Supplementary Materials

Vince Buffalo and Andrew Kern

December 26, 2022

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.

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`.

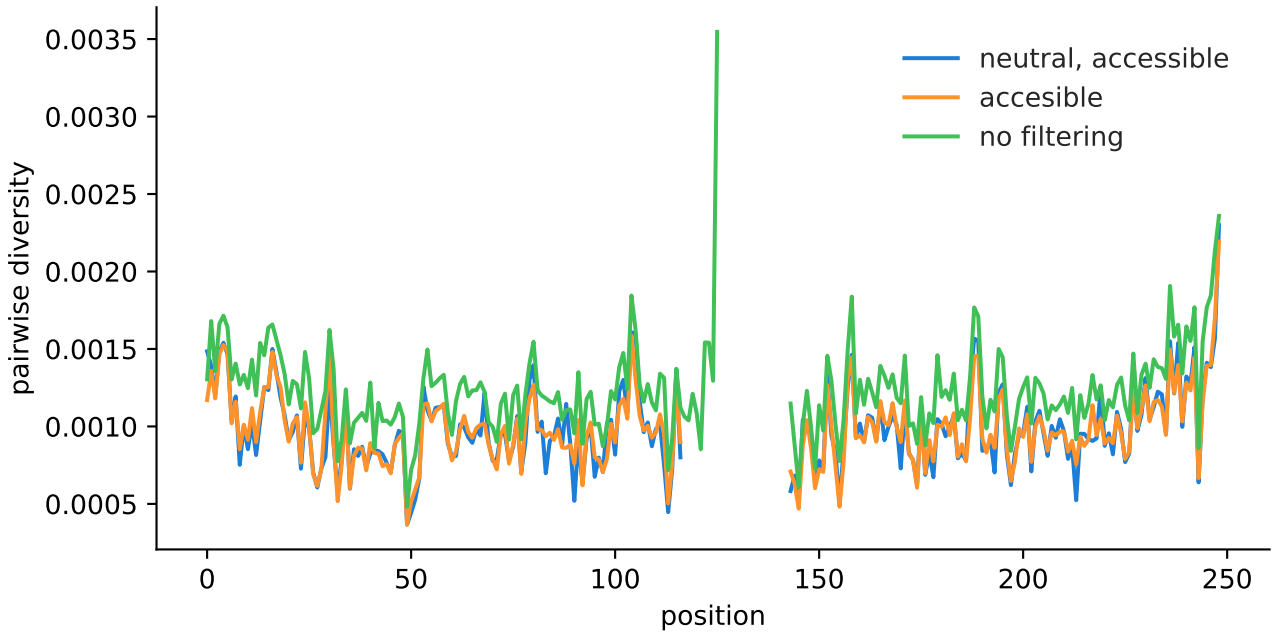

Figure 1: Estimates of chromosome 1 YRI diversity in non-overlapping megabase windows, under different filtering criteria. The filtering criteria are, (1) "neutral, accessible" which includes only putatively neutral sites, and ignores regions masked as inaccessible, (2) "accessible" which is only ignores 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.

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

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

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

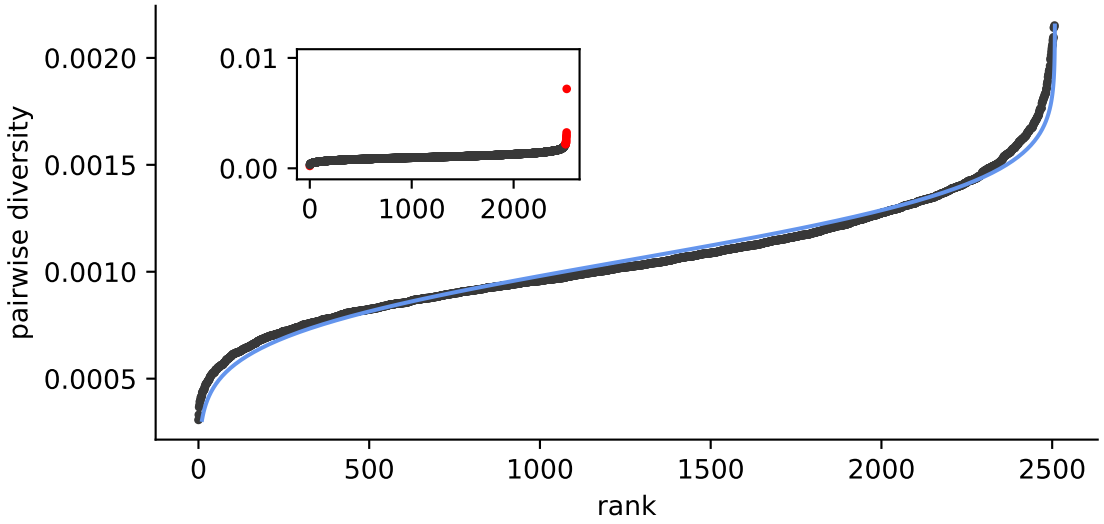

Figure 2: The distribution of diversity across the genome at the megabase scale, with outliers trimmed. The blue line is the normal CDF, fit with MLE parameters XXX. The inset figure is the untrimmed data, with trimmed points shown in red.

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).

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.

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,

$$\pi(x) = B(x, \Phi)\pi_0 \tag{6}$$

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,

$$\log \mathcal{L}(\Psi) = \sum_{v \in \mathcal{V}} \sum_{i \neq j \in \mathcal{S}} \log(P(O_{i,j}(v)|\Psi)) \tag{7}$$

(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

$$P(O_{i,j}(v)|\Psi) = \begin{cases} \pi(v), & O_{i,j}(v, \Psi) = 1 \\ 1 - \pi(v), & O_{i,j}(v, \Psi) = 0 \end{cases}. \tag{8}$$

(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,

$$\ell_v(\Psi) = \log(\pi(v, \Psi))n_{\mathrm{D},v} + \log(1 - \pi(v, \Psi))n_{\mathrm{S},v} \tag{9}$$

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,

$$\ell(\Psi, h) = \sum_b \sum_{v \in \mathcal{V}_b} \ell_v(\Psi) \tag{10}$$

$$= \sum_b \left[\log(\bar{\pi}(b, \Psi)) \sum_{v \in \mathcal{V}_b} n_{\mathrm{D},v} + \log(1 - \bar{\pi}(b, \Psi)) \sum_{v \in \mathcal{V}_b} n_{\mathrm{S},v} \right] \tag{11}$$

$$= \sum_b [\log(\bar{\pi}(b, \Psi)) Y_{\mathrm{D},b} + \log(1 - \bar{\pi}(b, \Psi)) Y_{\mathrm{S},b}] \tag{12}$$

$$\tag{13}$$

where h is the bandwidth or window size, b are physical-scale window indices, $\bar{\pi}(b|\Psi)$ 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. We fit h using out-sample cross validation.

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

$$Y_{\mathrm{D},b} \sim \mathrm{Binom}(\bar{\pi}(b, \Psi), Y_{\mathrm{D},b} + Y_{\mathrm{S},b}). \tag{14}$$

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.

1.8 The Background Selection Reduction Factor

Under the background selection model, the diversity in window b is $\pi(b, \Psi) = \bar{B}(b, \Psi)\pi_0$. 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 $\bar{B}(x|\Psi)$ at fixed sites x across the genome, and take the average of the fixed sites B values within widow b for the average $\bar{B}(b|\Psi)$.

There are two ways to parameterize background selection models. Both models use a fixed q-grid of selection coefficients ranging $s \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ (here, $q = 6$) and interpolate over a grid of mutation rates. First, for the *free-mutation* parameterization, each annotation feature class has a free mutation rate parameter for each of these selection coefficients. 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,

$$\widehat{w}_{i,k} = \frac{\widehat{\mu}_{i,k}}{\sum_i \widehat{\mu}_{i,k}}. \tag{15}$$

With this parameterization, there are $1 + K \times q$ for π_0 and the free DFE 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 $10^{-5} \leq \pi_0 \leq 10^{-1}$.

$$W = \begin{bmatrix} \mu_{10^{-6},\text{CDS}} & \mu_{10^{-6},\text{UTR}} & w_{10^{-6},\text{phastcons}} \\ \mu_{10^{-5},\text{CDS}} & \mu_{10^{-5},\text{UTR}} & w_{10^{-5},\text{phastcons}} \\ \vdots & \vdots & \vdots \\ \mu_{10^{-1},\text{CDS}} & \mu_{10^{-1},\text{UTR}} & w_{10^{-1},\text{phastcons}} \end{bmatrix} \tag{16}$$

The second parametrization is the *simplex* parametrization, which has a single mutation rate across all features, and the DFE is parameterized by a matrix W

which has $2 + K \times (q-1)$ free parameters.

In practice, this is site-specific. We can write the reduction at any neutral site v in the genome as the product of Bs across all segments,

$$B(v|\theta) = \exp\left(-\sum_g \int f(\mu(\mathcal{A}(g)), s, S_g) w(s|\mathcal{A}(g)) ds\right) \tag{17}$$

where S_g is exogenous genomic data about the segment, $S_g = \{L_g, r_g, \rho(|v - p_g|)\}$, where L_g is the segment's length, r_g is the recombination rate per basepair in the segment, and $\rho(|v - p_g|)$ is the recombination distance between the focal site v and the segment position p_g (we approximate, and use the nearest end position to the neutral site). Additionally, $w(s|\mathcal{A}(g))$ is the distribution of selection coefficients for segment g, if segment g is a member of annotation class $\mathcal{A}(g)$.

We can think about the DFE as the conditional distribution of a particular selection coefficient given a mutation occurs, for a particular annotation class. The BGS function $f(\cdot)$ only depends on μ through the introduction of deleterious alleles with selection coefficient s at rate $\omega(s|\mathcal{A}(g)) = \mu(\mathcal{A}(g))w(s|\mathcal{A}(g))$. Thus, we can write,

$$B(v|\theta) = \exp\left(-\sum_g \int f(\omega(s|\mathcal{A}(g)), s, S_g) ds\right) \tag{18}$$

which we can discretize as,

$$B(v|\theta) = \exp\left(-\sum_g \sum_s f(\omega(s|\mathcal{A}(g)), s, S_g)\right). \tag{19}$$

Next, note that there are a finite number of annotation classes, $\mathcal{A} \to \{a_1, a_2, \ldots, a_k\}$, so we can further partition this as

$$B(v|\theta) = \exp\left(-\sum_{\{g\,:\,\mathcal{A}(g)=a_1\}} \sum_s f(\omega(s|a_1), s, S_g) + \sum_{\{g\,:\,\mathcal{A}(g)=a_2\}} \sum_s f(\omega(s|a_2), s, S_g) + \ldots\right) \tag{20}$$

Let us define the $d_\omega \times d_s \times d_g$ multidimensional array $\mathbf{F}$, and the $d_g \times d_a$ feature classification matrix $\mathbf{A}$.

2 Calculation of B Maps

Each B map is determined by the annotation track of putatively conserved segments, which includes (1) the segment positions and lengths, (2) the type of feature of each segment (e.g. exons, UTRs, phastcons conserved elements, etc.), as well as (3) the recombination map, and (3) the deleterious selection coefficient $s(i)$ and mutation rate at which these deleterious enter the population for feature type i.

The B maps are computed at fixed, evenly spaced sites.

3 Ratchet Rate Prediction

$$R = \int \tag{21}$$

References

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.

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.