation of Contamination and Endogenous Mitochondrial Consensus Calling for Ancient DNA. *Genome Biology*, 16(1):224, 2015.

A Rényi. On the Central Limit Theorem for the Sum of a Random Number of Independent Random Variables. *Acta Mathematica Hungarica*, 11(1-2):97–102, 1960.

Harald Ringbauer, Graham Coop, and Nicholas H Barton. Inferring Recent Demography from Isolation-By-Distance of Long Shared Sequence Blocks. *Genetics*, 205(3):1335–1351, 2017.

JJ Robledo-Arnuncio and Francois Rousset. Isolation by Distance in a Continuous Population under Stochastic Demographic Fluctuations. *Journal of Evolutionary Biology*, 23(1):53–71, 2010.

Nadin Rohland, Eadaoin Harney, Swapan Mallick, Susanne Nordenfelt, and David Reich. Partial uracil–DNA–glycosylase Treatment for Screening of Ancient DNA. *Phil. Trans. R. Soc. B*, 370(1660):20130624, 2015.

Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman. Genetic Structure of Human Populations. *Science*, 298(5602):2381–2385, 2002.

Noah A Rosenberg, Saurabh Mahajan, Sohini Ramachandran, Chengfeng Zhao, Jonathan K Pritchard, and Marcus W Feldman. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genet*, 1(6):e70, 2005.

François Rousset. *Genetic Structure and Selection in Subdivided Populations (MPB-40)*. Princeton University Press New Jersey, 2004.

Stanley Sawyer. Results for the Stepping Stone Model for Migration in Population Genetics. *The Annals of Probability*, pages 699–728, 1976.

Susanna Sawyer, Johannes Krause, Katerina Guschanski, Vincent Savolainen, and Svante Pääbo. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. *PloS One*, 7(3):e34131, 2012.

Konrad Scheffler, Darren P Martin, and Cathal Seoighe. Robust Inference of Positive Selection from Recombining Coding Sequences. *Bioinformatics*, 22(20):2493–2499, 2006.

Stephan Schiffels and Richard Durbin. Inferring Human Population Size and Separation History from Multiple Genome Sequences. *Nature Genetics*, 46(8):919–925, 2014.

Joshua G Schraiber and Joshua M Akey. Methods and Models for Unravelling Human Evolutionary history. *Nature Reviews Genetics*, 16(12):727, 2015.

Daniel R Schrider, Jonathan N Hourmozdi, and Matthew W Hahn. Pervasive Multinucleotide Mutational Events in Eukaryotes. *Current Biology*, 21(12):1051–1054, 2011.

Gernot Segelbacher, Samuel A Cushman, Bryan K Epperson, Marie-Josée Fortin, Olivier Francois, Olivier J Hardy, Rolf Holderegger, Pierre Taberlet, Lisette P Waits, and Stéphanie Manel. Applications of Landscape Genetics in Conservation Biology: Concepts and Challenges. *Conservation Genetics*, 11(2):375–385, 2010.

B Shapiro and Michael Hofreiter. A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. *Science*, 343(6169):1236573, 2014.

Sara Sheehan, Kelley Harris, and Yun S Song. Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics*, 194(3):647–662, 2013.

Jiang-Cheng Shen, William M Rideout III, and Peter A Jones. The Rate of Hydrolytic Deamination of 5-Methylcytosine in Double-Stranded DNA. *Nucleic Acids Research*, 22 (6):972–976, 1994.

Yuichi Shiraishi, Georg Tremmel, Satoru Miyano, and Matthew Stephens. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLoS Genetics*, 11(12):e1005657, 2015.

Daniel Shriner, David C Nickle, Mark A Jensen, and James I Mullins. Potential Impact of Recombination on Sitewise Approaches for Detecting Positive Natural Selection. *Genetics Research*, 81(2):115–121, 2003.

Pontus Skoglund, Helena Malmström, Ayça Omrak, Maanasa Raghavan, Cristina Valdiosera, Torsten Günther, Per Hall, Kristiina Tambets, Jüri Parik, Karl-Göran Sjögren, et al. Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers. *Science*, 344(6185):747–750, 2014a.

Pontus Skoglund, Bernd H Northoff, Michael V Shunkov, Anatoli P Derevianko, Svante Pääbo, Johannes Krause, and Mattias Jakobsson. Separating Endogenous Ancient DNA from Modern Day Contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, 111(6):2229–2234, 2014b.

Pontus Skoglund, Erik Ersmark, Eleftheria Palkopoulou, and Love Dalén. Ancient Wolf Genome Reveals an Early Divergence of Domestic Dog Ancestors and Admixture into High-Latitude Breeds. *Current Biology*, 25(11):1515–1519, 2015.

Stephanie J Spielman and Claus O Wilke. The Relationship between dN/dS and Scaled Selection Coefficients. *Molecular Biology and Evolution*, 32(4):1097–1108, 2015.

Matthew Stephens. Bayesian Analysis of Mixture Models with an Unknown Number of Components, an Alternative to Reversible Jump Methods. *Annals of Statistics*, pages 40–74, 2000.

Matthew Stephens, Nicholas J Smith, and Peter Donnelly. A New Statistical Method for Haplotype Reconstruction from Population Data. *The American Journal of Human Genetics*, 68(4):978–989, 2001.

Jay F Storz, Amy M Runck, Stephen J Sabatino, John K Kelly, Nuno Ferrand, Hideaki Moriyama, Roy E Weber, and Angela Fago. Evolutionary and Functional Insights into the Mechanism Underlying High-Altitude Adaptation of Deer Mouse Hemoglobin. *Proceedings of the National Academy of Sciences*, 106(34):14450–14455, 2009.

Yoshiyuki Suzuki and Takashi Gojobori. A Method for Detecting Positive Selection at Single Amino Acid Sites. *Molecular Biology and Evolution*, 16(10):1315–1328, 1999.

Matt Taddy. On Estimation and Selection for Topic Models. In *International Conference on Artificial Intelligence and Statistics*, pages 1184–1193, 2012.

Fumio Tajima. Evolutionary Relationship of DNA Sequences in Finite Populations. *Genetics*, 105(2):437–460, 1983.

Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and Scalable Inference of Population History from Hundreds of Unphased Whole Genomes. Technical report, Nature Research, 2016.

The GeoNames Geographical Database. Geonames. Queried on July 2018, `http://www.geonames.org/`.

The World Bank. World development indicators, 2016. data retrieved from World Development Indicators, `https://data.worldbank.org/indicator/SP.POP.TOTL`.

Joseph W Thornton and Bryan Kolaczkowski. No Magic Pill for Phylogenetic Error. *Trends in Genetics*, 21(6):310–311, 2005.

Monica G Turner, Robert H Gardner, Robert V O'neill, et al. *Landscape Ecology in Theory and Practice*, volume 401. Springer, 2001.

GA Watterson. On the Number of Segregating Sites in Genetical Models without Recombination. *Theoretical population biology*, 7(2):256–276, 1975.

George H Weiss and Motoo Kimura. A Mathematical Analysis of the Stepping Stone Model of Genetic Correlation. *Journal of Applied Probability*, 2(1):129–149, 1965.

Jon F Wilkins. A Separation-of-Timescales Approach to the Coalescent in a Continuous Population. *Genetics*, 168(4):2227–2244, 2004.

Gregory A Wilson and Bruce Rannala. Bayesian Inference of Recent Migration Rates using Multilocus Genotypes. *Genetics*, 163(3):1177–1191, 2003.

William H Womble. Differential systematics. *Science*, 114(2961):315–322, 1951.

Wendy SW Wong, Ziheng Yang, Nick Goldman, and Rasmus Nielsen. Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. *Genetics*, 168(2):1041–1051, 2004.

Sewall Wright. Isolation by Distance. *Genetics*, 28(2):114, 1943.

Ziheng Yang. Among-site rate Variation and its Impact on Phylogenetic Analyses. *Trends in Ecology & Evolution*, 11(9):367–372, 1996.

Ziheng Yang. Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution. *Molecular Biology and Evolution*, 15(5):568–573, 1998.

Ziheng Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

Ziheng Yang. *Molecular Evolution: a Statistical Approach*. Oxford University Press, 2014.

Ziheng Yang and Joseph P Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12):496–503, 2000.

Shozo Yokoyama, Takashi Tada, Huan Zhang, and Lyle Britt. Elucidation of Phenotypic Adaptations: Molecular Analyses of Dim-Light Vision Proteins in Vertebrates. *Proceedings of the National Academy of Sciences*, 105(36):13480–13485, 2008.

Weiwei Zhai, Rasmus Nielsen, Nick Goldman, and Ziheng Yang. Looking for Darwin in Genomic Sequences Validity and Success of Statistical Methods. *Molecular Biology and Evolution*, 29(10):2889–2893, 2012.

Shancen Zhao, Pingping Zheng, Shanshan Dong, Xiangjiang Zhan, Qi Wu, Xiaosen Guo, Yibo Hu, Weiming He, Shanning Zhang, Wei Fan, et al. Whole-Genome Sequencing of Giant Pandas Provides Insights into Demographic History and Local Adaptation. *Nature Genetics*, 45(1):67, 2013.

Andreas Zimmermann, Johanna Hilpert, and Karl Peter Wendt. Estimations of Population Density for Selected Periods between the Neolithic and AD 1800. *Human Biology*, 81(3): 357–380, 2009.

Derrick J Zwickl and David M Hillis. Increased Taxon Sampling Greatly reduces Phylogenetic Error. *Systematic Biology*, 51(4):588–598, 2002.

# APPENDIX: SUPPLEMENTARY FIGURES

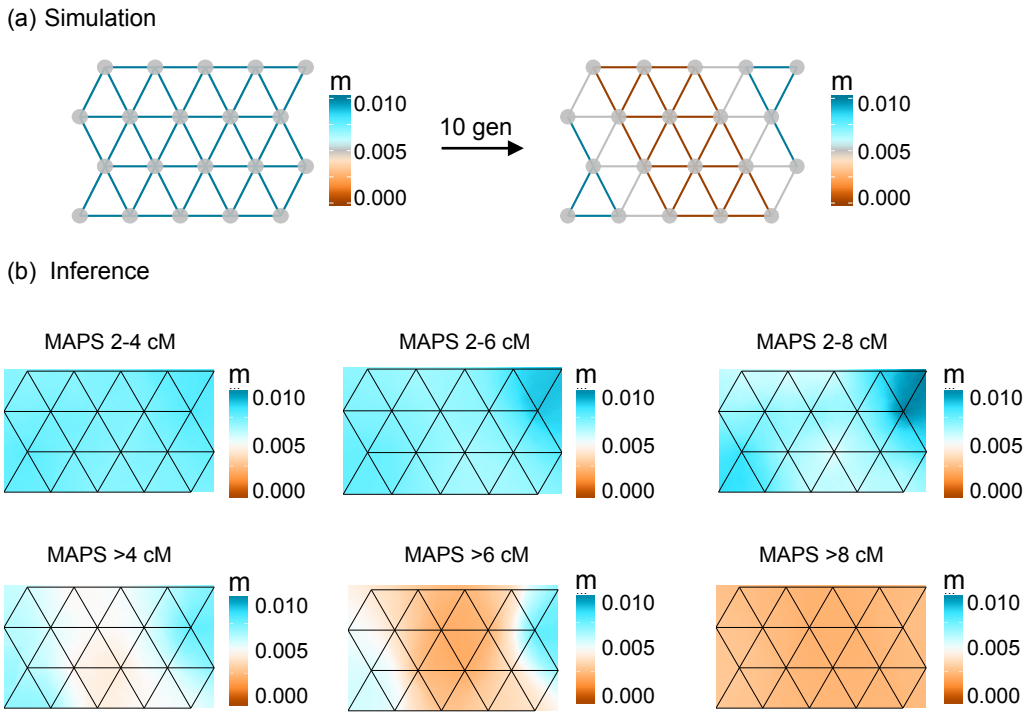



Figure S1: **The performance of MAPS on a recent barrier scenario under different PSC length bins.** Here, we investigate the ability of MAPS to detect a recent barrier (< 10 generations) for various PSC length bins (a) Simulation scenario. Population sizes were set to 10,000 per deme and 10 diploids were sampled per deme, replicating the conditions in Figure 2.2b. (b) Results for different PSC length bins. Length bins that encompass shorter segments (2-4cM 2-6cM 2-8cM) recover the higher uniform migration surface; while length bins with longer segments (>4, >6, >8) recover the recent ancestral barrier. For the last length scale (> 8cM), the signature of low migration extends across the habitat. The variation in migration rates is missed presumably because of the small number of shared segments at this length scale.

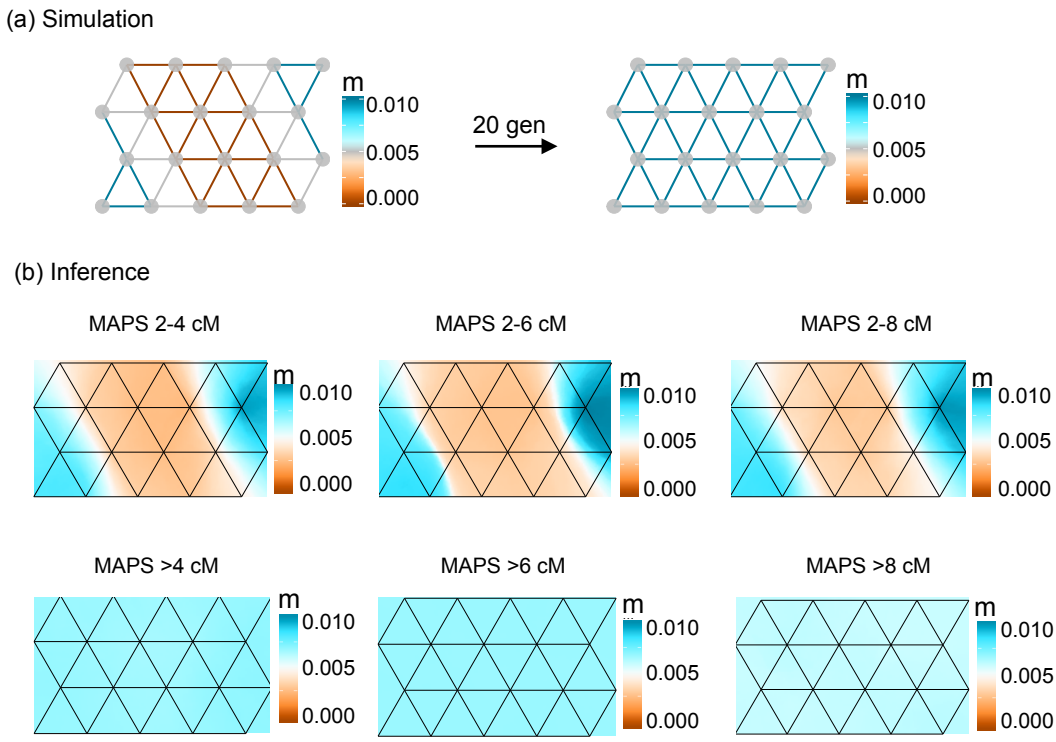



Figure S2: **The performance of MAPS on a past barrier scenario under different PSC length bins.** a) Simulation scenario. Population sizes were set to 10000 per deme and 10 diploids were sampled per deme, replicating the conditions in Figure 2.2c. (b) Results for different PSC length bins. Length bins that encompass shorter segments (2-4cM, 2-6cM, 2-8cM) recover the ancestral barrier; while length bins with longer segments (>4, >6, >8) recover the recent constant migration surface.

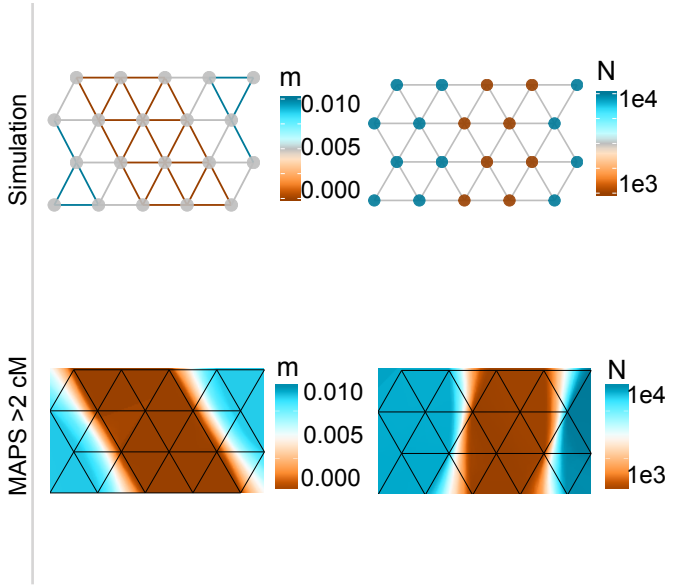



Figure S3: **The performance of MAPS under a jointly heterogeneous migration rate and population size surface.** a) Simulation Scenario. Heterogeneous population-sizes and migration rates (as shown) were simulated, and 10 diploid individuals were sampled per deme. (b) Results for PSC segments greater than 2cM are shown.

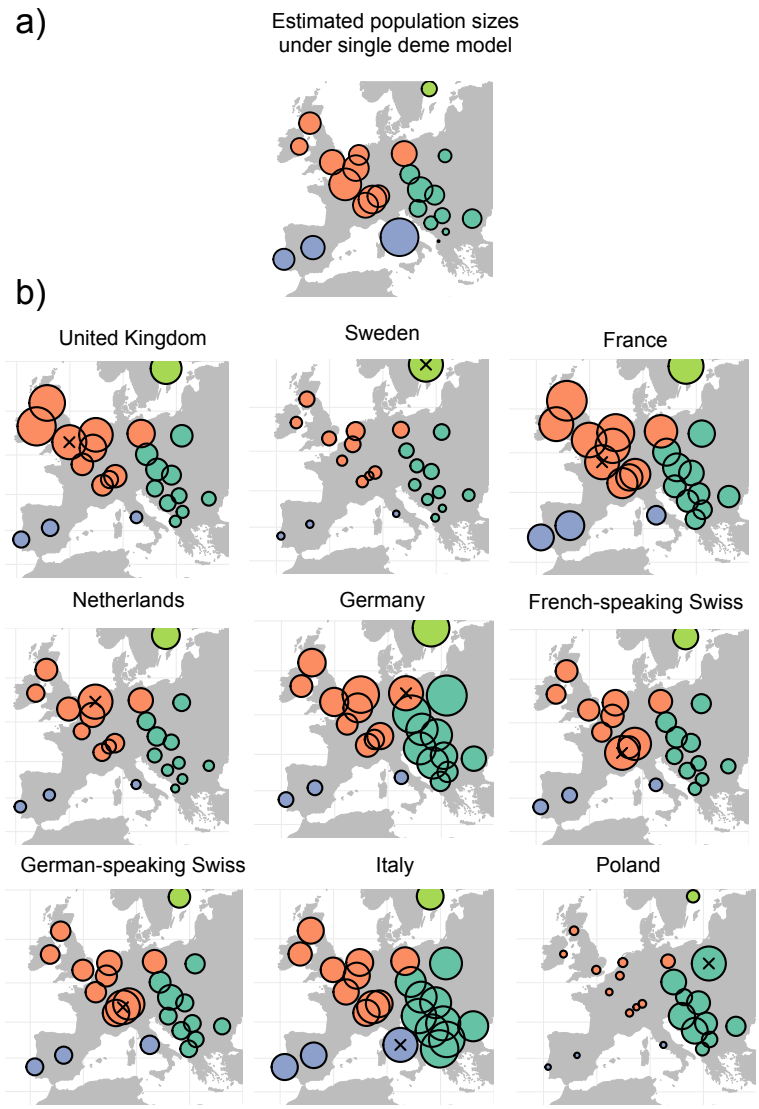



Figure S4: **Visualizing normalized sharing of PSC segments that are 1-5cM.** The color scheme is the same as used in Ralph and Coop [2013] where the colors give categories based on the regional groupings: W Western Europe, S Southern Europe, and E Eastern Europe (a) The average sharing within each sample locale is transformed to population sizes using the simple single deme estimator by Palamara et al. [2012]. This transformation can be roughly summarized as to say that $N_\alpha \propto \frac{1}{\bar{x}_{\alpha,\alpha}}$ where $N_\alpha$ is the effective population size in deme $\alpha$ and $\bar{x}_{\alpha,\alpha}$ is the average pairwise PSC sharing between individuals in deme $\alpha$. (b) Similar to Ralph and Coop [2013], for each focal population (marked with an x), we plot the normalized average pairwise sharing between that population and all others (normalized by the average sharing within the focal population), i.e. if $\alpha$ is the focal population, we show $\frac{\bar{x}_{\alpha,\beta}}{\bar{x}_{\alpha,\alpha}}$ for each other country $\beta$.

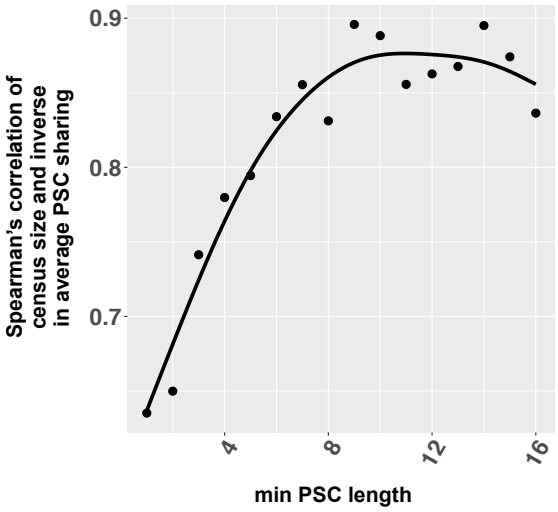



Figure S5: **The correlation between census size and inverse average PSC sharing as a function of minimum PSC length considered.** We use census size compiled from the The World Bank [2016] and National Records of Scotland [2011]. The smooth black curve denotes the loess fit. Longer PSC segments correlate more strongly with census size than shorter PSC segments

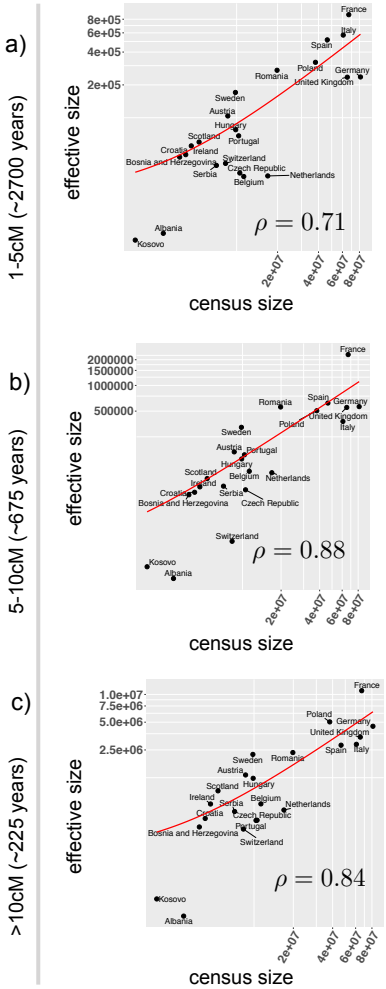



Figure S6: **Census size versus MAPS estimated population sizes.** Using the MAPS output, we estimate a total size per population by summing the estimated deme-level sizes across the area of each respective country (whether's a deme's location falls within a country was determined by querying The GeoNames Geographical Database). Finally, we plot the results on a log10 scale for different length scales (a) 1-5cM, (b) 5-10cM, and (c) >10cM. The red curve denotes the linear fit on the absolute scale.

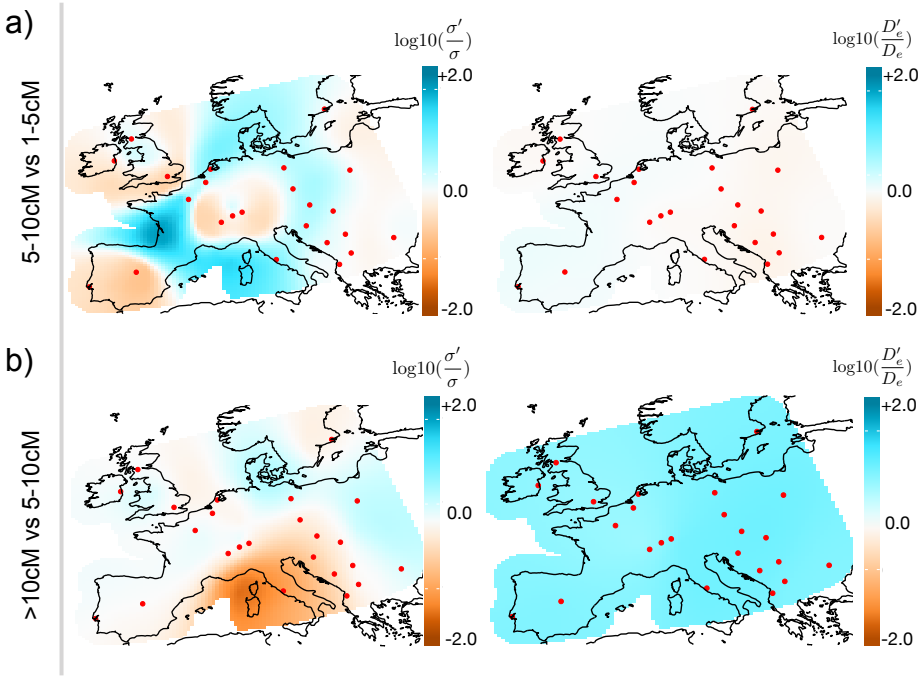



Figure S7: **Plots of estimated average log10 differences in demographic parameters between adjacent time scales**. (a) We plot estimates of $E[\log10(\frac{\sigma'}{\sigma})]$ and $E[\log10(\frac{D_e{}'}{D_e})]$ across the spatial habitat where $\sigma'$ ($D_e'$) denotes the dispersal rates (population densities) in the 5-10cM length bin and $\sigma$ ($D_e$) denotes the dispersal rates (population densities) in the 1-5cM length bin. (b) The results here are similarly plotted as above, however, the adjacent length scales are given by: 5-10cM and >10cM. The log10 differences are estimated in such a way so that the mean log10 difference is shrunk to zero. For example, for estimating dispersal in 5-10cM, we assume $\log10(\sigma') = E[\log10(\sigma)] + \epsilon$ where $E[\log10(\sigma)]$ is estimated using PSC segments 1-5cM and $\epsilon \sim N(0, \omega^2)$ is estimated from PSC segments 5-10cM. Consequently, the log ratio between dispersal rates from the two lengths bins is constructed to have mean zero *apriori* (i.e. $E[\log10(\frac{\sigma'}{\sigma})] = 0$).

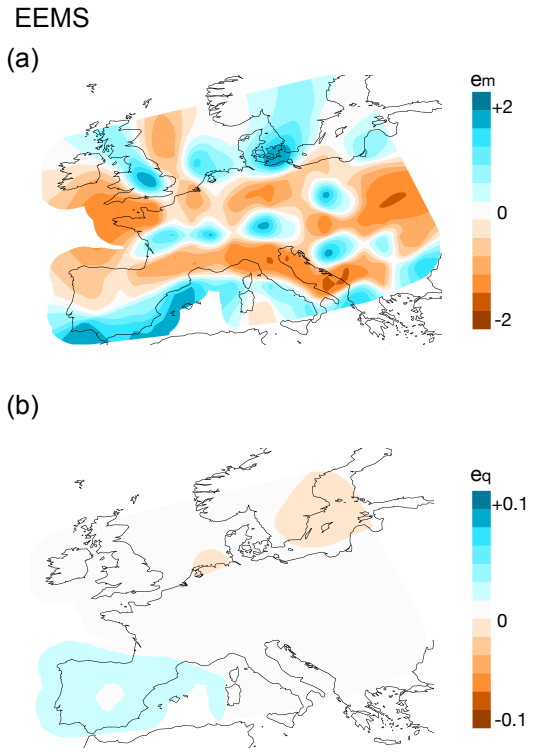
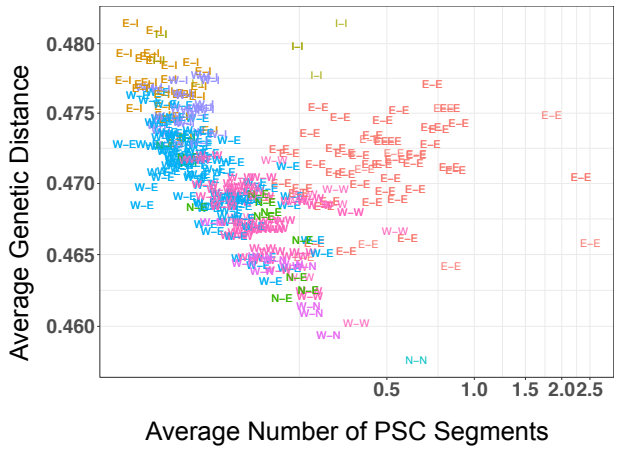



Figure S8: **EEMS applied to the POPRES dataset.** We apply EEMS to the same set of individuals as used in Figure 2.4 (see Methods). (a) The effective migration rates (b) The effective diversity rates. Here, we ran EEMS with 200 demes (as in Figure 2.4) with default parameters and averaged over 10 independent replicate chains. Each chain ran with 50e6 MCMC iterations, 25e6 set as burn-in, and we thinned every 5000 iterations.

(a)




(b)

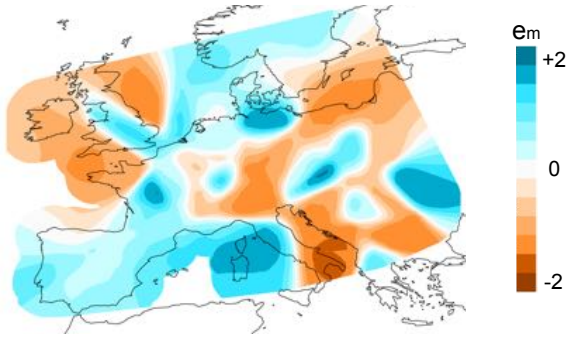



Figure S9: **Genetic distance vs PSC sharing** (a) The averaged genetic distance (as used in EEMS) is plotted against the average number of PSC segments (> 1cM) for each pair of populations. Each point denotes a pair, the symbols represent groupings from Ralph and Coop [2013] (W Western Europe, S Southern Europe, and E Eastern Europe), and the colors represent the pair of regions. We see a negative correlation between the two summary statistics (Pearson's $\rho$ = -0.38, p-value = 7e-11), with the largest deviations occurring in comparisons between Eastern European populations. (b) EEMS results on PSC data transformed to a distance matrix. First, we encoded the PSC sharing statistics into a similarity matrix $S$ such that $S_{i,j}$ is the number of shared PSC segments between samples i and j and $S_{i,i}$ is the maximum number of shared segments in the dataset (which we denote as $c$) to ensure $S$ is a similarity matrix. Next, we transformed $S$ to a genetic distance matrix $D$ such that $D = c11^T - S + E$ where $E \approx 0$ is a random genetic distance matrix of normal vectors with mean 0 and standard deviation of 0.01 added to ensure $D$ is full rank. Finally, we applied EEMS to the distance matrix $D$. Though this procedure is heuristic, we see shared features between this surface and the MAPS dispersal surface shown in Figure 2.4.

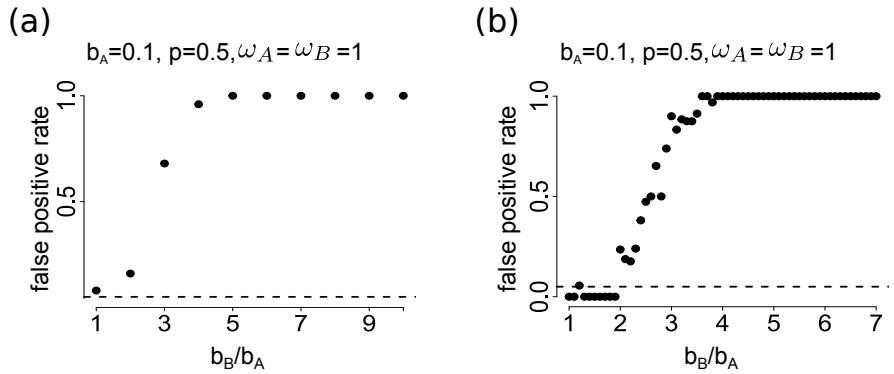



Figure S10: **The Sites Test is biased towards false positives under different implementations and variations.** (a) Rate heterogeneity causes false positive inferences by the sites test as implemented in Hyphy. Sequences were simulated under the evolutionary conditions shown and then analyzed using the Nielsen-Yang method as implemented in the Hyphy software package. The false positive rate out of 25 repetitions per conditions is plotted as a function of $b_B/b_A$. (b) Rate heterogeneity causes high false positives in the sites test when more complex models are used. Sequences were simulated under the conditions shown and analyzed using the sites test and models M7 vs. M8 as implemented in PAML software. Each false positive rate is calculated out of 40 repetitions.

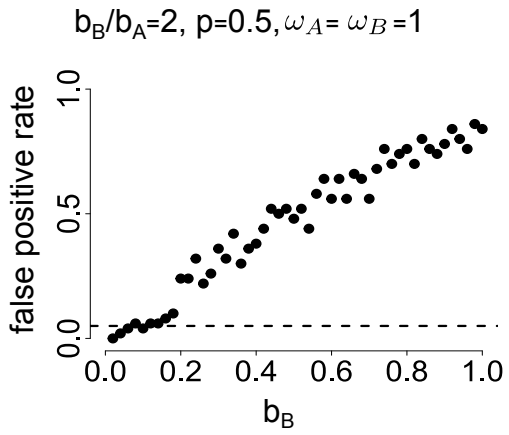



Figure S11: **Rate heterogeneity with longer branches increases the sites tests false positive inference rate.** The rate of false inferences of positive selection is shown with constant rate heterogeneity $b_B/b_A$ and increasing branch lengths in both partitions $b_B$ under the conditions shown. The false positive rate is calculated as the number of positive inferences out of 50 replicate alignments analyzed for each conditions

| Codon state pattern category (nucleotide differences) | Likelihood of submodel | | | | |
|---|---|---|---|---|---|
| | Class 0 (ω=0) | | Class 1 (ω=1) | | Class 2 (ω=25) |
| | M1a(*p*=0.18) | M2a (0.09) | M1a (0.82) | M2a (0.82) | M2a (0.09) |
| 0 | 0.80 | 0.85 | 0.41 | 0.52 | 0.03 |
| 0N,1S | 0.20 | 0.15 | 0.11 | 0.10 | 0.03 |
| 1N,0S | - | - | 0.32 | 0.28 | 0.15 |
| 2 | 0.002 | 0.001 | 0.15 | 0.09 | 0.44 |
| 3 | - | - | 0.02 | 0.01 | 0.35 |

Figure S12: **Site-specific likelihoods of submodels comprised by M1a and M2a. For each category of state pattern, the likelihoods of each mixture models submodels are shown, summed over all instances in the alignment.** Each column represents submodel 0, 1, or 2 in models M1a or M2a. Values of $\omega$ and the mixing proportion $p$ for each submodel are shown. The generating conditions are the rate-heterogeneous conditions with $\omega = 1$ as specified in Figure 3.5a. A dash indicates likelihood $< 0.001$.

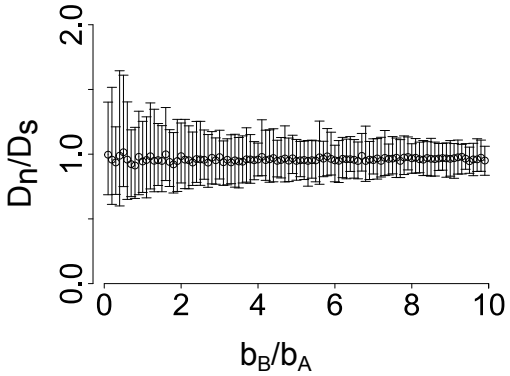



Figure S13: **The non-parametric counting method of Nei & Gojobori, 1986 is robust to model violation.** 50 replicate alignments at each condition were generated by evolutionary simulation with $p$=0.5, $\omega = 1$, $N$=996, and number of taxa=4. We implemented and applied the dN/dS inference method described by Nei & Gojobori, 1986. Confidence intervals were estimated by bootstrapping.

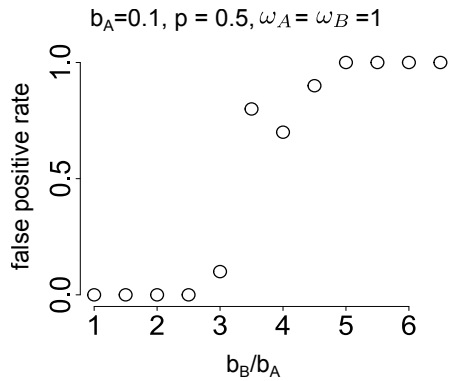



Figure S14: **Rate heterogeneity causes false inferences of positive selection in the BUSTED test of positive selection.** 10 replicate alignments at each condition were generated by evolutionary simulation with $p$=0.5, $\omega$=1, $N$=996, $b_A$=0.1, and number of taxa=4. We applied MEME to test for gene-wide positive selection. The false positive increases as the degree of rate hetereogeneity ($b_B/b_A$) increases.

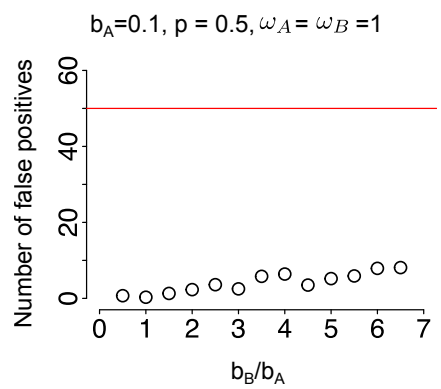



Figure S15: **Rate heterogeneity does not cause an elevated false positive rate in the MEME test of positive selection. 10 replicate alignments at each condition were generated by evolutionary simulation with $p$=0.5, $\omega$=1, $N$=996, $b_A$=0.1, and number of taxa=4.** We applied MEME to test for positive selection at each site. Here, we plot the average (over 10 replicates) number of sites inferred to be under positive selection at a significant level of 0.05. Under the null hypothesis, we expect $996 * 0.05 \approx 50$ sites to have a p-value $< 0.05$ (red line). The number of sites at that significance level increases as the degree of rate heterogeneity ($b_B/b_A$) increases but is far below the 50.

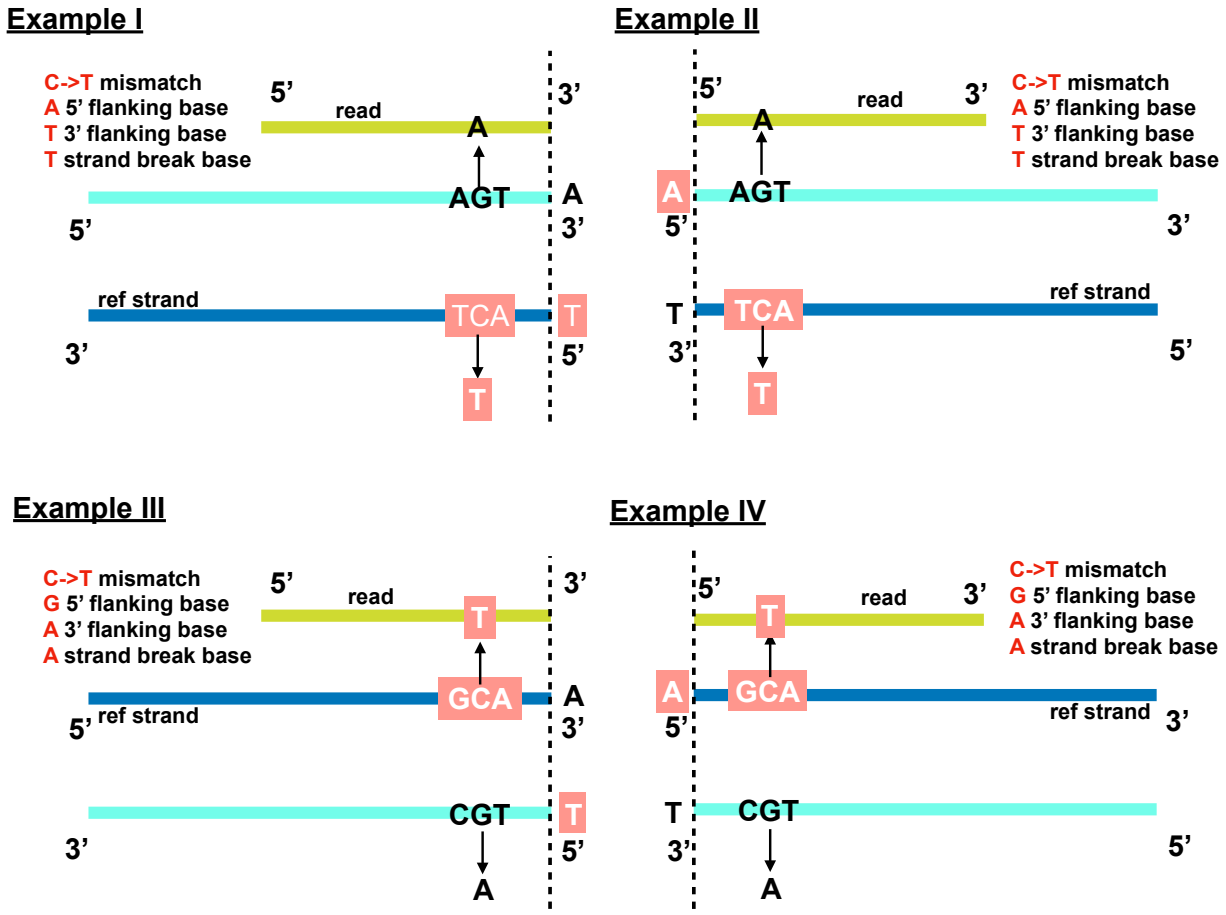



Figure S16: **Four examples of how the reference strand for a mismatch is designated.** The dark yellow line denotes the mapped read, and the blue and teal line represent the reference genome at two different strand orientations. The reference strand is always designated as the strand that contains the C mismatch or T mismatch (latter not shown).

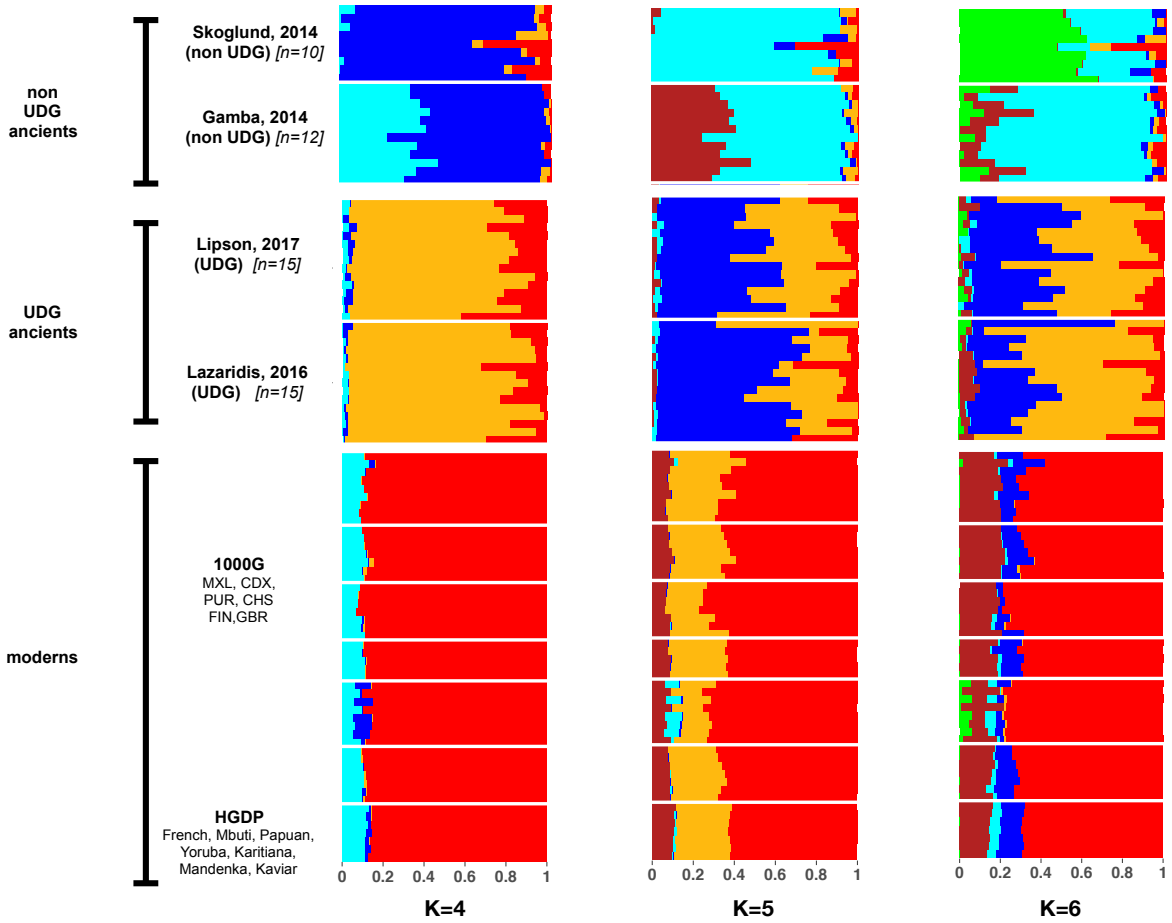



Figure S17: `aRchaic` **grades of membership for the example in Fig 4.1 corresponding to 3 different values of** $K$ **(**$K = 4, 5, 6$**).** Higher values of $K$ distinguish among the ancient studies, reflecting lab and study specific biases.

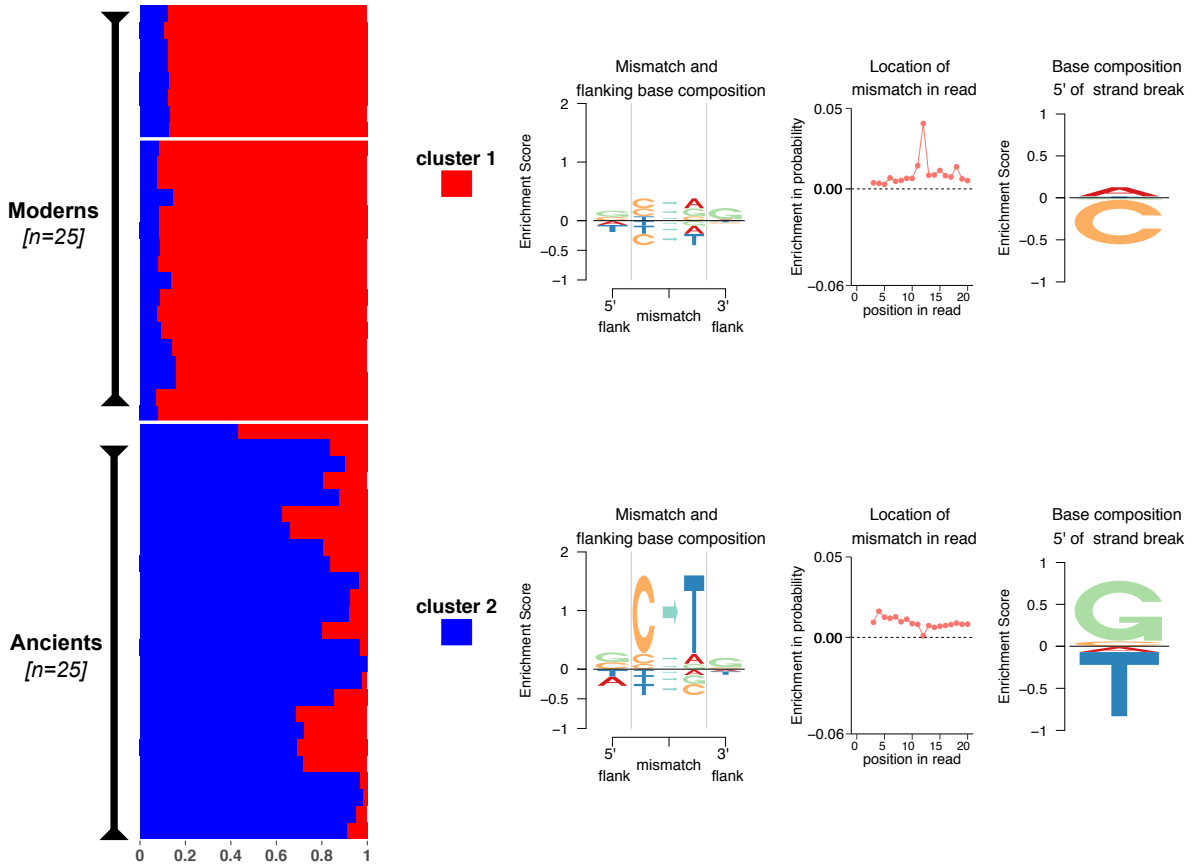



Figure S18: `aRchaic` **plot for** $K = 2$ **on the combined data of 25 moderns and 25 ancients from Lindo et al. [2016].** `aRchaic` clearly distinguishes the moderns from the ancients. The ancients are primarily presented by the blue cluster. This cluster shows an enrichment of C-to-T mismatches and depletion of T-to-C mismatches with respect to modern background, as well as enrichment of G and depletion of T at the 5' strand break. The red cluster shows a blip at 12th position from the end of the read, the explanation for which is provided in Figure S20.

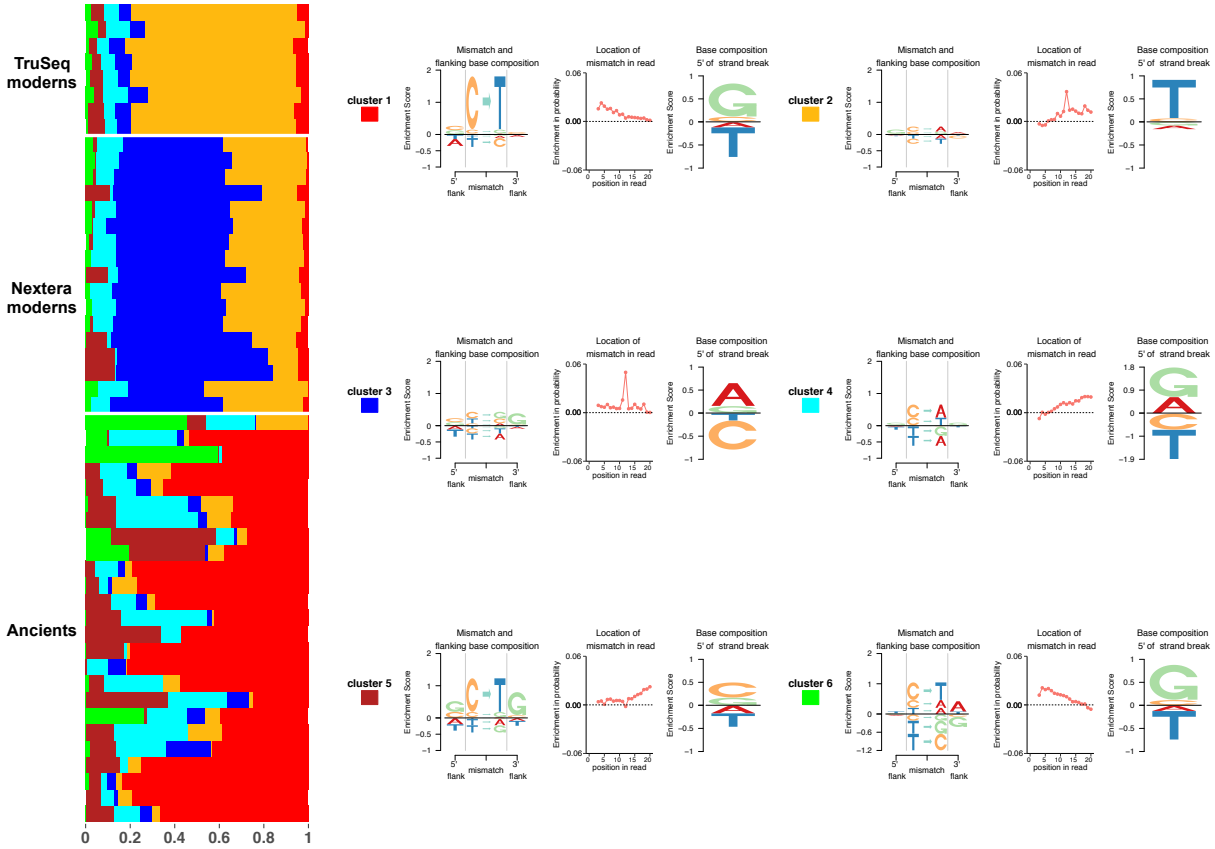



Figure S19: **We apply `aRchaic` with $K = 6$ on the data from Fig 4.4**. In addition to separating out the ancients from the moderns, `aRchaic` now distinguishes between moderns individuals based on library kit (Nextera vs Tru-seq.)

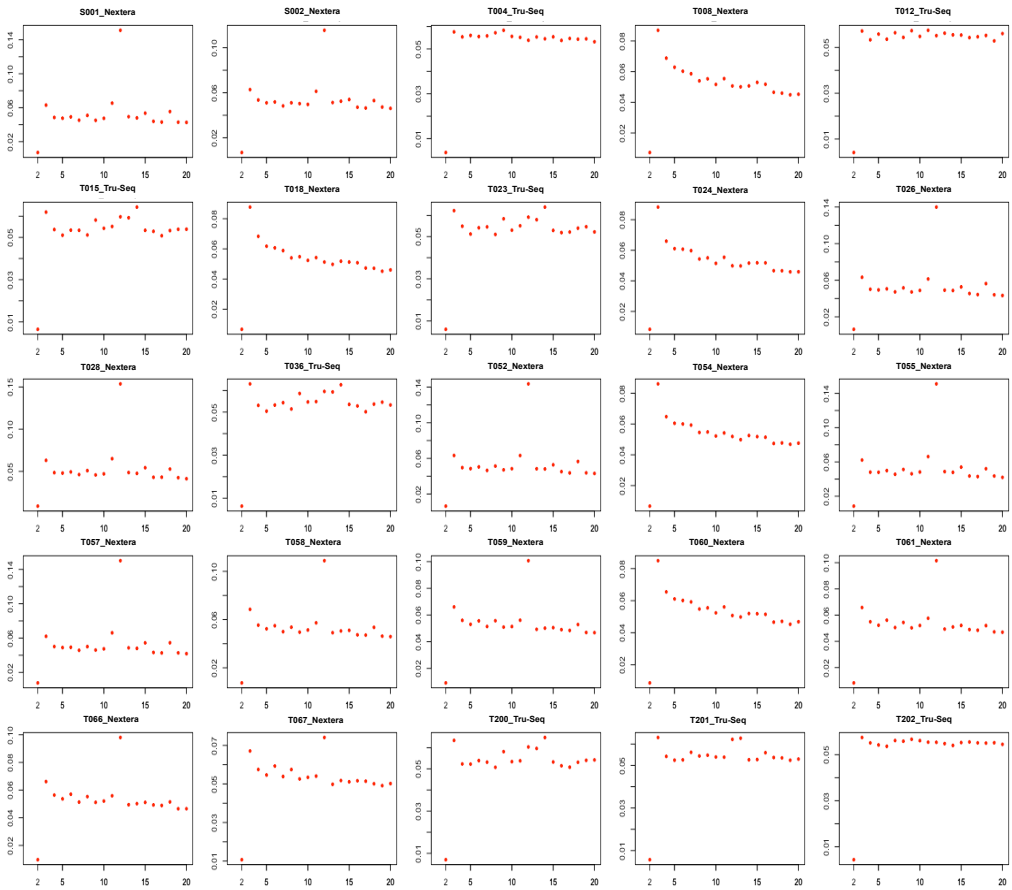



Figure S20: **The frequency of all mismatch types plotted against the position of the read (from the 5' end) for each of the 25 moderns samples in Lindo et al. [2016].** Each sample was prepared by one of two library kits: Nextera and TruSeq. Most of the samples prepared with the Nextera kit show a spike in frequency at the 12th position from the 5' end of the read