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, genetics 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 diploid structures for sugarcane, and many other large 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 covering 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 are not always known, such as the genome size, repeat composition or heterozygosity rate. While experimental methods are available for determining some of these characteristics, they can be expensive and laborious to perform.

Validation

We use a mixture model to quantify the quality of an assembly, or the expected number of heterozygous bases in the genome before mapping any variants. The model is based on the distribution of k-mers that are shared between multiple genomes or genomic regions.

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 50x 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 a diploid F1 genome. 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 reference results from a string-gapped comparison approach using NGM, a mapper specifically designed for heterozygous genomes, and the SAMtools SNP calling pipeline.

GenomeScope: Fast genome analysis from unassembled short reads

We have since used GenomeScope on dozens of genome projects to guide downstream analysis ranging from small microbial genomes to giant 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: Genome Analysis in Seconds

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 determine the 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 understanding 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.