BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 13 2013, pages 1679–1681
doi:10.1093/bioinformatics/btt198

Genome analysis

Advance Access publication April 25, 2013

RSVSim: an R/Bioconductor package for the simulation of structural variations

Christoph Bartenhagen* and Martin Dugas

Institute of Medical Informatics, University of Münster, Münster

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: RSVSim is a tool for the simulation of deletions, insertions, inversions, tandem duplications and translocations of various sizes in any genome available as FASTA-file or data package in R. The structural variations can be generated randomly, based on user-supplied genomic coordinates or associated to various kinds of repeats. The package further comprises functions to estimate the distribution of structural variation sizes from real datasets.

Availability: RSVSim is implemented in R and available at http://www.bioconductor.org. A vignette with detailed descriptions of the functions and examples is included.

Contact: christoph.bartenhagen@uni-muenster.de

Received on November 15, 2012; revised on March 22, 2013; accepted on April 22, 2013

1 INTRODUCTION

Next-generation sequencing (NGS) has accelerated the detection of (large) deletions, insertions, inversions, tandem duplications and translocations tremendously during the past years (Alkan *et al.*, 2011; Kidd *et al.*, 2008; Mills *et al.*, 2011). Various methods dealing with the comparison, annotation, visualization and especially the detection of structural variations (SVs) were published (Lam *et al.*, 2010; Nielsen *et al.*, 2010; Xi *et al.*, 2010). The simulation of SVs is a powerful, quick and inexpensive approach to assess their performance and correctness.

Evaluating the performance of programs for SV detection, for instance, requires a large number of validated variations of various types with high-breakpoint precision. Depending on the scope of the application, certain parts of a genome might be of special interest like coding regions, a subset of genes or regions of low complexity. Variant detection from real data may not always comply with these requirements because of limitations of the experiment design, few validated SVs or low-breakpoint resolution. A simulation, however, can generate a base exact ground truth consisting of (almost) any number of SVs, which can then be used to test the sensitivity and precision of SV detections (Bruno *et al.*, 2013; Rausch *et al.*, 2012). The output of the algorithm can be compared with the set of simulated SVs to see not only how many SVs were correctly identified but also the SVs missed and falsely predicted because of limitations of the algorithms rather than laboratory issues. A typical workflow for the assessment of an SV algorithm for SV detection consists of

SV simulation ⇒ Read simulation ⇒ SV algorithm ⇒ Evaluation

A FASTA-file with the simulated, rearranged genome can be used by most read simulators (e.g. Hu *et al.*, 2012; Huang *et al.*, 2011) to generate NGS datasets from various sequencing platforms, which can then be used to assess an SV algorithm.

A comprehensive simulation of different SV types of various sizes combined with a variety of read simulations with different numbers of reads, insert-sizes (for paired-end reads) or read lengths can give valuable information for the design of sequencing experiments.

Currently, most publications of SV detection methods restrict the SV simulation to deletions or insertions and implement a known set of SVs, taken, for example, from studies of the 1000 Genome project or Venter's genome or a random set into a reference genome (Jiang *et al.*, 2012; Marshall *et al.*, 2012; Rausch *et al.*, 2012). FUSIM (Bruno *et al.*, 2013) is a more sophisticated approach, but specialized on the simulation of fusion transcripts for RNA-Seq analysis. To our knowledge, RSVSim is the first toolkit that covers a wide range of SVs, which are detectable by DNA-Seq data, and that incorporates knowledge about size and formation mechanisms of SVs for realistic modelling of their breakpoints.

2 AVAILABLE FUNCTIONALITY

RSVSim is implemented in R and works with any kind of genome that is available as FASTA file or BSgenome package in R.

2.1 Types of structural variations

RSVSim can simulate five common types of SVs: deletions, insertions, inversions, tandem duplications and translocations. For deletions, a sequence is removed from the genome and the ends are joined together. Insertions remove or duplicate a segment from one place and insert it into the same or another chromosome. In case of inversions, one segment is cut out and its reverse complement is inserted back at the same position without any loss of sequence. Tandem duplication repeats a sequence several times one after the other. Translocation breaks two chromosomes into two parts each and exchanges the loose ends. Translocations can be simulated in a balanced or unbalanced fashion.

2.2 Size of structural variations

The size of every single SV, except translocations, can be set by the user to fixed or arbitrary values. According to studies from

*To whom correspondence should be addressed.

the 1000 Genomes Project, for deletions, insertions and duplications, the amount of SVs decreases rather quickly as their size increases (Mills *et al.*, 2011). RSVSim provides a function to estimate the distribution from real data by assuming a beta distribution of the values. Its shape can be derived from a given set of SVs. This function provides default parameters for all SV types except translocations, which were estimated from SVs between 500 bp and 10 kb from all sequencing studies available in the Database of Genomic Variants (DGV) release March 29, 2012 (Iafrate *et al.*, 2004): 1.129 deletions, 490 insertions, 202 inversions and 145 tandem duplications in total. Figure 1A compares the distribution of SV sizes of deletions and insertions from the DGV to the same number of values drawn from the fitted β distribution.

2.3 Breakpoint simulation

Structural variation formation in the human genome is not a random process but rather the result of mechanisms, such as non-allelic homologous recombination (NAHR), non-homologous recombination (NHR), variable number of tandem repeats (VNTRs) and transposable element insertions (TEIs) (Mills *et al.*, 2011; Pang *et al.*, 2013). These mechanisms can further be associated with repeat elements, such as LINEs, SINEs, Micro-/Minisatellites and segmental duplications (Lam *et al.*, 2010). Hence, SVs often overlap with regions of high sequence homology and/or sequence repeats. For the hg19, RSVSim uses the coordinates from the UCSC RepeatMasker track (Meyer *et al.*, 2013; Smit *et al.*, 1996–2010) to overlap both breakpoints (NAHR) or an extensive part of the SVs (NHR, VNTR and TEI) with repeat regions. Each SV type is associated differently to NAHR, NHR, VNTR and TEI events, based on the studies with SVs 1 kb by Chen *et al.*, 2008; Mills *et al.*, 2011; Ou *et al.*, 2011 and Pang *et al.*, 2013. The events themselves are biased towards certain kinds of repeats according to the enrichment analysis in Lam *et al.*, 2010 (Fig. 1B). The weights of these biases can be adjusted by the user or turned off to achieve a uniform breakpoint distribution.

Furthermore, SV breakpoint formation often co-occurs with smaller mutations. The user can also influence the amount of indels and SNPs that are generated randomly close to the breakpoints.

In general, SV breakpoints are placed in a non-overlapping manner. Unannotated regions or assembly gaps, denoted by the letter 'N', are detected automatically and excluded from the simulation.

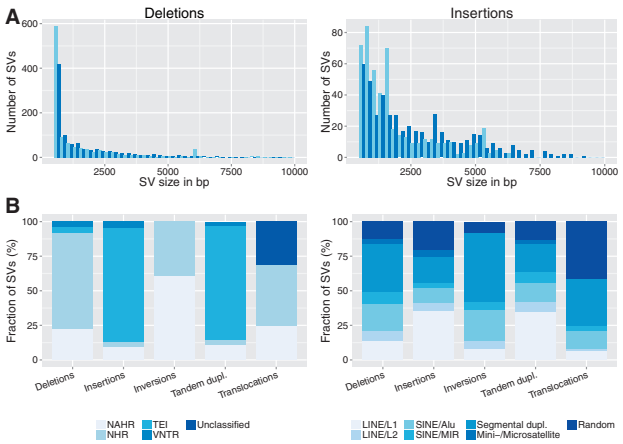

Fig. 1. (**A**) Distribution of SV sizes from studies in the DGV (light blue) and sizes drawn from a fitted beta distribution (dark blue). (**B**) Association of breakpoint formation with mutational mechanisms (left) and repeats (right) for a set of ~2.000 simulated SVs

2.4 Simulation within a subset of the genome

By default, the simulation is carried out across the whole genome. Some applications or special test cases demand the occurrence of SVs in certain subsets of the genome only, like exome sequencing experiments or SV callers specialized for variations within certain repeats or low-complexity regions. Therefore, the simulation can be restricted to a set of user-defined regions. These restrictions can be set individually for every SV type. The implementation in R and available R-packages facilitate the access of genome annotations from Ensembl or UCSC databases to extract coordinates of exons, introns or transcripts for hg19 and further regions of interest.

Another typical use case is the evaluation of an algorithm on a set of known, previously validated SVs from other studies or own experiments. For those cases, RSVSim allows to implement a given set of SVs at pre-defined positions.

These region parameters can further be used to generate heterozygous SVs. Repeating the simulation with a subset of the previously generated SVs creates two rearranged genomes, where only one of them holds certain variations. Both genomes can then be combined, e.g. for simulation of NGS reads. Related genomes with recurrent SVs, e.g. from paired samples (healthy/diseased), can be generated in a similar manner.

2.5 Output

RSVSim reports every SV with its location and size. Some algorithms for SV detection, e.g. those working with split-read mappings, are able to report the SV breakpoint with single-nucleotide resolution. Hence, the output of RSVSim further provides the breakpoint sequence, i.e. the sequence up- and downstream of the breakpoint in the rearranged genome, for comparison with the sequence predicted by SV detection.

All this information can be saved as tables in CSV format. The rearranged genome can be exported as a FASTA file.

The package includes a function to compare the output of the simulation with a set of SV detections in BED- or BEDPE-format. It computes the overlap between the breakpoints or approximate breakpoint regions, aligns the breakpoint sequences (if available) and calculates the sensitivity and precision.

3 CONCLUSION

The R package RSVSim provides functionalities to simulate the five common types of structural variations within any kind of genome. It enables the user to adapt the simulation to the scope of his own application or experiments in terms of type, size and position of every single SV. Artificially rearranged genomes

generated with RSVSim are useful for the performance evaluation of NGS algorithms that deal with SVs or genomes in general.

Conflict of Interest: none declared.

REFERENCES

Alkan,C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev .Genet.*, **12**, 363–376.

Bruno,A.E. *et al.* (2013) FUSIM: a software tool for simulating fusion transcripts. *BMC Bioinformatics*, **14**, 13.

Chen,W. *et al.* (2008) Mapping translocation breakpoints by next-generation sequencing. *Genome Res.*, **18**, 1143–1149.

Hu,X. *et al.* (2012) pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, **28**, 1533–1535.

Huang,W. *et al.* (2011) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Iafrate,A.J. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Jiang,Y. *et al.* (2012) PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*, **28**, 2576–2583.

Kidd,J.M. *et al.* (2011) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.

Lam,H.Y. *et al.* (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.*, **28**, 47–55.

Marshall,T. *et al.* (2012) CLEVER: clique-enumerating variant finder. *Bioinformatics*, **28**, 2875–2882.

Meyer,L.R. *et al.* (2013) The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, 64–69.

Mills,R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Nielsen,C.B. *et al.* (2010) Visualizing genomes: techniques and challenges. *Nat. Methods*, **3 (Suppl.)**, S5–S15.

Ou,Z. *et al.* (2011) Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Res.*, **21**, 33–46.

Pang,A.W. *et al.* (2013) Mechanisms of formation of structural variation in a fully sequenced human genome. *Hum. Mutat.*, **34**, 345–354.

Rausch,T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.

Smit,A. *et al.* (1996–2010) RepeatMasker Open-3.0. http://www.repeatmasker.org (21 April 2013, date last accessed).

Xi,R. *et al.* (2010) Detecting structural variations in the human genome using next generation sequencing. *Brief. Funct. Genomics*, **9**, 405–415.