GenomeScope: Fast reference-free genome profiling from short reads

Gregory W. Vurture[†], Fritz J. Sedlazeck[†], Maria Nattestad, Charles J. Underwood, Han Fang, James Gurtowski and Michael C. Schatz

Supplementary Notes and Figures

Other Supplementary Files:

Supplementary Table 3: Full results from simulated genomes
Supplementary Table 4: Full results from synthetic heterozygous E. coli genomes
Supplementary Table 5: Full results from genuine heterozygous genomes

Supplementary Note 1: Modeling and Analysis

1.1 Heterozygous K-mer Profiles

GenomeScope analyzes the *k-mer profile* (also called *k-mer spectrum*) of unassembled, unaligned (i.e. raw) reads to compute the genome and read characteristics. The *k-mer* profile counts how many *k-mers*, substrings of length *k*, occur in the sequence data. Throughout the analysis we use "canonical *k-mers*" where the counts for a *k-mer* and its reverse complement are counted together since the sequencing reads may originate from either strand of the DNA.

A key advantage of the *k-mer* based approach is that the *k-mer* profile can be computed without a reference genome by directly scanning the sequencing reads, so that it works well for novel genomes. Several programs are now available for quickly computing the *k-mer* profile including Jellyfish (Marcais and Kingsford, 2011) and KMC2 (Deorowicz, et al., 2015) that can process billions of reads per hour on a modern server, making it practical to run on virtually any genome with minimal compute requirements. It is also possible to estimate the *k-mer* profile using a variety of streaming techniques (Chikhi and Medvedev, 2014; Melsted and Halldorsson, 2014).

The shape of the *k-mer* profile reflects the complexity of the genome sequenced. For example, if a homozygous repeat-free genome has been sequenced without errors or biases to a given average coverage level (e.g. 100x coverage), then the *k-mer profile* will be a Poisson distribution centered near that coverage (Kelley, et al., 2010; Lander and Waterman, 1988). More precisely, the center of the distribution will be centered at the average *k-mer* coverage available, which is defined as

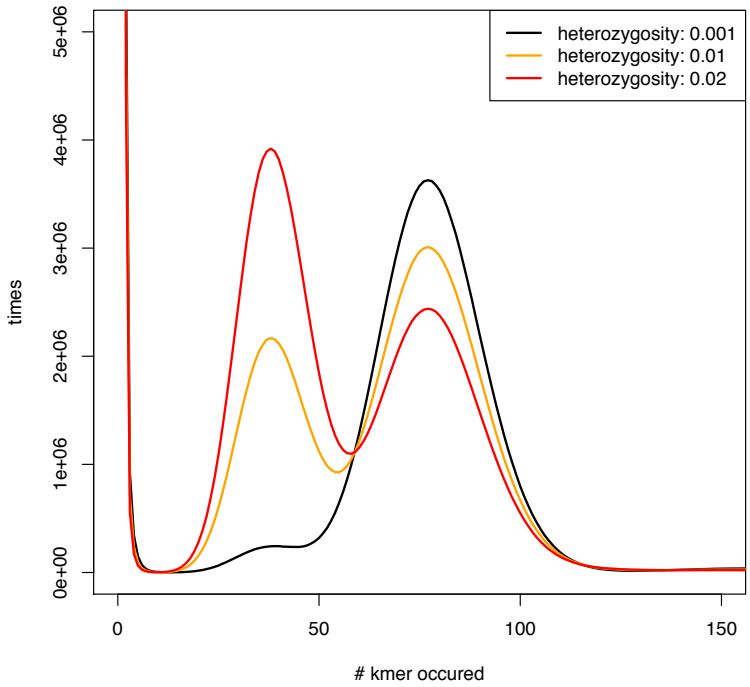
$$C_{kmer} = \left(\frac{L-k}{L}\right) * \left(\frac{N_{reads} * L}{G}\right) \quad \text{(Eq. 1)}$$

Here C_{kmer} is the average *k-mer* coverage, N_{reads} is the number of reads, L is the average read length, k is the length of the *k-mer*, and G is the genome size. This effectively shifts the peak down by k/L percent from the mean nucleotide coverage.

If instead the genome is heterozygous, then the *k-mer* profile will show a characteristic two-peak profile (Kajitani, et al., 2014). The two peaks will be centered at X and 2X (e.g. 50x and 100x) representing heterozygous *k-mers* that have been sequenced to half the coverage (50x) as the homozygous *k-mers* sequenced equally from both alleles (100x). The relative heights of the peaks will be determined by the overall rate of heterozygosity: low heterozygosity causes the first peak to be relatively short and higher rates of heterozygosity causes the first peak to grow taller and taller (**Supp. Fig 1**). Since each heterozygous base can create 2*k heterozygous *k-mers*, the height of the heterozygous peak grows very quickly and matches the height of the homozygous peak at around only 1.2% heterozygosity for k=21.

Repeats have the opposite effect and cause *k-mers* to be present more than 2X times, proportional to the number of copies of that repeat in the genome. Sequencing errors and read duplications distort the distribution by creating false *k-mers* that only occur a

few numbers of times or artificially creating duplicated *k-mers*, which reduces the effective average coverage and increases the variance of the distributions.



Supplementary Figure 1. Impact of heterozygosity on the *k-mer* profile. *K-mer* profiles were draw from 100x sequencing coverage of simulated reads with 0.1%, 1% and 2% heterozygosity embedded into the *D. melanogaster* reference genome.

1.2 GenomeScope Heterozygosity Model

Many of the above observations have been previously reported (Chikhi and Medvedev, 2014; Kajitani, et al., 2014; Liu, et al., 2013; Melsted and Halldorsson, 2014; Simpson, 2014), and related techniques have been implemented within a few genome assemblers (Bankevich, et al., 2012; Gnerre, et al., 2011). However, here we propose a novel, more complete model of heterozygosity and an easy to use automated tool to infer the characteristics of a genome from a *k-mer* profile that accounts for heterozygosity, repeats, and over-dispersion in the sequence data. It also robustly evaluates the sequencing error of the reads, and avoids over-represented *k-mers* that commonly occur as technical artifacts that distort the genome size estimation.

Before presenting the full GenomeScope heterozygosity model, it is illustrative to first consider the simpler scenario when a diploid genome contains no repetitive sequences so that every k-mer in the genome is distinct. For this analysis, consider that both the maternal and paternal haplotypes are each G nucleotides long and that heterozygous bases are substitutions in the paternal haplotype. If the diploid genome contains no heterozygous bases, then the *k-mer* profile of the diploid genome will contain exactly G *k-mers*, and the *k-mer* profile of the read sequences will show a single peak centered at twice the average haploid coverage level λ.

If instead the diploid genome has a non-zero heterozygosity rate `r`, then those heterozygous bases will create additional *k-mers* beyond the original `G` *k-mers*. Note that if r is the probability that a given base is heterozygous, then `1-r` is the probability that a given base is not heterozygous (i.e. homozygous). Furthermore, $(1-r)^k$ is the probability that a given *k-mer* is homozygous, and $1-(1-r)^k$ is the probability that a *k-mer* is heterozygous in at least once nucleotide. As a result, there will be $G*(1-r)^k$ homozygous *k-mers* and $2*G*(1-(1-r)^k)$ heterozygous *k-mers*. Of the heterozygous k-mers, $G*(1-(1-r)^k)$ will originate on the maternal haplotype and an additional $G*(1-(1-r)^k)$ *k-mers* will originate on the paternal haplotype. Consequently, the total number of *k-mers* present in the diploid genome will no longer be `G`, but rather will depend on the rate of heterozygosity and equal $(1+(1-(1-r)^k)*G$. At high rates of heterozygosity near 100%, the total number of *k-mers* present in the diploid genome will equal `2*G` meaning that that every k-mer in the maternal and paternal haplotypes is different.

Under this scenario, the k-mer profile of the reads will show two distinct peaks with heights proportional to the rate of heterozygosity. For example, assume the homozygous sequence "GTA" would occur 80 times in the reads, but because it is heterozygous as T or A, creates the *k-mers* "GTA" and "GAA" that each occur 40 times on average. As a result the homozygous *k-mer* distribution in the reads (mean=2λ=80) will be centered at twice the coverage of heterozygous *k-mers* (mean=λ=40). The heterozygous k-mer peak will have twice the amplitude as the number of heterozygous positions because heterozygous bases create a pair of heterozygous k-mers.

The full GenomeScope model builds on this analysis to consider the interplay between heterozygosity, repeats, sequencing depth, and sequencing biases. Central to our method is the following equation to describe the impact of heterozygosity and repeats on the *k-mer* distribution:

$$f(\mathbf{X};\alpha,\beta,\gamma,\delta,\lambda,\rho,G) = G*(\alpha\mathcal{NB}(X;\lambda,\lambda/\rho)+\beta\mathcal{NB}(X;2\lambda,2\lambda/\rho)+\gamma\mathcal{NB}(X;3\lambda,3\lambda/\rho)+\delta\mathcal{NB}(X;4\lambda,4\lambda/\rho))$$

where

G is a scaling parameter w.r.t to the genome size,

$\alpha,\beta,\gamma,\delta$ are the mixing weights for each distribution,

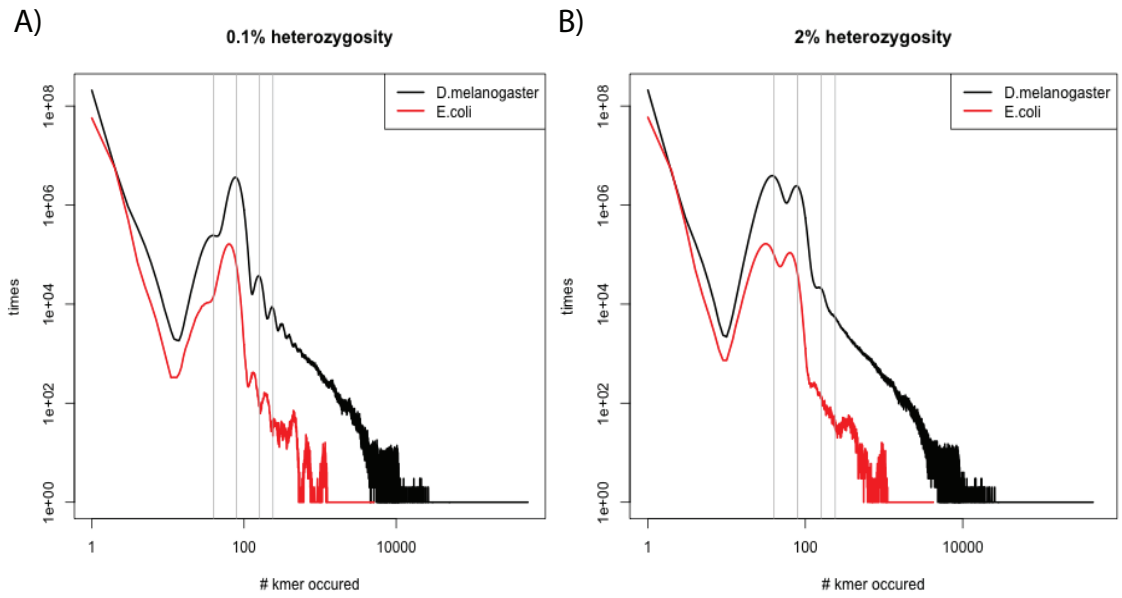
λ is the mean of a distribution,

ρ is the variance parameter of a distribution

(Eq. 2)

The equation describes the shape of the *k-mer* profile f(X) as a mixture model of four negative binomial terms where *`NB(x,mu,size)`* is the negative binomial distribution given a vector of coverage values, *`x`*, the mean, *`mu`*, and the dispersion parameter, *`size`*. We use a mixture of negative binomial model terms rather than Poisson terms because real sequencing data is often over-dispersed compared to a Poisson distribution and the `size` term can independently control the variance (Miller, et al., 2011). In the model, we determine the `size` term using a single term ρ that accounts for the overall rate of read duplications relative to the mean sequence coverage for that peak. Finally, the mixture model is scaled by a constant G, proportional to the haploid genome size (also see below for more details on genome size estimation).

Each of the four negative binomial distributions is scaled by a separate coefficient characterizing its respective peak: α for unique heterozygous *k-mers* that occur once in the diploid genome; β for unique homozygous *k-mers* that occur twice, γ for duplicated heterozygous *k-mers* that occur three times; and δ for duplicated homozygous *k-mers* that occur four times **(Supp. Figure 2)**. We model four peaks rather than two to account for heterozygous mutations occurring within repeats, which yields more accurate results. In principle, the modeling could be further extended to consider even higher-order repeats. However, only a small proportion of a genome sequence typically occurs in higher level repeats, since copy number typically falls off quickly in real genomes following a Zeta distribution (Kelley, et al., 2010) (Also see **Supplemental Figure 5**). It also becomes increasingly difficult to assign high frequency *k-mers* to their copy number state, since the variance in coverage associated with a copy number state becomes broader at higher coverage levels.



Supplementary Figure 2. *K-mer* profiles of simulated heterozygous genomes. In the figures, we plot the *k-mer* profiles of simulated shotgun sequencing data of *E. coli* (red) and *D. melanogaster* (black) with an overall heterozygosity rate of either 0.1% (A) or 2.0% (B). The vertical gray lines mark λ, 2*λ, 3*λ, and 4*λ, which correspond to the mean coverage values for unique heterozygous *k-mers*, unique homozygous *k-mers*, 2-copy repetitive heterozygous *k-mers*, and 2-copy repetitive homozygous *k-mers*, respectively.

We further determined the coefficients α, β, γ, and δ were related to the underlying genomic properties through the following system of equations (See also **Supplementary Figure 3)**:

$$\alpha = 2(1-d)(1-(1-r)^k) + 2d(1-(1-r)^k)^2 + 2d((1-r)^k)(1-(1-r)^k)$$
$$\beta = (1-d)((1-r)^k) + d(1-(1-r)^k)^2$$
$$\gamma = 2d((1-r)^k)(1-(1-r)^k)$$
$$\delta = d(1-r)^{2k}$$

(Eq. 3)

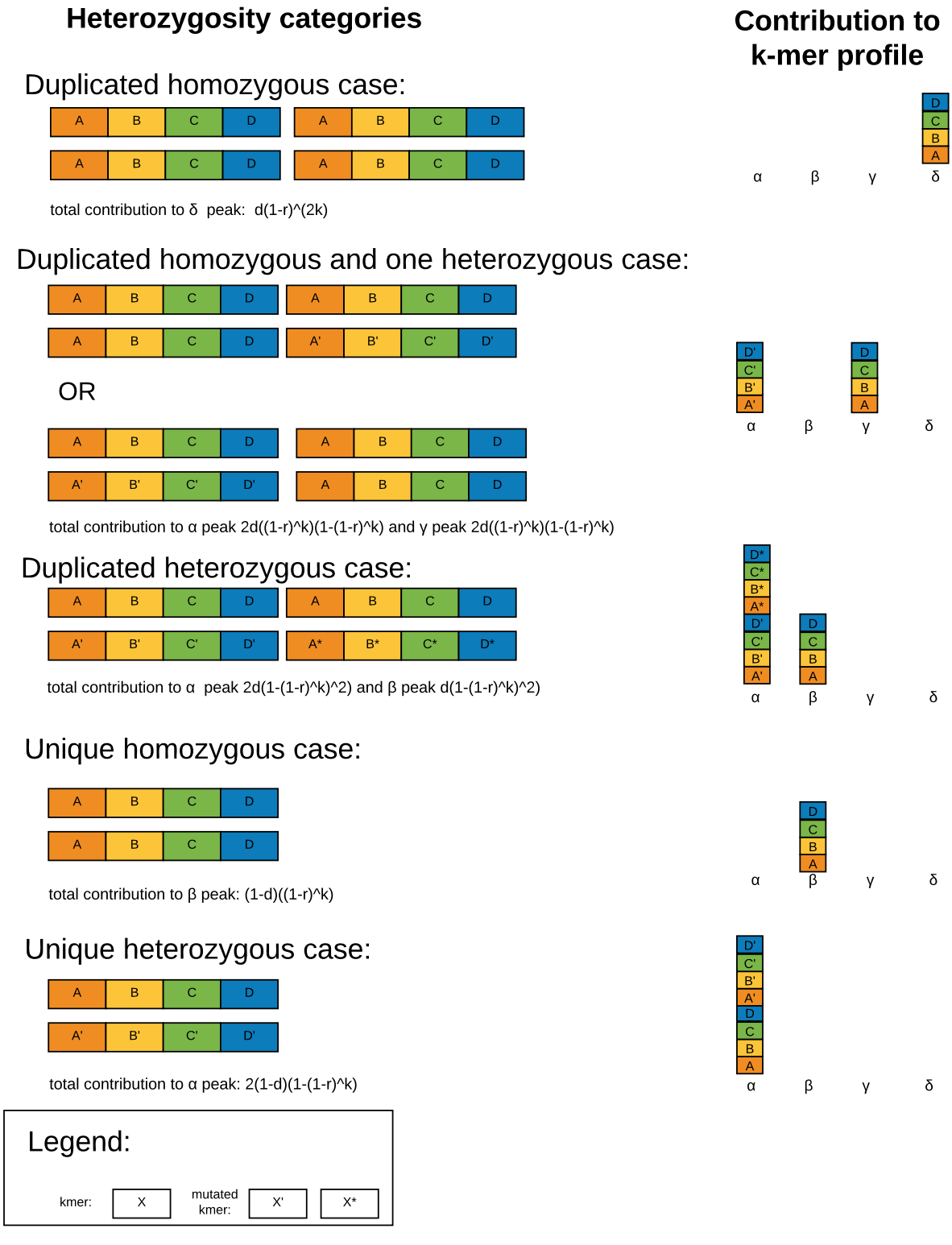
In this system of equations, k is the *k-mer* length used when constructing the *k-mer* profile, r is the rate of heterozygosity between sets of chromosomes (e.g., the percent of bases that are specific to one of the two homologous chromosomes), and d represents the percentage of the genome that is a two-copy repeat. Note that $1-d$ is the percentage of the genome that has not been duplicated (i.e. is unique), as $1-r$ is the probability that a given base is not heterozygous (i.e. is homozygous) as before.

As Eq. 3 states, the ***homozygous repeat coefficient δ*** accounts for the proportion of *k-mers* that are duplicated but had no heterozygous bases introduced for all $2k$ positions. Consequently there will be 4 copies of these *k-mers* in the diploid genome and have 4 times the *k-mer* coverage in the sequencing reads as individual heterozygous unique *k-mers*. Therefore, δ scales the negative binomial term centered at 4λ.

The ***heterozygous repeat coefficient γ*** accounts for *k-mers* that are duplicated in the e.g. maternal genome but are heterozygous in one of the two copies in the paternal genome but not both. Consequently there will be three copies of these *k-mers* in the diploid genome, so γ scales the negative binominal term centered at 3λ. In addition to these three k-mers, a new heterozygous *k-mer* on the paternal genome will also be introduced that will be counted by the α peak as explained below.

The ***homozygous unique coefficient β*** consists of two terms. The first term, $(1-d)((1-r)^k)$, accounts for the portion of the genome that is unique and had no heterozygous positions as described in repeat-free scenario above. The second term, $d(1-(1-r)^k)^2$, accounts for *k-mers* that were duplicated and then both of those duplicated *k-mers* are heterozygous in the paternal haplotype. These newly created heterozygous *k-mers* are accounted for by the α coefficient.

Finally, the ***heterozygous peak coefficient α*** is computed as the sum of three terms. The first term, $2(1-d)(1-(1-r)^k)$, accounts for heterozygous bases in the unique portion of the genome as described above for the repeat-free genome. The remaining two terms account for heterozygous bases that occur in repetitive portions of the genome. The second term, $2d(1-(1-r)^k)^2)$, counts the two new *k-mers* that are formed from β when a duplicated k-mer is heterozygous in both copies forming two new *k-mers*. The third term, $2d((1-r)^k)(1-(1-r)^k)$ counts the new *k-mers* that are formed from the heterozygous duplicated *k-mers* in γ.



Supplementary Figure 3. ***GenomeScope Heterozygosity Categories.*** The figure shows how duplications and heterozygosity impacts the *k-mer* profile by contributing *k-mers* to the four possible peaks.

The sum of α + β + γ + δ will be greater than 1 if there is a non-zero rate of heterozygosity. This is because introducing heterozygosity will create new *k-mers* relative to the haploid genome length `G` similar to what is described above for repeat-free genomes. The maximum value of the sum may be as large as `3*G` at extreme rates of duplication and heterozygosity, consisting of G k-mers from the (duplicated) maternal haplotype, and 2*G heterozygous *k-mers* from the paternal haplotype. The four coefficients can be scaled by 1/(α + β + γ + δ) so that they will sum to 1 and form a proper probability distribution for the mixture model. Equivalently, GenomeScope infers a value for G which has been scaled by 1/(α + β + γ + δ).

We stress that we discuss the *k-mers* as if they had a particular evolutionary history (e.g., first duplicated and then mutated on the paternal genome) only for convenience, but there are any number of possible mutational histories that could have lead to the observed genome. The model does not depend nor discriminate on those possible histories, only the number of occurrences of a *k-mer* in the diploid genome sequence. The model also assumes heterozygous mutations are equally likely to occur at any position in the genome and occur independently, as is commonly assumed by many variant calling algorithms (Li, 2011). These simplifying assumptions are not biologically valid, as different regions will be under different selective pressures and entire blocks may be mutated at once. However, we find the inferred properties are nevertheless highly accurate when analyzing real data from a variety of genomes with different sizes and complexities.

1.3 Model Fitting and Scoring

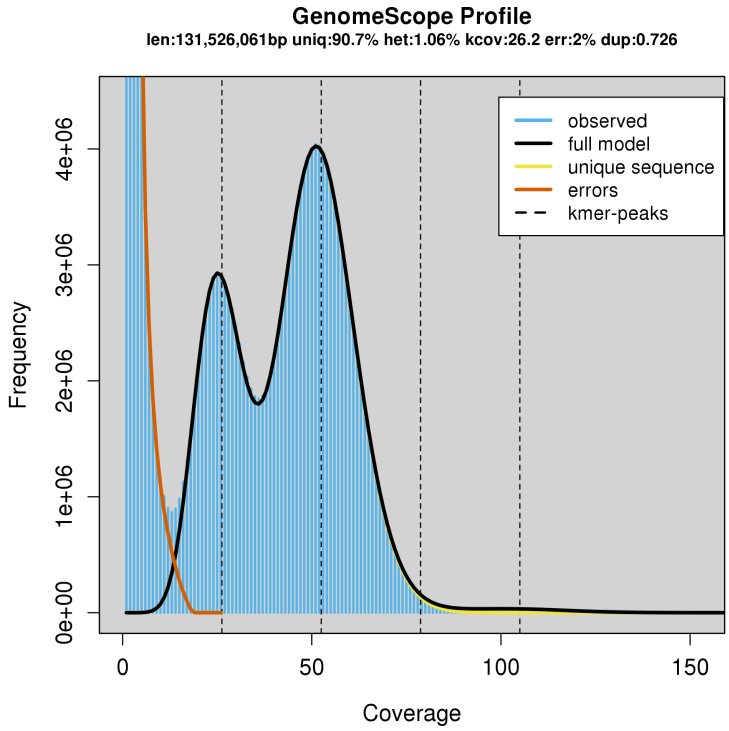
GenomeScope fits the model in Eq. 2 to the observed *k-mer* profile using a non-linear least squares estimate implemented by the `nls` function in R. Note there are five free parameters that must be inferred: `G`, `d`, `r`, λ, and ρ. The other coefficients α, β, γ, and δ can be derived from those five parameter values. GenomeScope initializes `nls` using `d=0`, `r=0`, ρ=0.5, λ=`estKmerCov` and `G=estGenomeSize` where `estKmerCov` is a naive estimate of the average *k-mer* coverage as the coverage with maximum height in the *k-mer* profile after excluding low coverage sequencing errors. It also uses this naive estimate to estimate the total genome size (`estGenomeSize`) by summing the total number of *k-mers* observed and dividing by this `estKmerCov`. Note that these values are only used as rough approximations as `nls` will optimize the individual parameters more precisely (also see below).

A major challenge analyzing these profiles on novel genomes is that it can be ambiguous if the tallest peak corresponds to the homozygous or heterozygous *k-mers*. To make the model fitting more robust it first assumes the maximum identified at estKmerCov corresponds to the unique heterozygous peak, and then performs the modeling a second time dividing estKmerCov in half in case the maximum was actually the unique homozygous peak. The final results are selected by picking the model parameters that better fit to the data (see below).

1.3.1 Sequencing Errors

GenomeScope does not attempt to explicitly model the sequencing error distribution, since we observe markedly different distributions for different datasets, due to different rates of sequencing errors, PCR duplicates, GC biases, contaminating sequences, and other causes. Instead the erroneous *k-mers* are determined empirically as any residual *k-mers* not explained by the model up to λ amount of coverage **(Supplementary Figure 4)**. It also iteratively attempts the modeling fitting several times (default: 4 iterations), each time excluding more of the low coverage *k-mers* that are presumably derived from sequencing errors (default: 5x coverage strides).

The estimate of erroneous *k-mers* is then used to estimate the percentage of erroneous bases in the reads. This is important because a single erroneous base can create as many as `k` erroneous *k-mers*, although single base errors within `k` bases of each other within a single read will create proportionally fewer erroneous *k-mers*. Consequently GenomeScope allows for multiple bases to be erroneous within a *k-mer* at a certain uniform rate `e`. The value for `e` is determined by fitting the observed number of erroneous *k-mers* to a binomial distribution allowing for the possibility of up to k errors to be present in each *k-mer* using the `uniroot` function in R.

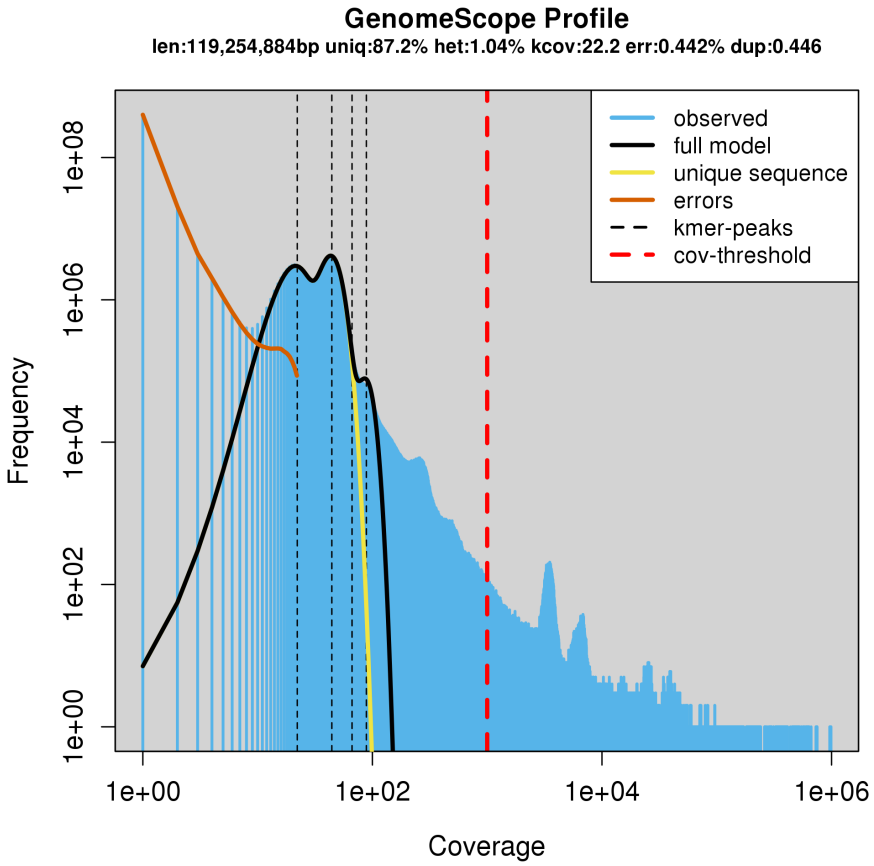


Supplementary Figure 4. Modeling results on D. melanogaster. The sequencing errors are identified by low coverage *k-mers* not explained by the model (shown in orange). This way a single cutoff value does not need to be used nor does it assume a particular shape to the distribution of the error *k-mers*. See below for more details on the *D. melanogaster* analysis.

1.3.2 Genome Size Estimation

The genome length parameter `G` in Eq 2. is not displayed to the user, since it does not include higher frequency repeats in the genome. This estimate is revised by summing the total number of *k-mers*, except presumptive sequencing errors identified as in section 1.3.1, and dividing by the `2*λ`, the estimated coverage for homozygous *k-mers*. GenomeScope also estimates the unique genome length by considering just the total number of *k-mers* that are present under the first two peaks (α + β) to the model (the unique heterozygous and unique homozygous sequences).

We introduce an additional parameter `CovMax`, which is the maximum *k-mer* coverage to be used for estimating the genome size (default 1,000x coverage). This is necessary because we observed that several datasets were "contaminated" by artificial high frequency *k-mers* such as phi-X spike-in sequences or, in the case of plant genomes, very high coverage of the chloroplasts genome **(Supplementary Figure 5)**. While these sequences are typically much smaller than the chromosomes, they can occur in many hundreds to thousands of copies per cell and their corresponding *k-mers* many contribute tens of megabases of sequence to the genome size estimate.



Supplementary Figure 5. High-frequency *k-mer* peak exclusion in genome size estimate. This shows the *k-mer* profile from the real Arabidopsis thaliana F1 sequencing described below. We investigated the unexpectedly high abundance *k-mers* occurring beyond 1000x coverage (shown with the dotted red line). We found the peaks near

3500x and 7000x coverage were highly enriched for organelle sequences: 85.8% of those *k-mers* align with bowtie (Langmead, et al., 2009) to the reference chloroplast genome and 2.9% align to the reference mitochondrial genome despite those sequences being more than 200 times smaller than the chromosomes. The later peaks beyond 10,000x coverage are enriched for phi-X (23.9% of *k-mers* align), which was used for calibrating the Illumina sequencing run.

1.3.3 Model Scoring

The final model parameters are decided by selecting the model parameters that have the smallest residual sum of square error (RSSE) between the predicted values and the observed values, excluding presumptive sequencing errors with low coverage below E, that is determined by the position where the model results first intersects the observed *k-mer* frequencies. (Eq. 4). In the event of a near tie in RSSE, the algorithm selects the model parameters that have higher rates of heterozygosity and smaller genome size rather than lower heterozygosity and larger genome size. The RSSE is reported to the user, as well as a more interpretable score representing the percentage of *k-mers* that are not captured by the model.

$$RSSE = \sum_{x=E}^{\infty} (kmer_{obs}[x] - kmer_{pred}[x])^2$$

Eq. 4

After deciding the final model parameters, the original *k-mer* profile is plotted, along with the inferred distribution from the model **(Figure 1B, Supplementary Figure 4 and below)**. A summary table is presented showing the interfered values, as well as the model score.

Supplementary Note 2: Simulated Data Sets and Analysis

To investigate the correctness of the GenomeScope predictions given different conditions such as SNP rate, sequencing error or read duplication rate we simulated 100bp Illumina-like reads from 324 simulated genomes varying in heterozygosity (0.1%, 1%, 2%), average rate of read duplication (1, 2, 3), sequencing error rate (0.1%, 1%, 2%), coverage (100x, 50x, 25x, 15x) and organism (*E. coli*, *A. thaliana*, and *D. melanogaster*)

The simulation is done in two stages:

First, we simulate SNPs along the reference genome by mutating the sequence at random locations at the specified rate. This simulation creates two copies of the reference genome to simulate the diploid genome and thus guarantee heterozygous SNPs.

Second, reads are simulated by randomly subsampling from either of the two (altered or unaltered) reference genomes. To simulate PCR duplications, a read can be subject to "duplication events" where multiple copies of the same subsequence are reported. The number of copies is determined according to a Poisson distribution with the user specified rate determining the average number of copies of each read. Each copy is further mutated separately to simulate a sequencing error at the user specified rate with a uniform error model. The reads are then stored in a fasta file.

The simulated reads are then processed with Jellyfish (Marcais and Kingsford, 2011) using kmer size of 21 (-m), counting both strands (-C) and a 1000000 (-s) limit on the histo function. The k-mer profiles are then analyzed using GenomeScope using default parameters. The unique portion of the reference genome was also determined with Jellyfish and is the number of k-mers (k=21) in the reference genome that occur exactly 1 time.

2.1 Estimating heterozygosity with read mapping and variation calling

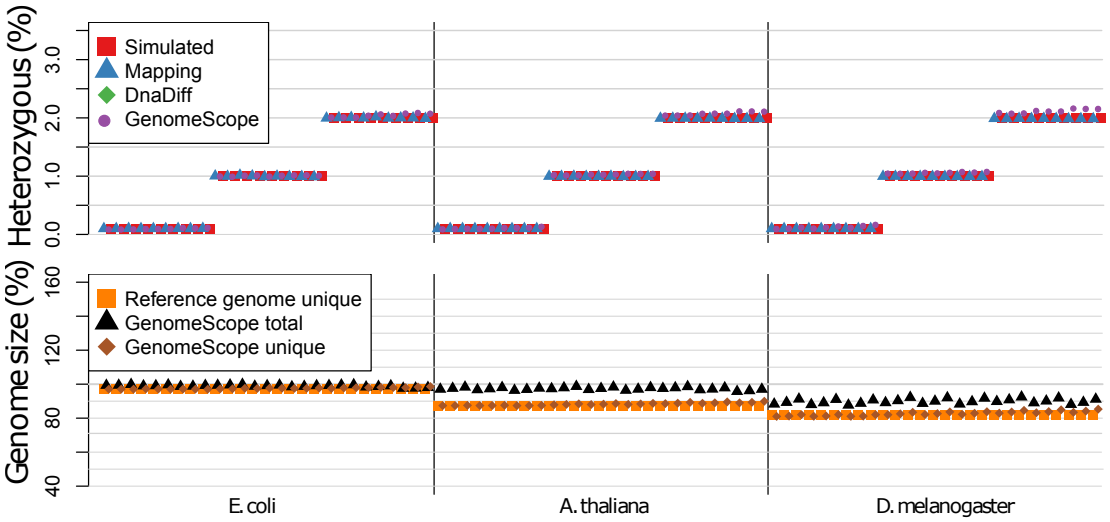
Reads were mapped using BWA-MEM (version: 0.7.12-r1039) (Li, 2013) to the individual reference genomes. Subsequently, SNPs were called with samtools mpileup (version 1.1) (Li, et al., 2009). Regions were ignored if they contained reads with a mapping quality < 10. We then counted the number of SNPs (including indels) given a minor allele frequency larger then 20%. The so obtain number was divided by the corrected reference size, excluding the regions with low mapping quality reads, to obtain the rate of heterozygosity.

2.2 Simulated Data Results

The complete results of the simulation are available in **Supplementary Table 3.** The heterozygosity and genome size estimates are shown in Supplementary **Figure 6** (selected values also shown in **Figure 1**).

With 100x and 50x sequence coverage, the GenomeScope estimates were very accurate over all conditions, with an median error of only 7.4% for heterozygosity and only 3.1% for genome size. At 25x coverage, the results were generally accurate with four instances not converging, a median error of 31.9% for heterozygosity and 5.8% in genome size. However, samples with high sequencing error rate and high read duplication had relatively poor performance (max error in heterozygosity of 0.1% simulated but 0.85% estimated and 17.6% error in genome size).

At only 15x coverage, the model often did not converge (36 of 81 trials) or if it did converge produced relatively poor estimates of the genome characteristics. Fortunately, poor results were easily recognized by the poor fit of the model to the observed sequence data and we strongly recommend users provide at least 50x sequencing coverage for robust results.



Supplementary Figure 6. Simulated results with 100x. For each of the simulated datasets, the simulated parameters and GenomeScope results are plotted for the rate of heterozygosity (top) and estimated genome size (bottom). The ordering of points within each set is displayed in increasing complexity from low error and duplicates rates to high sequencing error and duplication rates. Specific values for all 4 datasets (100x, 50x, 25, 15x coverage) are available in **Supplementary Table 3**.

Supplementary Note 3: Synthetic Heterozygous E. coli Analysis

To assess the performance of GenomeScope with real data we downloaded five *E. coli* data sets listed in **Supplementary Table 1**. All reads were trimmed to 100bp and equal numbers of reads from the pairs of accessions were sampled together to simulate 10 different heterozygous populations (E1-E10). The advantage of this approach is the finished reference genomes for each data set is available so that heterozygosity can be precisely measured. The GenomeScope analysis used default parameters, and Jellyfish was used to measure the unique reference genome size by counting the number of unique k-mers for the reference genome of the pair. The full results of the analysis are available in **Supplementary Table 4** and displayed in **Figure 1**.

Genome	Read Accession	Reference Sequence
E. coli ATCC 8739	ERR351256	gi|169752989|gb|CP000946.1|
E. coli O157:H7 str. Sakai	SRR530851	gi|15829254|ref|NC_002695.1|
E. coli REL606	SRR733088	gi|253972022|gb|CP000819.1|
E. coli BL21(DE3)	SRR941832	gi|387823261|ref|NC_012892.2|
E. coli UTI89	ERR687900	gi|91209055|ref|NC_007946.1|

Data set	Accession 1	Accession 2	Genome mapped to
E1	ERR351256	ERR687900	gi|91209055|ref|NC_007946.1|
E2	ERR351256	SRR530851	gi|15829254|ref|NC_002695.1|
E3	SRR530851	ERR687900	gi|91209055|ref|NC_007946.1|
E4	SRR733088	ERR351256	gi|169752989|gb|CP000946.1|
E5	SRR733088	ERR687900	gi|91209055|ref|NC_007946.1|
E6	SRR733088	SRR530851	gi|15829254|ref|NC_002695.1|
E7	SRR941832	ERR351256	gi|169752989|gb|CP000946.1|
E8	SRR941832	ERR687900	gi|91209055|ref|NC_007946.1|
E9	SRR941832	SRR530851	gi|15829254|ref|NC_002695.1|
E10	SRR941832	SRR733088	gi|253972022|gb|CP000819.1|

Supplementary Table 1. Synthetic E. coli populations. Datasets used for constructing the synthetic heterozygous E. coli samples.

3.1 Estimating heterozygosity with read mapping and variation calling

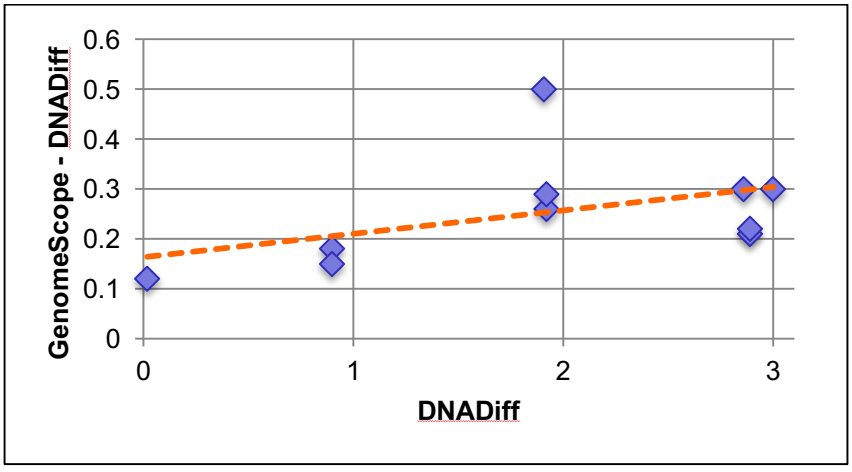
Reads were mapped using BWA-MEM and SAMTools as above, using the listed genomes as the reference genome for mapping (accession listed in **Supplemental Table 1**).

The results between GenomeScope and the mapping approach were related, although the GenomeScope estimated rate of heterozygosity was consistently higher than the results determined from the mapping results. We investigated this, and determined a major reason for this was BWA-MEM would fail to map reads to highly divergent regions between the two genomes and/or regions only occurring in one of the two genomes. Across all samples, BWA-MEM had an average mapping rate of only 92.2% (89.8% for MQ>10). Consequently, the variant caller would not be able to identify any heterozygous variants in those regions, thus artificially reducing the effective rate of heterozygosity.

3.2 Estimating heterozygosity with whole genome alignment

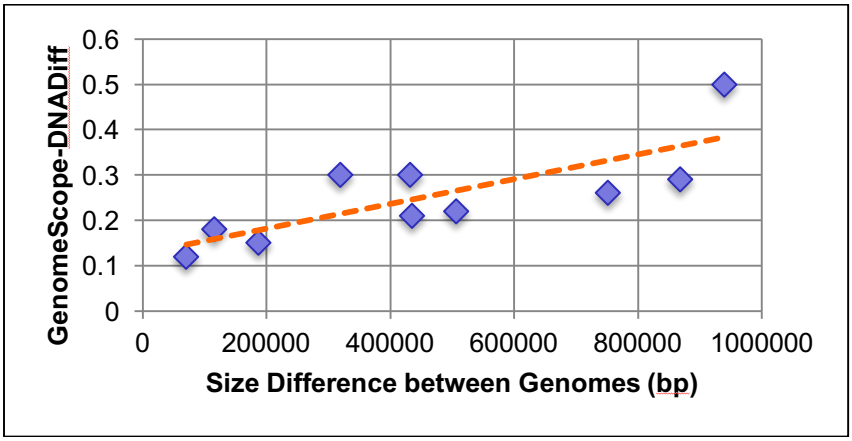
We aligned the two *E. coli* source genomes using the whole genome alignment algorithm MUMmer (Kurtz, et al., 2004) and used the associated tool dnadiff (Phillippy, et al., 2008) to compute the overall similarity between the genomes using the reported 1-to-1 matching identity. From this we converted the identity rate to the heterozygosity rate as 1 – (identity rate).

Interestingly, we noted that while the GenomeScope results were very similar to the estimates from MUMmer/dnadiff and more similar than the mapping results, GenomeScope would often estimate the rate of heterozygosity to be higher than estimate from MUMmer (0.1% to 0.5% higher). We investigated this relationship and noted that the difference in estimated heterozygosity rate increased as a function of the dnadiff estimated rate (**Supplementary Figure 7**).



Supplementary Figure 7. Difference between GenomeScope and DNADiff heterozygosity estimates. For each of the 10 samples (EC1-EC10), we plotted the difference in estimated heterozygosity rate to the result reported by dnadiff (blue diamonds). The orange line shows the linear regression of those points.

We further investigated this relationship and noted that the difference was even more strongly associated with the size difference between the two reference genomes **(Supplementary Figure 8).** We conclude that MUMmer/dnadiff is underestimating the rate of heterozygosity because it does not include regions that do not align in its calculation. Most significantly, if one genome is appreciably larger than the other, those extra bases cannot possibly align in a 1-to-1 manner, and are therefore not considered. In contrast, the full genome sequences will be included in the GenomeScope even if one "haplotype" is larger than the other or in regions of extreme sequence diversity.



Supplementary Figure 8. Difference between GenomeScope and DNADiff heterozygosity estimates. Blue diamonds show the difference in estimated heterozygosity as a function of the size difference between the genomes. The orange line shows the linear regression of those points.

Supplementary Note 4: Genuine Heterozygous Genome Analysis

Our final analysis was to study the performance of GenomeScope on several real datasets of large heterozygous diploid genomes. The accession numbers of the sequencing data used are listed in **Supplemental Table 2**, as well as the reference genomes used for read mapping. We focused on datasets with sufficiently deep coverage (>30x) and higher levels of heterozygosity (~0.5% or greater), such as heterozygous wild types or crosses of divergent lines, in order to challenge the inference algorithm over a wide range of heterozygosity rates.

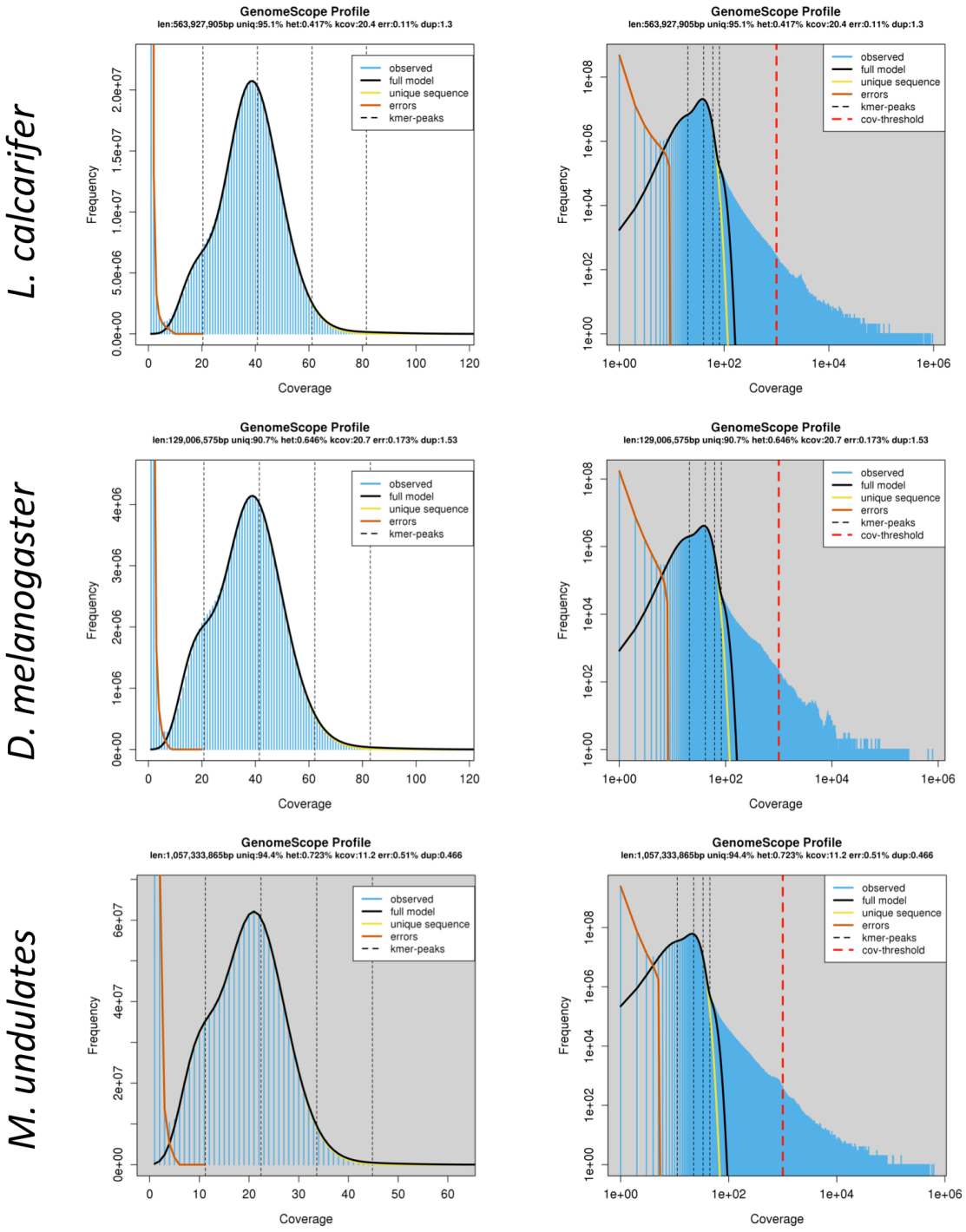
Sample	Species	Accession & Citation	Reference Genome Contig N50 Size
L. cal	*L. calcarifer* (Asian seabass)	SRP069219 (Vij, et al., 2016)	LLXD00000000 1.06 Mbp
D. mel	*D. melanogaster* (Fruit fly)	120317_I247_FCD0RDBACXX_L2_SZAXPI006700-32 (Keightley, et al., 2014)	Dmel6 21.4 Mbp
M. und	*M. undulates* (Budgerigar)	ERR244146 (Bradnam, et al., 2013)	MelUnd6.3 55.6kbp
A. tha	*A. thaliana Col-0 x Cvi-0* (Thale cress)	SRX1865253 (Chin, et al., 2016)	TAIR10 11.1 Mbp
P. bre	*P. bretschneideri* (Asiatic Pear)	SRP016889 (Wu, et al., 2013)	AJSU00000000 19.4 kbp
C. gig	*C. gigas* (Pacific oyster)	SRP045757 (Zhang, et al., 2012)	AFTI00000000 35.7 kbp

Supplemental Table 2. Heterozygous genome datasets. Genuine heterozygous sequencing data and reference genomes used in the analysis.

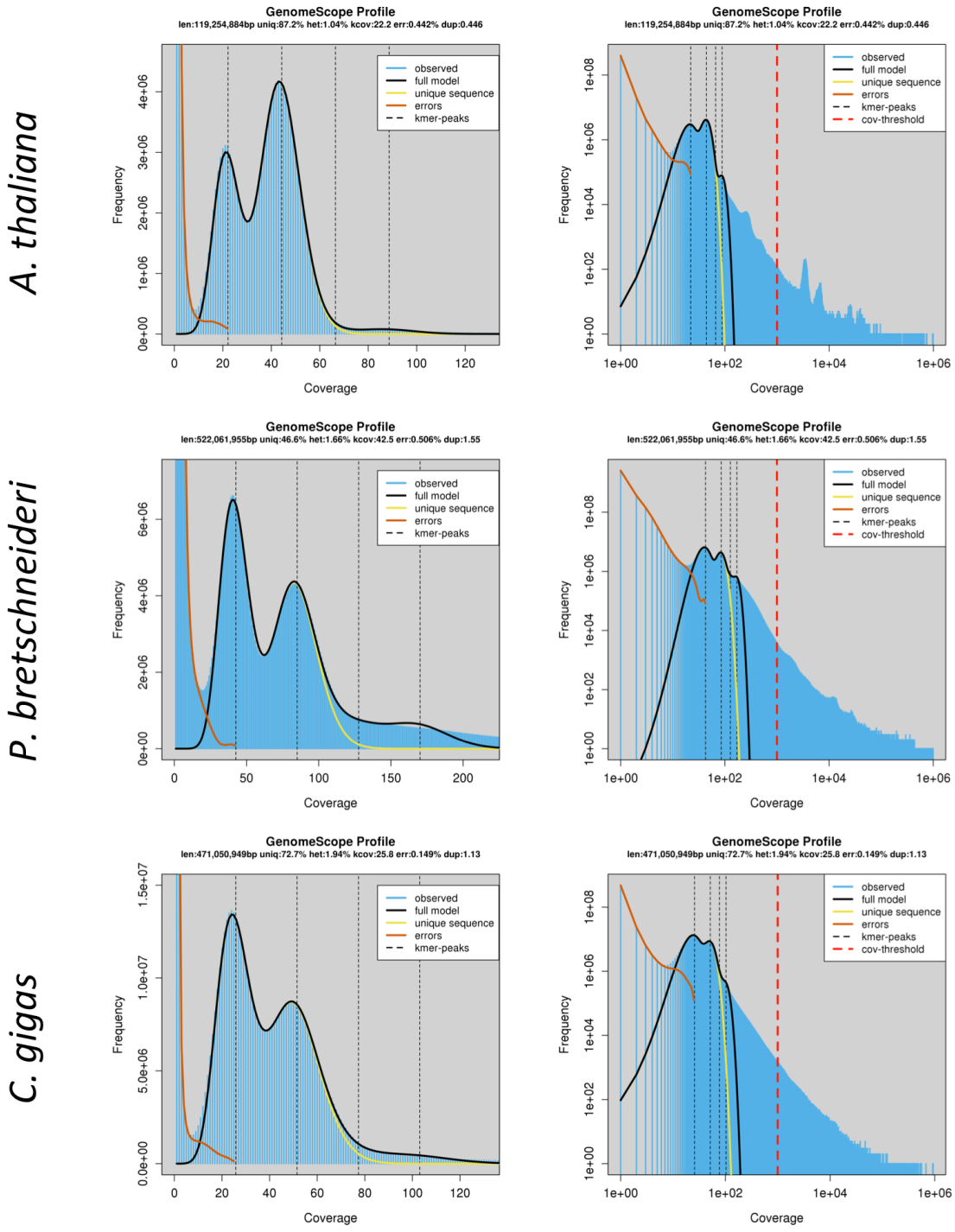
4.1 Estimating heterozygosity with read mapping and variation calling

The read mapping and SNP calling was performed as described in section 2.1. **Supplementary Table 5** has the full results comparing the GenomeScope results to the samtools results and **Supplementary Figure 9a and 9b** shows the k-mer profiles and model fits for all 6 genomes. As was observed in the above datasets, the GenomeScope results generally agree with the samtools results, although GenomeScope identifies a higher rate of heterozygosity in some of the genomes.

The genomes with the largest reported difference were *C. gigas* and *P. bretschneideri,* most likely because these were the lowest quality genomes with contig N50 sizes of only 19.4kbp and 35kbp, respectively, compared to >1Mbp for the model species *A. thaliana* or *D. melanogaster*. In particular, draft de novo assemblies of diploid genomes are known to have false "segmental duplications" where heterozygous regions are assembled separately (Kelley and Salzberg, 2010). In these regions, the mapping based approach would show no heterozygous variants, even though the entire sequence may be heterozygous. Overall, we are confident in the results computed, especially after filtering out the very high frequency k-mers in Arabidopsis profile distorted the genome size estimates (See **Supplemental Section 1.3.2**).



Supplementary Figure 9a. GenomeScope results in linear (left) and log (right) coordinates on the genuine sequencing datasets.



Supplementary Figure 9b. GenomeScope results in linear (left) and log (right) coordinates on the genuine sequencing datasets.

Supplementary References

Bankevich, A., *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19(5):455-477.

Bradnam, K.R., *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2013;2(1):10.

Chikhi, R. and Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 2014;30(1):31-37.

Chin, C.S., *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 2016;13(12):1050-1054.

Deorowicz, S., *et al.* KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics* 2015;31(10):1569-1576.

Gnerre, S., *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 2011;108(4):1513-1518.

Kajitani, R., *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* 2014;24(8):1384-1395.

Keightley, P.D., *et al.* Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. *Genetics* 2014;196(1):313-320.

Kelley, D.R. and Salzberg, S.L. Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biol* 2010;11(3):R28.

Kelley, D.R., Schatz, M.C. and Salzberg, S.L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol* 2010;11(11):R116.

Kurtz, S., *et al.* Versatile and open software for comparing large genomes. *Genome Biol* 2004;5(2):R12.

Lander, E.S. and Waterman, M.S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 1988;2(3):231-239.

Langmead, B., *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10(3):R25.

Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27(21):2987-2993.

Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints* 2013.

Li, H., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.

Liu, B., *et al.* Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv* 2013;1308.2012

Marcais, G. and Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;27(6):764-770.

Melsted, P. and Halldorsson, B.V. KmerStream: streaming algorithms for k-mer abundance estimation. *Bioinformatics* 2014;30(24):3541-3547.

Miller, C.A., *et al.* ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* 2011;6(1):e16327.

Phillippy, A.M., Schatz, M.C. and Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol* 2008;9(3):R55.

Simpson, J.T. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* 2014;30(9):1228-1235.

Vij, S., *et al.* Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. *PLoS Genet* 2016;12(4):e1005954.

Wu, J., *et al.* The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Res* 2013;23(2):396-408.

Zhang, G., *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 2012;490(7418):49-54.