Bioinformatics, 35(17), 2019, 3148–3150
doi: 10.1093/bioinformatics/btz028
Advance Access Publication Date: 14 January 2019
Applications Note

Genome analysis

PSiTE: a Phylogeny guided Simulator for Tumor Evolution

Hechuan Yang [1,2], Bingxin Lu[2], Lan Huong Lai[2], Abner Herbert Lim[2], Jacob Josiah Santiago Alvarez[2] and Weiwei Zhai[1,2,3,4,5,*]

[1]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, P.R.China, [2]Human Genetics, Genome Institute of Singapore, A*STAR, Singapore 138672, Singapore, [3]National Cancer Centre Singapore, Singapore 169610, Singapore, [4]School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore and [5]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, P.R.China

*To whom correspondence should be addressed.
Associate Editor: Russell Schwartz

Received on August 15, 2018; revised on December 10, 2018; editorial decision on January 4, 2019; accepted on January 8, 2019

Abstract

Summary: Simulating realistic clonal dynamics of tumors is an important topic in cancer genomics. Here, we present Phylogeny guided Simulator for Tumor Evolution, a tool that can simulate different types of tumor samples including single sector, multi-sector bulk tumor as well as single-cell tumor data under a wide range of evolutionary trajectories. Phylogeny guided Simulator for Tumor Evolution provides an efficient tool for understanding clonal evolution of cancer.

Availability and implementation: PSiTE is implemented in Python and is available at https://github.com/hchyang/PSiTE.

Contact: zhaiww1@gis.a-star.edu.sg

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

1 Introduction

Tumors are often a mixture of multiple cell populations (i.e. clones) (Andor *et al.*, 2016; Jamal-Hanjani *et al.*, 2017). Understanding tumor heterogeneity and performing clonal decomposition of tumors have become a very important topic for the field (Schwartz and Schäffer, 2017). Despite rapid progress, realistic simulators of tumor evolution that can help benchmarking clonal decomposition methods are still lacking.

Currently, there are two main kinds of tumor simulators. One type of simulators focus on simulating tumor sequencing data through adding synthetic mutations to BAM files (Ewing *et al.*, 2015; Samadian *et al.*, 2017) or mixing sequencing reads from normal and tumor genomes (Mu *et al.*, 2015). Even though explicitly mixing different clones with user-specified parameters is possible in more recent packages (Ivakhno *et al.*, 2017; Nabavi, 2018; Qin *et al.*, 2015), the evolutionary relationship of these clones has to be specified *apriori* and can be quite subjective.

The other type of simulators focuses only on simulating the clonal relationship of the tumor population. Both forward simulators based on branching processes (Chowell *et al.*, 2018) and coalescent simulators (Beerenwinkel *et al.*, 2015) can be used to simulate population dynamics of cancer cells. Even though they can simulate a broad range of evolutionary histories of tumor evolution, this type of simulators cannot generate sequencing data of tumor. In order to leverage the strength of both of these two types of simulators, we developed Phylogeny guided Simulator for Tumor Evolution (PSiTE), a flexible tool that can simulate next-generation sequencing data of single, multi-sector bulk tumor samples and single-cell tumor samples under a wide variety of evolutionary histories.

2 Results

2.1 Overview of PSiTE and its core module *phylovar*

PSiTE takes advantage of the second kind of tumor simulators by first taking in the phylogenetic relationship of the tumor sample and then simulating genomes of tumor clones within the sample. These genomes were subsequently used to simulate next-generation

sequencing data of the tumor samples. There are six modules in PSiTE: (i) *vcf2fa* (producing the germline genome of the cancer patient), (ii) *phylovar* (simulating somatic events along the history of tumor evolution), (iii) *chain2fa* (producing genomes of individual tumor clones), (iv) and (v) *fa2wgs* and *fa2wes* (generating whole genome/exome next-generation sequencing data) and (vi) *allinone* (a convenient wrapper combining all the previous five modules) (Fig. 1). A simple tutorial and evaluation of the computational performance of PSiTE can be found in the Supplementary Material.

Given the ancestral relationship (i.e. the phylogenetic tree) of tumor samples, the core module *phylovar* simulates both somatic single nucleotide variants (SNVs) and copy number variants (CNVs, including aneuploidies) along the evolutionary history of tumor cells according to a Poisson process with user-specified rate parameters. The size of a CNV follows either a parametric distribution (e.g. an exponential distribution) or an empirical distribution specified by users. In simulating an amplification, PSiTE first 'copies' the subtree beneath the focal branch where the CNV occurs and then continues simulating mutational events concurrently on both the original and duplicated subtree (Fig. 1). In the case of deletions, local copy numbers are reduced and all previously simulated SNVs in the deleted segment are removed. This 'copy-and-edit' approach allows efficient joint simulation of SNVs and CNVs (tandem duplications or randomly placed CNVs), in which the phasing information between SNVs and CNVs can be explicitly simulated and traced.

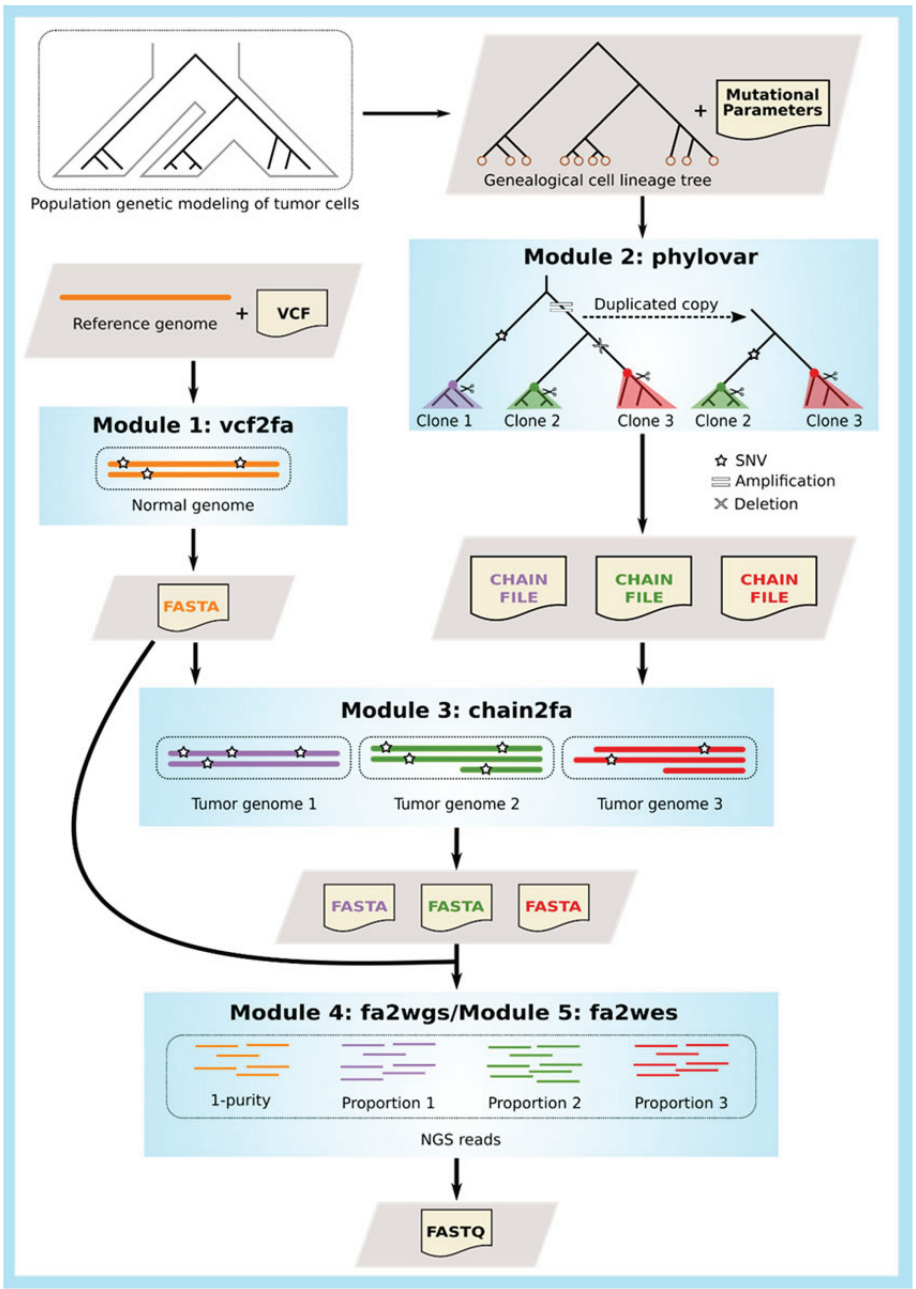

Fig. 1. The flow chart of PSiTE. Starting from the germline genome of the individual, PSiTE utilizes *vcf2fa, phylovar, chain2fa, fa2wgs* and *fa2wes* simulating multiple tumor clones in a patient cancer (see main text as well as the PSiTE manual). Aside from these five basic modules, PSiTE also provides a convenient wrapper *allinone* (module 6), which runs all the simulation steps in one command

2.2 Tree pruning and multiple tumor data types

Since many somatic variants are only present in a small proportion of samples and are often undetectable by sequencing (Wang *et al.*, 2013), *phylovar* allows users to prune the tree at varying frequency cutoffs prior to simulating mutational events (see PSiTE manual for details, Fig. 1). Tree pruning allows PSiTE to greatly increase the computational efficiency of the simulation.

PSiTE utilizes existing sequence simulators to obtain short reads from normal and tumor genomes. For whole genome sequencing data, PSiTE employs ART (Huang *et al.*, 2012) as the short-read generator. For whole exome sequencing data, two simulators are available, namely Wessim (Kim *et al.*, 2013) and CapGem, a hybrid simulator created by us that combines Wessim and Capsim (Cao *et al.*, 2018). As single-cell and multi-sector data become increasingly popular for surveying tumor heterogeneity (Shapiro *et al.*, 2013), PSiTE also allows users to generate mutational profiles and short reads of multiple sectors or single-cell data. The multi-sector data are implemented by incorporating an affiliation file which records the membership relation between tumor cells and tumor sectors. After simulating multiple tumor clones, short reads will be distributed across sectors according to the membership relation and their respective frequencies.

2.3 Detailed output empowering the benchmarking of clonal decomposition methods

One strength of PSiTE is that it can output very detailed information about the history of tumor evolution. In addition to files recording true frequencies of SNVs and CNVs, PSiTE also outputs a phylogenetic tree in the NHX format where the mutational information is recorded along the evolutionary history (Zmasek and Eddy, 2001). Together with the affiliation file, users can examine in fine details at different parts of the evolutionary history and inspect the genomic profile of subclones at any scale (down to single-cell resolution). In addition, PSiTE also provide a simple version where users can directly simulate the input files for clonal decomposition methods without simulating the raw sequencing data (see Supplementary Material). Therefore, PSiTE provides a powerful means to benchmark clonal decomposition methods.

3 Conclusions

In summary, we have developed PSiTE, which can generate realistic tumor samples with complex evolutionary histories. Despite a few potential limitations including limited mutational types and imperfect single-cell data, the simulated datasets can provide gold standard benchmarks for clonal decomposition methods, as well as other tools (e.g. variant caller) in studying tumor heterogeneity and evolution (see Supplementary Material for a discussion and demo).

Funding

This work was supported by Genome Institute of Singapore; and Institute of Zoology, Chinese Academy of Sciences. W.Z. is supported in part by National Key R&D program of China grant [2018YFC0910400, 2018YFC1406902].

Conflict of Interest: none declared.

References

Andor,N. *et al.* (2016) Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.*, **22**, 105–113.

Beerenwinkel,N. *et al.* (2015) Cancer evolution: mathematical models and computational inference. *Syst. Biol.*, **64**, e1–25.

Cao,M.D. *et al.* (2018) Simulating the dynamics of targeted capture sequencing with CapSim. *Bioinformatics*, **34**, 873–874.

Chowell,D. *et al.* (2018) Modeling the subclonal evolution of cancer cell populations. *Cancer Res.*, **78**, 830–839.

Ewing,A.D. *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods*, **12**, 623–630.

Huang,W. *et al.* (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Ivakhno,S. *et al.* (2017) tHapMix: simulating tumour samples through haplotype mixtures. *Bioinformatics*, **33**, 280–282.

Jamal-Hanjani,M. *et al.* (2017) Tracking the Evolution of Non–Small-Cell Lung Cancer. *N. Engl. J. Med.*, **376**, 2109–2121.

Kim,S. *et al.* (2013) Wessim: a whole-exome sequencing simulator based on in silico exome capture. *Bioinformatics*, **29**, 1076–1077.

Mu,J.C. *et al.* (2015) VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, **31**, 1469–1471.

Nabavi,S. (2018) VarSimLab: A complete command-line pipeline to simulate genomic variations. Github (https://github.com/NabaviLab/VarSimLab), last access: Aug 1, 2018.

Qin,M. *et al.* (2015) SCNVSim: somatic copy number variation and structure variation simulator. *BMC Bioinformatics*, **16**, 66.

Samadian,S. *et al.* (2017) Bamgineer: introduction of simulated allele-specific copy number variants into exome and targeted sequence data sets. *PLoS Comput. Biol.*, **14**, e1006080.

Schwartz,R. and Schäffer,A.A. (2017) The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.*, **18**, 213–229.

Shapiro,E. *et al.* (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, **14**, 618–630.

Wang,Q. *et al.* (2013) Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med.*, **5**, 91.

Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.