TE Annotation Benchmark Proposal *(Template)*

1. *Name*

Drosophila melanogaster genome.

2. *Authors*

Emmanuelle Lerat

Lab. Biométrie et Biologie Evolutive, Université Lyon 1, France

emmanuelle.lerat@univ-lyon1

3. *Description*

The *Drosophila melanogaster* genome has been published in 2000. Since then, it has been the subject of throughout annotation process, in particular to describe the content of transposable elements (TEs). To date, the last version of TE insertion annotation is available on flybase (version 5.57; http://flybase.org/) and an exhaustive list of reference sequences are described in the Repbase database (http://www.girinst.org/).

4. *Specification*

	Description	Comments
Type (real, modified real*, simulated, other types?)	Real and modified real	Assembled genomic data
Primary Uses (to measure sensitivity? specificity? other metrics?)	Sensitivity, specificity	Used to test the performance of various *de novo* programs
Additional Uses	--	--
Taxa	*Drosophila melanogaster*	--
Source	flybase	--
Documentation	--	--
Version	--	--
Other	--	--

* e.g., modified real = real + modeled evolution

5. *Details*

* data description

The genome of *D. melanogaster* corresponds to about 140 Mb (Adams et al. 2000; Smith et al. 2007). The content of repeats has been estimated to 7% of the euchromatin (Bergman et al. 2006; Smith et al. 2007), and 77% of the heterochomatin (Smith et al. 2007). TE insertions have been annotated and represent a total of 5,409 insertions (version 5.57; http://flybase.org/). However, some reference TE are also to be taken into account, that are indicated in the Repbase database (http://www.girinst.org/). It is thus possible to still find some unannotated copies in this genome, although the vast majority has been described. Concerning the insertions, it has been estimated that less than a thousand are nested (Bergman et al. 2006; Smith et al. 2007).

* performance of signature based programs

To compare the ability of signature based approach programs developed to detect specifically LTR-retrotransposons, the X chromosome of *D. melanogaster* was used as a benchmark (Lerat 2010). Considering the annotations, this chromosome is supposed to contain 225 LTR-retrotransposon insertions among which 96 are full-length elements, which are the type of sequences that these specific programs are able to detect. It was thus possible to compute the sensitivity (TP/(TP+FN)) of each program.

* performance of a read based *de novo* program

To test the ability for a read based *de novo* program, simulated read data were obtained using the program ART (Huang et al. 2012), at 10% of coverage and with various read lengths (80, 150 and 250 nts) based on the genome sequence of *D. melanogaster*, with and without taking into account of the chromosomes U and Uextra. This dataset can be used conjointly with the set of reference sequences described in Repbase.

6. *References*

Adams MD et al. (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185-2195.

Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. Genome Biol. 7:R112.

Huang W, Li L, Myers JR, Marth GT (2012) ART: a next-generation sequencing read simulator. Bioinformatics 28:593-594.

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104:520-533.

Smith CD, Shu S, Mungall CJ, Karpen GH (2007) The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. Science 316:1586-1591.