# GenomeScope: Fast genome analysis from unassembled short reads

Greg Vurture[1,2], Fritz J. Sedlazeck[1,3], Maria Nattestad[1], Charles Underwood[1], Han Fang[1,4], James Gurtowski[1], Michael C. Schatz[1,3]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
[2]Department of Mathematics, New York University, New York, NY
[3]Department of Computer Science, Johns Hopkins University, Baltimore, MD
[4]Department of Applied Mathematics, Stony Brook University, Stony Brook, NY

## Abstract

Current developments in de novo assembly technologies have been focused on relatively simple genomes. Even the human genome, with a heterozygosity rate of only ~0.1% and 2n diploid structure, is significantly simpler than many other species, especially plants. However, genomics is rapidly advancing towards sequencing more complex species such as pineapple, sugarcane, or wheat that have much higher rates of heterozygosity (>1% for pineapple), much higher ploidy (8n for sugarcane), and much larger genomes (16Gbp for wheat).

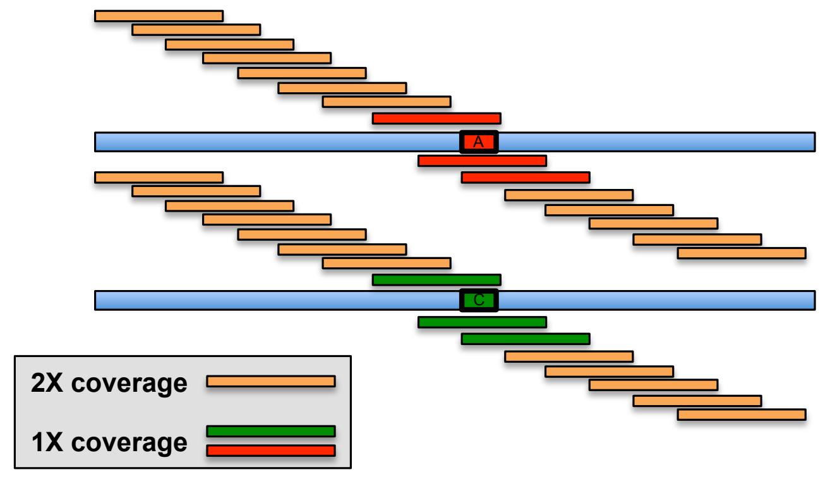
One of the first goals when sequencing a new species is determining the overall characteristics of the genome structure, including the genome size, abundance of repetitive elements, and the rate of heterozygosity. These features are needed to study trends in genome evolution, and can inform the parameters that should be used for the individual assembly steps. They can also serve as an independent quality control during any analysis, such as quantifying the quality of an assembly, or measuring the expected number of heterozygous bases in the genome before mapping any variants.

We have developed an analytical model and open-source software package GenomeScope that can infer the global properties of a genome from unassembled sequenced data. GenomeScope uses the k-mer count distribution, e.g. from Jellyfish, and within seconds produces a report and several informative plots describing the genome properties. We validate the approach on simulated heterozygous genomes, as well as synthetic crosses of related strains of microbial and eukaryotic genomes with known reference genomes. GenomeScope was also applied to study the characteristics of several novel species, including pineapple, pear, the regenerative flatworm *Macrostomum lignano*, and the Asian sea bass.

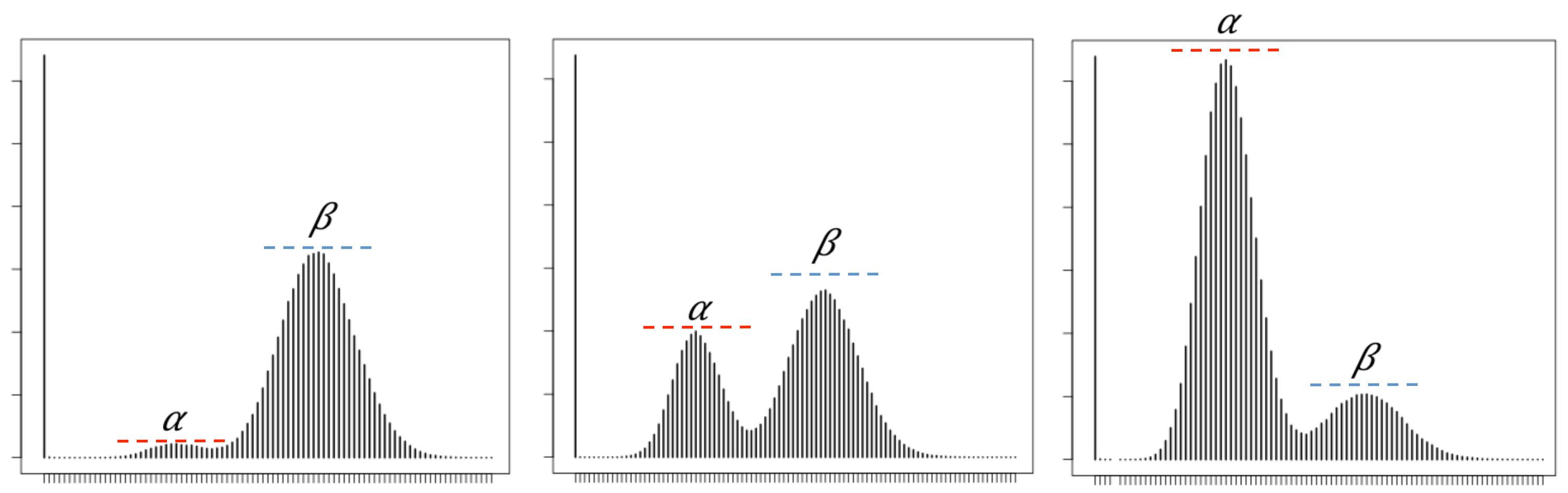
## Background

The advent of high throughput sequencing enables the assessment of novel genomes and the resequencing of known genomes on a daily bases. However, even the most basic characteristics of the genomes, such as the genome size, repeat composition or heterozygosity rate, may not be known, making it difficult to select the most appropriate analysis methods e.g. mapper, de novo assembler, or SNP caller. Furthermore, establishing these characteristics in advance can reveal if the analysis methods are not capturing the full complexity of the genome, such as underreporting the expected number of variants or not assembling a significant fraction of the genome. While experimental methods are available for determining some of these characteristics, they can be expensive and laborious to perform.

The analysis of k-mers (substrings of length k) is a powerful "reference-free" technique to study genomes from unassembled raw sequencing data. A number of software packages, including Jellyfish, KMC, and others are able to rapidly scan hundreds to thousands of gigabases of sequence data per hour to count the number of occurrences of every k-mer present in the dataset. Once these counts are available, GenomeScope analyzes the distribution of k-mers frequencies to infer the genome characteristics.



For example, if a genome has been sequenced to 50x coverage, then the k-mers should occur about 50 times on average, minus a shift proportional to the k-mer length divided by the read length. Absent any sequencing biases, this distribution will be well characterized by a Poisson distribution centered at the average k-mer frequency. This allows for the genome size to be estimated by identifying the peak of the distribution relative to the total amount of sequencing done. Repetitive sequences are also easily recognized as k-mers occurring more frequently than expected by this peak. However, if a diploid genome has an appreciable rate of heterozygous variants, the k-mer coverage distribution will be bimodal with a second peak at half the coverage of the first representing the heterozygous k-mers. The ratio of the height of the heterozygous first peak, labeled $\alpha$, to that of the homozygous second peak, labeled $\beta$, is directly proportional to the heterozygosity rate in a predictable manner.



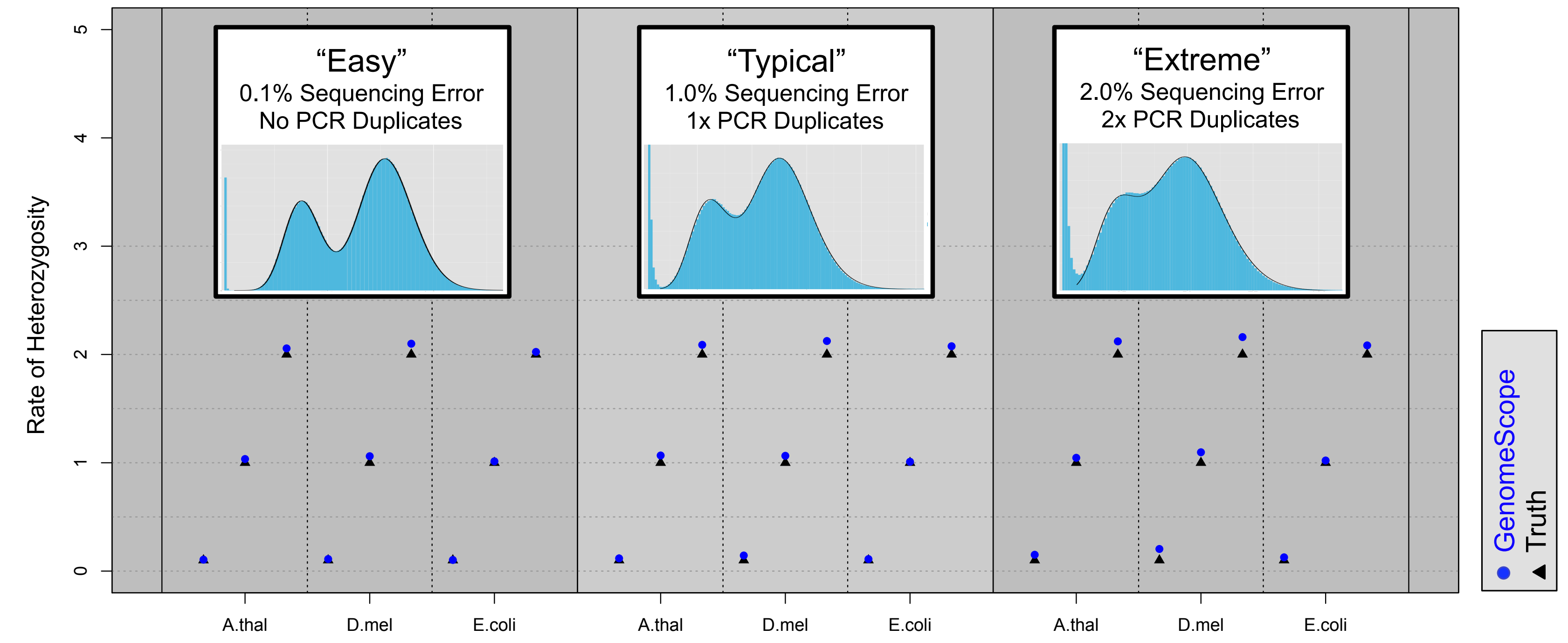
0.1% heterozygosity 1% heterozygosity 5% heterozygosity

Using this relationship, GenomeScope can quickly infer the heterozygosity rate and other genome characteristics from the k-mer distribution using a mixture model composed of 4 negative binomial (*NB*) terms scaled by the genome size G. The first two terms capture the heterozygous and homozygous k-mers for the unique regions of the genome, and the last two terms model the k-mer coverage of repetitive elements. The model parameters are determined using non-linear least squares regression (NLS) implemented in R.

$$f(x) = G\Big\{\alpha NB(x,\lambda,\lambda/\rho) + \beta NB(x,2\lambda,2\lambda/\rho) + \gamma NB(x,3\lambda,3\lambda/\rho) + \delta NB(x,4\lambda,4\lambda/\rho)\Big\}$$

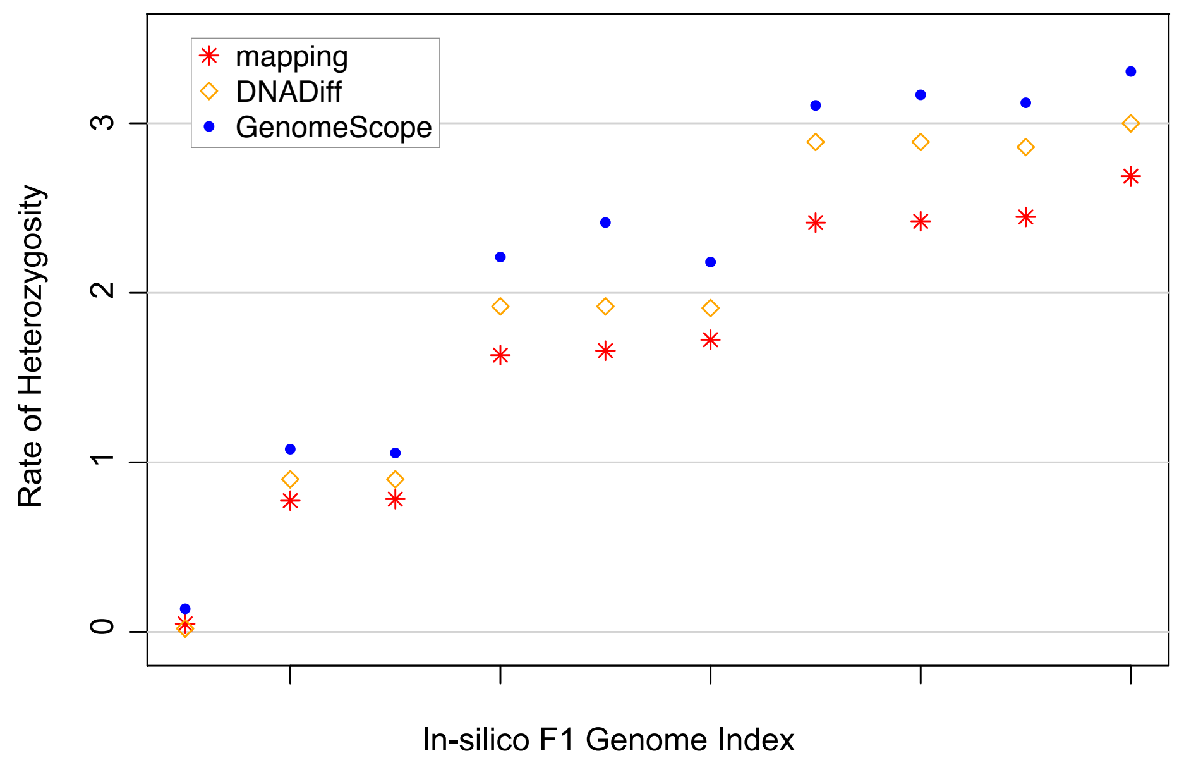
## Validation

We first validate GenomeScope with extensive simulation embedding heterozygous variants at a known rates into the genomes of 3 important model organisms: *A. thaliana*, *D. melanogaster*, and *E. coli*. From each diploid genome, we then sample 100x coverage of 100bp Illumina-like reads with different amounts of sequencing error and PCR duplicates. Higher rates of PCR duplicates increase the variance of the coverage distributions, necessitating modeling the coverage as a negative binomial rather than a Poisson distribution. The true heterozygosity rate and the GenomeScope estimates are presented below, and show that GenomeScope can accurately infer the rate of heterozygosity under a wide range of errors and sequencing biases. The estimates of the genome size and repetitive sequence content were also very accurate.

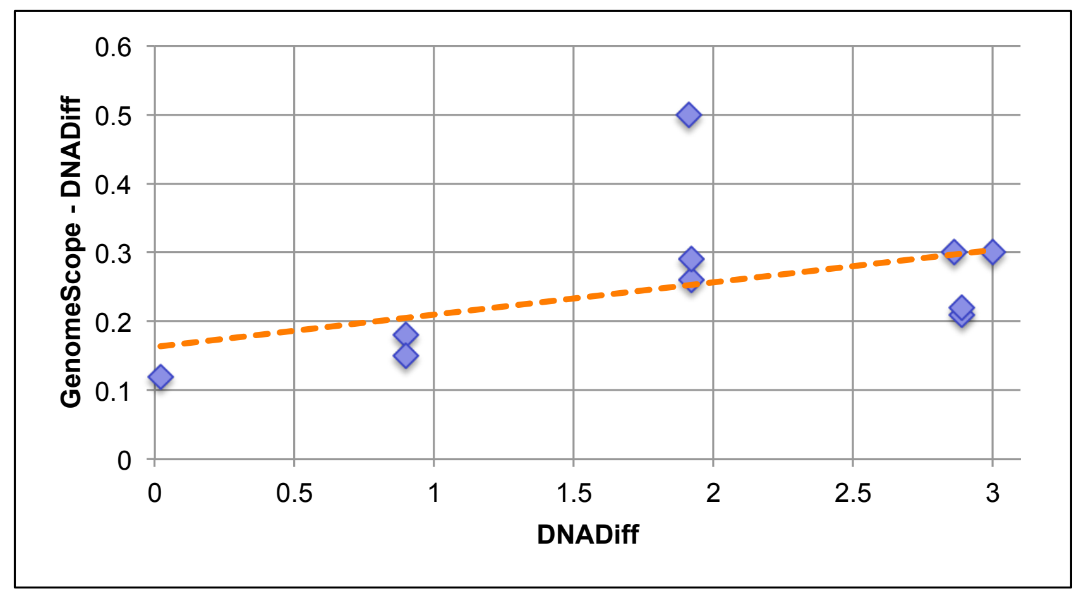


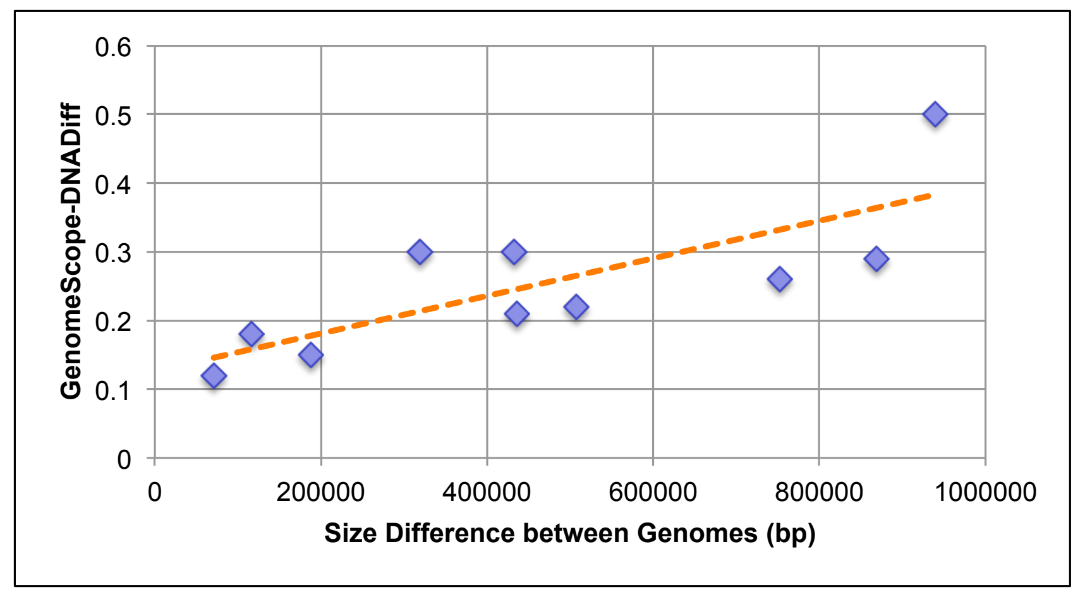
## In silico F1 sequencing

We further validate GenomeScope by analyzing genuine Illumina sequencing data (100bp paired-end reads) from 10 *in silico* F1 genomes. In each experiment, we combine 50x coverage from 2 different strains of *E. coli*, each of which has a finished genome available. This allows us to systematically explore a wide range of heterozygosity rates (from 0.1% to over 3%). It also allows us to compare the GenomeScope results to those computed by whole genome alignment of the reference genomes using DNADiff from the MUMmer package. We also computed the results from a standard SNP calling approach using NGM, a mapper specifically designed for heterozygous genomes, and the SAMtools SNP calling pipeline.



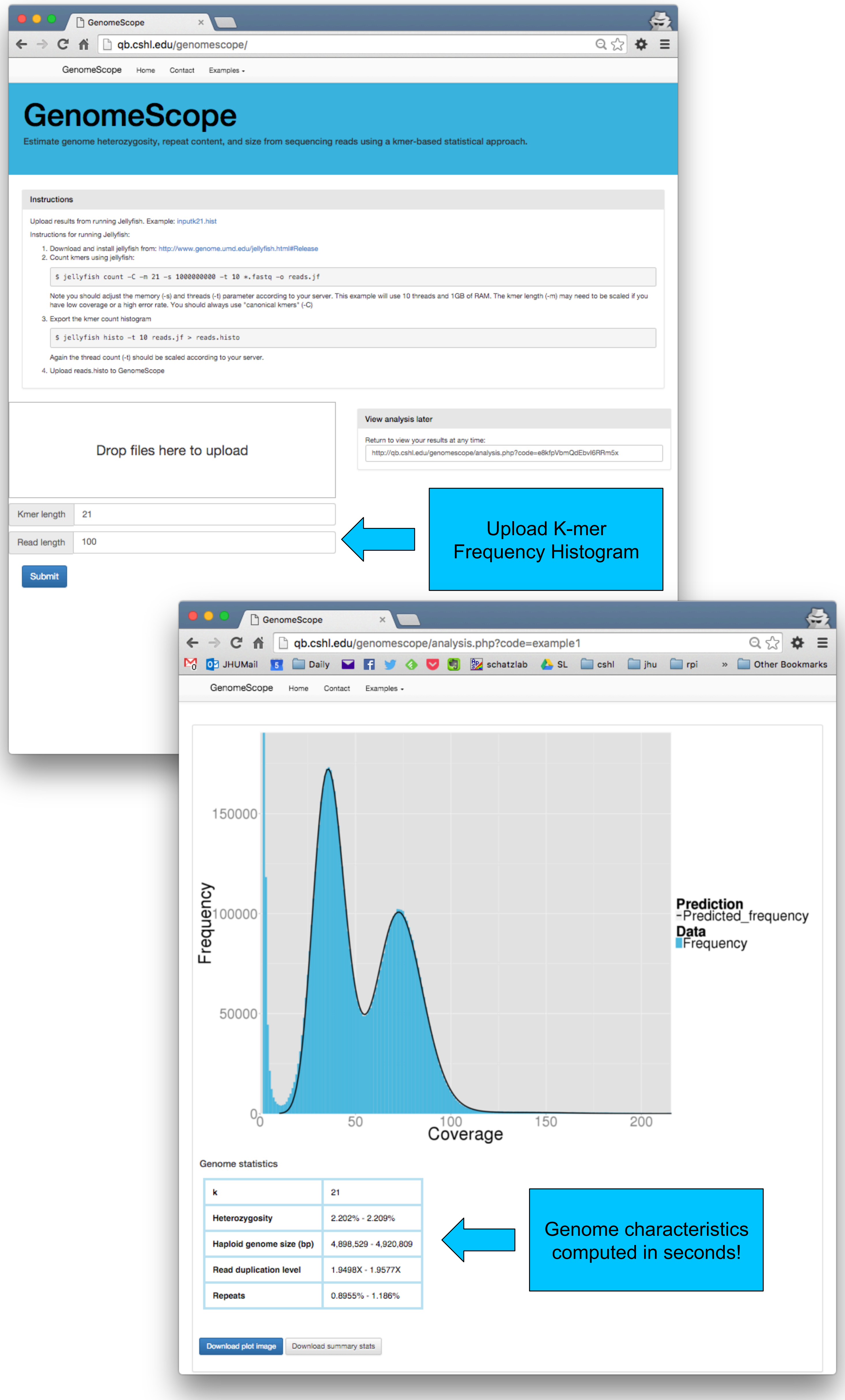
We observe a strong correlation between GenomeScope, DNADiff, and the mapping results in their estimated rate of heterozygosity, although GenomeScope is consistently higher than DNADiff measures, which is itself higher than the mapping results. We further determined the difference in heterozygosity rate between GenomeScope and DNADiff to be related to the rate of heterozygosity computed by DNADiff (bottom left), and further related to the size difference between the reference genomes used in the study (bottom right). From this we conclude that DNADiff is underestimating the true rate of heterozygosity present because it does not consider portions of the genomes that do not align to each other, while GenomeScope performs an unbiased genome-wide analysis. For similar reasons, the heterozygosity results determined by read mapping are consistently below the other approaches, because it can only identify variants within highly mappable regions, i.e. non-repetitive regions with relatively low rates of differences compared to the reference where reads can be mapped.





## GenomeScope: Genome Analysis in Seconds

We have since used GenomeScope on dozens of genome projects to guide downstream analysis ranging from small microbial genomes to gigabase mammalian and plant genomes. To simplify its operation, we have packaged the statistical modeling as a web app so that can be used with any sequencing project. To use it, simply upload your k-mer frequency histogram from Jellyfish or other tools and seconds later it will display a plot of the histogram with the results of the statistical modeling.



GenomeScope

Estimate genome heterozygosity, repeat content, and size from sequencing reads using a kmer-based statistical approach.

Instructions

Upload results from running Jellyfish. Example: inputk21.hist

Instructions for running Jellyfish:

1. Download and install jellyfish from: http://www.genome.umd.edu/jellyfish.html#Release
2. Count kmers using jellyfish:

```
$ jellyfish count -C -m 21 -s 1000000000 -t 10 *.fastq -o reads.jf
```

   Note you should adjust the memory (-s) and threads (-t) parameter according to your server. This example will use 10 threads and 1GB of RAM. The kmer length (-m) may need to be scaled if you have low coverage or a high error rate. You should always use "canonical kmers" (-C)

3. Export the kmer count histogram

```
$ jellyfish histo -t 10 reads.jf > reads.histo
```

   Again the thread count (-t) should be scaled according to your server.

4. Upload reads.histo to GenomeScope

Drop files here to upload

View analysis later

Return to view your results at any time:

http://qb.cshl.edu/genomescope/analysis.php?code=e8kfpVbmQdEbvl6RRm5x

Kmer length: 21

Read length: 100

Upload K-mer Frequency Histogram

Genome statistics

| | |
|---|---|
| k | 21 |
| Heterozygosity | 2.202% - 2.209% |
| Haploid genome size (bp) | 4,898,529 - 4,920,809 |
| Read duplication level | 1.9498X - 1.9577X |
| Repeats | 0.8955% - 1.186% |

Genome characteristics computed in seconds!

***Try GenomeScope Today!***
***http://qb.cshl.edu/genomescope/***