SimulaTE: simulating complex landscapes of transposable elements of populations

Robert Kofler
Institute für Populationsgenetik, Vetmeduni Vienna, 1210 Wien, Austria

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.

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
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 ($\chi^2_{exp} (300, 400)$; $\chi^2_{obs} (299.5, 403.4)$; 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: $exp = 3\%$, $obs = 4.1\%$; 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/simulatex/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 Gonzalez,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. Mobile DNA, 6, 1–9.

Koller,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.

Koller,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.

Schlotterer,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 Wadell,S. (2017) Resolving the prevalence of somatic transposition in Drosophila. eLife, 6, e28297.