BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 3 2013, pages 389–390
doi:10.1093/bioinformatics/bts697

*Genome analysis*

Advance Access publication December 10, 2012

# RetroSeq: transposable element discovery from next-generation sequencing data

Thomas M. Keane*, Kim Wong and David J. Adams

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

Associate Editor: Martin Bishop

**ABSTRACT**

**Summary:** A significant proportion of eukaryote genomes consist of transposable element (TE)-derived sequence. These elements are known to have the capacity to modulate gene function and genome evolution. We have developed RetroSeq for detecting non-reference TE insertions from Illumina paired-end whole-genome sequencing data. We evaluate RetroSeq on a human trio from the 1000 Genomes Project, showing that it produces highly accurate TE calls.

**Availabilty:** RetroSeq is open-source and available from https://github.com/tk2/RetroSeq.

**Contact:** tk2@sanger.ac.uk

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

Received on July 12, 2012; revised on November 29, 2012; accepted on December 1, 2012

## 1 INTRODUCTION

Transposable elements were first discovered in maize by Barbara McClintock in the early 20th century and have since been found in almost every organism (McClintock, 1950). They are often referred to as genomic parasites and most are relics of ancient viral infections. In large eukaryote genomes such as human and mouse, TEs make up almost half of the genome (Gogvadze and Buzdin, 2009). There are two distinct classes of TEs: class I retroelements that move by a 'copy and paste' fashion and the less prevalent class II DNA transposons that operate by a 'cut and paste' mechanism. Within the retroelements, there are two distinct classes, the long terminal repeat (LTR)-bound elements and the non-LTR elements. In the human genome, there are two main types of non-LTR elements, namely the short interspersed nuclear elements (SINE) and long interspersed nuclear elements (LINE). Within these classes, the Alu and L1 subfamilies are known to remain functionally active and polymorphic. In laboratory mice, the LTR-bound elements (also known as endogenous retroviral elements—ERVs) can be divided into several subfamilies and are known to be responsible for up to 10% of spontaneous mutations (Maksakova *et al*., 2006).

With the advent of next-generation sequencing technologies, it has become feasible to catalogue all types of molecular variation including insertions of large sequences such as TEs. Previously, Hormozdiari *et al*. (2010) developed VariationHunter, Quinlan *et al*. (2011) developed Hydra and Lee *et al*. (2012) developed Tea for finding non-reference TE insertions. Several other authors have used unpublished pipelines for finding non-reference TEs in human samples (Stewart *et al*., 2011; Ewing and Kazazian, 2011). Furthermore, a number of authors have developed TE insertion site junction sequencing assays and computational methods to detect non-reference TEs (Akagi *et al*., 2008; Iskow *et al*., 2010).

In this article, we present our software, RetroSeq, which can be used to discover non-reference TE insertions from whole genome sequencing data with high accuracy. Previously, we used RetroSeq to create a comprehensive catalogue of just over one hundred thousand polymorphic SINE, LINE and ERV elements across 17 mouse strains (Nellaker *et al*., 2012). Using data from a trio of northern and western European ancestry (CEU) from the 1000 Genomes Project, we show how RetroSeq can be used to create an accurate set of TE calls.

## 2 METHODS, RESULTS, DISCUSSION

The input to RetroSeq is a binary alignment file (BAM) file, a reference genome and a library of mobile element sequences or a BED file of the locations of known TE elements in the reference genome. The BAM file should contain both the mapped pairs and the pairs with one end unmapped. RetroSeq is implemented in Perl and uses SAMtools (Li *et al*., 2009) to access the BAM files. RetroSeq has been tested with alignments derived from both MAQ (Li *et al*., 2008) and BWA (Li and Durbin, 2009). RetroSeq operates in two phases, the first being the discovery phase where discordant mate pairs are detected and assigned to a TE class (Alu, SINE, LINE, etc.) using either the annotated TE elements in the reference and/or aligned with Exonerate (Slater and Birney, 2005) to the supplied library of transposable element sequences. The calling phase uses the anchoring mates of the TE candidate reads from the previous step and clusters these based on their genomic location, and the strand to which they are aligned to (Supplementary Fig. S1). Forward- and reverse-strand clusters are created from the anchor reads and the clusters are then merged into regions around putative break points. RetroSeq profiles the density of the matched forward and reverse clusters and uses any available soft-clipped reads to refine the break points of the TE insertion (see Supplementary Methods).

To evaluate the performance of RetroSeq, we obtained high depth (>75×) Illumina HiSeq data produced at the Broad Institute (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120117_ceu_trio_b37_decoy/) for a CEU trio (father NA12891, mother NA12892 and the female offspring NA12878) from the 1000 Genomes Project and used Retroseq, Tangram (Marth group, unpublished data) and Tea (Lee *et al*., 2012) to

*To whom correspondence should be addressed.

**Table 1.** Comparison of TE calls for the CEU trio for RetroSeq, Tangram (Marth group, unpublished data) and Tea (Lee *et al*., 2012)

| | | RetroSeq | | Tangram | | Tea | |
|---|---|---|---|---|---|---|---|
| Type | Sample | Total | PCR | Total | PCR | Total | PCR |
| Alu | NA12891 | 1038 | 0.97 | 1192 | 0.98 | 1127 | 0.92 |
| | NA12892 | 1046 | 0.98 | 1185 | 0.98 | 1078 | 0.92 |
| | NA12878 | 1078 | 0.98 | 1326 | 0.99 | 1038 | 0.89 |
| L1 | NA12891 | 121 | 0.81 | 190 | 0.81 | 286 | 0.81 |
| | NA12892 | 127 | 0.88 | 219 | 0.88 | 262 | 0.76 |
| | NA12878 | 174 | 0.82 | 227 | 0.87 | 168 | 0.84 |

The 'Total' column is the number of calls predicted by each caller and the 'PCR' column indicates the sensitivity of the methods relative to the PCR-validated calls from Stewart *et al*. (2011).

find Alu and L1 insertions in each individual (Table 1). This trio was previously part of a survey of Alu and L1 elements in a 1000 Genomes pilot project follow-up study (Stewart *et al*., 2011); however, the sequencing data available at the time provided lower Illumina sequencing coverage for each genome (9–16×), which hindered the sensitivity of Alu and L1 detection (Chip Stewart, personal communication). For Alu elements, the sensitivity of RetroSeq and Tangram is >97% for all three individuals, with Tea slightly lower (see Table 1). For L1 elements, the sensitivity of all the methods is uniformly lower, with RetroSeq and Tangram performing best. However, when we look at the RetroSeq false negative rates by TE type in the trio, we do not see a significant difference in the rates for L1 (6.3%) over Alu (6.8%) calls.

We can estimate an upper false discovery rate in the child by examining the calls relative to the expected inheritance patterns. If we consider the calls private to the child as false positives, the false discovery rate of the callers varies significantly (Supplementary Table S1), with RetroSeq having the lowest overall rate (7.7%), followed by Tangram (12.1%) and Tea (14.3%). If we take the calls shared by the parents and not found in the offspring, we can estimate the upper false negative rate for RetroSeq in the offspring at 6.7%. We can use the PCR-validated calls with precise break points to examine the accuracy of the break points estimated by RetroSeq. Supplementary Figs S2–S4 show the distribution of the break points found by RetroSeq around the PCR-validated break points. In NA12878, the vast majority (92%) of the break points are within ±50 bp of the PCR break points, with 40% being within 10 bp (Supplementary Fig. S4).

The coverage for these samples is extremely high (>75×), so it is useful to ask what is the effect on the sensitivity of TE calling when the sequencing depth is lower. Therefore, we sub-sampled the data from sample NA12878 at various depths and plotted the sensitivity relative to (i) the PCR-validated calls and (ii) the intersection of the computational calls from Stewart *et al*., 2011 and RetroSeq. Supplementary Fig. S5 shows that there is a significant drop off in sensitivity at depths lower than 20×, with the sensitivity of the computational calls >90% at 40× coverage. Thus, in the context of TE calling in low coverage populations, data from multiple individuals could be pooled to increase the sensitivity of TE discovery.

## ACKNOWLEDGEMENTS

We would like to acknowledge Binnaz Yalcin, Wayne Frankel and Christoffer Nellaker for their help in evaluating early versions of the software. We gratefully acknowledge Alice Eunjung Lee, Peter J. Parker, Gabor Marth and Jiantao Wu for providing callsets for the CEU trio comparison.

*Funding*: This work was supported by the Medical Research Council, UK and the Wellcome Trust. D.J.A. is supported by Cancer Research-UK and the Wellcome Trust.

*Conflict of Interest*: none declared.

## REFERENCES

Akagi,K. *et al.* (2008) Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. *Genome Res*., **18**, 869–880.

Ewing,A.D. and Kazazian,H.H. (2011) Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res*., **21**, 985–990.

Gogvadze,E. and Buzdin,A. (2009) Retroelements and their impact on genome evolution and functioning. *Cell Mol. Life Sci*., **66**, 3727–3742.

Hormozdiari,F. *et al.* (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, i350–i357.

Iskow,R.C. *et al.* (2010) Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, **141**, 1110–1112.

Lee,E. *et al.* (2012) Landscape of somatic retrotransposition in human cancers. *Science*, **337**, 967–971.

Li,H. *et al.* (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*., **18**, 1851–1858.

Maksakova,I. *et al.* (2006) Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. *PLoS Genet*., **2**, e2.

McClintock,B. (1950) The origin and behavior of mutable loci in maize. *Proc. Natl Acad. Sci. USA*, **36**, 344–355.

Nellaker,C. *et al.* (2012) The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol*., **13**, R45.

Quinlan,A.R. *et al.* (2011) Genome sequencing of mouse induced pluripotent stem cells reveals retroelement stability and infrequent DNA rearrangement during reprogramming. *Cell Stem Cell*, **9**, 366–373.

Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Stewart,C. *et al.* (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet*., **7**, e1002236.