RetroSeq: Transposable element discovery from next-generation sequencing data

Thomas M. Keane[1,*], Kim Wong[1] and David J. Adams[1]
[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK.

1 Supplementary Methods

1.1 BAM Files

The BAM files used in the analysis were produced by the Broad institute and obtained from the 1000 genomes ftp site:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120117_ceu_trio_b37_decoy/

Briefly, the reads were aligned with BWA (v0.5.9-r16) to the hs37d5 1000 genomes human reference, reads were realigned around known indel sites from dbSNP 132, and the quality values were recalibrated using GATK. It should be noted that these extra processes (indel realignment and base quality recalibration) are **NOT** necessary for RetroSeq as it only uses the mapped location of reads.

RetroSeq has only been tested on BAM files obtained from BWA and Maq alignments. It may work with other aligners but it does depend on how the other aligners represent discordantly mapped reads.

1.2 Repeat Library

The Alu and L1 BED files were derived directly from Repeatmasker (v3.3) with Repeat Library 20120124:

http://www.repeatmasker.org/genomes/hg19/RepeatMasker-rm330-db20120124/hg19.fa.out.gz

Copies of the repeat annotation files used are posted here:

ftp://ftp-mouse.sanger.ac.uk/other/tk2/RetroSeq/hg19/

Also, we used Alu and L1 sequence files to increase the sensitivity of the discover stage (-align and -refTEs parameter). These were derived directly from Repbase (v16):

http://www.girinst.org/server/RepBase/index.php

Copies of the actual sequence files used are posted here:

ftp://ftp-mouse.sanger.ac.uk/other/tk2/RetroSeq/hg19/hg19_probes.tgz

1.3 Discovery Phase

RetroSeq scans the BAM file to identify discordantly mapped reads (0x2 flag unset) OR reads where the mate is unmapped. All discordantly mapped reads where one-end is mapped within the library of repeat elements are recorded. The sequence of unmapped mates is aligned with exonerate (-m affine:local --bestn 5) to the library of repeat element sequences and reads that align with greater the min percent identity (-id parameter,

default 80) and greater than the minimum length (-len parameter, default 36bp) are recorded.

In each case, the location and orientation of the mate of these reads is recorded in the output file (-output parameter) if the mapping quality is above the specified threshold (-q parameter, default 30). For the rest of this document, these reads are referred to as anchor reads.

1.4 Phase
A BED file of anchor supporting reads is produced from the discover phase.

The forward and reverse orientated anchor reads are then clustered separately by start position taking only one read per base position to avoid considering possible library PCR duplicate reads. Each cluster with greater than a specified number of reads (-reads parameter, default: 5) is recorded with a maximal allowable gap of 120bp between reads in a cluster.

1.5 Merging forward/reverse clusters
RetroSeq uses bedtools window command (part of bedtools: http://code.google.com/p/bedtools/) to carry out all possible merges of the forward and reverse clusters and outputs a new BED file with the possible merged clusters and the number of supporting reads per merged cluster along with the outer co-ordinates of the forward/reverse clusters. These are then treated as putative regions with a TE insertion call.

1.6 Breakpoint determination
The average read-depth in the regions surrounding the merged region clusters is computed and regions where the depth is greater than a specified cutoff (-depth parameter, default: 200) are excluded. For each region, RetroSeq starts at the 5' end and for each base position computes the cumulative number of forward and reverse discordant reads along with any soft-clipped reads (penalizing the total count where there are gaps of 100bp without supporting reads). The breakpoint is determined as the position where these two cumulative distributions maximise. Finally the breakpoint is characterised by the number of supporting discordant reads, the ratio of forward to reverse discordant reads 5' of the breakpoint and 3' of the breakpoint, and the distance between the last 5' forward supporting read and first 3' reverse supporting read. These annotations are included in the final VCF file so they can be used for final filtering.

1.7 Proximity to reference elements
All calls where the breakpoint is predicted to be in close proximity to a reference annotated TE insertion are filtered out. For the CEU analysis, the Alu calls were filtered using bedtools window (-v -w 100 options) compared to the Alu elements in the Repeatmasker reference annotation.

For L1 calls, we first removed L1 calls that directly overlapped with reference Alu elements using bedtools intersect (-v). And the resulting calls were then filtered for proximity to reference L1 elements using bedtools window (-v -200).

1.8 VCF Output
The final TE calls from RetroSeq are in VCF format (http://vcftools.sourceforge.net/). The calls are annotated with information on number of supporting reads (GQ tag). The FL tag ranges from 1-8 and gives information on the breakpoint with 8 being the most confident calls and lower values indicating calls that don't meet the breakpoint criteria for reasons such as lack of 5' or 3' supporting reads. The example in Fig. 1 shows two

human Alu calls, one with a GQ of 15 and FL of 5 and the other has a GQ of 37 and FL of 8 (highly confident).

For the final calls in Table 1, we selected calls from the VCF file with the following INFO tags based on ROC curves with existing PCR validated calls for these individuals:

FL=6 & GQ>=28
FL=7 & GQ>=20
FL=8 & GQ>=20

```
##fileformat=VCFv4.0
##source=RetroSeq v0.2
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=MEINFO,Number=4,Type=String,Description="Mobile element info of the form NAME,START,END,POLARITY">
##ALT=<ID=INS:ME,Type=String,Description="Insertion of a mobile element">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=FL,Number=1,Type=Integer,Description="Call Status - for reference calls a flag to say if the call failed a particular filter. Filters are
ordered by priority in calling (higher number indicates closer to being called). 1 - depth too high in region, 2 - not enough reads in cluster, 3 - not
enough total flanking reads, 4 - not enough inconsistently mapped reads, 5 - neither side passes ratio test, 6 - one side passes ratio test, 7 - distance
too large at breakpoint, 8 - PASSED all filters">
##INFO=<ID=NOT_VALIDATED,Number=0,Type=Flag,Description="Not validated experimentally">
##INFO=<ID=1000G,Number=0,Type=Flag,Description="Overlaps with 1000G MEI call">
##INFO=<ID=REPEATMASKER,Number=0,Type=Flag,Description="Overlaps with a reference ME element">
#CHROM   POS      ID       REF      ALT      QUAL     FILTER   INFO     FORMAT   NA12891
10       135035407.        G        <INS:ME> 15       .        MEINFO=Alu,135035407,135035408,NA;NOT_VALIDATED;SVTYPE=INS      GT:GQ:FL 1/1:15:5
11       98774710 .        A        <INS:ME> 37       .        MEINFO=Alu,98774710,98774711,NA;NOT_VALIDATED;SVTYPE=INS       GT:GQ:FL 1/1:37:8
```

Fig. 1: Example of first 20 lines of the RetroSeq VCF output for NA12891