# SEX-DETector
# User Manual

Aline Muyle

January 2, 2017

## Contents

# 1 SEX-DETector: why, when and how?

## 1.1 What does SEX-DETector do?

SEX-DETector is a likelihood-based method that can infer sex-linked genes (genes located in the non-recombining region of sex chromosomes) using RNA-seq data in a cross (parents and progeny sequenced). In species where a cross cannot be obtained, an empirical method using males and females from an inbred line is proposed here, but note that this method is less powerful than the model-based SEX-DETector that uses cross data (see publication).

## 1.2 In what cases can SEX-DETector be used?

SEX-DETector can be used on species with separated sexes where a cross can be sequenced. The method works on any sex chromosome type (XY, ZW, UV) and can even assess the presence and type of sex chromosomes using a model comparison strategy. SEX-DETector will work best on species with sex chromosomes that are either young or of intermediate age. In old sex chromosome systems, SEX-DETector will work on recently evolved strata and will only return X copies for old strata (if appropriate SNPs are present). Indeed, high levels of X-Y divergence prevent the coassembly of X and Y alleles.

## 1.3 In what cases will inferences be difficult, and what are the solutions?

- If RNA-seq is used, genes that are not expressed in the sampled tissue cannot be identified. This can be overcome by the use of DNA-seq data, or the combination of multiple tissues for RNA-seq data. Note that file formats might require to be adapted in the case of DNA-seq data.

- SEX-DETector cannot identify Y alleles unless they coassemble with their X counterpart. Y copies of sex-linked genes will therefore be missed in old sex chromosome systems where X-Y divergence prevents coassembly of X and Y copies. To try and identify missed Y contigs, X-hemizygous genes can be blasted onto male-specific contigs, which may represent the diverged Y copy.

- In old systems, only the X copies of sex-linked genes can be identified, using X SNPs. However, this identification of X copies will be impossible in RNA-seq data if only one non-random X is expressed in females, as is the case in non-random X chromosome inactivation or imprinting. This problem will not be as severe in the case of random X chromosome inactivation where, depending on the cell, one or the other X is expressed, so that at the level of

the tissue both Xs are expressed. In moderately young systems this will not be a problem because X/Y SNPs can be used for inferences.

- Because only one family is sequenced, only a few recombination events can be tracked in the dataset. Therefore, genes located in the pseudoautosomal regions (PAR) flanking the non-recombining region of sex chromosomes, can be wrongly inferred as fully sex-linked, when in fact they are only partially sex-linked. This will however make many analyses conservative (such as Y degeneration). In case PAR genes need to be filtered out, increasing the number of progenies sequenced will help as it increases the number of meiosis events studied. Otherwise, males and females from different wild populations can be sequenced by RNA-seq in order to check that inferred Y alleles are always and only present in males.

### 1.4 Can SEX-DETector be used on other genetic elements than sex chromosomes?

The SEX-DETector pipeline could also be used on other systems than sex chromosomes, such as mating type loci, B chromosomes, incompatibility loci, supergenes and any other type of dominant loci associated with a phenotype, for which a cross between a heterozygous and a homozygous individual is possible and for which both alleles are expressed at the transcript level in heterozygous individuals.

## 2 Presentation of the workflow

The workflow goes from further assemby after trinity to contigs segregation type inference, see Figure 1.

## 3 Generation of SEX-DETector input files

### 3.1 Transcriptome assembly using Trinity

We recommend assembling both male and female transcriptomes together after trimming the reads.

### 3.2 Further assembly using Cap3

As sex chromosomes diverge due to recombination supression, it is common that X and Y copies are assembled into different contigs. As the identification of X/Y alleles rely on both copies being assembled in the same contig we recommend running

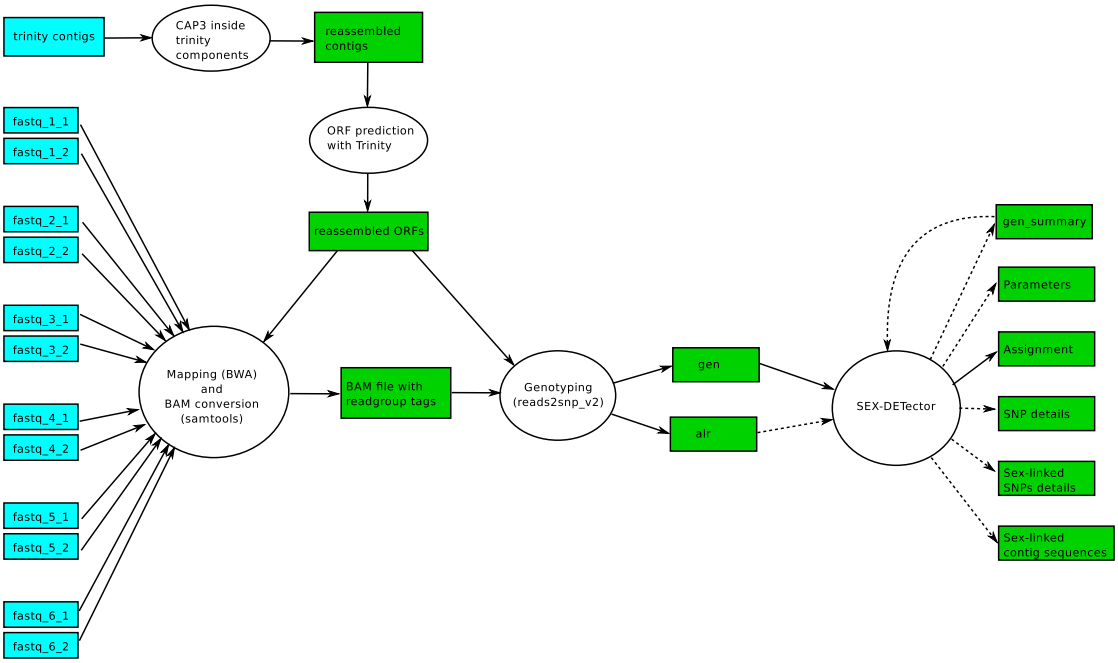


Figure 1: Workflow of a SEX-DETector analysis on RNAseq data. Blue boxes: input files; green boxes: files produced during the analysis; ellipses: software tools. Dashed arrows indicate optional input or output files. When using the Galaxy workflow, the fastq files need first to be listed using a file selector wrapper.

CAP3 inside of trinity components in order to further assemble contigs together. The command line requires the trinity assembly file name as input (**-input *trinity_assembly_name.fasta***), the output name (**-output *output.fasta***) and the parameter for cap3 assembly threshold (the minimum percent of similarity between sequences to assemble them into a single contig, **-sim *integer***) this parameter should be adapted to the hypothesized sex chromosomes age (90 for young sex chromosomes, 70 for old heteromorphic sex chromosomes):

```
./wrapper_galaxy_CAP3.pl –input input.fasta –output output.fasta –sim 70
```

The CAP3 programm must be in /bin and /bin in $PATH.

The input file should contain the trinity assembly in the following format (trinity 2014):

```
>c1_g1_i1 len=319 path=[1271:0-318]
ATACTTACGGAGGCAATGTATGAGGTACTGACGAATGACAAATTAACTCATCAACCAAAG
TTTGTTTTTTACTTTTCCTTTTTTTTGGGTTTTAAGGGAGGAACGAGAGCAAAATTGATG
...
```

The output is a fasta file for which contigs were assembled inside trinity components, as well as contigs that couldn't be assembled (singlets). The contigs assembled by CAP3 have the name of one of the original contigs.

## 3.3 ORFs prediction using Trinity

```
transcripts_to_best_scoring_ORFs.pl -t input.fasta
```

## 3.4 Mapping of reads onto the assembly using BWA

Raw reads can be mapped on the assembly after the CAP3 step. First the assembly file must be indexed :

```
bwa index -a is assembly.fasta
```

Then the reads of each individual can be mapped onto the assembly (here is the case of paired end, allowing for 5% mismatches between reads and contigs):

```
bwa aln -n 5 assembly.fasta sample_read_1.fastq > sample_read_1.sai
bwa aln -n 5 assembly.fasta sample_read_2.fastq > sample_read_2.sai
bwa sampe -r "@RG\tID:individual_name" assembly.fasta sample_read_1.sai sample_read_2.sai
 sample_read_1.fastq sample_read_2.fastq > sample.sam
```

The result is a sam formated alignment of reads onto the assembly.

## 3.5 SAMtools

The sam files of each individual can be compressed into bam files using the SAMtools program after indexing the assembly file (only mapped reads are conserved in the bam file here):

```
samtools faidx assembly.fasta
samtools view -t assembly.fasta.fai -F 4 -h -S -b -o sample.bam sample.sam
```

Then each bam file of each individual needs to be ordered, using samtools:

```
samtools sort sample.bam sample_sorted
```

the result is a sorted bam formated alignment of reads onto the assembly.

## 3.6 Generation of the alr file and if necessary the gen file

The gen file is not compulsory for data without cross as genotypes can be inferred from read counts.

This step is carried on by reads2snp, a genotyper designed for non model organisms that allows alleles to have different expression levels (option **-aeb**), which is useful for sex-chromosomes as the Y copy is less expressed than the X copy. The minimum number of reads to be taken into account can be set to 3 (option **-min 3**)By default reads2snp removes paralogous positions with the paraclean program, this should be disabled (option **-par 0**) because X and Y copies look like paralogs. The number of threads should be set to 1 (option **-nbth 1**) so that contigs are shown in the same order in the alr and the gen file. For neat SNP calling base quality can be filtered to be higher than 20 (option **-bqt 20**) and mapping quality higher than 10 (option **-rqt 10**). Input files must be the assembly fasta file (**-bamref *assembly_name.fasta***) and a text file listing all sorted bam file paths (one per line, **-bamlist *bamlist_name***) and the output base name must be provided (**-out *output_name***):

```
./reads2snp_2.0 -min 3 -par 0 -nbth 1 -aeb -bqt 20 -rqt 10 -bamlist bam_list_file -
bamref assembly.fasta -out outputs_base_name
```

The alr file is tab delimited and shows each position of each contig on a line with the majoritary allele (most common allele in number of reads), then M if the position is monomorphic and P if the position shows different alleles, then for each individual the total read number if the position is monomorphic or if the position is polymorphic the total read number followed by the number of reads for each base [A/C/G/T]:

```
>contig_name
maj        M/P        Species|individual1        Species|individual2
T        M        8        1
C        P        17[0/0/0/17]        14[0/1/0/13]
```

The gen file is tab delimited and shows each position of each contig on a line with the position number starting from 1, then for each individual the inferred genotype followed by a pipe and the posterior probability of the inferred genotype:

```
>contig_name
position        Species|individual1        Species|individual2
1        TT|1        NN|0
2        TT|1        TT|0.999989
```

## 3.7 Generation of the gen_summary file if necessary

Unnecessary for data without cross.
This step allows SEX-DETector to run faster. The command line requires a gen

file (**-alr_gen** ***gen_file_name***), then a gen_summary file name for the output (**-alr_gen_sum** ***gen_summary_file_name***). The names of the homogametic progeny individuals separated by commas (**-hom** ***string***), the names of the heterogametic progeny individuals separated by commas (**-het** ***string***), the homogametic parent name (**-hom_par** ***string***) and the heterogametic parent name (**-het_par** ***string***):

```
./SEX−DETector_prepare_file.pl input.gen output.alr_gen_summary −hom homogametic_
progeny_name1,homogametic_progeny_name2,... −het heterogametic_progeny_name1,
heterogametic_progeny_name2,... −hom_par homogametic_parent_name −het_par
heterogametic_parent_name
```

The alr_gen_summary file is similar to the gen file except that it shows only one occurence of each possible SNP in the dataset and shows the number of time it happens in the first column instead of the position number of the SNP:

```
>dataset.alr_gen_summary
number_time_line_happens        Species|individual1        Species|individual2
1         CC         AA
34         TT         CT
```

### 3.7.1 X/Y or Z/W systems

The command line for X/Y or Z/W systems requires the input gen file name (first argument), the output gen_summary file name (second argument), the names of the homogametic progeny individuals separated by commas (option -hom), the names of the heterogametic progeny individuals separated by commas (option -het), the homogametic parent name (option -hom_par) and the heterogametic parent name (option -het_par):

```
./SEX−DETector_prepare_file.pl input.alr_gen output.alr_gen_summary −hom homogametic_
progeny_name1,homogametic_progeny_name2,... −het heterogametic_progeny_name1,
heterogametic_progeny_name2,... −hom_par homogametic_parent_name −het_par
heterogametic_parent_name
```

### 3.7.2 U/V system

The command line for U/V systems requires the input gen file name (first argument), the output gen_summary file name (second argument), the names of the female progeny individuals separated by commas (option -female), the names of the male progeny individuals separated by commas (option -male) and the parent name (option -par):

```
./SEX-DETector_UV_prepare_file.pl input.alr_gen output.alr_gen_summary -female female_progeny_name1,female_progeny_name2,... -male male_progeny_name1,male_progeny_name2,... -par parent_name
```

# 4 SEX-DETector output files

## 4.1 Default outputs

### 4.1.1 Parameters file

This file is a default output only when a cross is used.

**X/Y or Z/W system** The Parameters file is in a tab delimited text format with a header and for each line the inferred parameters values for a given iteration of the EM (Expectation Maximization) algorithm:

- iteration number
- dataset global autosomal segregation probability
- dataset global X/Y segregation probability
- dataset global X-hemizygous segregation probability
- Y genotyping error probability
- other genotyping error probability
- global genotypes probabilities for the homozygous sex and for autosomal segregation in the heterozygous sex (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT)
- global genotypes probabilities for X/Y segregation in the heterozygous sex (AC, CA, AG, GA, AT, TA, CG, GC, CT, TC, GT, TG)
- global genotypes probabilities for X-hemizygous segregation in the heterozygous sex (A, C, G, T)
- for each line in the case the debug option is set or for the last line in case the L option is set: value of Q, H and L the likelihood of the model, the number of free parameters, the sample size (number of SNPs positions used to estimate parameters) and the BIC value of the model.

**U/V system** The Parameters file is in a tab delimited text format with a header and for each line the inferred parameters values for a given iteration of the EM (Expectation Maximization) algorithm:

- iteration number
- dataset global autosomal segregation probability
- dataset global U/V segregation probability
- genotyping error probability
- global genotypes probabilities for autosomal segregation in the parent (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT)
- global genotypes probabilities for U/V segregation in the parent (AC, CA, AG, GA, AT, TA, CG, GC, CT, TC, GT, TG)
- for each line in the case the debug option is set or for the last line in case the L option is set: value of Q, H and L the likelihood of the model, the number of free parameters, the sample size (number of SNPs positions used to estimate parameters) and the BIC value of the model.

#### 4.1.2 Assignment file

**With a cross**

**X/Y or Z/W system** The assignment file is in a tab delimited text format with a header and for each line the information about each contig of the input file(s):

- contig name
- contig weighted mean probability to be autosomal
- contig weighted mean probability to be X/Y (or Z/W)
- contig weighted mean probability to be X-hemizygous (or Z-hemizygous)
- contig assignment to a segregation type (a contig is considered sex-linked if the sum of the X/Y plus the X-hemizygous probability is higher than the autosomal probability and if at least one X/Y or X-hemizygous SNP was found without genotyping error and without aberrant reads, a contig is considered autosomal if the autosomal probability is higher than the X/Y plus the X-hemizygous probability and if at least one autosomal SNP was

found without genotyping error, for other cases the contig is assigned to a "lack-information" category). Aberrant reads are cases where homogametic individuals carry Y or W reads, or heterogametic individuals carry reads with the paternal X or maternal Z allele (note that this filter will be applicable only if an alr file is provided as input).

- total number of SNP inferred without genotyping error
- number of autosomal SNP inferred without genotyping error
- number of autosomal SNP inferred with genotyping errors
- number of X/Y SNP inferred without genotyping error
- number of X/Y SNP inferred with genotyping errors
- number of X-hemizygous SNP inferred without genotyping error
- number of X-hemizygous SNP inferred with genotyping errors
- if alr file is provided the number of X/Y or Z/W SNP inferred without genotyping errors that do not have Y (or W) allele expression in the homogametic sex nor paternal X / maternal Z allele in heterogametic individuals.
- if alr file is provided the number of X-hemizygous or Z-hemizygous SNP inferred without genotyping errors that do not have paternal X / maternal Z allele bearing reads in heterogametic individuals.

**U/V system** The assignment file is in a tab delimited text format with a header and for each line the information about each contig of the input file(s):

- contig name
- contig weighted mean probability to be autosomal
- contig weighted mean probability to be U/V
- contig assignment to a segregation type (a contig is considered sex-linked if the U/V probability is higher than the autosomal probability and if at least one U/V SNP was found without genotyping error, a contig is considered autosomal if the autosomal probability is higher than the U/V probability and if at least one autosomal SNP was found without genotyping error, for other cases the contig is assigned to a "lack-information" category)
- total number of SNP inferred without genotyping error

- number of autosomal SNP inferred without genotyping error
- number of autosomal SNP inferred with genotyping errors
- number of U/V SNP inferred without genotyping error
- number of U/V SNP inferred with genotyping errors.

**Without cross** The assignment file is in a tab delimited text format with a header and for each line the information about each contig of the alr input file:

- contig name
- contig assignment to a segregation type (a contig is assigned to a sex-linked segregation type if it has at least one X/Y or Z/W SNP)
- number of X/Y (or Z/W) SNPs
- number of other SNPs
- number of SNPs having more than two alleles.

## 4.2 Optional outputs

### 4.2.1 SNPs detail file

**With a cross**

**X/Y or Z/W system** The SNPs detail file is in a tab delimited text format with a header and for each line the information about each SNP of each contig of the input file:

- contig name
- SNP position in contig
- SNP autosomal posterior probability
- SNP X/Y posterior probability
- SNP X-hemizygous posterior probability
- inferred heterogametic and homogametic parent genotypes (separated by a comma) for autosomal segregation type

- SNP type for autosomal segregation (either monomorphic, not informative if the heterogametic sex is homozygous or if both parents are heterozygous for the same genotype, informative otherwise)
- inferred heterogametic and homogametic parent genotypes (separated by a comma) for X/Y segregation type
- SNP type for X/Y segregation (either monomorphic, not informative if the heterogametic sex is homozygous or if both parents are heterozygous for the same genotype, XY, XX, XXY,XXX, XXXY depending on the number of different X and Y alleles)
- inferred heterogametic and homogametic parent genotypes (separated by a comma) for X-hemizygous segregation type
- SNP type for X-hemizygous segregation (either monomorphic, not informative if the heterogametic sex is homozygous or if both parents are heterozygous for the same genotype, XX0, XXX0 depending on the number of different X alleles)
- number of genotyping errors in the case of the autosomal segregation type
- number of genotyping errors in the case of the X/Y segregation type
- number of genotyping errors in the case of the X-hemizygous segregation type
- number of Y genotyping errors
- if alr file provided, the number of individuals presenting abnormally high levels (>2 %) of aberrant reads (homogametic individuals showing Y reads or heterogametic individuals with reads bearing the paternal X or maternal Z allele).
- heterogametic parent observed genotype
- if alr file provided, heterogametic parent expression (number of reads for A, C, G, T separated by '/')
- homogametic parent observed genotype
- if alr file provided, homogametic parent expression (number of reads for A, C, G, T separated by '/')

And then for each progeny individual, starting with the homogametic sex:

- individual observed genotype
- if alr file provided, individual expression (number of reads for A, C, G, T separated by '/')

**U/V system** The SNPs detail file is in a tab delimited text format with a header and for each line the information about each SNP of each contig of the input file:

- contig name
- SNP position in contig
- SNP autosomal posterior probability
- SNP U/V posterior probability
- inferred parent genotype for autosomal segregation type
- SNP type for autosomal segregation (either monomorphic or informative)
- inferred parent genotype for U/V segregation type
- SNP type for U/V segregation (either monomorphic or informative)
- number of genotyping errors in the case of the autosomal segregation type
- number of genotyping errors in the case of the U/V segregation type
- parent observed genotype
- if alr file provided, parent expression (number of reads for A, C, G, T separated by '/')

And then for each progeny individual, starting with females:

- individual observed genotype
- if alr file provided, individual expression (number of reads for A, C, G, T separated by '/')

**Without cross** The SNPs detail file is in a tab delimited text format with a header and for each line the information about each SNP of each contig of the input file:

- contig name
- SNP position in contig
- the SNP type (either "XY", "other" or "mul" if more than 2 alleles present)
- the inferred Y allele in the case of a X/Y SNP (NA otherwise)
- the inferred X allele in the case of a X/Y SNP (NA otherwise)

And then for each individual, starting with the homogametic sex:

- individual observed genotype
- if alr file provided, individual expression (number of reads for A, C, G, T separated by '/')

### 4.2.2 Sex-linked SNPs detail file

**With a cross**

**X/Y or Z/W system** The sex-linked SNPs detail file is in a tab delimited text format with a header and for each line the information about each sex-linked SNP of each sex-linked contig of the input file (a SNP is considered to be sex-linked if the X/Y or the X-hemizygous posterior probability is higher than the two other segregation type probabilities):

- contig name
- SNP position in contig
- SNP segregation type (either X/Y or X-hemizygous) posterior probability
- SNP type (either XY, XX, XXY, XXX, XXXY, XX0 or XXX0, 0 indicating the absence of a Y expressed allele)
- the homogametic parent inferred X1 (or Z1) allele (NA if individual is homozygous)
- if alr file provided, the homogametic parent read number for the X1 (or Z1) allele (NA if individual is homozygous)

- the homogametic parent inferred X2 (or Z2) allele (NA if individual is homozygous)
- if alr file provided, the homogametic parent read number for the X2 (or Z2) allele (NA if individual is homozygous)
- the homogametic parent inferred single X (or Z) allele if the individual is homozygous at this position
- if alr file provided, the homogametic parent total read number at this position
- the heterogametic parent inferred X (or Z) allele (NA if individual is homozygous)
- if alr file provided, the heterogametic parent read number for the X (or Z) allele (NA if individual is homozygous)
- the heterogametic parent inferred Y (or W) allele (NA if individual is homozygous)
- if alr file provided, the heterogametic parent read number for the Y (or W) allele (NA if individual is homozygous)
- the heterogametic parent inferred single allele if the individual is homozygous at this position
- if alr file provided, the heterogametic parent total read number at this position

And then for each progeny individual of the homogametic sex:

- the homogametic individual inferred X1 (or Z1) allele (NA if individual is homozygous)
- if alr file provided, the homogametic individual read number for the X1 (or Z1) allele (NA if individual is homozygous)
- the homogametic individual inferred X2 (or Z2) allele (NA if individual is homozygous)
- if alr file provided, the homogametic individual read number for the X2 (or Z2) allele (NA if individual is homozygous)
- the homogametic individual inferred single X (or Z) allele if the individual is homozygous at this position

- if alr file provided, the homogametic individual total read number at this position

And then for each progeny individual of the heterogametic sex:

- the heterogametic individual inferred X (or Z) allele (NA if individual is homozygous)
- if alr file provided, the heterogametic individual read number for the X (or Z) allele (NA if individual is homozygous)
- the heterogametic individual inferred Y (or W) allele (NA if individual is homozygous)
- if alr file provided, the heterogametic individual read number for the Y (or W) allele (NA if individual is homozygous)
- the heterogametic individual inferred single allele if the individual is homozygous at this position
- if alr file provided, the heterogametic individual total read number at this position

When the genotype is unknown for an individual (due to lack of data), the alleles are noted as "N". When the allele is followed by a "*" it means it was obtained from read counts and not from the reads2snp genotyping. Alleles written as "?" indicate incompatibility between the inferred segregation type by SEX-DETector, the reads2snp inferred genotype and the read numbers of different alleles. Y alleles for X hemizygous SNPs and homozygous alleles for heterozygous individuals are noted as 'NA'.

**U/V system** The sex-linked SNPs detail file is in a tab delimited text format with a header and for each line the information about each sex-linked SNP of each sex-linked contig of the input file (a SNP is considered to be sex-linked if the U/V posterior probability is higher than the autosomal one):

- contig name
- SNP position in contig
- SNP posterior U/V posterior probability
- SNP type (either XY, XX, XXY, XXX, XXXY, XX0 or XXX0, 0 indicating the absence of a Y expressed allele)

- the homogametic parent inferred X1 (or Z1) allele
- if alr file provided, the homogametic parent read number for the X1 (or Z1) allele
- the homogametic parent inferred X2 (or Z2) allele
- if alr file provided, the homogametic parent read number for the X2 (or Z2) allele
- the homogametic parent inferred single X (or Z) allele if the individual is homozygous at this position
- if alr file provided, the homogametic parent total read number at this position
- the heterogametic parent inferred X (or Z) allele
- if alr file provided, the heterogametic parent read number for the X (or Z) allele
- the heterogametic parent inferred Y (or W) allele
- if alr file provided, the heterogametic parent read number for the Y (or W) allele
- the heterogametic parent inferred single allele if the individual is homozygous at this position
- if alr file provided, the heterogametic parent total read number at this position

And then for each progeny individual of the homogametic sex:

- the homogametic individual inferred X1 (or Z1) allele
- if alr file provided, the homogametic individual read number for the X1 (or Z1) allele
- the homogametic individual inferred X2 (or Z2) allele
- if alr file provided, the homogametic individual read number for the X2 (or Z2) allele
- the homogametic individual inferred single X (or Z) allele if the individual is homozygous at this position

- if alr file provided, the homogametic individual total read number at this position

And then for each progeny individual of the heterogametic sex:

- the heterogametic individual inferred X (or Z) allele
- if alr file provided, the heterogametic individual read number for the X (or Z) allele
- the heterogametic individual inferred Y (or W) allele
- if alr file provided, the heterogametic individual read number for the Y (or W) allele
- the heterogametic individual inferred single allele if the individual is homozygous at this position
- if alr file provided, the heterogametic individual total read number at this position

When the genotype is unknown for an individual (due to lack of data), the alleles are noted as "N". When the allele is followed by a "*" it means it was obtained from read counts and not from the reads2snp genotyping. Alleles written as "?" indicate incompatibility between the inferred segregation type by SEX-DETector, the reads2snp inferred genotype and the read numbers of different alleles. Homozygous unique alleles for heterozygous individuals are noted as 'NA'.

**Without cross** The sex-linked SNPs detail file is in a tab delimited text format with a header and for each line the information about each X/Y (or Z/W) SNP of each sex-linked contig of the input file :

- contig name
- position
- inferred Y (or W) allele
- inferred X (or Z) allele
- the number of X (or Z) reads in each homogametic individual, tab delimited
- the number of Y (or W) then X (or Z) reads in each heterogametic individual, tab delimited

#### 4.2.3 SEX-linked contig sequences

This is a fasta format with sequences called by the contig name followed by Y, W, X, Z or X1, X2, X3 depending in the number of X or Z sequences for the contig (the alleles attribution to a sequence is random and certainly does not represent haplotypes).

# 5 SEX-DETector

## 5.1 SEX-DETector for data with a cross

### 5.1.1 X/Y system

The command line requires a gen file (**-alr_gen *gen_file_name***) and a gen_summary file (**-alr_gen_sum *gen_summary_file_name***). The alr file (**-alr *alr_file_name***) should only be specified if the user wants an output for sex-linked sequences or wants expression levels to be shown in outputs. The command line also requires the base name for output files (**-out *output_file_name***), the system type (xy or zw, **-system *string***, this is only used for output file headers), the names of the homogametic progeny individuals separated by commas (**-hom *string***), the names of the heterogametic progeny individuals separated by commas (**-het *string***), the homogametic parent name (**-hom_par *string***) and the heterogametic parent name (**-het_par *string***):

```
./SEX−DETector.pl −alr_gen input.gen −alr_gen_sum input.alr_gen_summary −out output_
name −hom homogametic_progeny_name1,homogametic_progeny_name2,... −het heterogametic
_progeny_name1,heterogametic_progeny_name2,... −hom_par homogametic_parent_name −het_
par heterogametic_parent_name −system xy [options]
```

Options :

**-h** display help message

**-seq** outputs the optional sequence file for sex-linked contigs (requires -alr *alr_file_name*).

**-detail** outputs the optional detail file for every SNP in the dataset.

**-detail-sex-linked** outputs the optional detail file for all sex-linked SNPs in the dataset (requires -alr *alr_file_name*).

**-alr *alr_file_name*** expression levels are shown in SNP detail outputs.

**-L** outputs Likelihood and BIC values for the last iteration of the EM algorithm in the parameters output file.

**-debug** outputs Likelihood and BIC values for each iteration of the EM algorithm in the parameters output file.

**-SEM** SEM algorithm (Stochastic step for the first ten iterations).

**-no_sex_chr** the model will only have autosomal segregation type, this is useful to test for the presence of sex chromosomes in a species by comparing BIC values.

**-thr *float*** threshold for contig posterior segregation probability, only contigs with autosomal or sex-linked probability higher than this threshold are shown. Default value 0.8.

**-param *control_file_name*** provide starting parameter values for the EM algorithm in a control text file that consists of a line equivalent to a line of the parameters output file. This is useful in case the EM algorithm stopped before the end and the user wants to launch it from where it stopped.

**-skip_opt** avoid optimization of parameters with the SEM algorithm, the input parameters are directly used for inferences.

**-p *float*** input parameter p (Y genotyping error). Default value 0.1.

**-E *float*** input parameter $\epsilon$ (genotyping error). Default value 0.01.

**-pi_1 *float*** input parameter $\pi_1$ (autosomal segregation probability). Default value 1/3

**-pi_2 *float*** input parameter $\pi_2$ (X/Y segregation probability). Default value 1/3

**-pi_3 *float*** input parameter $\pi_3$ (X hemizygous segregation probability). Default value 1/3

### 5.1.2 U/V system

The command line requires a gen file (**-alr_gen *gen_file_name***) and a gen_summary file (**-alr_gen_sum *gen_summary_file_name***). The alr file (**-alr *alr_file_name***) should only be specified if the user wants an output for sex-linked sequences or wants expression levels to be shown in outputs. The command line also requires the base name for output files (**-out *output_file_name***), the names of the female progeny individuals separated by commas (**-female *string***), the names of the male progeny individuals separated by commas (**-male *string***), the parent name (**-par *string***):

```
./SEX−DETector_UV.pl −alr_gen input.gen −alr_gen_sum input.alr_gen_summary −out
output_name −female female_progeny_name1,female_progeny_name2,... −male male_progeny_
name1,male_progeny_name2,... −par parent_name [options]
```

Options :

**-h** display help message

**-seq** outputs the optional sequence file for sex-linked contigs (requires -alr *alr_file_name*).

**-detail** outputs the optional detail file for every SNP in the dataset.

**-detail-sex-linked** outputs the optional detail file for all sex-linked SNPs in the dataset (requires -alr *alr_file_name*).

**-alr *alr_file_name*** expression levels are shown in SNP detail outputs.

**-L** outputs Likelihood and BIC values for the last iteration of the EM algorithm in the parameters output file.

**-debug** outputs Likelihood and BIC values for each iteration of the EM algorithm in the parameters output file.

**-SEM** SEM algorithm (Stochastic step for the first ten iterations).

**-no_sex_chr** the model will only have autosomal segregation type, this is useful to test for the presence of sex chromosomes in a species by comparing BIC values.

**-thr *float*** threshold for contig posterior segregation probability, only contigs with autosomal or sex-linked probability higher than this threshold are shown. Default value 0.8.

**-param *control_file_name*** provide starting parameter values for the EM algorithm in a control text file that consists of a line equivalent to a line of the parameters output file. This is useful in case the EM algorithm stopped before the end and the user wants to launch it from where it stopped.

**-skip_opt** avoid optimization of parameters with the SEM algorithm, the input parameters are directly used for inferences.

**-E *float*** input parameter $\epsilon$ (genotyping error). Default value 0.01.

**-pi_1 *float*** input parameter $\pi_1$ (autosomal segregation probability). Default value 1/3

**-pi_2 *float*** input parameter $\pi_2$ (X/Y segregation probability). Default value 1/3

## 5.2 SEX-DETector for data without cross

This version of SEX-DETector is for datasets without cross, either brothers and sisters from an inbred line, or males and females from a same population.
The command line requires either a gen file (**-alr_gen *gen_file_name***) or an alr file (**-alr *alr_file_name***), if the gen file is provided genotypes from reads2snps are used, otherwise genotypes are inferred using read counts in the alr file (by default an allele needs to be reprensented by at least 2% of the total read count of an individual, and more than three reads to be taken into account), if the gen file is provided the alr file should only be specified if the user wants an output for sex-linked sequences or wants expression levels to be shown in outputs. The command

line also requires the base name for output files (**-out *output_file_name***), the system type (xy or zw, **-system *string***, this is only used for output file headers), the names of the homogametic individuals separated by commas (**-hom *string***), the names of the heterogametic individuals separated by commas (**-het *string***):

```
./SEX−DETector_inbred.pl (−alr input.alr | −alr_gen input.gen) −out output_name −system
xy −hom homogametic_individual_name1,homogametic_individual_name2,... −het
heterogametic_individual_name1,heterogametic_individual_name2,... [options]
```

Options :

**-h** display help message

**-seq** outputs the optional sequence file for sex-linked contigs (requires -alr *alr_file_name*).

**-detail** outputs the optional detail file for every SNP in the dataset.

**-detail-sex-linked** outputs the optional detail file for all sex-linked SNPs in the dataset (requires -alr *alr_file_name*).

**-alr *alr_file_name*** expression levels are shown in SNP detail outputs.

**-min *integer*** minimum number of read required to take an allele into account (necessary only when genotyping is done from read counts), default 3.

**-err *float*** minimum proportion of total read count required to take an allele into account for an individual (necessary only when genotyping is done from read counts), default 0.02.

**-thr *integer*** minimum number of individuals of each sex required with a defined genotype in order to consider a position in a contig, default 3 (3 males and 3 females).

# 6 Galaxy wrappers and workflow

## 6.1 Overview

The three versions of SEX-DETector can be used through the graphical interface provided by the Galaxy project. In addition to the wrappers for SEX-DETector, wrappers for the steps upstream of SEX-DETector are included in an example workflow (except the assembly by Trinity, which is very computer-intensive).

These wrappers have been tested with Galaxy changesets 29ce93a13ac7 and 75efcf774765.

## 6.2 Installation

Like many Galaxy wrappers, this wrapper consists of two files:

- the script `SEX-DETector_wrapper.sh`
- `SEX-DETector_wrapper.xml` that links the script to Galaxy's interface

Put both files at a place where Galaxy can find them.

The actual work is done by the perl scripts for SEX-DETector, which should be put at a place where the shell can find them (e.g. in `/usr/bin/`). Alternatively, indicate the exact paths of these scripts in lines 9-13 of the wrapper script `SEX-DETector_wrapper.sh`.

For some options, the script also needs samtools, the path of which can be indicated in line 14.

## 6.3 Usage of SEX-DETector under Galaxy

The SEX-DETector galaxy wrapper is written to launch one of the three versions of SEX-DETector, based on the user's choices. SEX-DETector needs the output of Reads2snp, a genotyper for non-model organisms, designed in the context of the PopPhyl project (`http://kimura.univ-montp2.fr/PopPhyl/`). Its galaxy wrapper is available there. In most cases, the genotype file produced by Reads2snp is a necessary input, whereas the alr file is optional, but the population version of SEX-DETector can run with only the alr file.

In all cases, the user has to indicate which individuals are to be considered as heterogametic or homogametic respectively, and when data from a cross are used, which individuals are the parents or the offspring respectively. The wrapper provides two possibilities to assign the individuals to a type. First, the names of the individuals can be entered in text boxes, in the same format as one would use for SEX-DETector's command line. These names can be found in Reads2snp's output files, or in the "List of read groups" file produced by the "BWA (multi)" wrapper (see below). Alternatively, if one used one bam file per individual as an input for Reads2snp, the user can choose those bam files from the Galaxy history, and assign the individual type to the individual represented by the file. The individual names are either contained in the identifier (ID) or the sample (SM) tag of the readgroup (@RG) in the bam file.

Other common options for SEX-DETector can be changed through the Galaxy wrapper; exceptions are the options -h (display help), -param (provide initial parameter values), -skip_opt (avoid parameter optimization), -debug, and the parameter value options -p, -E, -pi_1, -pi_2 and -pi_3.

## 6.4 Workflow

A complete workflow, starting from reads assembled by Trinity, and the individual fastq read files, can be constructed in Galaxy, using the following steps:

1. Starting from a Trinity assembly, components can be further assembled together by CAP3. The "CAP3 on trinity assembly" wrapper performs this task, with one ajustable parameter which is the overlap percent identity cutoff for CAP3.

2. The reads in the individual fastq files then need to be aligned to these reference contigs, which can be done using the "BWA (multi)" wrapper. This wrapper needs a list of fastq files as input, which can be contructed from fastq files available in the Galaxy history using the "File selector" (for single-end sequencing) or the "Paired-end Fastq File selector" (for paired-end sequencing). It is advised to provide names for each individual, as otherwise the rather uninformative file names given by Galaxy are used as individual names. The "BWA (multi)" wrapper is specially designed for non-model organisms, for which no external reference genome is available. It allows to build a new index from the Trinity/CAP3 contigs only once, aligning all reads in the individual files to this index (the standard BWA wrapper requires that the index be rebuilt for each individual, which requires much more time). It is furthermore advised to use merge all individuals in one bam file, wherein the individuals are identified by their read group tags (@RG). The list of tags can be found in the output "List of read groups".

3. The single bam file containing all individuals and contigs serves as input for Reads2snp. Be sure to tell Reads2snp to use only one thread, as otherwise the order of the contigs might be changed. Also, disable "paraclean usage" and check "Account for allelic expression bias (model 2 in Tsagkogeorga et al 2012)".

4. SEX-DETector can now be run on the Reads2snp output. For the versions of SEX-DETector that use known parents and progeny, a gen_summary file needs to be produced to speed up calculations. If, however, one already has produced such a file in a previous run, and the types of the individuals (homo- or heterogametic, parent or offspring) have not been changed, one can choose to reuse the previously produced file. Exactly how to assign types to individuals is explained above.

You can find the source codes, detailed installation instructions and examples on the web page dedicated to SEX-DETector at `http://lbbe.univ-lyon1.fr/-SEX-DETector-.html`