Robust estimates of heterozygosity from low coverage sequencing data

Katarzyna Bryc[1,2], Nick Patterson[2], David Reich[1]

Contact: kasia1@gmail.com

1) Department of Genetics, Harvard Medical School, Boston, MA; 2) Broad Institute of Harvard and MIT, Cambridge, MA

Motivation

- Heterozygosity, or the fraction of nucleotides within an individual that differ between their two parental chromosomes, can be informative about population demographic history and natural selection.
- In practice, estimating heterozygosity from sequence data is confounded by sequencing errors, mapping errors, and imperfect power to detect SNPs.
- Obtaining an unbiased estimate of heterozygosity is especially difficult for ancient genomes (such as Neandertal or Denisova) where the sequences are expected to have a high error rate, or with low-coverage sequence data where there is low power to detect heterozygous SNPs.

Method summary

We present a method that leverages the genome-wide **joint information** across sequence reads **from a panel of individuals**, to estimate the heterozygosity of an individual of interest.

We assume **independence** between the observed alleles of the individual at a particular locus and the combination of observed alleles in a panel of individuals, **conditional** on the true underlying genotype.

We then apply an Expectation-Maximization (EM) algorithm to estimate the most likely distribution of counts across the unknown underlying genotypic states, from which we obtain an estimate of the proportion of loci that are heterozygous in that individual.

Proof of Principle

We find that we are able to obtain accurate estimates of heterozygosity across a variety of coverage levels (5X to 30X) on simulated sequence data (see Figure 1). We note a small bias of about 0.3% from the true rate for 5X coverage read data, but with higher coverage this bias goes to 0.

Simulations We simulated data using the *MaCS* program [1] by based on the demographic history of Wall *et al.* (2009) [2] and adding a chimpanzee outgroup. We then simulated sequence read data, adding errors at each base position at a rate of 0.2% in our individual of interest, 0.01% in the panel individuals, and 0.1% in the chimpanzee outgroup used to assign ancestral and derived alleles.

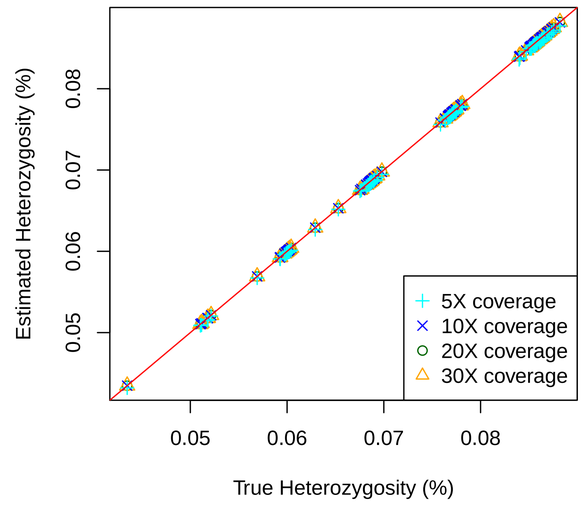

Figure 1: Estimated genome-wide heterozygosity rates versus true rates for simulated data.

Results on 11 human genomes from diverse world-wide populations

To avoid introducing a population bias in aligning to the human reference genome, we aligned the sequence data to the chimpanzee reference genome.

We observe consistent estimates of heterozygosity independent of the number of alleles sampled (Figure 2, *symbol*). However, we find a strong effect of sequencing coverage on heterozygosity estimates, with higher estimates of heterozygosity for regions of the genome covered much less (< 10X) or much more (>30X) than the mean (∼20X).

Estimates of heterozygosity for each of the 11 individuals are presented in Table 1.

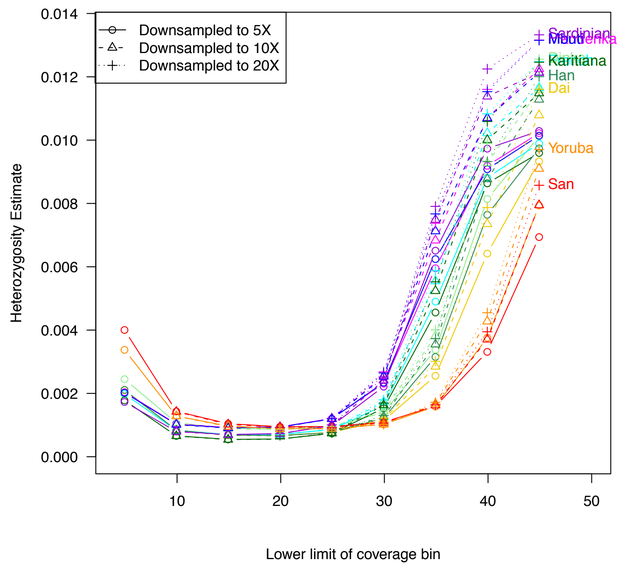

Figure 2: Estimates (*y-axis*) using 5, 10 and 20 alleles (shown by *symbol*). Heterozygosity is strongly biased by genomic coverage (*x-axis*).

Individual	Heterozygosity rate
San	0.0008103
Yoruba	0.0007421
Mandenka	0.0007273
Mbuti	0.0007299
Dinka	0.0007175
French	0.0005369
Sardinian	0.0005329
Dai	0.0005178
Han	0.0005162
Karitiana	0.0003934

Table1: Heterozygosity estimates for the 20X coverage bin.

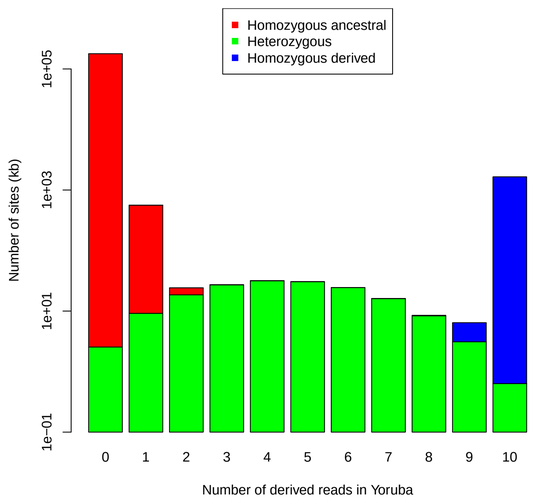

Figure 3: Distribution of number of derived alleles, colored by estimated true state for a Yoruban individual.

Removing CNV regions reduces coverage bias

We applied a filter to the data, provided by the Eichler lab, filtering out regions that were likely to contain of copy-number variants. This resulted in a strong reduction in the increase in heterozygosity of higher-coverage genomic regions (Figure 4).

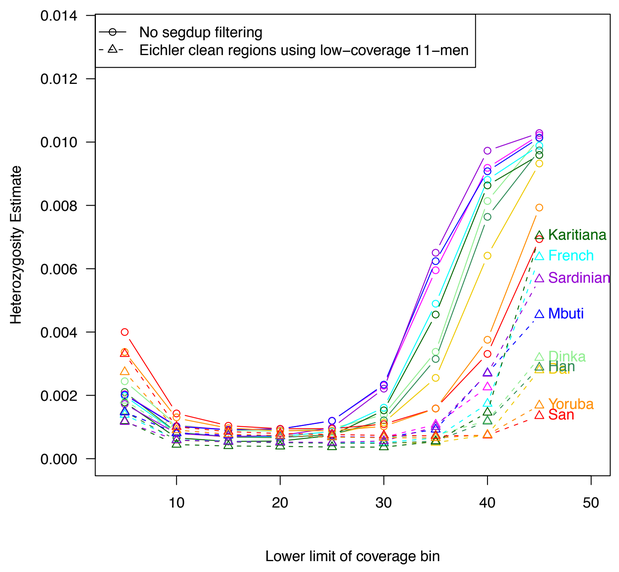

Figure 4: Applying a CNV-free filter to the sequence data reduces the increase of heterozygosity at high coverage.

Conclusions

The reads from the panel of individuals increases the information we have about the genotypic state at our individual of interest in much the same way as using multi-sample SNP calling.

Absolute heterozygosity is not a well-defined quantity in the analysis of genomic data, as it depends on the particular filters that are use to identify the regions being analyzed.

Because the regions of the genome that pass our filters are likely to be lower in complexity and substantially biased towards lower diversity due to alignment biases, we expect that the absolute values of these estimates are a lower bound on the true estimates of heterozygosity.

References

[1] Gary Chen, Paul Marjoram, and Jeffrey Wall. Fast and flexible simulation of dna sequence data. *Genome research*, 19(1):136–142, 2009.

[2] Jeffrey D Wall, Kirk E Lohmueller, and Vincent Plagnol. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol*, 26(8):1823–7, Aug 2009.

Mathematical Model

Let a be the unknown diploid genotype of our target individual at some position in the genome, and c be the aligned chimpanzee allele. Then the allele distribution in the panel of individuals will depend on $g = (a, c)$.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be the vector of alleles by taking one randomly sampled read from each of the n individuals.

Let $\mathbf{w}$ be the observed alleles from the reads for our individual. Both $\mathbf{w}$ and $\mathbf{x}$ are observed quantities for a given position in the genome for our individual, and we are interested in modeling the joint probability of $\mathbf{w}$, $\mathbf{x}$ as the sum of the joint probabilities conditional on g.

We will assume conditional independence of $\mathbf{w}, \mathbf{x}$ on the true unobserved genotype g. Then from this it follows that:

$$P(\mathbf{w}, \mathbf{x}|g) = P(\mathbf{w}|g)P(\mathbf{x}|g) \quad (1)$$

$$P(\mathbf{w}, \mathbf{x}, g) = P(\mathbf{w}|g)P(\mathbf{x}|g)P(g) \quad (2)$$

We use the chimpanzee, or outgroup, allele to define the ancestral (0) and derived (1) alleles. We create a count matrix N of dimension $||\mathbf{w}|| \times ||\mathbf{x}||$. From this matrix we wish to estimate the true values of $P(g)$, $P(\mathbf{w}|g)$, and $P(\mathbf{x}|g)$ using the EM algorithm.

Let Y_{obs} be the observed counts of alleles in the matrix $N_{\mathbf{w},\mathbf{x}}$ So the counts of the number of loci where the individual is $\mathbf{w}$ and the other individuals form the alleles $\mathbf{x}$ is represented by the corresponding row and column entry in the matrix N.

Let Y_{mis} be $N_{\mathbf{w},\mathbf{x},g}$, the missing or unobserved counts of the alleles with the true parameter state g. Then the likelihood of the data is:

$$\mathcal{L} = \sum_{\mathbf{w},\mathbf{x}} N_{\mathbf{w},\mathbf{x}} \log P(\mathbf{w}, \mathbf{x}) \quad (3)$$

If the hidden variable g were observed, the log-likelihood would be

$$\mathcal{L}' = \sum_{\mathbf{w},\mathbf{x},g} N_{\mathbf{w},\mathbf{x},g} \log P(\mathbf{w}, \mathbf{x}, g) \quad (4)$$

We can reduce the number of parameters we must fit in equation (4) by relying on our conditional independence from equation (2) above.

The Expectation-Maximization (EM) algorithm is an appropriate choice to maximize our likelihood over the parameter space under the conditional independence constraints from (1).

The estimates for $\hat{P}$ and $\hat{N}_{\mathbf{w},\mathbf{x},g}$, updating the parameters:

$$\hat{P}(\mathbf{w}|g) = \frac{\sum_{\mathbf{x}} \hat{N}_{\mathbf{w},\mathbf{x},g}}{\sum_{\mathbf{w},\mathbf{x}} \hat{N}_{\mathbf{w},\mathbf{x},g}}$$

$$\hat{P}(\mathbf{x}|g) = \frac{\sum_{\mathbf{w}} \hat{N}_{\mathbf{w},\mathbf{x},g}}{\sum_{\mathbf{w},\mathbf{x}} \hat{N}_{\mathbf{w},\mathbf{x},g}}$$

$$\hat{P}(g) = \frac{\sum_{\mathbf{w},\mathbf{x}} \hat{N}_{\mathbf{w},\mathbf{x},g}}{N}$$

$$\hat{N}_{\mathbf{w},\mathbf{x},g} = N_{\mathbf{w},\mathbf{x}} \cdot \frac{\hat{P}(\mathbf{w}, \mathbf{x}, g)}{\hat{P}(\mathbf{w}, \mathbf{x})} = N_{\mathbf{w},\mathbf{x}} \cdot \frac{\hat{P}(\mathbf{w}|g)\hat{P}(\mathbf{x}|z)\hat{P}(g)}{\sum_z \hat{P}(\mathbf{w}|g)\hat{P}(\mathbf{x}|g)\hat{P}(g)}$$

By basic EM theory these re-estimated values of $\hat{P}$ will generate a non-decreasing sequence of values for the log likelihood $\mathcal{L}$.

Stabilization of likelihood surface

To further stabilize our estimates using the EM, we parameterize the distribution of the number of derived alleles found in our individual as a Beta-Binomial, and use MOM estimates to update the parameters.