A review of methods for detecting non-reference transposable element insertions from high throughput genome resequencing data.

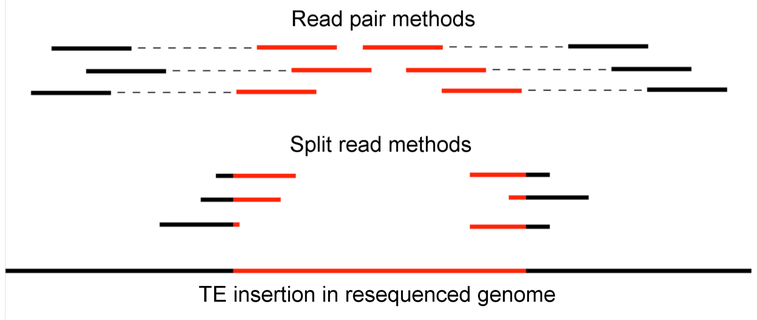
Michael G. Nelson and Casey M. Bergman (Michael Smith Building, Faculty of Life Sciences, University Of Manchester, UK)

The University of Manchester

Introduction

The declining cost of DNA sequencing has opened the door for many new applications in genomics, including detection of transposable elements (TEs) in resequenced genomes. Several methods have recently been developed to detect TE insertions in high throughput sequencing data for a number of different applications including cancer genomics, bacterial genomics and evolutionary biology. These methods fall into two basic categories:

1. Searching for **read pairs**, where one read aligns to a unique reference genomic sequence and the other does not align to the reference or aligns to a distant location in the reference.
2. Searching for **split reads**, reads that partially contain TE sequence and partially contain unique reference genomic DNA.



We have undertaken a study to evaluate the performance of bioinformatic methods to detect TEs in resequencing data. Currently we are focusing on the technical aspects of installation and use, and a more detailed evaluation study is in progress.

Methods

Five methods were selected for this study that had software available and could be adapted for use with different TE families and genomes. They have been tested on a subset of eight *Drosophila melanogaster* TEs; *Doc, jockey, opus, Burdock, roo pogo, hobo, P-element*.

Table 1. List of the software methods tested. Strand refers to whether a method predicts the orientation of an insertion; TSD refers to whether the TSD is annotated; and Reference TEs refers to whether the software can detect insertions shared with the reference genome.

Method	Type	Reference	Strand	TSD	Reference TEs
ngs_te_mapper	Split read	Linheiro *et al.* (2012)	✔	✔	
RelocaTE	Split read	Robb *et al.* (2013)	✔	✔	✔
RetroSeq	Read pair	Keane *et al.* (2013)			
PoPoolationTE	Read pair	Kolfer *et al.* (2012)			✔
TE-locate	Read pair	Platzer *et al.* (2012)	✔		✔

Dependency Analysis

Table 2. Summary of software required to run each of the TE detection methods.

		ngs_te_mapper	RelocaTE	RetroSeq	PoPoolationTE	TE-locate
System requirements	Unix system	✔	✔	✔	✔	✔
	Perl		✔	✔	✔	✔
	R	✔				
	BioPerl		✔			
Toolkits	RepeatMasker				✔	
	RMBlast				✔	
	TRF				✔	
	BEDTools			✔		
	SAMTools		✔	✔	✔	
	BCFTools			✔		
Mappers	Blat	✔	✔			
	Exonerate			✔		
	Blast (Legacy)		✔			
	Bowtie		✔			
	BWA			✔	✔	✔

Table 3. A summary of the input files and formats required to complete each TE detection pipeline. * indicates a modified version of the file is needed. RelocaTE requires some information about the TSD of a TE. RetroSeq requires the sequences to be stored in individual files with a custom format file to identify the locations of the sequence files.

	ngs_te_mapper	RelocaTE	RetroSeq	PoPoolationTE	TE-locate
Reference genome (Fasta)	✔	✔	✔	✔	✔
Consensus TE sequences (Fasta)	✔	✔*	✔*	✔	
Annotation of reference TEs (GFF)				✔	✔
Annotation of reference TEs (custom format)		✔			
NGS data (Fastq)	✔	✔			
NGS data (paired-end Fastq)				✔	
Alignment (BAM)			✔		
Alignment (sorted SAM)					✔
RepeatMasker RepBase Library				✔	
TE hierarchy (custom format)				✔	

Supported by wellcome trust

Results

Figure 1. Annotation framework of the tested methods showing the annotation of one *hobo* element predicted by all methods. The two split read methods annotate the range of TSD of a TE insertion. The read pair methods annotate a single base or range of two neighbouring base pairs.

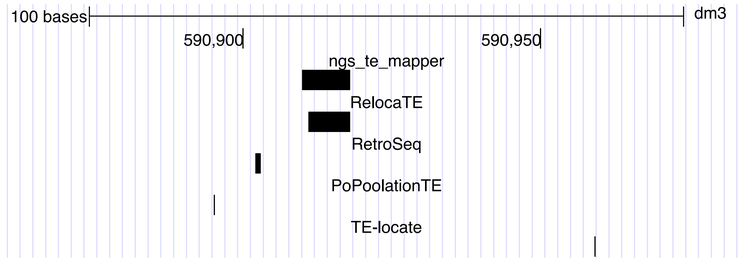


Figure 2. Overview of predictions for eight TE families on *D. melanogaster* chromosome arm 2L.

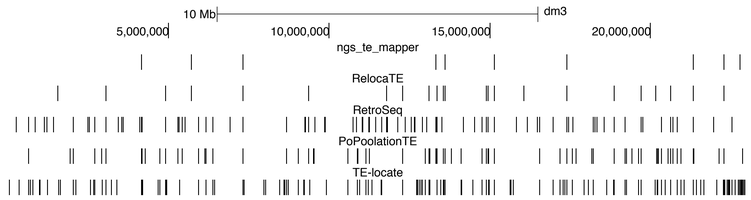


Table 4. Concordance of TE predictions for eight *D. melanogaster* TE families. Numbers (above diagonal) and percentages (below) of concordant TE predictions (defined as same family, within 100bp). The split read methods are generally concordant when both predicting the same TE. The read pair methods, as they are not directly detecting a breakpoint, will annotate an insertion in the same general location but which sometimes can be outside of the 100bp cut off tested here. The diagonal shows the total number of predictions made by each method.

	ngs_te_mapper	RelocaTE	RetroSeq	PoPoolationTE	TE-Locate
ngs_te_mapper	55	51	48	18	11
RelocaTE	92.73	172	132	42	27
RetroSeq	87.27	76.77	284	117	52
PoPoolationTE	32.73	24.42	41.20	314	113
TE-Locate	20	15.70	18.31	35.99	437

References

Keane, T.M., Wong, K., *et al.* (2012). "RetroSeq: transposable element discovery from illumina paired-end sequencing data." *Bioinformatics (Oxford, England)*. PMID: 23233656
Kofler, R., Betancourt, A.J., et al. (2012). "Sequencing of pooled DNA samples (pool-seq) uncovers complex dynamics of transposable element insertions in drosophila melanogaster." PLoS Genet, 8(1) PMID: 22291611
Linheiro, R.S. and Bergman, C.M. (2012). "Whole genome resequencing reveals natural target site preferences of transposable elements in drosophila melanogaster." PLoS ONE, 7(2). PMID: 22347367
Platzer, A., Nizhynska, V., et al. (2012). "TE-Locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data." Biology, 1(2):395410.
Robb, S.M.C., Lu, L., *et al.* (2013). "The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice." *G3*, 3(6): 949–957. PMID: 23576519.