Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Genome Analysis

GenomeScope: Fast reference-free genome profiling from short reads

Gregory W. Vurture[1,†], Fritz J. Sedlazeck[2,†], Maria Nattestad[1], Charles J. Underwood[1], Han Fang[1,3], James Gurtowski[1] and Michael C. Schatz[1,2,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, 11724, NY, USA, [2]Johns Hopkins University, Baltimore, 21218, MD, USA [3]Stony Brook University, Stony Brook, 11794, NY, USA

*To whom correspondence should be addressed. †The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: GenomeScope is an open-source web tool to rapidly estimate the overall characteristics of a genome, including genome size, heterozygosity rate, and repeat content from unprocessed short reads. These features are essential for studying genome evolution, and help to choose parameters for downstream analysis. We demonstrate its accuracy on 324 simulated and 16 real datasets with a wide range in genome sizes, heterozygosity levels, and error rates.

Availability and Implementation: http://qb.cshl.edu/genomescope/, https://github.com/schatzlab/genomescope.git

Contact: mschatz@jhu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High throughput sequencing enables the sequencing of novel genomes on a daily basis. Nevertheless, even their most basic characteristics, such as their size or heterozygosity rate, may be initially unknown, making it difficult to select appropriate analysis methods e.g. read mapper, *de novo* assembler, or SNP caller (Smolka, et al., 2015). Determining these characteristics in advance can reveal if an analysis is not capturing the full complexity of the genome, such as underreporting the number of variants or failure to assemble a significant fraction of the genome. Experimental methods are available for measuring some of these properties, although can require significant cost and labor. Simpson (2014) proposed a computational method to measure some of these properties from sequencing reads using *de novo* assembly techniques. However, this method is computationally intensive and can be difficult to interpret as results are reported relative to the assembly graph, such as the variant-induced branch rate rather than the rate of heterozygosity.

2 Methods

Here we introduce *GenomeScope* to estimate the overall genome characteristics (total and haploid genome length, percentage of repetitive content, and heterozygosity rate) as well as overall read characteristics (read coverage, read duplication, and error rate) from raw short read sequencing data. The estimates do not require a reference genome and they can be inferred via a statistical analysis of the *k-mer profile.* The profile measures how often *k-mers,* substrings of length *k*, occur in the sequencing reads and can be computed very quickly using tools such as Jellyfish (Marcais and Kingsford, 2011). The profiles reflect the complexity of the genome: homozygous genomes have a simple Poisson profile while heterozygous ones have a characteristic bimodal profile. Repeats add additional peaks at even higher *k-mer* frequencies, while sequencing errors and read duplications distort the profiles with false *k-mers* and increased variances (**Supplementary Note 1**).

Aware of these possible complexities, *GenomeScope* fits a mixture model of four negative binomial distributions to the *k-mer* profile to measure the relative abundances of heterozygous and homozygous, unique and two-copy sequences. Errors and higher copy repeats are identified by *k-mers* falling outside the model range. *GenomeScope* then infers the genome and read characteristics, along with an error measure to indicate how well the model represents the data. The modeling framework is available as a standalone R application and also as an easy-to-use web application, both generating publication quality plots and summary tables in seconds (See **Supplementary Note** 1 for details and data requirements).

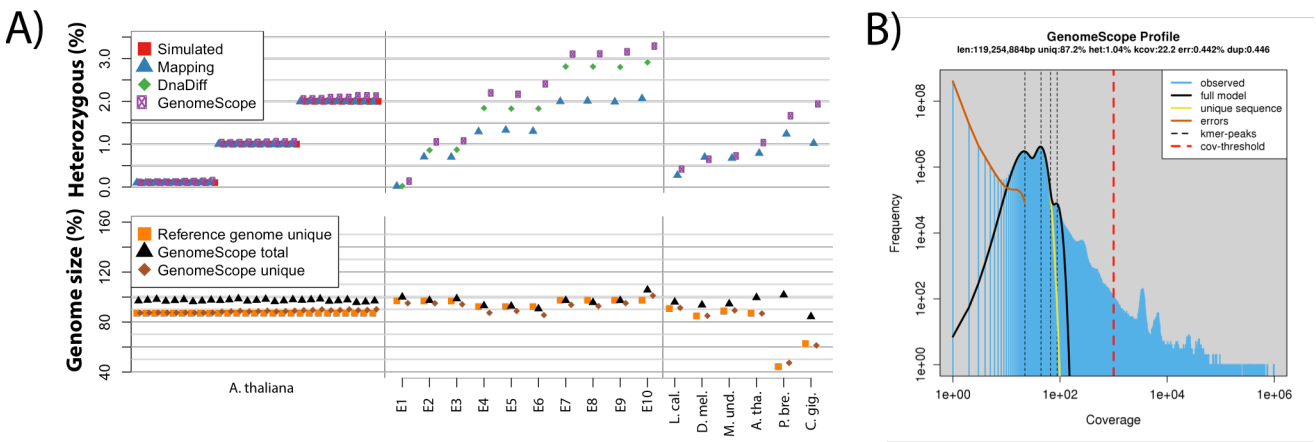

Figure 1. (A) *GenomeScope* heterozygosity, total genome size, and unique genome size estimates: (left) twenty seven simulated *A. thaliana* datasets with vary amounts of heterozygosity, sequencing error or read duplications; (middle) ten synthetic mixtures of real *E. coli* sequencing data; and (right) six genuine plant and animal sequencing datasets: *L. calcarifer* (Asian seabass), *D. melanogaster* (fruit fly), *M. undulates* (budgerigar), *A. thaliana* Col-Cvi F1 (thale cress), *P. bretschneideri* (pear), C. gigas (Pacific oyster). Also displayed are the true simulated values (Simulated), the results from a mapping and variant calling pipeline (Mapping), and a whole genome alignment (DnaDiff) where available. **(B)** GenomeScope k-mer profile plot of the *A. thaliana* dataset showing the fit of the GenomeScope model (black) to the observed k-mer frequencies (blue). The unusual peak of very high frequency k-mers (~10,000x coverage) were determined to be highly enriched for organelle sequences.

3 Results

We first applied GenomeScope to analyze 324 simulated data sets 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*, *D. melanogaster*) (**Supplementary Table 3, Supplementary Note 2**). A subset of the results for *A. thaliana* are displayed in **Figure 1A** (left), and show that the GenomeScope results are highly concordant with the true simulated rates over many conditions. The results were also highly concordant to a standard short-read variant analysis pipeline using BWA-MEM (Li, 2013) and SAMTools (Li, et al., 2009) or through whole genome alignment using DnaDiff (Phillippy, et al., 2008) of the original and mutated reference sequence. We next evaluated ten *E.coli* datasets where genuine sequencing reads from two divergent strains were synthetically mixed together (**Figure 1A**, middle). This allowed us to evaluate GenomeScope on real sequencing reads where the genome sequences, and hence their heterozygosity rates, were precisely known. We find high concordance to the results of the whole genome alignment of the reference genomes, although mapping the reads and calling variants resulted in artificially lower rates of heterozygosity because the short reads failed to map over the most heterozygous and repetitive regions (**Supplementary Note 3).**

Finally, we applied GenomeScope to six different genuine plant and animal data sets up to 1.1Gbp in size with significant levels of heterozygosity and an assembled reference genome (**Supplementary Note 4**). Since the available references were haploid, it was not possible to validate the results with whole genome alignment but could compare to the short read mapping results. The results are generally concordant, although the GenomeScope heterozygosity estimates were consistently higher than those from read mapping, similar to the *E. coli* results caused by short read mapping deficiencies, and most discrepant for the lowest quality draft genomes. We added a parameter to exclude extremely high frequency *k-mers* (default: 1,000x or greater), since those represented organelle sequences occurring hundreds to thousands of times per cell in *A. thaliana* that artificially inflated the genome size (**Figure 1b; Supplemental Note 1.3.2**). After accounting for high copy sequences, the inferred genome sizes were 99.7% accurate as confirmed by orthogonal technologies, such as the established reference genomes or flow cytometry when available **(Supplementary Note 4).**

4 Discussion

We have shown on 340 data sets that *GenomeScope* is a fast, reliable and accurate method to estimate the overall genome and read characteristics of data sets without a reference genome. Using the web application, users can upload their *k-mer* profile and seconds later *GenomeScope* will report the genomic properties and generate high quality figures and tables. As such, we expect *GenomeScope* to become a routine component of all future genome analysis projects.

Acknowledgements

We would like to thank Arthur Delcher, Rachel Sherman, and Steven Salzberg for their helpful discussions and testing.

Funding

NSF [DBI-1350041 and IOS-1237880]; NIH [R01-HG006677]

Conflict of Interest: none declared.

References

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.

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.

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.

Smolka, M., *et al.* Teaser: Individualized benchmarking and optimization of read mapping results for NGS data. *Genome Biol* 2015;16:235.