Bioinformatics, 34(8), 2018, 1419–1420
doi: 10.1093/bioinformatics/btx772
Advance Access Publication Date: 27 November 2017
Applications Note

Genetics and population analysis

SimulaTE: simulating complex landscapes of transposable elements of populations

Robert Kofler

Institute für Populationsgenetik, Vetmeduni Vienna, 1210 Wien, Austria

Associate Editor: Oliver Stegle

Received on July 20, 2017; revised on October 11, 2017; editorial decision on November 24, 2017; accepted on November 25, 2017

Abstract

Motivation: Estimating the abundance of transposable elements (TEs) in populations (or tissues) promises to answer many open research questions. However, progress is hampered by the lack of concordance between different approaches for TE identification and thus potentially unreliable results.

Results: To address this problem, we developed SimulaTE a tool that generates TE landscapes for populations using a newly developed domain specific language (DSL). The simple syntax of our DSL allows for easily building even complex TE landscapes that have, for example, nested, truncated and highly diverged TE insertions. Reads may be simulated for the populations using different sequencing technologies (PacBio, Illumina paired-ends) and strategies (sequencing individuals and pooled populations). The comparison between the expected (i.e. simulated) and the observed results will guide researchers in finding the most suitable approach for a particular research question.

Availability and implementation: SimulaTE is implemented in Python and available at https://sourceforge.net/projects/simulates/. Manual https://sourceforge.net/p/simulates/wiki/Home/#manual; Test data and tutorials https://sourceforge.net/p/simulates/wiki/Home/#walkthrough; Validation https://sourceforge.net/p/simulates/wiki/Home/#validation.

Contact: robert.kofler@vetmeduni.ac.at

1 Introduction

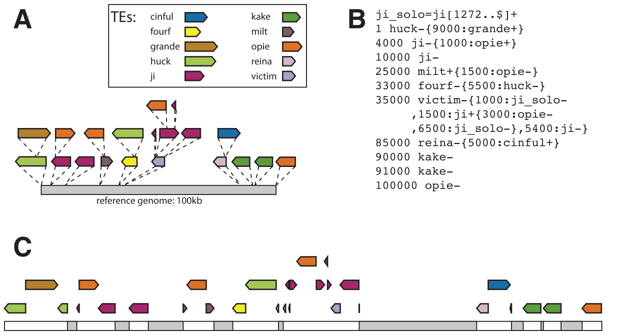
Transposable elements (TEs) are short DNA sequences that selfishly spread within genomes. They are responsible for diverse phenomena ranging from genome evolution to human disease (Kazazian, 2004). The advent of NGS enabled the study of TE dynamics within populations (or tissues, a ‘population’ of cells), which promises to shed light on many open research questions such as the amount of positively selected TEs (Casacuberta and González, 2013), the evolution of TE activity (Kofler *et al.*, 2012) and the role of TEs in brain development (Erwin *et al.*, 2014). However, progress in the field is hampered by the lack of concordance among different approaches for estimating TE abundance in populations (Nelson *et al.*, 2017; Ewing, 2015). The problem is even exacerbated for species with highly repetitive genomes (e.g. whole genome duplications) or genomic resources of low quality (e.g. reference genome, database of known TEs). Thus, it is often simply not known whether a given approach for TE identification yields suitable results for a particular question. Therefore, we developed SimulaTE: a tool for simulating reads (Illumina or PacBio) based on a population with an arbitrarily complex TE landscape. A reference contig and sequences of the TEs of interest are required as input, thus even non-model organism may be used. The comparisons between the expected (i.e. simulated) and the observed TE landscapes enable researchers to assess the suitability of the available genomic resources as well as the targeted approach for TE identification.

2 Approach

SimulaTE proceeds in three steps: first, the TE landscape of a population is outlined using a simple syntax; second, a genome is built for every individual in the population and third, reads are simulated using the genomes of all individuals as a template. As a main feature

© The Author 2017. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com



```
ji_solo=ji[1272..$]+
1 huck-{9000:grande+}
4000 ji-{1000:opie+}
10000 ji-
25000 milt+{1500:opie-}
33000 fourf-{5500:huck-}
35000 victim-{1000:ji_solo-
     ,1500:ji+{3000:opie-
     ,6500:ji_solo-},5400:ji-}
85000 reina-{5000:cinful+}
90000 kake-
91000 kake-
100000 opie-
```

Fig. 1. Reconstructing a classic TE landscape with SimulaTE (SanMiguel et al., 1996). (**A**) The complex TE landscape near the *Adh1-F* locus in maize described by SanMiguel et al. (1996). Dashed lines indicate TE insertion sites and arrows the strand of the insertion. (**B**) Syntax to create the TE landscape of SanMiguel et al. (1996) using our DSL. This code describes the sequence of a single individual. Sequences for more individuals may be generated by simply adding more columns. (**C**) Annotation of the resulting sequence using RepeatMasker. Note that the annotated sequence agrees with the simulated one (compare to A)

we developed a simple domain specific language (DSL; a programming language custom tailored to a specific task) that enables describing even complex TE landscapes using a simple syntax. In particular, our DSL allows specification of the following properties of TE landscapes: (i) position, (ii) family, (iii) strand, (iv) population frequency, (v) target site duplication, (vi) haplotype (i.e. linkage of TE insertions), (vii) sequence divergence, (viii) truncations (internal and external) and (ix) nested insertions, including recursively nested insertions. To demonstrate the utility of our DSL we reconstructed a classic example of a complex TE landscape, the nested TE insertions near the *Adh1-F* locus in maize (Fig. 1A and B; SanMiguel *et al.*, 1996). Next, SimulaTE interprets the DSL code and generates the genomes of all individuals within a population, where individuals may either be haploid or diploid. Finally reads may be simulated using the genomes of all individuals in the population as a template. SimulaTE allows simulation of Illumina paired-end, Illumina single-end and PacBio reads. Chimeric reads may be simulated for Illumina paired-ends (Treiber and Waddell, 2017). Furthermore, individuals of a population may be sequenced either separately or as a pool [Pool-Seq: (Schlötterer *et al.*, 2014)].

3 Validation

We evaluated whether the complex TE landscape described by SanMiguel *et al.* (1996) (Fig. 1A) was accurately built by SimulaTE. We annotated the obtained sequence with RepeatMasker and found that the observed TE landscape (Fig. 1C) agrees with the expected one (99.98% overlap at the nucleotide level; Fig. 1A).

Next we tested whether SimulaTE also accurately creates complex TE landscapes for populations. We simulated a population with 100 individuals having 1000 TE insertions with random position, family, strand and population frequency ($0.1 \geq f \geq 0.9$). We simulated sequencing of a pooled population and paired-end reads (2 × 100 bp). TE insertions were identified with PoPoolationTE2 (Kofler *et al.*, 2016). All 1000 simulated TE insertions were identified. Also the simulated position (average deviation 8.5 bp) and population frequency (average deviation 3%; Spearman's rank correlation of simulated and observed values $\rho = 0.99$; $P < 2.2 \times 10^{-16}$) were accurately reproduced.

Finally we tested the properties of the simulated reads. We simulated Illumina paired-end reads for a sequence of 1 Mb, aligned the reads back to the reference with bwa mem (Li and Durbin, 2010) and computed quality metrics with Picard (http://broadinstitute.github.io/picard.). We found that the error rate ($exp = 1\%$; $obs = 0.99\%$; only base substitutions were simulated), the distribution of the fragment size ($\mathcal{N}_{\text{exp}}(300, 400)$, $\mathcal{N}_{obs}(299.5, 403.4)$; Chi-squared test $\chi^2 = 91.399$, $P = 1$) and the coverage ($exp = 100$; $obs = 99.99$) were accurately simulated. We also simulated long reads (e.g. PacBio) with a mean read length of 11 kb, a bimodal read length distribution and an error rate of 10% (half insertions, half deletions). Reads were again aligned with bwa mem (Li and Durbin, 2010). The read length distribution (Two-sample Kolmogorov-Smirnov test $D = 0.0074$, $P = 0.63$) and the coverage were accurately reproduced ($exp = 1118.2$; $obs = 1118.4$). However the observed number of indels was to low (deletions: $exp = 5\%$, $obs = 4.1\%$; insertions $exp = 5\%$, $obs = 4.1\%$), which is likely due to difficulties of aligning reads with many indels, where usually base substitutions are given preference over indels. In agreement with this we found 1.5% base substitutions despite none being simulated.

For details of the validation see https://sourceforge.net/p/simulates/wiki/Home/#validation.

Acknowledgements

We thank Nayuta Yamashita for critical reading of the manuscript and all members of the Institute of Population Genetics for helpful comments.

Funding

This work was supported by an Austrian Science Foundation (FWF) grant (P29016).

Conflict of Interest: none declared.

References

Casacuberta,E. and González,J. (2013) The impact of transposable elements in environmental adaptation. *Mol. Ecol.*, **22**, 1503–1517.

Erwin,J.A. *et al.* (2014) Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat. Rev. Neurosci.*, **15**, 497–506.

Ewing,A.D. (2015) Transposable element detection from whole genome sequence data. *Mobile DNA*, **6**, 1–9.

Kazazian,H.H. (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.

Kofler,R. *et al.* (2012) Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genet.*, **8**, e1002487.

Kofler,R. *et al.* (2016) PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. *Mol Biol Evol.*, **33**, 2759–2764.

Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **26**, 589–595.

Nelson,M.G. *et al.* (2017) McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole Genome Shotgun Sequencing Data. *G3 Genes Genomes Genet.*, **7**, 2763–2778.

SanMiguel,P. *et al.* (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765–768.

Schlötterer,C. *et al.* (2014) Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.*, **15**, 749–763.

Treiber,C.D. and Waddell,S. (2017) Resolving the prevalence of somatic transposition in Drosophila. *eLife*, **6**, e28297.