Supplementary Materials

Vince Buffalo and Andrew Kern

June 13, 2023

1 Human Genomic Data

1.1 YRI Samples

In order to try to ensure accurate estimation of pairwise diversity, which is a ratio estimator that is sensitivity to its denominator, we use the complete per-basepair genotype calls (gVCF) produced by Illumina's DRAGEN pipeline (Illumina, Inc. 2020). The original samples were from 178 Yoruban individuals sequenced to 30x by the New York Genome Center (Byrska-Bishop et al. 2022). This allows for filtering to be applied to the entire genome at once, rather than just variants, so the denominator does not need to be estimated separately.

The full list of samples is available in TSV format in the GitHub repository (`data/h1kg/yri_samples.tsv`).

1.2 Filtering gVCFs

gVCFs were filtered using a custom Python tool, `gvcf2counts.py` (in `tools/gvcf2counts.py`), which reads the gVCFs, filters them according to the criteria below, and outputs a Numpy `.npy` file of reference and alternative allele counts for each chromosome (hereafter, "allele counts matrices").

Genotypes are included in the allele count if and only if:

1. The variant call is set to `PASS` in the VCF.
2. The `QUAL > 50`.
3. The `GQ > 30` (or `RGQ > 30` for invariant sites).

Because the data underlying the counts files are per-basepair resolution gVCFs, each chromosome's allele counts matrix is of size $l \times 2$, where l is the total chromosome length. Basepairs that fail these filtering requirements lead to a row of zero counts, e.g. no observed reference and alternative allele counts, and thus do not effect the data that goes into the binomial likelihood or π estimates used in figures.

1.3 Site-based Filtering of Counts

The allele counts matrices include many basepairs that may have allele counts that pass the genotype call filters, but are still need to be filtered out because the region of the genome may produce unreliable estimates. The following filters are applied based on masking regions:

1. **Non-accessible regions**: masks out centromeres (`acen` entries in `cytoBand.txt`), with 5Mbp padding on either side. The file of passing masks is `data/annotation/no_centro.bed`.

2. **Reference masking**: soft and hard-masked regions in the human GRCh38 reference genome are also masked. Soft-masked regions were determined by Ensembl (XXX), which uses Repeat Masker (XXX) and Dust (XXX).

3. **Non-"putatively" neutral regions**: Additionally, for fitting our likelihood and estimating observed pairwise diversity, we only consider . This masks out phastCons regions (from `phastConsElements100way.bed.gz`) and Ensembl gene regions (from annotation file `Homo_sapiens.GRCh38.107.chr.gff3.gz`). While introns are possibly under some weak selection, they collectively make up nearly 40% of the human genome and are included so genome-wide diversity can be estimated more precisely (possibly at the expense of some bias).

These files are all produced by the Snakemake file `data/annotation/Snakefile`. Note that the levels of accessible putatively neutral bases

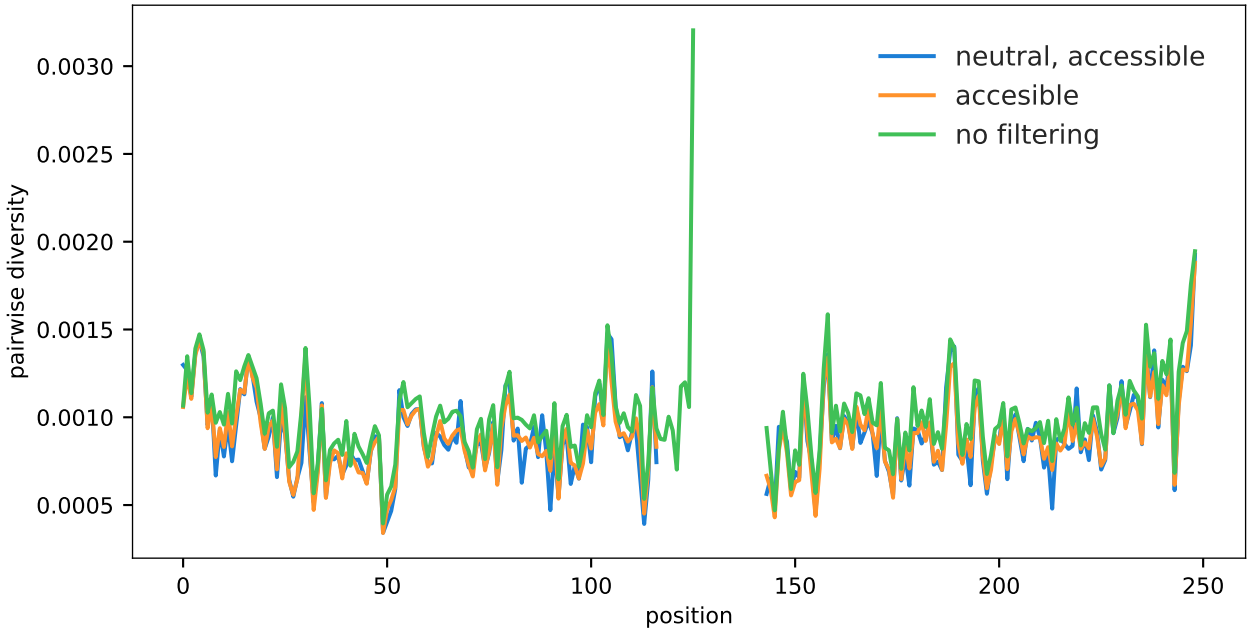

Figure 1: Estimates of chromosome 1 YRI diversity in non-overlapping megabase windows, under different filtering criteria. The filtering criteria are, (1) "neutral, accessible" which includes only putatively neutral sites, and ignores regions masked as inaccessible, (2) "accessible" which is only ignores sites masked as inaccessible, and (3) "no filtering" which uses all available data. Note that filtering changes the absolute level of diversity, but has more minor effects on regional patterns of diversity at the megabase scale.

chrom	accessible	neutral	both
chr1	42.4	64.4	22.9
chr2	46.8	58.7	23.7
chr3	44.1	62.9	24.8
chr4	44.0	59.3	21.5
chr5	44.1	58.0	21.4
chr6	45.0	58.0	22.1
chr7	44.4	65.8	26.4
chr8	43.8	61.2	23.1
chr9	39.2	64.5	20.3
chr10	44.6	62.8	25.3
chr11	42.7	63.5	23.4
chr12	41.7	62.7	22.7
chr13	39.7	62.0	18.3
chr14	37.6	65.4	19.0
chr15	37.0	69.4	20.3
chr16	39.4	63.8	20.9
chr17	40.4	64.0	22.3
chr18	40.9	59.8	19.8
chr19	30.9	69.8	18.0
chr20	36.9	61.6	19.8
chr21	31.6	68.6	18.2
chr22	26.9	75.0	16.4

1.4 Data Summary Matrices

Our underlying data for all likelihood and pairwise diversity estimates is the allele count matrix $\mathbf{C}$ with dimensions $L \times 2$, where L is the chromosome length. This is transformed to a pairwise summary matrix with identical dimensions, $\mathbf{Y}$. The first column of $\mathbf{Y}$ is the number of pairwise comparisons between chromosomes that are identical, and the second column is the number that are different. Both of these columns are combinatoric summaries of the raw allele counts matrix needed for the binomial likelihood and pairwise diversity estimates. Let $[c_1, c_2]$ be a row of $\mathbf{C}$ for basepair l (the l index is omitted for clarity), and $n = c_1 + c_2$. Then, the (1) total number of pairwise combinations of chromosomes n_T, (2) the number of pairwise with identical alleles n_S, and (3) the number of pairwise combinations with differing alleles n_D are respectively,

$$\begin{aligned} n_T &= \frac{n(n-1)}{2} \\ n_S &= \binom{c_1}{2} + \binom{c_2}{2} \\ n_D &= n_T - n_S \end{aligned}$$

which would be stored in row $\mathbf{Y}_l = [n_S, n_D]$. Note that the per-site $\mathbf{Y}$ handles non-polymorphic sites and missing data. Non-polymorphic sites have $n_S = \binom{n}{2}$ and $n_D = 0$, and missing data has $n_S = n_D = 0$.

1.5 Pairwise Diversity Estimates

The pairwise diversity at site l across the n sampled chromosomes can be calculated from row l of the $\mathbf{Y}$ matrix as follows,

$$\pi_l = \frac{n_D}{n_T} \tag{1}$$

which is identical to the more common expression of this estimator,

$$\pi_l = \frac{2}{n(n-1)} \sum_{i<j}^{n} k_{i,j} \tag{2}$$

where $k_{i,j}$ is 1 if the alleles at this site differ at site l, and 0 otherwise.
There are three ways to aggregate π_l across all sites. The first is,

$$\pi^{(1)} = \frac{1}{L} \sum_{i=1}^{L} \frac{n_{D,i}}{n_{T,i}} \tag{3}$$

which if the number of samples across loci is constant, simplifies to an unweighted average across sites. Second, one can take a weighted average, with weights determined by the total number of samples present at a site,

$$\pi^{(2)} = \frac{1}{\sum_{i=1}^{L} n_i} \sum_{i=1}^{L} n_i \frac{n_{D,i}}{n_{T,i}} \tag{4}$$

Third, one can weight by the number of pairwise comparisons at a site, $n_{T,i}$, rather than total number of samples, n_i, which leads to a ratio of sums,

$$\pi^{(3)} = \frac{\sum_{i=1}^{L} n_{D,i}}{\sum_{i=1}^{L} n_{T,i}}. \tag{5}$$

We predominantly use the estimator $\pi^{(3)}$, as it corresponds to how we summarize the matrix $\mathbf{Y}$ across windows for our likelihood. All methods have mean squared errors and biases very close to one another (TODO).

Note, however, that estimates of pairwise diversity often condition on the accessible bases, and thus treat this as fixed. However, the number of accessible bases varies across the chromosome; this can lead to a source of apparent bias during block-bootstrap estimates of uncertainty. In this case, pairwise diversity is a ratio estimator, and is thus biased, since by Jensen's inequality $\mathbb{E}(y/x) \geq \mathbb{E}(y)/\mathbb{E}(x)$ for random variables x and y.

1.6 Window-based Summaries and filtering

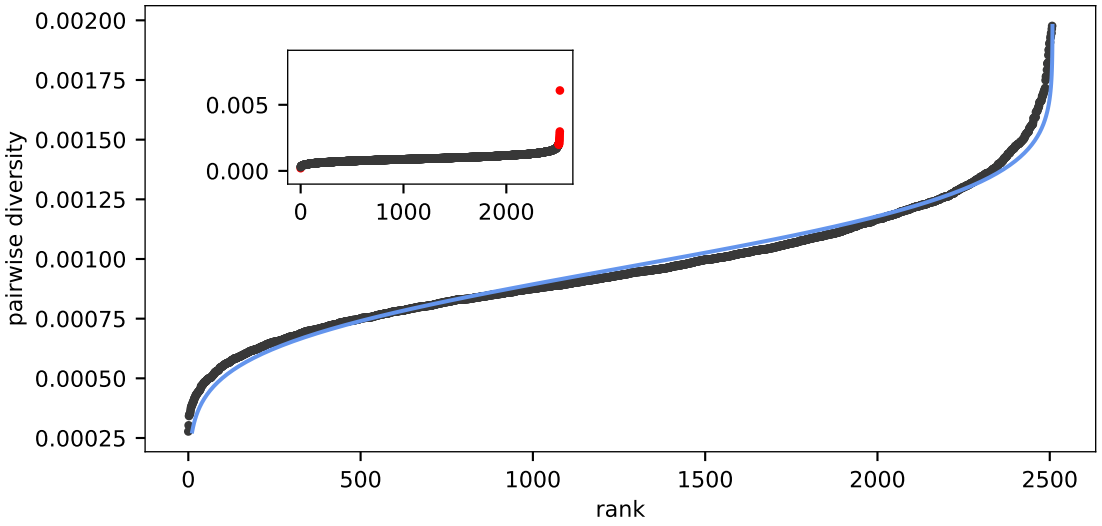

Figure 2: The distribution of diversity across the genome at the megabase scale, with outliers trimmed. The blue line is the normal CDF, fit with MLE parameters XXX. The inset figure is the untrimmed data, with trimmed points shown in red.

The likelihood method is fit to non-overlapping binned summaries of the allele counts matrix. Based on exploratory data analyses, some bins were outliers and excluded (XXX).

Table 1: R^2 and mutation rate estimates for all models. XXX note about repeats

Model	Track type	Pop.	B' R^2_{LOO}	B' R^2_{IS}	B R^2_{IS}	B' $\hat{\mu} \times 10^{-8}$	B $\hat{\mu} \times 10^{-8}$
phastcons>CDS>genes	sparse	YRI	**68.45**	68.16	65.23	1.8636	0.2975
phastcons>CDS>genes	full	YRI	68.19	68.12	63.67	1.7082	0.1666
CADD 6%	full	YRI	68.08	68.08	62.96	1.4612	0.1721
CADD 6%	sparse	YRI	68.04	68.03	68.01	2.0912	2.0716
CADD 8%	full	YRI	66.98	67.11	63.61	1.0126	0.1713
CADD 8%	sparse	YRI	66.67	67.01	66.98	1.5688	1.5473
CDS>genes>phastcons	full	YRI	64.90	65.51	61.23	4.1951	0.1620
CDS>genes>phastcons	sparse	YRI	64.90	65.51	63.16	4.1795	0.3070
phastcons>CDS>genes	sparse	CEU	**61.98**	63.02	60.02	2.0194	0.3036
phastcons>CDS>genes	full	CEU	61.78	63.01	58.57	2.3499	0.1679
CADD 6%	full	CEU	61.74	62.94	57.42	1.5365	0.1739
CADD 6%	sparse	CEU	61.66	62.86	62.84	2.1215	2.1084
CADD 8%	full	CEU	60.80	62.12	57.95	1.1572	0.1742
CADD 8%	sparse	CEU	60.59	62.06	62.04	1.6030	1.5876
CADD 6%	full	CHB	59.62	61.31	55.31	1.5083	0.1727
CADD 6%	sparse	CHB	59.54	61.21	61.20	2.1242	2.1170
phastcons>CDS>genes	sparse	CHB	**59.44**	61.08	57.88	2.1982	0.3060
phastcons>CDS>genes	full	CHB	59.29	61.08	56.71	2.5573	0.1694
CADD 8%	full	CHB	58.88	60.63	55.86	1.1128	0.1733
CADD 8%	sparse	CHB	58.65	60.57	60.55	1.6098	1.6001
CDS>genes>phastcons	sparse	CEU	58.58	60.30	58.41	4.1041	0.3110
CDS>genes>phastcons	full	CEU	58.58	60.30	56.75	4.1001	0.1648
CDS>genes>phastcons	sparse	CHB	56.48	58.73	56.60	3.7195	0.3136
CDS>genes>phastcons	full	CHB	56.48	58.73	54.95	3.7552	0.1654

1. **Fraction of inaccessible sites per window**. `mask_inaccessible_bins_frac`

2. **Outliers**: Based on exploratory analysis, there were some regions with very high diversity. The 0.05% right tail was excluded.

Additional Results

Model Comparisons

Table of R^2 Values

Derivation of Approximate Drift R^2

Our approach models the observed megabase-scale diversity y_i in window i as the predicted level under our negative selection model, $\pi_0 B_i$, plus some random residual error ε_i,

$$y_i = \pi_0 B_i + \varepsilon_i \tag{6}$$

where each $B_i = f(X_i|\Psi)$ for some data X_i (i.e. the recombinaton map and annotation model) and parameters Ψ. To compare our models and assess model goodness-of-fit, we use the out-sample R^2_{LOO} which is calculated by fitting the genome-wide model leaving one chromosome out, and calculating the residual variance between predictions and observed values for this out-sample chromosome.

In general, R^2 is calculated for a model fit across windows as,

$$R^2 = 1 - \frac{V_{\text{res}}}{V_{\text{tot}}} \tag{7}$$

where V_{res} is the mean squared error across windows and V_{tot} is the total variance in the predicator. Suppose we have estimates $\widehat{\pi}_0 \widehat{B}_i$ for $\pi_0 B_i$ from our model. Then,

$$V_{\text{res}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\pi}_0\widehat{B}_i)^2 \tag{8}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2}_{\text{irreducible error}} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}(\pi_0 B_i - \widehat{\pi}_0\widehat{B}_i)\right)^2}_{\text{bias squared}}. \tag{9}$$

The total variance in the predictor is,

$$V_{\text{tot}} = V_{\text{res}} + V_{\text{model}} + 2\widehat{\pi}_0 \operatorname{cov}_i(\varepsilon_i, \widehat{B}_i) \tag{10}$$

where

$$V_{\text{model}} = \frac{\widehat{\pi}_0^2}{n}\sum_{i=1}^{n}\left(\widehat{B}_i - \frac{1}{n}\sum_{i=1}^{n}\widehat{B}_i\right)^2. \tag{11}$$

Generally, fitting procedures act to minimize the squared deviation between predictors and observed values, which drives $\operatorname{cov}_i(\varepsilon_i, \widehat{B}_i)$ to zero XXX. Murphy et al. (2022) suggest XXX

Thus, our R^2 depends on (1) the bias of our predictors, (2) the covariance between residuals and predictors, and (3) the irreducible error due to variance in diversity within windows. However, it is of interest to approximate the irreducible error under theoretic models that assume all irreducible error is due entirely to the variance of coalescence times in a window, assuming that all of the effects of negative selection on the variance can be thought of as a local reduction in the effective population size to $B_i N_e$ and demographic effects amount to a simple rescaling of N_e. This also assumes mutation rates are constant across the genome, and do not contribute to the irreducible variance. We call this theoretic goodness-of-fit under drift R^2_{drift}, and it provides a rough approximation of the irreducible error in our model due to "coalescence noise" around the modeled reductions due to negative selection.

Then,

$$\mathbb{E}(\varepsilon_i^2) = \text{var}(\pi_i) \tag{12}$$

since the mean residual is zero. The variance in pairwise diversity $\text{var}(\pi_i)$ is equivalent to the variance in the number of segregating sites for a sample of two, $\text{var}(\pi) = \text{var}(S_2)$. The variance in coalescence times in a window of width w basepairs with population-scaled recombination rate $\rho = 4B_i N_e r w$ is,

$$\text{var}(S_2) = \theta + \theta^2 \frac{2}{\rho^2} \int_0^\rho (\rho - x) f_2(x) dx \tag{13}$$

(Wakeley 2009, eq. 7.20) where,

$$f_2(\rho) = \frac{\rho + 18}{\rho^2 + 13\rho + 18}. \tag{14}$$

Since this is a rough approximation, we set $\theta = w\pi_0 B_i$ and fix $\rho = w\gamma\pi_0 B_i$ where γ is the ratio of recombination to mutation rates, r/μ. In humans, $\gamma \approx 1$, but we explore different ratios to assess sensitivity of this calculation (Supplementary Figure 3).

This back-of-the-envelope calculation finds that the upper bound of R^2 would be around 74% for Yoruba, and 80-81% for European and Han Chinese models (Figure 3 This rough approximation assumes constant demography, which for bottlenecked out-of-Africa populations assumes that the variance in coalescence times is determined entirely by reducing the effective population size. However, the variance in coalescence times is *higher* in bottlenecked populations (CEU and CHB) compared to Yoruba, which we confirm with simple simulations using the `OutOfAfrica_3G09` model from `stdpopsim` (Adrion et al. 2020; Gutenkunst et al. 2009). This higher than expected variance would inflate SS_{res}, which would decrease

A more accurate approximate of the upper R^2 bound could rely on simulations with more realistic demography.

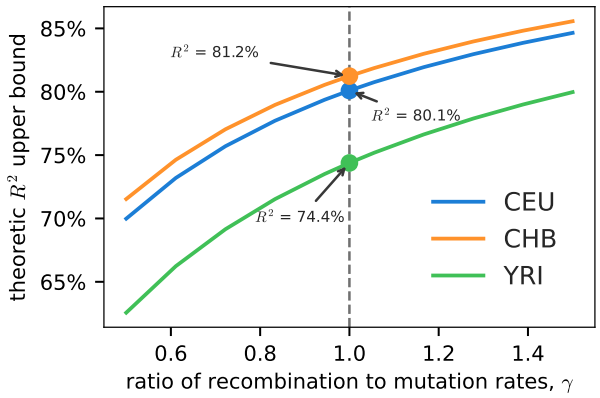

Figure 3: A rough approximation to the theoretic R^2 upper bound assuming constant demography.

Likelihood

Our model is essentially a generalized nonlinear model with a binomial link function. This is the form used by Elyashiv et al. (2016) and Murphy et al. (2022). These models fit the observed number of pairwise nucleotide differences in genomic windows to the expected pairwise diversity under some evolutionary model. We only consider BGS models here, so the mean function for position x is the product of the proportion by which BGS reduces diversity at position x, $B(x)$, and genome-wide neutral diversity π_0,

$$\pi(x) = B(x, \Phi)\pi_0 \tag{15}$$

Φ is the set of BGS parameters (i.e. the DFE for each feature type and mutation rate), and $\pi_0 = 4N_e\mu$ is determined by the genome-wide drift-effective population size N_e, set by only reproductive and demographic processes. Regional mutation rate heterogeneity can be accounted for with a regional mutation rate scaling function $m(x)$, $\pi(x) = B(x, \Phi)m(x)\pi_0$, but we do not explore mutation-rate heterogeneity, as it is unclear how to differentiate variance in mutation rate from the substitution rate heterogeneity we find across the genome.

The likelihood of the set of parameters $\Psi = \{\pi_0, \Phi\}$ can be written in the form of,

$$\log \mathcal{L}(\Psi) = \sum_{v \in \mathcal{V}} \sum_{i \neq j \in \mathcal{S}} \log(P(O_{i,j}(v)|\Psi)) \tag{16}$$

(c.f. Elyashiv et al. 2016; McVicker et al. 2009; Murphy et al. 2022) where $\mathcal{V}$ is the set of putatively neutral sites, $\mathcal{S}$ is the set of samples, and Ψ are the BGS parameters. The indicator variable $O_{i,j}(v)$ is 1 if samples i and j are different at putatively neutral site v, and zero otherwise. While the theoretic $\pi(v)$ gives the average number of pairwise differences, for small values, this is approximately the heterozygosity probability, so we can write

$$P(O_{i,j}(v)|\Psi) = \begin{cases} \pi(v), & O_{i,j}(v, \Psi) = 1 \\ 1 - \pi(v), & O_{i,j}(v, \Psi) = 0 \end{cases}. \tag{17}$$

(c.f. Elyashiv et al. 2016).

As in Section 1.5, the number of pairwise differences and the total number of pairwise comparison at a site are sufficient statistics for the likelihood. Then, the binomial log-likelihood for the data at site v is,

$$\ell_v(\Psi) = \log(\pi(v, \Psi))n_{\mathrm{D},v} + \log(1 - \pi(v, \Psi))n_{\mathrm{S},v} \tag{18}$$

1.7 The scale of processes

We can observe $\widehat{\pi}(x)$ at a per-basepair resolution. However, for a variety of reasons, we do not want to fit the composite likelihood model to the per-basepair scale of data. First, this would be computationally infeasible. Second, the mean function $\pi(x)$ varies on a natural scale that is itself a free parameter of the model. Our model can be written as,

$$\ell(\Psi, h) = \sum_b \sum_{v \in \mathcal{V}_b} \ell_v(\Psi) \tag{19}$$

$$= \sum_b \left[\log(\bar{\pi}(b, \Psi)) \sum_{v \in \mathcal{V}_b} n_{\mathrm{D},v} + \log(1 - \bar{\pi}(b, \Psi)) \sum_{v \in \mathcal{V}_b} n_{\mathrm{S},v} \right] \tag{20}$$

$$= \sum_b [\log(\bar{\pi}(b, \Psi)) Y_{\mathrm{D},b} + \log(1 - \bar{\pi}(b, \Psi)) Y_{\mathrm{S},b}] \tag{21}$$

$$\tag{22}$$

where h is the bandwidth or window size, b the bin index for windows of width h, $\mathcal{V}_b$ is the set of putatively neutral sites in bin b, $\bar{\pi}(b|\Psi)$ are the average diversity in bin b, and $Y_{\mathrm{S},b}$ and $Y_{\mathrm{S},b}$ are the sums across putatively neutral sites within a bin. Note that by binning, we sum the pairwise summaries of the data $\mathbf{Y}$ across sites, so the likelihood across bins is naturally weighted by the quantity of observed data.

This model corresponds to a binomial likelihood for the observed data summarized at genomic scale h. Thus an alternative way to express this model is as,

$$Y_{\mathrm{D},b} \sim \mathrm{Binom}(\bar{\pi}(b, \Psi), Y_{\mathrm{D},b} + Y_{\mathrm{S},b}). \tag{23}$$

Here, $\bar{\pi}(b, \Psi)$ is assumed to be the *probability* of sampling two different alleles, rather than the average *number* of pairwise differences; these are approximately equal when π is small. This corresponds to an identity link function; one could alternatively use a two-alleles finite sites model link function of the form, $\pi/(1 + 2\pi)$. We experimented with this and found there was little difference between these link functions, so we opted for the simpler identity link function.

2 Background Selection Models

The BGS models considered here both assume multiplicative fitness effects across sites. Thus, the total reduction due to segments across the genome is the product of the individual reductions,

$$B(x, \Psi) = \prod_g^S \exp\left(\int_0^1 b(\mu(s, k(g)), s, L_g, r_g, d(x, g)) ds \right) \tag{24}$$

$$= \exp\left(\sum_g^S \int_0^1 b(\mu(s, k(g)), s, L_g, r_g, d(x, g)) ds \right) \tag{25}$$

where, $b(\mu(s, k(g)), s, L_g, r_g, d(x, g))$ is the log reduction due to segment g, and there are S total segments. Here, $\mu(s, k(g))$ is the per-basepair per-genome rate that mutations with selection coefficient s enter segments with feature class $k(g)$, L_g and r_g are the length in basepairs and recombination rate per basepair of segment g, and the recombination distance between focal site x and segment g is $d(x, g)$. The functional form of b varies depending on whether the classic BGS model is used, or our form of Santiago and Caballero's (2016) equations (see Section XXX).

Under both background selection models used in our likelihood, the diversity in window b is $\bar{\pi}(b, \Psi) = \bar{B}(b, \Psi)\pi_0$. Here, $\bar{B}(b|\Psi)$ is the predicted reduction in diversity due to BGS in window

b, given background selection parameters Ψ. In practice, we pre-calculate $B(x|\Psi)$ at fixed sites x across the genome, and take the average of the fixed sites B values within widow b for the average $\bar{B}(b|\Psi)$.

2.1 Reductions Under the Classic BGS Model

The classic BGS model, which assumes deleterious mutations cannot fix, expresses the reduction for a focal site in the middle of a segment (Hudson and Kaplan 1995; Hudson and Kaplan 1994; Nordborg et al. 1996); here we consider a focal site directly to an arbitrary side of a segment under a deleterious mutations, and extend the model to handle a focal segment arbitrarily far from the segment. To simplify notation, we will only consider a single feature class in this and the next section. Using our notation from equation (36), the classic mutation-selection-balance BGS model is,

$$b(\mu, s, L, r, 0) = -\int_0^L \frac{2\mu(l)}{s(1+(1-s)r(l)/s)^2} dl \tag{26}$$

(c.f. Hudson and Kaplan 1995 equation 5), where $\mu(l)$ is the *haploid* per-basepair mutation rate at l and $r(l)$ is the recombination fraction between the focal site (at position 0) and basepair l. Assuming constant per-basepair recombination and mutation rates, $r(l) = r_{\mathrm{BP}} l$ and $\mu(l) = \mu$, and

$$b(\mu, s, L, r_{\mathrm{BP}}, 0) = -\frac{2\mu L}{s + (1-s) r_{\mathrm{BP}} L} \tag{27}$$

(c.f. Hudson and Kaplan 1995 equation 8, where the factor of two difference is due to the position of the focal site). Often per-region rates are defined $U = 2\mu L$ and $M = r_{\mathrm{BP}} L$ are the segment-wide mutation diploid mutation rates (mutations per diploid segment per generation) and recombination length (Morgans).

For computational efficiency, we adjust this model by setting $r(l) = d + r_{\mathrm{BP}} l$, where d is the recombination fraction between the focal site x and the start of segment g. This allows us to pre-compute the local effects of the segment, and substitute in b as the focal site changes. This integral substitution has a closed form,

$$b(\mu, s, L, r_{\mathrm{BP}}, d) = \frac{-2\mu L}{(d(1-s)+s)(s + (d + r_{\mathrm{BP}} L)(1-s))} \tag{28}$$

and terms can be collected in powers of d and pre-computed for all segments for the grid of μ and s. Our implementation pre-computes these segment components, and computes the final b value for an x and g by calculating the distance $d(x, g) = |M(x) - M(g)|$, where $M(x)$ is the cumulative recombination map length at position x in Morgans. Since $r(l)$ could vary by position, segments with differing recombination rates are split into segments of the same annotation class by their recombination rate. This has no effect on the calculation due to the multiplicative fitness, but ensures more accurate estimates of the reduction factor B.

2.2 Reduction Under the Modified Santiago and Caballero Model

When weak deleterious mutations can fix, causing substitutions, the classic background selection models break down. We find that the generalized background selection model of Santiago and Caballero (2016) works well (see Section XXX for our derivation). Unlike the classic BGS models which only produce B maps, that only estimate the reduction in effective population size, The generalized BGS model of Santiago and Caballero (2016) also predicts the deleterious substitution rate per generation per segment (the "ratchet" rate), which we call the R map. We describe how these are calculated substitution rates per segment, for each mutation rate and selection coefficient combination in the grids. We describe how these are calculated here.

First, the reduction factor experienced by a focal site within the segment g (and thus is similar to what selected alleles in g experience) under generalized background selection is calculated as

$$\widetilde{b}'(\mu, s, L, r_{\mathrm{BP}}) = \log\left(\frac{\widetilde{N}_{e,\infty}}{N}\right) \tag{29}$$

where $\widetilde{N}_{e,\infty}$ is the asymptotic effective population size, which is the solution to the following system of two equations (see XXX for derivation),

$$N_{e,\infty} = N\exp\left(-\frac{V}{2}Q_{\infty}^2\right) \tag{30}$$

$$R = \frac{4UsN_{e,\infty}}{\exp(4N_{e,\infty}s) - 1} \tag{31}$$

where

$$V = Us - 2Rs \tag{32}$$

$$Q_{\infty}^2 = \frac{2}{(1-Z)(2-(2-M)Z)} \tag{33}$$

$$Z = 1 - \frac{Us}{U-R} \tag{34}$$

and $U = 2\mu L$ and $M = r_{\mathrm{BP}}L$ as above. In this section, we assume μ is the mutation rate to variants with heterozygous selection coefficient s to simplify notation; later, μ refers to the genome-wide average mutation rate and the mutation rate is scaled by the DFE. Solving this system of two equations for each segment also gives the total substitution or "ratchet" rate per generation per segment ($\widetilde{R}$). The per-basepair substitution rate of a segment with mutation rate μ and selection coefficient s is $\lambda_d(\mu, s) = \widetilde{R}/L$

Under Santiago and Caballero's model, the correct effective population size to gauge reduction in heterozygosity or coalescent times is given by equation XXX. This is because selective processes can lead to non-constant pairwise coalescent rates, which impact pairwise coalescent rates and thus the reduction factor. We experimented with the corrections described in Section XXX, and found (1) they had little effect on the accuracy of the B maps, and (2) they were incredibly costly to

calculate, even with the approximations described in equations XXX. Overall, we solve for the asymptotic $\widetilde{N}_{e,\infty}$ for each segment, and use that to calculate b'.

Note that the asymptotic model above determines the reduction in effective population size experienced by a focal site in the middle of a L-basepair segment, but B maps require the reduction b' experienced by a focal site x at an arbitrary recombination distance apart. For each segment, we pre-compute the solution to the system of two equations, giving us V and Q_∞^2, and then use these values in the equation (30),

$$b'(\mu, s, L, r, d) = -\frac{V}{2}\left(\frac{1}{(1-(1-d)Z}\right)^2 \tag{35}$$

where $Z = 1 - Us/(U - \widetilde{R})$ and $V = Us - 2\widetilde{R}s$. TODO check d.

2.3 Discretization of Parameters and Annotation Feature Classes

In practice, to fit such a model, we must discretize both $\mu(s, k(g))$ and s in Equation (36). We define μ_i to be the rate of mutations entering the population per basepair, per generation with selection coefficient s_i (i.e. the product of the mutation rate and the distribution of deleterious fitness effects). Then, the reduction at position x is,

$$B(x, \Psi) = \exp\left(\sum_g^S \sum_i b\left(\mu_{i,\,k(g)}, s_i, L_g, r_g, d(x, g)\right)\right). \tag{36}$$

Because it is computationally-intensive to numerically solve the system of two non-linear equations to calculate b at each segment, these are pre-computed for an m_s-element grid of selection coefficients and an m_μ-element grid of mutation rates–DFE products. The number of annotation features K is variable and set by the annotation data specified. There are two ways to parameterize the mutation rates and DFE for all annotation classes in background selection models. First, for the *free-mutation* parameterization, each annotation feature class has a free mutation rate parameter for each of these selection coefficients (as in equation (36)).

In the free-mutation parameterization, there are $1 + K \times m_s$ for π_0 and the free mutation rate parameters, where K is the number of annotation features. The MLE optimization for this approach is unconstrained, though π_0 is bounded to be within $10^{-5} \leq \pi_0 \leq 10^{-1}$ and mutation rates are bounded within $10^{-11} \leq \mu \leq 10^{-7}$. The BGS model requires a mutation rate for each selection coefficient in the grid, for each annotation class (here, as an example, CDS, UTRs, and phastcons), which are stored in the $m_s \times K$ matrix $\mathbf{M}_\mathrm{F}$,

$$\mathbf{M}_\mathrm{F} = \begin{bmatrix} \mu_{10^{-6},\mathrm{CDS}} & \mu_{10^{-6},\mathrm{UTR}} & \mu_{10^{-6},\mathrm{PC}} \\ \mu_{10^{-5},\mathrm{CDS}} & \mu_{10^{-5},\mathrm{UTR}} & \mu_{10^{-5},\mathrm{PC}} \\ \vdots & \vdots & \vdots \\ \mu_{10^{-1},\mathrm{CDS}} & \mu_{10^{-1},\mathrm{UTR}} & \mu_{10^{-1},\mathrm{PC}} \end{bmatrix} \tag{37}$$

That is, the distribution of fitness effect (DFE) is implied by the total mutation rate across selection coefficients for an annotation class. We can normalize by the total mutation rate estimate

to get the estimated DFE, giving us the DFE weight for annotation class k and selection coefficient with index i,

$$\widehat{w}_{i,k} = \frac{\widehat{\mu}_{i,k}}{\sum_i \widehat{\mu}_{i,k}}. \tag{38}$$

The second parametrization is the *simplex* parametrization, which has a single mutation rate across all features, so it is a DFE weight matrix times the mutation rate μ,

$$\mathbf{M}_{\mathrm{S}} = \mu \begin{bmatrix} w_{10^{-6},\mathrm{CDS}} & w_{10^{-6},\mathrm{UTR}} & w_{10^{-6},\mathrm{PC}} \\ w_{10^{-5},\mathrm{CDS}} & w_{10^{-5},\mathrm{UTR}} & w_{10^{-5},\mathrm{PC}} \\ \vdots & \vdots & \vdots \\ w_{10^{-1},\mathrm{CDS}} & w_{10^{-1},\mathrm{UTR}} & w_{10^{-1},\mathrm{PC}} \end{bmatrix} \tag{39}$$

which has $2 + K \times (m_s - 1)$ free parameters, since each column must sum to one. This imposes the constraint that $\sum_i w_{i,k} = 1$, requiring constrained MLE optimization.

3 Ratchet Rate Prediction

After rescaling each segment N_e by the local predicted reduction $\hat{B}(x)$, the ratchet rates and B' values are re-calculated.

The rescaled B' calculation outputs a $m_\mu \times m_s \times S$ multidimensional array $\mathbf{R}$ of ratchet rates per segment. These are rescaled by the segment lengths $\mathbf{l}$, giving the per-basepair mutation rate array. The per-basepair ratchet rate given maximum likelihood estimates of $\hat{\mu}$ and $\widehat{\mathbf{W}}$ is used to predicted the ratchet rate for each segment g. The DFE estimate for feature k is a column vector of $\widehat{\mathbf{W}}$, i.e. $\widehat{\mathbf{w}}_k$

$$\lambda_D(g) = \tag{40}$$

The predicted ratchet rate for segment g $\lambda_d(g)$ is based on these values.

3.1 Numeric Optimization

mu bounds

bounds and time, bounds and convergence

Same optima, but took longer

References

Adrion, Jeffrey R et al. (2020). "A community-maintained standard library of population genetic models". en. In: *Elife* 9.

Byrska-Bishop, Marta et al. (2022). "High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios". en. In: *Cell* 185.18, 3426–3440.e19.

Elyashiv, Eyal et al. (2016). "A Genomic Map of the Effects of Linked Selection in Drosophila". en. In: *PLoS Genet.* 12.8, e1006130.

Gutenkunst, Ryan N, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante (2009). "Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data". en. In: *PLoS Genet.* 5.10, e1000695.

Hudson, R R and N L Kaplan (1995). "Deleterious background selection with recombination". en. In: *Genetics* 141.4, pp. 1605–1617.

Hudson, Richard R and Norman L Kaplan (1994). "Gene Trees with Background Selection". In: *Non-Neutral Evolution: Theories and Molecular Data*. Ed. by Brian Golding. Boston, MA: Springer US, pp. 140–153.

Illumina, Inc. (2020). *1000 Genomes Phase 3 Reanalysis with DRAGEN 3.5 and 3.7*. `https://registry.opendata.aws/ilmn-dragen-1kgp.`. Accessed: 2021-7-19.

McVicker, Graham, David Gordon, Colleen Davis, and Phil Green (2009). "Widespread genomic signatures of natural selection in hominid evolution". en. In: *PLoS Genet.* 5.5, e1000471.

Murphy, David A, Eyal Elyashiv, Guy Amster, and Guy Sella (2022). "Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements". In: *Elife* 11, e76065.

Nordborg, Magnus, Brian Charlesworth, and Deborah Charlesworth (1996). "The effect of recombination on background selection*". In: *Genet. Res.* 67.02, pp. 159–174.

Santiago, Enrique and Armando Caballero (2016). "Joint Prediction of the Effective Population Size and the Rate of Fixation of Deleterious Mutations". en. In: *Genetics* 204.3, pp. 1267–1279.

Wakeley, John (2009). *Coalescent Theory: An Introduction*. Roberts and Company Publishers.