SCSIM: Jointly simulating correlated single-cell and bulk next-generation DNA sequencing data

Collin Giguere†, Harsh Vardhan Dubey†, Vishal Kumar Sarsani†, Hachem Saddiki‡, Shai He§ and Patrick Flaherty*†

†Correspondence: flaherty@math.umass.edu
‡Department of Mathematics & Statistics, University of Massachusetts Amherst, 710 N. Pleasant St., 01003 Amherst, USA
*Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: Recently, it has become possible to collect next-generation DNA sequencing data sets that are composed of multiple samples from multiple biological units where each of these samples may be from a single cell or bulk tissue. Yet, there does not yet exist a tool for simulating DNA sequencing data from such a nested sampling arrangement with single-cell and bulk samples so that developers of analysis methods can assess accuracy and precision.

Results: We have developed a tool that simulates DNA sequencing data from hierarchically grouped (correlated) samples where each sample is designated bulk or single-cell. Our tool uses a simple configuration file to define the experimental arrangement and can be integrated into software pipelines for testing of variant callers or other genomic tools.

Conclusions: The DNA sequencing data generated by our simulator is representative of real data and integrates seamlessly with standard downstream analysis tools.

Keywords: single-cell DNA sequencing; simulator; DNA sequencing; hierarchical Dirichlet

Background

Simulation software is important for developing and improving statistical methodology for next-generation sequencing data [1]. There are currently 149 such genetic data simulators indexed by the National Cancer Institute [2], and four of these simulators produce DNA sequencing reads with single-nucleotide variants: GemSIM [3], NEAT [4], SInC [5], and CuReSim [6]. But none of these software tools simulate bulk and single-cell data or produces data from a hierarchically grouped sampling designs typical of experimental data [7, 8].

To address this need, we have developed a software package, single-cell NGS simulator (SCSIM), to allow researchers to simulate bulk and single-cell next-generation sequencing (NGS) data from a hierarchical grouped sampling design. Figure 1 shows a high level workflow diagram of the simulator.

The nested sampling structure is implemented using a truncated hierarchical Dirichlet process mixture model [9]. Errors induced by whole-genome amplification (WGA) of single-cells are simulated using the method described by Zafar et. al. [10]. Then, given the set of diploid reference sequences, NGS reads are simulated using dwgsim [11]. Finally, bulk NGS data is simulated by sampling without replacement from the set of reads from pure samples in proportions defined by
the truncated Dirichlet process mixture. The command-line software takes a single haploid reference sequence in FASTA format and a YAML configuration file, and produces FASTQ reads that can be used for downstream alignment and variant calling tasks.

### Implementation

The software is implemented in python and all input parameters are controlled through a single YAML configuration file. The source code can be downloaded from the github repository [https://github.com/flahertylab/scsim](https://github.com/flahertylab/scsim) and the implementation can be run from a docker container defined in the repository.

![SCSIM simulation workflow](image)

**Figure 1: SCSIM simulation workflow.** Inputs (shown in rounded boxes with green text) are the reference sequence and experiment design. Outputs (shown in cornered boxes with orange text) are the bulk and single cell FASTQ read files. Intermediate objects are shown in purple with no boxes.

**Mutated Synthetic Prototype Genome Simulation**

Mutated diploid sequences are generated from a single reference (FASTA). Given $N$—the number of mutated synthetic prototype genomes and $n$—the number of possible SNV locations, one-third of SNVs are shared across all mutated sequences, one-third of SNVs are shared across one-half of the mutated sequences, and the remaining one-third of SNVs are shared across a proportion of sequences chosen from a uniform distribution [10]. The locations of the SNVs are equally spaced across the region of interest in the reference FASTA. Given the status of each SNV location in each mutated sequence, the type of diploid mutation (heterozygous or homozygous) and base substitutions are generated according to a transition probability matrix derived from Pattnaik et. al. (S1NC) [5]. The transition probabilities, and SNV locations can be set by the user, and have default values derived from literature.
Hierarchical Sampling Structure

A hierarchical Dirichlet model is used to simulate correlation between related samples. First, the distribution over mutated synthetic prototype genomes $G_0$ in the population is sampled from a Dirichlet with parameter $\alpha$. Then, the distribution over mutated sequences for a biological unit $i$ is sampled from a Dirichlet distribution with parameter $\beta_i G_0$ where $\beta_i$ controls the concentration of the distribution of the biological unit around the population distribution $G_0$. Finally, the distribution over mutated sequences for sample $j$ in biological unit $i$ is distributed as a Dirichlet with parameter $\gamma_{ij} G_i$ where $\gamma_{ij}$ controls the concentration of the sample around the biological unit distribution $G_i$.

Single-cell Sample Simulation

For each of the $N_{sc}$ single-cell samples, a whole-genome amplification (WGA) model is applied to the corresponding mutated sequence. Allelic dropout (ADO) and false positive (FP) mutations from the WGA process were generated as done previously [10]. The ADO rate was set to 20% and the FP rate was set to $3.2 \times 10^{-5}$ [12]. The ADO and FP rates are calculated in reference to the entire length of the reference sequence. Finally, sequencing reads and corresponding FASTQ files were generated with $\text{dwgsim}$ for each of the $N_{sc}$ single-cell samples.

Bulk Sample Simulation

To generate bulk sample FASTQ files where the distribution over mutated synthetic prototype genomes follows $G_{ij}$, first $\text{dwgsim}$ was used to generate FASTQ reads for each of the $n$ mutated synthetic prototype genomes. Then, for each of the $N_{bu}$ bulk samples, the FASTQ reads from each of the mutated synthetic prototype genomes were mixed according to the distribution $G_{ij}$.

Results

Here we describe simulation results showing mutations from reads generated using SCSIM are accurately identified by two different variant calling tools.

Simulation protocol  In order to assess the accuracy and consistency of reads simulated by SCSIM, we extracted a 1 million base pair region from hg38 starting at chr20:1000000. We generated three diploid synthetic prototype genomes each with 100 potential SNVs spaced every 8080 base pairs. The zygosity of the SNVs was sampled according to the method described in the implementation section. The prior parameter for the Dirichlet distribution across mutated synthetic prototype genomes, $\alpha$ was set to $(0.1, 0.3, 0.6)$. Figure 2 shows the distribution of true SNVs across genomic position for each mutated synthetic prototype genome. The concentration parameter for the Dirichlet distribution at the biological unit level $\beta_i$, was set to 0.1 for all $i$. The concentration parameter for the Dirichlet distribution for samples within biological units, $\gamma_{ij}$ was set to 0.1 for all $(i,j)$. Two samples were generated for each unit with unit 1 and 4 having one bulk and one single-cell sample, unit 2 having two single-cell samples, and unit 3 having 2 bulk samples. The mean coverage level was set to $24x$. For bulk samples, we drew 1,000,000 reads according to the distribution over prototypes that was realized by the hierarchical
Dirichlet model. A total of 90 SNVs introduced across 4 single-cell and 4 bulk samples from the described simulation protocol are the true SNVs and served as the gold standard set.

![Figure 2: Simulated true SNV locations across mutated synthetic prototype genomes.](image)

**Visualization of simulated reads** The simulated FASTQ reads of single-cell and bulk samples across 4 biological units were mapped to the human genome assembly (hg38) using the Burrows-Wheeler alignment tool[13]. Figure 3 shows IGV visualizations of three genomic locations with varying proportions of samples with true SNVs. Mapped reads were used to call variants by monovar[10] and BCFtools[14], two popular SNV callers used for the single-cell data. Monovar and BCFtools were run with default parameter values on the BAM files of all single-cell and bulk data. BCFtools and Monovar called 154 and 156 SNVs respectively across 4 single-cell and 4 bulk simulated samples. Our analysis showed that out of 90 true SNVs, 88 were called by both BCFtools and Monovar. Figure 4 shows a Venn diagram of the concordance between true SNVs and called SNVs from BCFtools and Monovar.

**Conclusions**

Given the increase in the amount of single-cell next-generation DNA sequencing data there is a need for reproducible bioinformatics methods for performing statis-
tical inference on that data. To our knowledge there are no methods for jointly simulating bulk and single-cell sequencing data, yet these simulation tools are needed to test and validate inference methods. SCSIM jointly simulates bulk and single-cell next-generation sequencing data and generates correlated samples using a hierarchical truncated Dirichlet distribution for sampling the distribution over mutant sequences for bulk samples. Our implementation, using a docker container, allows it to be inserted in a bioinformatics pipeline without modifying existing dependencies.

**Availability and requirements**

**Project name** single-cell DNA sequencing data simulator

**Project home page** https://github.com/flahertylab/scsim

**Operating systems(s)** Any

**Programming language** Python

**Other requirements** docker

**License** MIT

**Competing interests**
The authors declare that they have no competing interests.

**Author’s contributions**
CG contributed to the development of the algorithm and implementation and drafting of the manuscript. HVD contributed to the implementation of the algorithm, analysis and interpretation of the data, and drafting of the manuscript. VKS contributed to the analysis and interpretation of the data and drafting of the manuscript. HS contributed to the development of the algorithm. SH contributed to the implementation of the algorithm and drafting of the manuscript. PF contributed to the study conception and design, development of the algorithm, analysis and interpretation of the data and drafting of the manuscript.

**Acknowledgements**
This research was funded in part by NIH award 1R01GM135931-01.
References

1. Escalona, M., Rocha, S., Posada, D.: A comparison of tools for the simulation of genomic next-generation sequencing data. Nature Reviews Genetics 17(8), 459–469 (2016). doi:10.1038/nrg.2016.57

2. NCI Division of Cancer Control & Population Sciences: Genetic Simulation Resources. https://popsmodels.cancercontrol.cancer.gov/gar/. Accessed: 2018-11-27 (2018)

3. McElroy, K.E., Luciani, F., Thomas, T.: GEMSIM: General, error-model based simulator of next-generation sequencing data. BMC Genomics 13(1), 74 (2012). doi:10.1186/1471-2164-13-74

4. Stephens, Z.D., Hudson, M.E., Mainzer, L.S., Taschuk, M., Weber, M.R., Iyer, R.K.: Simulating next-generation sequencing datasets from empirical mutation and sequencing models. PLOS ONE 11(11), e0167047 (2016). doi:10.1371/journal.pone.0167047

5. Pattnaik, S., Gupta, S., Rao, A.A., Panda, B.: SlInC: an accurate and fast error-model based simulator for SNPs, indels and CNVs coupled with a read generator for short-read sequence data. BMC Bioinformatics (2014). doi:10.1186/1471-2105-15-40

6. Caboche, S., Audebert, C., Lemoine, Y., Hot, D.: Comparison of mapping algorithms used in high-throughput sequencing: Application to ion torrent data. BMC Genomics 15(1), 264 (2014). doi:10.1186/1471-2164-15-264

7. Casasent, A.K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., Casasent, T., Meric-Bernstam, F., Edgerton, M.E., Navin, N.E.: Multiclonal invasion in breast tumors identified by topographic single cell sequencing. Cell 172(1-2), 205–21712 (2018). doi:10.1016/j.cell.2017.12.007

8. Zhou, J., Deng, Y., Shen, L., Wen, C., Yan, Q., Ning, D., Qin, Y., Xue, K., Wu, L., He, Z., Voordeekers, J.W., Nostrand, J.D.V., Buzzard, V., Mihailetz, S.T., Enquist, B.J., Weiser, M.D., Kaspari, M., Waide, R., Yang, Y., Brown, J.H.: Temperature mediates continental-scale diversity of microbes in forest soils. Nature Communications 7(1), 12083 (2016). doi:10.1038/ncomms12083

9. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006). doi:10.1198/016214506000000302

10. Zafar, H., Wang, Y., Nakhle, L., Navin, N., Chen, K.: Monovar: single-nucleotide variant detection in single cells. Nature Methods, 505–507 (2015). doi:10.1038/nmeth.3835

11. Homer, N.: Whole Genome Simulator for Next-Generation Sequencing. http://github.com/nh13/dwgsim. Accessed: 2018-11-27 (2018)

12. Leung, M.L., Wang, Y., Waters, J., Navin, N.E.: SNES: single nucleus exome sequencing. Genome Biology 16(1), 55 (2015). doi:10.1186/s13059-015-0616-2

13. Li, H.: A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (2011). doi:10.1093/bioinformatics/btr509