Supplementary Materials

Vince Buffalo and Andrew Kern

December 8, 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`.

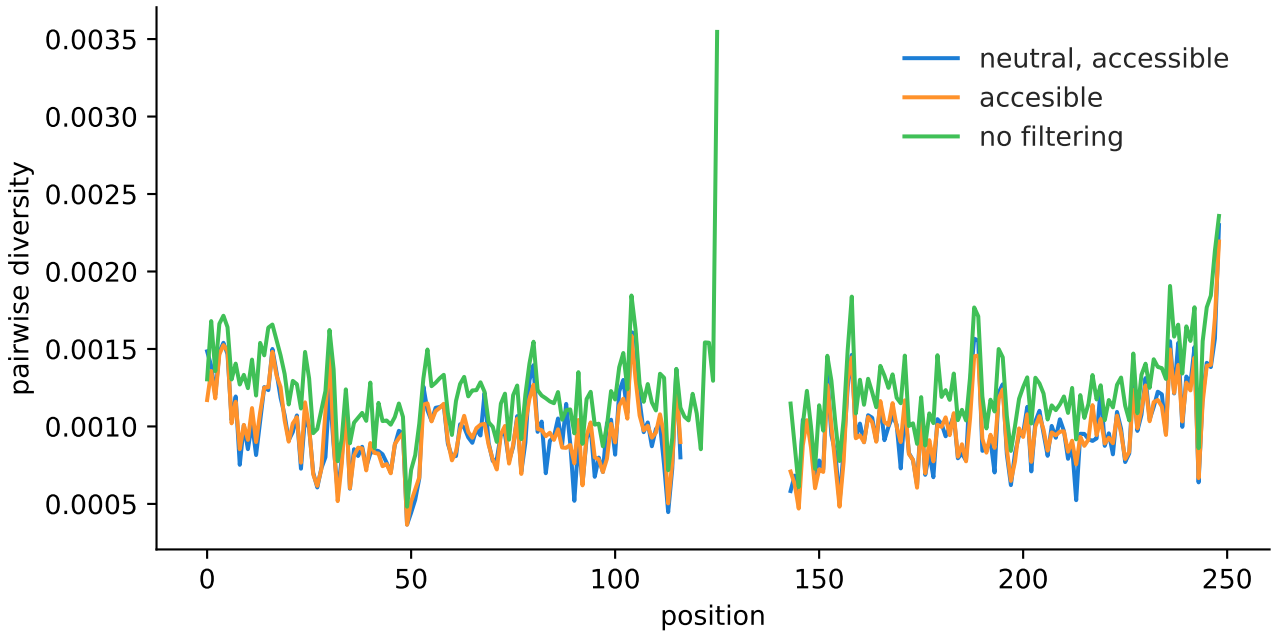

Figure 1: Estimates of chromosome 1 YRI diversity in non-overlapping megabase windows, under different filtering criteria. The filtering criteria are, (1) "neutral, accessible" which includes only putatively neutral sites, and ignores regions masked as inaccessible, (2) "accessible" which is only ignores 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 $\mathbf{Y}_l = [n_S, n_D]$.

1.5 Pairwise Diversity Estimates

The pairwise diversity at site l across the n sampled chromosomes can be calculated from 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, and 0 otherwise. Note that for sites with missing data, $c_1 = c_2 = 0$ and thus $n_T = 0$ and the division is invalid. Such cases were marked as missing with floating point values `NaN`s. Binned summaries of data are weighed by the number of complete cases, e.g. rows with $c_1 + c_2 > 0$.

This per-basepair diversity is then averaged over all callable bases, to give a genome-wide or window estimate of diversity of accessible bases in the set $\mathcal{A}$,

$$\bar{\pi}(\mathcal{A}) = \frac{\sum_{l \in \mathcal{A}} \pi(l)}{L_{\mathcal{A}}}.$$

where $L_{\mathcal{A}} = |\mathcal{A}|$ is the number of accessible bases. Note that the sum in the numerator is random over the sample of chromosomes sampled from the population. 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.

The genome-wide $\bar{\pi}$ can also be estimated by summing the columns of $\mathbf{Y}$, and calculating n_D/n_T.

1.6 Window-based Summaries and filtering

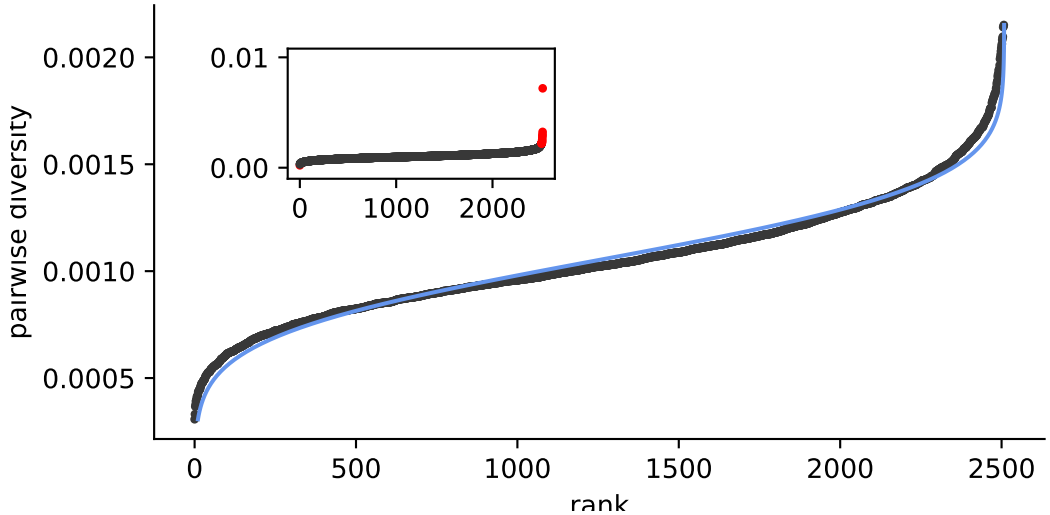

Figure 2: The distribution of diversity across the genome at the megabase scale, with outliers trimmed. The blue line is the normal CDF, fit with MLE parameters XXX. The inset figure is the untrimmed data, with trimmed points shown in red.

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.

2

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.

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.