A probabilistic method for identifying sex-linked genes using RNA-seq-derived genotyping data

Aline Muyle[1], Jos Käfer[1], Niklaus Zemp[2], Sylvain Mousset[1], Franck Picard[1]*, Gabriel AB Marais[1]*

[1] *Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS/Université Lyon 1, Villeurbanne, France*

[2] *ETH Zurich, Institute of Integrative Biology, Universitätstrasse 16, 8092 Zürich, Switzerland*

* equal contribution as senior authors

Corresponding authors: Aline Muyle, aline.muyle@univ-lyon1.fr, and Gabriel Marais, gabriel.marais@univ-lyon1.fr, LBBE – UMR 5558, CNRS/Université Lyon 1, Villeurbanne, France, Tel: (+33) (0) 4 72 43 29 09, Fax: (+33) (0) 4 72 43 13 88.

Running title: Identifying sex-linked genes using NGS

Keywords: sex chromosomes, XY, ZW, UV, sex-linked genes, non-model organisms, RNA-seq, Galaxy workflow

Abstract

The genetic basis of sex determination remains unknown for the vast majority of organisms with separate sexes. A key question is whether a species has sex chromosomes (SC). SC presence indicates genetic sex determination, and their sequencing may help identifying the sex-determining genes and understanding the molecular mechanisms of sex determination. Identifying SC, especially homomorphic SC, can be difficult. Sequencing SC is also very challenging, in particular the repeat-rich non-recombining regions. A novel approach for identifying sex-linked genes and SC consisting of using RNA-seq to genotype male and female individuals and study sex-linkage has recently been proposed. This approach entails a modest sequencing effort and does not require prior genomic or genetic resources, and is thus particularly suited to study non-model organisms. Applying this approach to many organisms is, however, difficult due to the lack of an appropriate statistically-grounded pipeline to analyse the data. Here we propose a model-based method to infer sex-linkage using a maximum likelihood framework and genotyping data from a full-sib family, which can be obtained for most organisms that can be grown in the lab and for economically important animals/plants. Our method works on any type of SC (XY, ZW, UV) and has been embedded in a pipeline that includes a genotyper specifically developed for RNA-seq data. Validation on empirical and simulated data indicates that our pipeline is particularly relevant to study SC of recent or intermediate age but can return useful information in old systems as well; it is available as a Galaxy workflow.

Introduction

Species with separate sexes (males and females) are common. They represent ~95% of animal species (Weeks 2012) and as much as 50-75% of the species in some land plant lineages (Ming et al. 2011). They are rarer in angiosperms; yet ~15,000 species with separate sexes (dioecious) have been found (Renner 2014). Many crops (e.g. papaya, strawberries, kiwi, spinach) are dioecious or derive from a dioecious progenitor (Ming et al. 2011). However, the mechanisms for sex determination remain unknown for most plant species and a number of animal species; in many cases it is not known whether sex chromosomes are present. Sex chromosomes of similar size and morphology (homomorphic) are particularly difficult to identify with cytology. Homomorphic sex chromosomes are probably frequent in groups such as angiosperms where many dioecious species have evolved recently from hermaphroditic ancestors and sex chromosomes are expected to be young and weakly diverged (Ming et al. 2011), in groups such as fish where sex determination mechanisms evolve quickly and the replacement of a pair of sex chromosomes by another (sex chromosome turn over) is high (Mank and Avise 2009), or in groups such as amphibians where occasional recombination limits sex chromosome divergence (Stock et al. 2013). Nevertheless, the exact frequency of homomorphic sex chromosomes remains unknown in those groups. In angiosperms, for example, dioecy has evolved probably >800 times independently (Renner 2014), but less than 40 sex chromosome pairs (including 20 homomorphic pairs) have been reported so far (Ming et al. 2011).

Obtaining sequences of sex chromosomes is also very difficult since they have non-recombining regions. Those regions can be very large and comprise most (or all) of the Y chromosome in the heteromorphic systems. The human Y chromosome, for instance, is largely non-recombining, i.e. does not make cross-over with the X during meiosis. Only two small regions of the human Y, called the pseudoautosomal regions can do so. Non-recombining regions are found in all sex chromosome types: the Y and W in diploid XY and ZW systems and also in both sex chromosomes in the UV haploid systems found in some mosses and algae (Bachtrog et al. 2011). The non-recombining regions of the genome are known to accumulate large amounts of repeats, including transposable elements (Charlesworth et al. 1994; Gaut et al. 2007). In some species, the X and Z chromosomes also accumulate repeats, although at a lesser extent than their Y/W counterparts, because selection against repeats is reduced on these chromosomes due to a smaller effective population size compared to the rest of the genome (Bellott et al. 2010; Gschwend et al. 2012).

This makes the sex chromosomes (particularly the non-recombining portions of the Y, W and U/V chromosomes) difficult to assemble, especially when using short-read sequencing technologies. Consequently, many genome projects have focused on individuals of the homogametic sex (XX or ZZ) to avoid this difficulty and also the problem of reduced coverage of the X and the Z when sequencing individuals of the heterogametic sex (discussed in Hughes and Rozen 2012). Y chromosomes have been sequenced using strategies relying on establishing a BAC map prior to sequencing. The single-haplotype iterative mapping and sequencing (SHIMS, described in Hughes and Rozen 2012) in particular has provided high-quality assemblies of the mammalian Y chromosome (Skaletsky et al. 2003; Hughes et al. 2010; Hughes et al. 2012; Bellott et al. 2014). These strategies are, however, labour-demanding and costly, which explains why only a handful of Y chromosomes have been fully sequenced to date (<15), many of which have small non-recombining regions: the livevort *Marchantia* (Yamato et al. 2007), the fish medaka (Kondo et al. 2006), the green alga *Volvox* (Ferris et al. 2010), the tree papaya (Wang et al. 2012) and the brown alga *Ectocarpus* (Ahmed et al. 2014).

Producing high-quality assembly is not always necessary and alternative, less expensive strategies have been recently developed for identifying sex chromosome sequences based on next-generation sequencing (NGS) data. A first category of approaches relies on the comparison of one male and one female genome. Identifying X-linked scaffolds can be done by studying the genomic male: female read coverage ratio along the genome: autosomal contigs will have a ratio of 1 while X-linked ones will have a ratio of 0.5 (Vicoso and Bachtrog 2011; Vicoso et al. 2013a; Vicoso et al. 2013b). The Y scaffolds are simply those that are exclusively present in the male genome. A more sophisticated analysis can be done by a prior exclusion of the repeats shared by the Y and the female genome (Carvalho and Clark 2013; Akagi et al. 2014). Also, a combination of RNA-seq and genome data of male and female individuals has been used to increase the number of known Y-linked genes in well-studied systems (Cortez et al. 2014). Similar analyses were done for ZW systems (Vicoso and Bachtrog 2011; Moghadam et al. 2012; Ayers et al. 2013; Vicoso et al. 2013a; Vicoso et al. 2013b). This approach, however, is suitable only if reasonably well-assembled reference genome is available, in the studied species or in a close relative.

A second category of approaches relies on studying how SNPs segregate among sexes. Full genome sequencing data of several male and female individuals can be used to genotype individuals of different sexes and study sex-linkage of single-nucleotide polymorphisms (SNPs) and scaffold from sex chromosomes can be ascertained (Al-Dous et al. 2011). If such genomic resources are lacking, as in many species with large genomes, complexity and size can be reduced by using transcriptomes instead of complete genomes. RNA-seq can be used instead of DNA-seq data to genotype individuals of different sexes and identify sex-linked SNPs and sex-linked genes. Sex chromosomes can thus be investigated in species with hitherto unknown genomes. One possibility is to sequence several male and female individuals of an inbred line and identify X/Y gene pairs by looking for SNPs showing Y-linkage (see Muyle et al. 2012; and current Supplementary Figure S1). On the other hand, sequencing the parents and a few offspring individuals of known sex from a specific cross allows the identification of sex-linked genes using both Y-linkage and X-linkage information (see Bergero and Charlesworth 2011; Chibalina and Filatov 2011; and current Figure 1.B). The approach based on RNA-seq derived genotypes has identified hundreds of new sex-linked genes in species where only a few were known before (Bergero and Charlesworth 2011; Chibalina and Filatov 2011; Muyle et al. 2012) or in species without previously known sex-linked genes or genomic resources available (Hough et al. 2014). Those are promising results, but genotyping and inference of sex-linkage were done without a statistical framework. The genotyping was performed (i) with tools that were developed for DNA-seq data, which do not account for uneven allele expression as may be the case for X/Y gene pairs or (ii) using reads number thresholds to distinguish true SNPs from sequencing errors, which were determined empirically for a given dataset, but may not be correct for another dataset as they depend on the sequencing coverage and the number of offspring that were used for sequencing. Arbitrary numbers of sex-linked SNPs per gene were used to classify genes as sex-linked or not. The lack of appropriate and statistically-grounded methods and pipelines clearly limits the application of the RNA-seq-based genotyping approach to more organisms despite its high potential.

Here we propose a model-based method (called SEX-DETector) for inferring sex-linkage from genotyping data from a full-sib family. Our model describes the genotypes of the parents and the F1 offspring for autosomal and sex-linked genes and accounts for genotyping errors. A likelihood-based approach is used to compute the posterior probabilities of being autosomal, X/Y and X-hemizygous (X-linked copy only) for each RNA-seq contig. The method is developed for any chromosome type (XY, ZW, UV), and the likelihood framework additionally offers the possibility to test for the presence and type of sex chromosomes in the data based on model selection. SEX-DETector is embedded in a pipeline including different steps from assembly to sex-linkage inference (Figure 1.A). Genotyping is done using a genotyper specifically designed for RNA-seq data (Tsagkogeorga et al. 2012; Gayral et al. 2013) that takes unequal allelic expression into account, which is relevant here as the Y copy of a X/Y gene pair tends to be less expressed than the X copy (reviewed in Bachtrog 2013).

We tested our pipeline on RNA-seq data of a family from *Silene latifolia*, a dioecious plant with relatively recent but heteromorphic sex chromosomes, for which some sex-linked and autosomal genes have been experimentally characterized. Our method could detect 83% of known sex-linked genes expressed in the tissue used to obtain our RNA-seq data (i.e. flower bud). To compare our pipeline to previous ones that rely on arbitrary thresholds, we computed sensitivity and specificity values on all known *S. latifolia* genes and our pipeline showed a much higher sensitivity (0.63 compared to 0.25-0.43) while specificity remained close to 1, which suggests that the number of sex-linked genes has been underestimated in previous work. Applying our pipeline to a comparable RNA-seq data from *Silene vulgaris*, a plant without sex chromosomes yielded no sex-linked genes as expected. We further tested the SEX-DETector method using simulations, which indicated good performance with a modest experimental effort and on different sex chromosome systems. The advantages, limits and potential solutions to those limitations of our method/pipeline and the approach based on RNA-seq derived genotyping data in general are discussed. Our method/pipeline is particularly relevant to study sex chromosomes of recent or intermediate age, for which the approaches relying on male vs. female genome comparisons are not adapted (as they require X and Y reads not to co-assemble). It can also provide useful information about old systems, which may complement that given by the approaches relying on male vs. female genome comparisons specifically devoted to study those systems.

For an easy use of the method, we developed a Galaxy workflow (including extra assembly, mapping, genotyping and sex-linked gene detection) that takes assembled contigs and the raw reads from any system (XY, ZW, UV) as input and returns a set of sex-linked gene sequences and allele-specific expression level estimates.

Material and Methods

Description of the probabilistic Model

The observed data consists of contigs containing offspring and parental genotypes (OG, PG respectively). From these genotyping data we want to infer the unknown segregation type (S) of each contig. We first suppose that S is a Multinomial random variable M(1, π_1, π_2, π_3) such that π_1, π_2, π_3 are the probabilities for one contig of being autosomal, X/Y (or Z/W), or X (or Z) hemizygous, respectively. Our strategy relies on the introduction of genotyping errors that may concern offspring as well as parent genotypes, either on the Y (YGE ~ B(p)) contigs, or on the other alleles (GE ~ B(ε)). We further introduce true homogametic and heterogametic parental genotypes (TMG for true mother genotype and TFG for true father genotype respectively), which are also unknown. Variable TMG is supposed to be Multinomial with parameters (α_1, …α_M), α_m standing for the probability for genotype m, and TFG is also multinomial, with parameters depending on the segregation type TFG|S_j ~ M(1, β_1..., β_{Nj}). Finally, when there is no genotyping error, if the segregation type and true parental genotypes were known, the conditional distribution of variables (OG, PG) is Multinomial, with parameters fully determined by segregation tables (see Supplementary Table S1). Parameters are estimated by maximum likelihood using an Expectation-Maximisation (EM) algorithm that includes a Stochastic step (SEM algorithm) to deal with initialization issues. The outputs of the method are maximum likelihood estimates of parameters π, p, ε, α, β, and the posterior probabilities for hidden variables (TMG, TFG, YGE, GE, and S) given the observed data (OG,PG). The maximum *a posteriori* rule is used to infer parental genotypes, genotyping errors, and most importantly segregation types. Every notation and computation details are provided in Text S1 and S2, see also Supplementary Figure S5.

To infer contig status, we defined what we call informative SNPs, which are autosomal or X/Y positions for which the heterogametic parent is heterozygous and different from the homogametic parent (otherwise it is not possible to differentiate between X/Y and autosomal segregation). Only informative SNPs are considered for computing a contig average segregation type, where SNPs are weighted by their posterior genotyping error probability (lower weight for contigs with higher posterior genotyping error probability). Contigs are assigned as sex-linked if they have at least one informative sex-linked (X/Y or X-hemizygous) SNP without genotyping error, and if the average sex-linked posterior probability is higher than the autosomal one and higher than a chosen threshold. Accordingly, a contig is inferred as autosomal if it has at least one autosomal SNP without genotyping error and if the autosomal posterior probability is higher than the sex-linked one and higher than the given threshold. A posterior segregation type probability threshold of 0.8 was chosen here. This parameter can be changed by the user. The code of SEX-DETector was written in Perl.

Data analysis

Plant material and sequencing: RNA-seq data were generated from a cross in the dioecious plant *S. latifolia*, which has sex chromosomes and from a cross in the gynodioecious plant *S. vulgaris*, which does not have sex chromosomes. We used the following RNAseq libraries that were used in previous studies: Leuk144-3_father, a male from a wild population; U10_37_mother, a female from a ten-generation inbred line (Muyle et al. 2012); and their progeny (C1_01_male, C1_3_male, C1_04_male, C1_05_male, C1_26_female, C1_27_female, C1_29_female, C1_34_female). For *S. vulgaris* the hermaphrodite father came from a wild population (Guarda_1), the female mother from another wild population (Seebach_2) and their hermaphrodite (V1_1, V1_2, V1_4) and female (V1_5, V1_8, V1_9) progeny.

Individuals were grown in a temperature-controlled greenhouse. The QiagenRNeasy Mini Plant extraction kit was used to extract total RNA two times separately from four flower buds at developmental stages B1–B2 after removing the calyx. Samples were treated additionally with QiagenDNase. RNA quality was assessed with an Aligent Bioanalyzer (RIN.9) and quantity with an Invitrogen Qubit. An intron-spanning PCR product was checked on an agarose gel to exclude the possibility of genomic DNA contamination. Then, the two extractions of the same individual were pooled. Individuals were tagged and then pooled for sequencing. Samples were sequenced by FASTERIS SA on an Illumina HiSeq2000 following an Illumina paired-end protocol (fragment lengths 150–250bp, 100 bp sequenced from each end).

A normalized 454 library was generated for *S. latifolia* using bud extracts from 4 different developmental stages.

Assembly: Adaptors, low quality and identical reads were removed. The transcriptome was then assembled using TRINITY (Haas et al. 2013) on the combined 10 individuals described previously as well as the 6 individuals from (Muyle et al. 2012) and the normalized 454 sequencing that was transformed to illumina using 454-to-illumina-transformed-reads. Then, isoforms were collapsed using /trinity-plugins/rsem-1.2.0/rsem-prepare-reference. PolyA tails, bacterial RNAs and ribosomal RNAs were removed using ribopicker. ORFs were predicted with trinity transcripts_to_best_scoring_ORFs.pl.

In order to increase the probability of X and Y sequences to be assembled in the same contig, ORFs were further assembled using CAP3 (cap3 -p 70, Version Date: 10/15/07, Huang and Madan 1999) inside of TRINITY components.

Mapping, genotyping and segregation inference: Illumina reads from the 10 individuals of the cross were mapped onto the assembly using BWA (version 0.6.2, bwa aln -n 5 and bwa sampe, Li and Durbin 2009). The libraries were then merged using SAMTOOLS (Version 0.1.18, Li et al. 2009). The obtained alignments were locally realigned using IndelRealigner (GATK, McKenna et al. 2010; DePristo et al. 2011) and were analysed using reads2snp (Version 3.0, -fis 0 -model M2 -output_genotype best -multi_alleles acc -min_coverage 3 -par false, Tsagkogeorga et al. 2012) in order to genotype individuals at each loci while allowing for biases in allele expression and not cleaning for paralogous SNPs as X/Y SNPs tend to be filtered out by paraclean (the program that removes paralogous positions, Gayral et al. 2013). SEX-DETector was then used to infer contigs segregation types after estimation of parameters using an EM algorithm. Posterior segregation types probabilities were filtered to be higher than 0.8. See pipeline in Figure 1.A.

The tester set in *S. latifolia*: For various tests, we used 209 genes with previously known segregation type : 129 experimentally known autosomal genes, 31 experimentally known sex-linked genes (X/Y or X-hemizygous) and 49 X CDS from BAC sequences (Supplementary Table S2).

The sequences of these 209 genes were blasted (blast -e 1E-5) onto the *de novo* assembly in order to find the corresponding ORF of each gene. Blasts were filtered for having a percentage of identity over 90% and an alignment length over 100bp and manually checked. Multiple RNA-seq contigs were accepted for a single gene if they matched different regions of the gene. If multiple contigs matched the same region of a gene, only the contig with the best identity percentage was kept. The gene was considered inferred as sex-linked if at least one of his matching contig was sex-linked. The inferred status of the genes by SEX-

DETector was then used to compute specificity and sensitivity values.

The same approach was used to compute sensitivity and specificity values for three previous studies that inferred *S. latifolia* RNA-seq contigs segregation patterns (Bergero and Charlesworth 2011; Chibalina and Filatov 2011; Muyle et al. 2012).

BIC test for the presence of sex chromosomes in *S. latifolia* and *S. vulgaris*: The ML framework of the method allows the use of statistical tests, for instance testing for the actual presence of sex-linked genes in the dataset. A model with the three possible segregation types can be compared to a model with only autosomal segregation type using BIC.

$$BIC(M) = -2 \log L + k \log n$$

Where BIC(M) is the BIC value of model M, L is the likelihood of the model, k is the number of free parameters and n is the sample size (number of polymorphic positions used to estimate parameters in the EM algorithm). The model with the lower BIC value is chosen. It is also possible to test for a X/Y versus a Z/W system by comparing both BIC values. In case a model with sex chromosomes fits best the data but no sex-linked genes are inferred, then it means there are no sex chromosomes in the dataset.

Simulations

Sequences were simulated for two parents (or a single parent in the case of a UV system) using ms to generate a coalescent tree (Hudson 2002, see Supplementary Figure S2) and then seq-gen to generate sequences using the ms tree and molecular evolution parameters (version 1.3.2x, seq-gen -mHKY -l contig_length -f 0.26 0.21 0.23 0.3 -t 2 -s theta, Rambaut and Grassly 1997). Different types of sequences were generated: either autosomal (ms 4 1 -T) or X/Y (ms 4 1 -T -I 2 3 1 -n 2 0.25 -n 1 0.75 -ej XY_divergence_time 2 1 -eN XY_divergence_time 1) or X-hemizygous (same parameters as X/Y but no Y sequence drawn) or U/V (ms 2 1 -T -I 2 1 1 -n 2 0.5 -n 1 0.5 -ej UV_divergence 2 1 -eN UV_divergence 1). Then, allele segregation was randomly carried on for a given number of progeny of each sex, using the segregation pattern determined when generating sequences with ms and seq-gen (see Supplementary Table S1 for segregation tables).

$\theta=4N_e\mu$ was set to 0.0275 as estimated in *S. latifolia* by (Qiu et al. 2010). μ was set to 10^{-7}, which implies that N_e was equal to ~70,000. Contig lengths were randomly attributed from the observed distribution of contigs lengths of the *S. latifolia* assembly presented previously. Equilibrium frequencies used for seq-gen were retrieved from SEX-DETector inferences on the observed *S. latifolia* data. The transition to transversion ratio was set to 2 as inferred by PAML (Yang 2007) on *S. latifolia* data (Käfer et al. 2013). The rate of genotyping error (ε) was set to 0.01 and the rate of Y genotyping error (p) was set to 0.13 as inferred by SEX-DETector on the observed *S. latifolia* data.

Five types of datasets were simulated, with ten repetitions for each set of parameters and 10,000 contigs were simulated for each dataset:

- Effect of X-Y divergence: Five different X-Y divergence times in units of $4N_e$ generations were tested, either *S. latifolia* X-Y divergence time (4.5My) or 10 times or 100 times older or younger. The proportion of X-hemizygous contigs among sex-linked contigs was set accordingly to X-Y divergence time: 0.002, 0.02, 0.2, 0.6 and 1 for respectively 45,000 years, 450,000 years, 4.5 My, 45 My and 450 My divergence time. As well as the proportion of Y genotyping error (since Y expression is known to decrease with X-Y divergence): 0, 0.01, 0.13, 0.2 and 1 respectively. Four offspring of each sex were simulated. The proportion of sex-linked contigs was set to 10%.

- Effect of the number of sex-linked contigs: Five different proportions of sex-linked contigs (X/Y pairs or X hemizygous) were tested: 30% (3000 sex-linked contigs out of 10,000), 5%, 1%, 0,1% and 0,01%. Four offspring of each sex were simulated and X-Y divergence was set to 4.5 My.

- Effect of theta: Three different $\theta=4N_e\mu$ (polymorphism) were tested: 0.000275, 0.00275 and 0.0275. Five offspring of each sex were simulated and X-Y divergence was set to 4.5 My, the X-Y divergence time in unit of $4N_e$ generations varied accordingly to the value of theta. The proportion of sex-linked contigs was set to 10%.

- Effect of the number of individuals in Z/W and X/Y systems: Nine different numbers of offspring individuals of each sex were tested for the X/Y system: 2, 3, 4, 5, 6, 7, 8, 12 or 16 individuals of each sex. Sex chromosome size was set to 10% and X-Y/Z-W divergence to 4.5 My.

- Effect of the number of individuals in U/V systems: Eight different numbers of offspring individuals of each sex were tested for the U/V system: 1, 2, 3, 4, 5, 6, 7 or 8 individuals of each sex. Sex chromosome size was set to 10% and U-V divergence to 4.5 My.

For each simulated dataset, segregation types were inferred using SEX-DETector and were compared to the true segregation types in order to compute sensitivity and specificity values.

Galaxy workflow

A Galaxy workflow has been developed (see user guide and source codes at http://lbbe.univ-lyon1.fr/-SEX-DETector-.html).

Empirical method without a cross

The method for inbred brothers and sisters (or males and females sampled from the same population) is not model-based and relies on empirical filtering of SNPs: individuals are first genotyped using read counts, an allele is retained if it represents over 2% of the total read count at the position, a position is considered if it has more than three reads (the thresholds can be changed). Another possibility is to genotype individuals using reads2snp. Then, SNPs are filtered to retrieve cases where males are all heterozygous and females all homozygous in the case of an XY system (as in Muyle et al. 2012; and see current Supplementary Figure S1). A contig is considered sex-linked if it shows at least one such SNP.

Data access

The RNA-seq data have been submitted to the GEO database under the series XXXX.

Results

A probabilistic method for inferring sex-linkage from family genotyping data

SEX-DETector is based on the genotypes of two parents and their progeny from which we infer the segregation type of each contig. The model considers that SNPs can be transmitted to the progeny by three segregation modes: (i) autosomal, (ii) sex-linked with both X and Y (or Z and W) alleles present and (iii) X (or Z) hemizygous, i.e. sex-linked with only the X (or Z) allele present (the Y or W allele being inactivated, lost, too weakly expressed or in a different contig due to X/Y or Z/W divergence, see Discussion). We also considered the case of UV sex chromosomes, similar to the one for XY and ZW, without hemizygous segregation and with only one parent (the sporophyte). Our method relies on the segregation of SNPs (see Supplementary Table S1) that are fully described in the progeny when both parent genotypes and segregation type are known (see Figure 1.B for an example). We use a likelihood-based framework to assess the *posterior* probability of each segregation type for each informative SNP given the observed genotype data.

A strong advantage of our model-based approach compared to empirical methods is the amount of information captured from the data thanks to a hierarchical probabilistic model. Observed genotypes of parents and progeny were incorporated in a model using genotype probabilities (see Material and Methods). We accounted for discrepancies that may exist between observed and true genotypes (because of genotyping errors) by introducing two genotyping error parameters: one for any type of genotyping error, and one specific to the Y (or W) as genotyping errors are more frequent on the Y (or W) allele due to reduced expression and low RNA-seq read coverage. The UV model does not contain any specific genotyping error for the non-recombining sex chromosome as both U and V are non-recombining. Our model accounts for genotyping errors that are likely to be present on parental genotypes as well. These steps (i.e. estimating genotype probabilities and genotyping errors) are essential as each true parental genotype has a different probability to occur in the dataset due to the level of heterozygosity and the base composition of a given species, and they will ensure that the method can apply to different species. Then, for each SNP and individual, we compute the *posterior* probabilities of genotyping errors, which allows us to compute the *posterior* probabilities of observing the true parental genotypes and then the segregation types *posterior* probabilities for each SNP. The segregation type of each contig is finally inferred by averaging informative

SNP posteriors. All X-linked SNPs are informative whereas informative X/Y SNPs are positions for which the heterogametic parent is heterozygous and different from the homogametic parent (otherwise it is not possible to distinguish X/Y and autosomal segregation). Each SNP posterior is weighted by its posterior probability of genotyping error (so that SNPs with higher genotyping error posteriors have less effect on the final inference about a contig segregation type). The contig inferred segregation type corresponds to the one with the highest *posterior* (which corresponds to the maximum *a posteriori* rule).

A pipeline for analysing RNA-seq data from a family

An important step in the pipeline is read assembly. To be able to detect Y-linked SNPs as well as XY gene pairs, the reads from X and Y transcripts must co-assemble into a single RNA-seq contig. This is achieved using Trinity and Cap3 for further assembly: Trinity will produce groups of contigs including alternative transcripts and alleles; joining the alleles into one contig is done with Cap3 (see Material and Methods). Note that the detection of the X-linked SNPs will not depend on the efficiency of the X and Y read co-assembly, and it will still be possible to identify sex-linked genes in case of low or no X and Y read co-assembly (see Discussion). Most genotypers have been developed for analysing genomic data, not transcriptomic data. A major difference between these two types of data is that coverage can significantly differ among alleles in transcriptomic data because of differences in expression level among alleles. For X/Y gene pairs, such differences are frequent, with the Y copy being less expressed than the X one (reviewed in Bachtrog 2013). Genotypers for genomic data will typically consider the less expressed alleles as sequencing errors as we have experienced and corrected manually in previous work (Muyle et al. 2012). To solve this problem, we used a genotyper specifically developed for RNA-seq data, called reads2snp, which allows differences in expression level among alleles (Tsagkogeorga et al. 2012; Gayral et al. 2013). Our genotype inferences were different compared to standard genotypers when X and Y copies had different expression levels (data not shown). We developed Galaxy wrappers for SEX-DETector and used available wrappers for other tools (including reads2snp) to prepare a Galaxy workflow.

Testing our pipeline's performance using a *Silene latifolia* dataset

The SEX-DETector pipeline (Figure 1.A) was run on a *Silene latifolia* dataset. *S. latifoli*a is a dioecious plant species with well-studied XY chromosomes that had several interesting characteristics for benchmarking our method/pipeline: (1) *S. latifolia* genome and sex chromosomes are quite large (the genome is 3Gb, the X is 400 Mb and the Y is 550 Mb); (2) no reference genome is available in this species; (3) *S. latifolia* sex chromosomes are relatively recent (~5 MY old; Rautenberg et al. 2010) but clearly heteromorphic; X-Y synonymous divergence ranges from 5 to 25% (Bergero et al. 2007); *S. latifolia* thus represents a system of intermediate age; (4) a tester set of 209 genes for which segregation type has been established is available in this species (Supplementary Table S2). The dataset consists of a cross (two parents and four offspring of each sex). RNA-seq data was obtained for each of these individuals tagged separately and the reads were assembled using Trinity and then Cap3, the final assembly included 46,178 ORFs (Table 1). RNA-seq reads were mapped onto this assembly (see Supplementary Table S3 for library sizes and mapping statistics) and genotyping was done for each individual using reads2snp. SEX-DETector was run on the genotyping data to infer autosomal and sex-linked genes (Table 1). For further analysis, only contigs having at least one SNP without genotyping error and showing a *posterior* probability ≥ 0.8 (of being autosomal or sex-linked) were retained. Figure 2.A-D shows examples from the tester set. For some genes, all SNPs show clearly the same correct segregation type (Figure 2.A-C), whereas in some genes mixed segregation patterns were inferred, which we attribute to co-assembly of recent paralogs or other assembly/mapping problems (see Figure 2.D and Discussion).

We used our tester set to test the performance of our pipeline, i.e. estimate its sensitivity (the capacity to detect true sex-linked genes) and specificity (the capacity not to assign autosomal genes as sex-linked). 83% of the known sex-linked genes expressed in the RNA-seq data used here (i.e. flower bud) were detected, indicating a high sensitivity. We obtained a specificity of 99% for this dataset as one gene, OxRZn, was supposedly wrongly assigned as a sex-linked gene by SEX-DETector. However, this gene was earlier assessed as autosomal on the basis of the absence of male specific alleles (Marais et al. 2011) and SEX-DETector assigned it to a sex-linked category because of two clear X-hemizygous SNPs, without genotyping error. It is therefore likely that OxRZn is in fact a true positive and more research on that gene is required.

Comparing our pipeline to others using a *S. latifolia* dataset

We compared the performance of our pipeline to those used in previous work on inferring sex-linkage with RNA-seq data in *S. latifolia* (Bergero and Charlesworth 2011; Chibalina and Filatov 2011; Muyle et al. 2012). Those pipelines differ in many ways, and the data themselves can be different. In previous work, offspring individuals of the same sex were sometimes pooled before sequencing (Bergero and Charlesworth 2011; Chibalina and Filatov 2011). We used again the tester set of 209 *S. latifolia* genes with known segregation types, which we blasted onto each dataset to find the corresponding contigs and their inferred segregation type (for details see Supplementary Table S2). Because the different pipelines require different types of data (pooled progeny versus individually-tagged offspring) and with different read coverages, we computed sensitivity on all known genes (expressed or not). Our pipeline outperformed other pipelines in terms of sensitivity (Table 2), while specificity was comparable (Table 2), which indicates that the number of sex linked genes in *S. latifolia* has previously been under-estimated.

As further analysis showed, this under-estimation was due to overly conservative filtering in previous work. To exclude false positives, genes with at least 5 sex-linked SNPs were retained in previous studies. More filtering was done by excluding contigs with autosomal SNPs (Bergero and Charlesworth 2011; Hough et al. 2014). As shown in Figure 3, keeping only contigs with at least 5 sex-linked SNPs removes nearly half of the contigs inferred as sex-linked by SEX-DETector, many of which have a high posterior probability. Excluding further those with autosomal SNPs (keeping those with sex-linked SNPs only) removes 74% of the contigs (Figure 3.B). Comparatively, SEX-DETector removes 12% of contigs when filtering for a posterior probability higher than 0.8 (Table 1), as most genes have a very high *posterior* segregation type probability which indicates a strong signal in the data and illustrates the benefits of using a model-based approach.

Simulations show that SEX-DETector requires a modest experimental effort and works on different sex chromosome systems

We simulated genotypes for a cross (parents and progeny) by generating a coalescent tree with autosomal or sex-linked history (Supplementary Figure S2) and generated the parental sequences using that tree and molecular evolution parameters. Progeny genotypes were obtained by random segregation of alleles from the parents and a genotyping error layer was added (see Material and Methods). 10,000 contigs were simulated for each dataset.

In order to know how many offspring of each sex should be sequenced to achieve the best sensitivity and specificity trade-off using SEX-DETector, we varied the number of progeny individuals in the simulations. For an X/Y or Z/W system, optimal results were obtained when sequencing five progeny individuals of each sex (Figure 4.A); sequencing more progeny individuals did not improve the results further. This suggests that sequencing 12 individuals (two parents and five progeny individuals of each sex) may be sufficient to achieve optimal performances with SEX-DETector on an X/Y or Z/W system. For a U/V system, two progeny individuals of each sex seems sufficient to obtain optimal SEX-DETector performance (Figure 4.B), which suggests that sequencing five individuals (the sporophyte parent and two progeny of each sex) may be enough in the case of a U/V system. Our simulations thus suggest that SEX-DETector requires a modest experimental effort to reliably identify expressed sex-linked genes.

In order to assess the applicability of SEX-DETector to different types of sex chromosomes (old versus young, homomorphic versus heteromorphic) and species (highly versus weakly polymorphic), we used the same simulation procedure and tested the effect of one parameter at a time on SEX-DETector sensitivity and specificity. In our simulations, the degree of polymorphism within species had no influence on the performance of our method (Supplementary Figure S3.A). As for the influence of the size of the non-recombining region (homomorphic or heteromorphic sex chromosomes), it was tested using different % of sex-linked genes in a genome with no effect on the performance of SEX-DETector (Supplementary Figure S3.B). The limit of detection of a sex-linked contig was reached only when 1 sex-linked contig out of 10,000 contigs was present. Finally, the simulations indicated that our method is robust to the X-Y divergence, as young and old sex chromosomes were evenly detected (Supplementary Figure S3.C).

SEX-DETector identifies unknown sex chromosomes using model selection

It is common that species with separated sexes have unknown sex determination system, i.e. it is unknown whether they have sex chromosomes and if they have, of what type (Z/W versus X/Y). The likelihood-based framework of SEX-DETector allows us to test for these assumptions by comparing models' fit to the data using the Bayesian Information Criterion (BIC, see Material and Methods). In species for which sex determination is unknown, it is possible to compare models with and without sex chromosomes, and then if sex chromosomes are detected, it is possible to compare models with X/Y or Z/W system. This was tested on real data and simulated data.

In the *S. latifolia* dataset, the best model inferred by SEX-DETector was a model with sex chromosomes as expected, with 1357 sex-linked contigs (which represents 9% of the contigs with a posterior probability higher than 0.8). In the *Silene vulgaris* dataset (a species without sex chromosomes), no sex-linked contigs were inferred, the best model fit to the data was thus a model without sex chromosomes as expected (see Material and Methods).

In order to know from which proportion of sex-linked genes sex chromosomes can be detected, we compared models on simulated data with varying numbers of sex-linked contigs out of 10,000 simulated contigs (Table 3 and Supplementary Table S4). When no sex-linked contigs were simulated, as expected the best model was the one without sex chromosomes. This was also the case when a single sex-linked contig was simulated. In this case, SEX-DETector could not detect it due to lack of information in the dataset. When ten or more sex-linked contigs were simulated, the best model was the one with sex chromosomes as expected. Thus, ten sex-linked contigs out of 10,000 provide sufficient information for SEX-DETector (i.e. 1 sex-linked gene out of 1000 genes can be detected).

Once the presence of sex chromosomes has been inferred, it can be tested whether the system is X/Y or Z/W. The model comparison between X/Y and Z/W systems worked on both real and simulated data: the best model for *S. latifolia* was, as expected, the X/Y system (Table 3 and Supplementary Table S4).

Discussion

Strengths and limits of the SEX-DETector pipeline

Our pipeline offers a number of interesting features. Both real *S. latifolia* data and simulations suggest that very few individuals need to be sequenced: twelve individuals (two parents plus five male and five female offprings) in the case of a X/Y or Z/W system, and five individuals (the sporophytic parent plus two male and two female gametophytic offprings) for a U/V system appears to be enough to get very good performance from our pipeline. This makes the strategy very accessible given the cost of RNA-seq. Our pipeline is user-friendly thanks to a Galaxy workflow that goes from further assembly of contigs using Cap3 to sex-linked genes and allelic expression levels (Figure 1.A).

New sex chromosomes can thus be characterised in species that have separated sexes and for which sex-determination has remained unknown. Indeed, the method allows one to test for the presence of sex chromosomes in the data, and then test for an X/Y versus a Z/W system, using the BIC.

Sensitivity results (Table 2) showed that the SEX-DETector pipeline is more powerful for detecting sex-linked genes compared to previous pipelines relying on empirical methods (Bergero and Charlesworth 2011; Chibalina and Filatov 2011; Muyle et al. 2012). The use of a cross allows identification of both X/X and X/Y SNPs unlike in (Muyle et al. 2012). Individually tagged progeny individuals give more information than the pools (that were used in Bergero and Charlesworth 2011; Chibalina and Filatov 2011). The reads2snp genotyper suited for RNA-seq data with allelic expression biases prevents the weakly expressed Y alleles to be wrongly classified as sequencing errors. Further, the probabilistic framework allows to filter contigs on posterior segregation type probability rather than the number of sex-linked SNPs as done in other studies (Bergero and Charlesworth 2011; Hough et al. 2014), allowing to preserve more true sex-linked genes without increasing false positive inferences (Table 2, Figure 3).

SEX-DETector infers both sex-linked and autosomal genes. In earlier approaches, contigs were inferred as either sex-linked or not sex-linked, and the latter consisted of a mix of autosomal and undetected sex-linked genes. In SEX-DETector, contigs for which no segregation type inference was possible, for example because of a lack of informative SNPs, are classified as undetermined and are not merged with autosomal genes. Having reliable inferences on autosomal genes is highly useful for comparative studies between autosomal and sex-linked genes.

RNA-seq derived genotyping data approaches, including ours, have limitations. The use of RNA-seq suffers from the limitation that some genes may not be expressed, or may be too weakly expressed in the data for these approaches to infer segregation types, which implies that using approaches based on RNA-seq genotyping data necessarily underestimates the number of truly sex-linked genes. Also, no expression of a Y copy in the studied tissue of a X/Y gene pair will result in the X copy being ascertained as X-hemizygous. Extracting RNA from tissues with complex transcriptomes (many expressed genes) will attenuate these problems, for example flower buds in plants or reproductive organs (e.g. testis) in animals. Using not a single but several tissues/organs/development stages may also help solving these problems, but will increase the cost. Pooling the tissues for each individual before sequencing may be a way to avoid such extra-cost.

Our pipeline includes a *de novo* assembly, which may lead to co-assembly of very recent paralogs into chimeric contigs. This problem is common to all available approaches to obtain sex chromosome sequences except the SHIMS approach (Hughes and Rozen, 2012). However, and most importantly, our method will be able to identify the genes where such problems might have occurred when paralogs are on different chromosomes (as these genes will show a mixture of sex-linked and autosomal SNPs). These genes can be excluded from further analysis by the user. Paralogs from the same chromosome will be more difficult to be handled by our method (and nearly all others, except the SHIMS).

Our simulations suggest that SEX-DETector should work on different systems; its performance was indeed excellent in a wide range of situations even when introducing genotyping errors (Supplementary Figure S3). The simulations were manned to assess the SEX-DETector method performance and not that of the whole pipeline. We therefore directly simulated genotypes and did not include all the possible errors occurring upstream the genotyping steps (assembly and mapping). In particular, the failure to co-assemble X and Y reads and assembly errors were not simulated, which may have implications on the applicability of the pipeline to old systems.

Highly divergent X/Y genes are expected to assemble into separate X and Y contigs (Muyle et al. 2012). The X-Y divergence threshold beyond which co-assembly will not be possible is not known. *S. latifolia* does not have a particularly low X-Y synonymous divergence since it ranges from 5% to 25%

(Bergero et al. 2007). By comparison, most X/Y gene pairs in humans exhibit a divergence lower than 30% (mean X-Y synonymous divergence for strata 3, 4 and 5 is respectively 30%, 10% and 5%, Skaletsky et al. 2003). When running SEX-DETector on our tester set, none of the known X/Y gene pairs was inferred as X-hemizygous, whereas several are the oldest *S. latifolia* stratum (with X-Y synonymous divergence of 20-25%, Bergero et al. 2007). The good performance of our methods on *S. latifolia* suggests that the failure to co-assemble X and Y reads will not be an issue for systems with moderate X-Y divergence. A recent study using the RNA-seq-based segregation approach on *Rumex hastatulus*, an older system (~15 MY), returned hundreds of sex-linked genes and suggests this approach works on even more divergent systems than *S. latifolia* (Hough et al. 2014).

Moreover, as already explained, in the case of failure to co-assemble X and Y reads, the X contigs will still be identified (through X-linked SNPs) and the SEX-DETector analysis will return X-hemizygous genes. The Y contig will not be identified as SEX-DETector does not detect Y contigs alone (they cannot be distinguished from autosomal genes that are exclusively expressed in males, i.e. that have male-limited expression). To identify the Y contigs that have been missed, other strategies need to be investigated, for example testing whether the X-hemizygous genes match to male-specific contigs, which may represent the divergent Y contigs. Note that when this was done for our 332 inferred X-hemizygous genes, only 5 of them had a significant match with a male-specific contigs, suggesting that our set of inferred X-hemizygous genes includes only a few wrongly inferred X/Y gene pairs with a divergent Y expressed in flower bud.

Our simulations also suggested that SEX-DETector might work similarly on homomorphic (few sex-linked genes) and heteromorphic (many sex-linked genes) systems. In homomorphic systems, the sex chromosomes typically include one or two large pseudoautosomal regions (PARs). The genes close to the pseudoautosomal boundary may exhibit partial linkage. It is likely that 10 offsprings or so will not be enough to tell apart the partially sex-linked from the fully sex-linked genes (Supplementary Figure S4). For some analyses, it will not be a limitation, it can even be useful to have as many genes as possible from the sex chromosomes including those in the PARs. But for others, it may be important to distinguish genes in the sex-specific regions form genes in the PARs. In this case, the number of sequenced offsprings should be increased as it is expected that partial sex-linkage should vanish when analysing many individuals. To avoid increasing too much the experimental costs, bulk sequencing of the offspring could be a solution as SEX-

DETector offers the possibility of analysing pooled data.

A difficulty for the identification of X-linked SNPs may be the presence of X chromosome inactivation, or any instance of dominance effects in which only one X-linked transcript is present in RNA-seq data. X chromosome inactivation is the inactivation of one of the two X chromosomes in females. If in the studied tissue, the same X chromosome is consistently inactivated, heterozygous mothers will appear homozygous while different alleles are found in their sons, and this will make the detection of sex-linked SNPs using X-linkage information more difficult. The method will not give erroneous results but basically loose power: X-hemizygous genes may not be detected and even X/Y gene pairs may be less easy to detect (with only Y-linkage information). The consequences should not be too serious for recent / intermediate systems. In old systems where the method will identify mostly the X-linked genes, what can we expect? At present X chromosome inactivation is only known in mammals and it is not known whether it exists in other taxa. In the case of random X chromosome inactivation (as in placentals) tissues will include a mixture of cells with one X or the other inactivated, both transcripts will be present in the RNA-seq data and the problem will vanish. In marsupials, the paternal X is inactivated but X-inactivation is incomplete (Al-Nadaf et al. 2010), not all of the genes are fully inactivated, which leaves some room to find some X-linked genes, even in this extreme case.

Importantly, SEX-DETector will also work on genotyping data derived from genome sequencing. In some cases, it might be more efficient to use DNA-seq data instead of RNA-seq data to genotype individuals. This alternative approach will however scale with genome size, and may be costly for species with large genome. For example, if (1) the sex-linked genes have very different expression patterns so that RNA-seq of many tissues would be required to identify many sex-linked genes, (2) the sex-specific region of the sex chromosomes is expected to be very small and all genes might not be expressed in all tissues so that the number of expressed sex-linked genes might be outside the range of SEX-DETector power (<10 genes), (3) X chromosomes inactivation (especially non-random) is suspected, one could collect DNA instead of RNA sequences from a family for solving these potential problems.

What strategy is best for detecting sex-linked genes?

Strategies relying on sequencing male and a female genome/transcriptome have successfully been used in several organisms (Vicoso and Bachtrog 2011; Carvalho and Clark 2013; Vicoso et al. 2013a; Vicoso et al. 2013b; Moghadam et al. 2012; Ayers et al. 2013; Cortez et al. 2014). They are expected to work better for small genomes and require the sex chromosomes to be divergent enough. In species with very large and complex genomes with many repeats, the assembly of NGS genomic data will be challenging (if at all possible), which may harden the male/female genome comparison and may result in obtaining highly fragmented and incomplete sex-linked gene catalogues. In young and weakly diverged sex chromosome systems, X and Y (or Z and W, or V and U) sequences will assemble together, which will prevent their identification by these strategies.

Strategies relying on getting RNA-seq data to perform a segregation analysis are expected to be insensitive to genome size. This has some practical aspects as the sequencing costs will not scale with genome size. The approach will thus remain affordable also for species with large genomes. It concentrates on the transcribed part of the sex chromosomes and will directly give the sequences of many sex-linked genes, the primary material for studies of sex chromosomes. It also provides expression data that can be used to address various questions about the evolution of gene expression on sex chromosomes. As discussed in the previous section, these approaches are ideal for weakly to moderately diverged X/Y gene pairs whose reads will co-assemble. In more diverged systems, X and Y copy co-assembly may be more difficult, and these approaches will mainly return X-hemizygous genes (even if a Y copy exists). Both types of strategies are thus complementary and may be used hand-in-hand in some systems (e.g. sex chromosome systems showing different levels of divergence).

Conclusions - Perspectives

Our SEX-DETector method/pipeline requires family data and will work optimally on young / intermediate systems (although returning useful information for old systems) as discussed above. Family data can be obtained in many organisms (e.g. all those that have genetic map), typically all the organisms that can be grown in the lab. Such data are also available in many agronomically important animals/crops. As mentioned in the introduction, it is likely that many sex chromosome systems that remain to be characterized are homomorphic. The applicability of SEX-DETector is thus broad.

SEX-DETector along with the other methods recently developed for obtaining sex-linked sequences using NGS at low cost will render feasible the large-scale comparative analysis of different sex chromosome systems. In particular, the comparison of systems from closely related species becomes possible, which will give much more information than previously performed comparisons of phylogenetically highly distantly related systems.

To study sex chromosomes, in some organisms for which family cannot be obtained easily, a method to infer sex-linkage from population data (males and females sampled from a population) would be very useful. Such a method has recently been proposed (Gautier 2014) but it has mainly been developed to sort markers (autosomal, X, Y, organelles) before performing population genetics analysis on NGS genomic data, and relies on ploidy levels to detect sex-linked contigs. This method is thus designed for old X/Y or Z/W sex chromosomes only. Our pipeline currently includes the empirical method used in (Muyle et al. 2012) for population data (Supplementary Figure S1), but the extension of a SEX-DETector version for population data (based on a population genetics model) is currently under development.

Finally, our pipeline could also be used, pending such adjustments, 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.

Reviewer links to deposited data

During the review process, the SEX-DETector galaxy workflow and associated test datasets are available on the public galaxy.prabi.fr server. The data as well as the tool interface are visible to anonymous users, but to use them, you should register for an account ("user → Register"), and import the data library "SEX-DETector" ("Shared Data → Data Libraries") into your history. More instructions can be found in the "readme" file in this dataset. The user manual for SEX-DETector is available here: https://lbbe.univ-lyon1.fr/Download-5251.html

Acknowledgements

We thank Alex Widmer for access to RNA-seq datasets and comments on the manuscript, Nicolas Galtier and Sylvain Glémin (ISEM-Montpellier) for useful discussions about reads2snp, Vincent Miele (LBBE) for SEX-DETector profiling and advice on code performance, Khalid Belkhir (ISEM-montpellier) for providing and adapting Galaxy wrappers for analyses used upstream of SEX-DETector (BWA, reads2snp) and Philippe Veber (LBBE) for help with Galaxy. This work was financially supported by Agence Nationale de la Recherche grants to GABM (grants numbers: ANR-11-BSV7-013, ANR-11-BSV7-024; ANR-14-CE19-0021) and SNF project to Alex Widmer (SNF 31003A_141260).

References

Ahmed, S., Cock, J.M., Pessia, E., Luthringer, R., Cormier, A., Robuchon, M., Sterck, L., Peters, A.F., Dittami, S.M., Corre, E., et al. (2014). A haploid system of sex determination in the brown alga *Ectocarpus* sp. *Curr Biol* 24(17): 1945-1957.

Akagi T, Henry IM, Tao R, Comai L. 2014. Plant genetics. A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. *Science* 346(6209): 646-650.

Al Nadaf S, Waters PD, Koina E, Deakin JE, Jordan KS, Graves JA. 2010. Activity map of the tammar X chromosome shows that marsupial X inactivation is incomplete and escape is stochastic. *Genome Biol* 11(12): R122.

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J et al. 2011. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat Biotechnol* 29(6): 521-527.

Ayers KL, Davidson NM, Demiyah D, Roeszler KN, Grützner F, Sinclair AH, Oshlack A, Smith CA. 2013. RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken and allows comprehensive annotation of the W-chromosome. *Genome Biol* 14(3): R26.

Bachtrog D, Kirkpatrick M, Mank JE, McDaniel SF, Pires JC, Rice W, Valenzuela N. 2011. Are all sex chromosomes created equal? *Trends Genet* 27(9): 350-357.

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* 14(2): 113-124.

Bellott D, Skaletsky H, Pyntikova T, Mardis E, Graves T, Kremitzki C, Brown L, Rozen S, Warren WC, Wilson RK et al. 2010. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* 466: 612-616.

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, Koutseva N, Zaghlul S, Graves T, Rock S et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* 508(7497): 494-499.

Bergero R, Charlesworth D. 2011. Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. *Curr Biol* 21(17): 1470-1474.

Bergero R, Forrest A, Kamau E, Charlesworth D. 2007. Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: evidence from new sex-linked genes. *Genetics* 175(4): 1945-1954.

Carvalho AB, Clark AG. 2013. Efficient identification of Y chromosome sequences in the human and Drosophila genomes. *Genome Res* 23(11): 1894-1907.

Charlesworth B, Sniegowski P, Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371: 215-220.

Chibalina MV, Filatov DA. 2011. Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr Biol* 21(17): 1475-1479.

Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* 508(7497): 488-493.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5): 491-498.

Ferris P, Olson B, Hoff PD, Douglass S, Casero D, Prochnik S, Geng S, Rai R, Grimwood J, J S et al. 2010. Evolution of an expanded sex-determining locus in Volvox. *Science* 328: 351-354.

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat Rev Genet* 8(1): 77-84.

Gautier M. 2014. Using genotyping data to assign markers to their chromosome type and to infer the sex of individuals: a Bayesian model-based classifier. *Mol Ecol Resour*.

Gayral P, Melo-Ferreira J, Glemin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N et al. 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genet* 9(4): e1003457.

Gschwend AR, Yu Q, Tong EJ, Zeng F, Han J, VanBuren R, Aryal R, Charlesworth D, Moore PH, Paterson AH et al. 2012. Rapid divergence and expansion of the X chromosome in papaya. *Proc Natl Acad Sci U S A* 109(34): 13716-13721.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8(8): 1494-1512.

Hough J, Hollister JD, Wang W, Barrett SC, Wright SI. 2014. Genetic degeneration of old and young Y chromosomes in the flowering plant Rumex hastatulus. *Proc Natl Acad Sci U S A*.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome Res* 9(9): 868-877.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2): 337-338.

Hughes JF, Rozen S. 2012. Genomics and genetics of human and primate y chromosomes. *Annu Rev Genomics Hum Genet* 13: 83-108.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* 483(7387): 82-86.

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, Daalen SKMv, Minx PJ, Fulton RS, McGrath SD, Locke DP, C.Friedman et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* 463(13 January)): 536-953.

Käfer J, Talianova M, Bigot T, Michu E, Gueguen L, Widmer A, Zluvova J, Glemin S, Marais GA. 2013. Patterns of molecular evolution in dioecious and non-dioecious Silene. *J Evol Biol* 26(2): 335-346.

Kondo M, Hornung U, Nanda I, Imai S, Sasaki T, Shimizu A, Asakawa S, Hori H, Schmid M, Shimizu N et al. 2006. Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. *Genome Research* 16: 815-826.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16): 2078-2079.

Mank JE, Avise JC. 2009. Evolutionary diversity and turn-over of sex determination in teleost fishes. *Sex Dev* 3(2-3): 60-67.

Marais GA, Forrest A, Kamau E, Käfer J, Daubin V, Charlesworth D. 2011. Multiple nuclear gene phylogenetic analysis of the evolution of dioecy and sex chromosomes in the genus Silene. *PLos One* 6(8): e21915.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshule D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20(9): 1297-1303.

Ming R, Bendahmane A, Renner S. 2011. Sex chromosomes in land plants. *Annual Review of Plant Biology* 62: 485-514.

Moghadam HK, Pointer MA, Wright AE, Berlin S, Mank JE. 2012. W chromosome expression responds to female-specific selection. *Proc Natl Acad Sci U S A* 109(21): 8207-8211.

Muyle A, Zemp N, Deschamps C, Mousset S, Widmer A, Marais G. 2012. Rapid *De Novo* Evolution of X Chromosome Dosage Compensation in *Silene latifolia*, a Plant with Young Sex Chromosomes. *PloS Biol* 10(4): e1001308.

Qiu S, Bergero R, Forrest A, Kaiser VB, Charlesworth D. 2010. Nucleotide diversity in *Silene latifolia* autosomal and sex-linked genes. *Proc Biol Sci* 277(1698): 3283-3290.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* 13(3): 235-238.

Rautenberg A, Hathaway L, Oxelman B, Prentice HC. 2010. Geographic and phylogenetic patterns in *Silene* section *Melandrium* (Caryophyllaceae) as inferred from chloroplast and nuclear DNA sequences. *Mol Phylogenet Evol* 57(3): 978-991.

Renner SS. 2014. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *Am J Bot* 101(10): 1588-1596.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423(19 June): 825 - 837.

Stock M, Savary R, Betto-Colliard C, Biollay S, Jourdan-Pineau H, Perrin N. 2013. Low rates of X-Y recombination, not turnovers, account for homomorphic sex chromosomes in several diploid species of Palearctic green toads (*Bufo viridis* subgroup). *J Evol Biol* 26(3): 674-682.

Tsagkogeorga G, Cahais V, Galtier N. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. *Genome Biol Evol* 4(8): 740-749.

Vicoso B, Bachtrog D. 2011. Lack of global dosage compensation in Schistosoma mansoni, a female-heterogametic parasite. *Genome Biol Evol* 3: 230-235.

Vicoso B, Emerson JJ, Zektser Y, Mahajan S, Bachtrog D. 2013a. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biol* 11(8): e1001643.

Vicoso B, Kaiser VB, Bachtrog D. 2013b. Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proc Natl Acad Sci U S A* 110(16): 6453-6458.

Wang J, Na JK, Yu Q, Gschwend AR, Han J, Zeng F, Aryal R, VanBuren R, Murray JE, Zhang W et al. 2012. Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc Natl Acad Sci U S A* 109(34): 13710-13715.

Weeks SC. 2012. The role of androdioecy and gynodioecy in mediating evolutionary transitions between dioecy and hermaphroditism in the animalia. *Evolution* 66(12): 3670-3686.

Yamato KT, Ishizaki K, Fujisawa M, Okada S, Nakayama S, Fujishita M, Bando H, Yodoya K, Hayashi K, Bando T et al. 2007. Gene organization of the liverwort Y chromosome reveals distinct sex chromosome evolution in a haploid system. *Proceedings of the National Academy of Sciences of the USA* 104: 6472-6477.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24(8): 1586-1591.

Tables

Table 1: Results of our pipeline on the *S. latifolia* dataset.

ORF Types	Numbers
ORFs in final assembly	46178
ORFs with enough coverage to be studied	43901
ORFs with enough informative SNPs to compute a segregation probability	17189
ORFs with posterior segregation probability over 0.8	15164
ORFs assigned to an autosomal segregation type	13807 (91 %)
ORFs assigned to a X/Y segregation type	1025 (7 %)
ORFs assigned to a X hemizygous segregation type	332 (2 %)

Table 2: comparison of sensitivity and specificity values obtained with different methods using 209 known *S. latifolia* genes.

	Sensitivity	Specificity
SEX-DETector	0.625 (0.509 – 0.731)	0.99 (0.958 – 0.999)
Muyle et al 2012	0.275 (0.181 – 0.386)	0.984 (0.945 – 0.998)
Bergero & Charlesworth 2011	0.25 (0.159 – 0.359)	0.992 (0.958 – 0.999)
Chibalina & Filatov 2011	0.425 (0.315 – 0.541)	1 (0.972 – 1)

Table 3: model comparison using SEX-DETector on real datasets (*S. latifolia* - with sex chromosomes and *S. vulgaris* - without sex chromosomes) and simulated X/Y datasets with varying number of sex-linked contigs out of 10,000 simulated contigs. The best model is chosen as the one having the lowest BIC value (see Supplementary Table S4 for details).

	dataset	models with sex chromosomes				model without sex chromosomes
		XY model		ZW model		
		BIC	number of sex-linked genes	BIC	number of sex-linked genes	BIC
Real datasets	*Silene latifolia* (XY system)	best	15164	-	9	-
	Silene vulgaris (no sex chromosomes)	-	0	best	0	-
Simulated datasets of 10,000 genes with different numbers of sex-linked genes (XY system)	0 sex-linked genes	-	0 – 1	-	0	best
	1 sex-linked gene	-	0	-	0	best
	10 sex-linked genes	best	16 – 57	-	0 – 1	-
	100 sex-linked genes	best	156 – 181	-	0 – 10	-
	500 sex-linked genes	best	592 – 624	-	23 – 40	-
	3000 sex-linked genes	best	3159 – 3200	-	1688 – 1807	-

Figure legends

Figure 1 : Schematic steps of our pipeline (**A**) and examples of family genotypes for the three types of segregation types in an X/Y system (**B**).

Figure 2: Results of the SEX-DETector pipeline for known *S. latifolia* genes. Only informative SNPs are shown: positions that are inferred as polymorphic, and for which, in case of autosomal or X/Y segregation, the heterogametic parent is heterozygous and different from the homogametic parent (otherwise it is not possible to differentiate between X/Y and autosomal segregation). Segregation type posterior probabilities are shown for each informative SNP (see legend on figure for colour code) and inferred number of genotyping errors for each segregation type are shown inside the bars (a genotyping error is inferred when its posterior probability is higher than 0.5). **A)** SlE72 (this gene is known to be autosomal), its weighted autosomal mean probability is 0.99. **B)** SlCypX (this gene is known to be X/Y), its weighted sex-linked mean probability is 0.96. **C)** WUS1 (this gene is known to be X-hemizygous), its weighted sex-linked mean probability is 0.99. **D)** BAC284N5-CDS13_SlX6a (this gene is known to be sex-linked), its weighted sex-linked mean probability is 0.82.

Figure 3: The number of SNPs without genotyping error was plotted against the posterior segregation type probability for each autosomal (**A**) and sex-linked (**B**) contigs of the *S. latifolia* dataset. The distributions of both variables are shown, and means for each category on the histograms are indicated by red dots. Sex-linked genes kept after the filter commonly used in empirical methods are shown in green (at least five sex-linked SNPs and no autosomal SNPs).

Figure 4: ROC curve showing the effect of the number of progeny sequenced on sensitivity (TPR, true positive rate) and specificity (1-FPR, false positive rate) in simulated data. A perfect classification of contigs would lead to a point having TPR equal to one and FPR equal to zero (top left corner of the graph). **A)** X/Y or Z/W sex determination system (all points overlap in the top left corner when over five progeny of each sex are used). **B)** U/V system (all points overlap in the top left corner when over two progeny of each sex are used).

Figure 1.

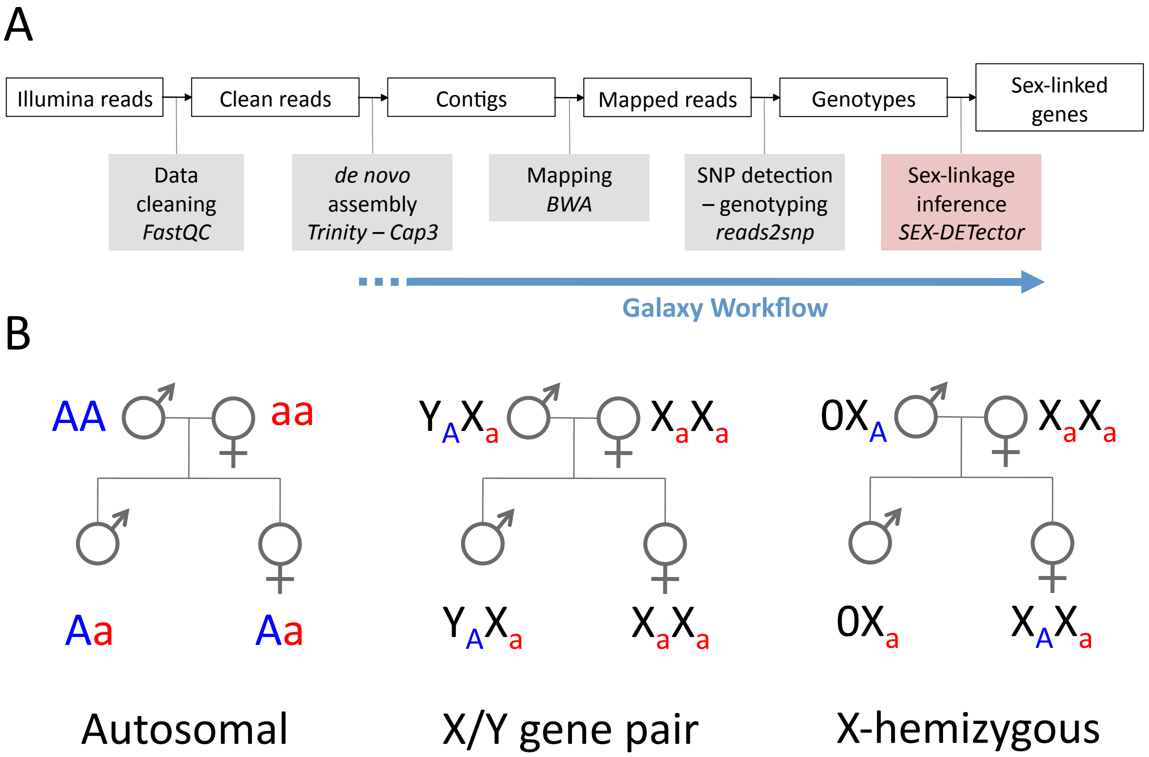

A
Illumina reads
Clean reads
Contigs
Mapped reads
Genotypes
Sex-linked genes
Data cleaning *FastQC*
de novo assembly *Trinity – Cap3*
Mapping *BWA*
SNP detection – genotyping *reads2snp*
Sex-linkage inference *SEX-DETector*
Galaxy Workflow
B
AA
aa
Aa
Aa
Y_AX_a
X_aX_a
Y_AX_a
X_aX_a
$0X_A$
X_aX_a
$0X_a$
X_AX_a
Autosomal
X/Y gene pair
X-hemizygous

Figure 2.

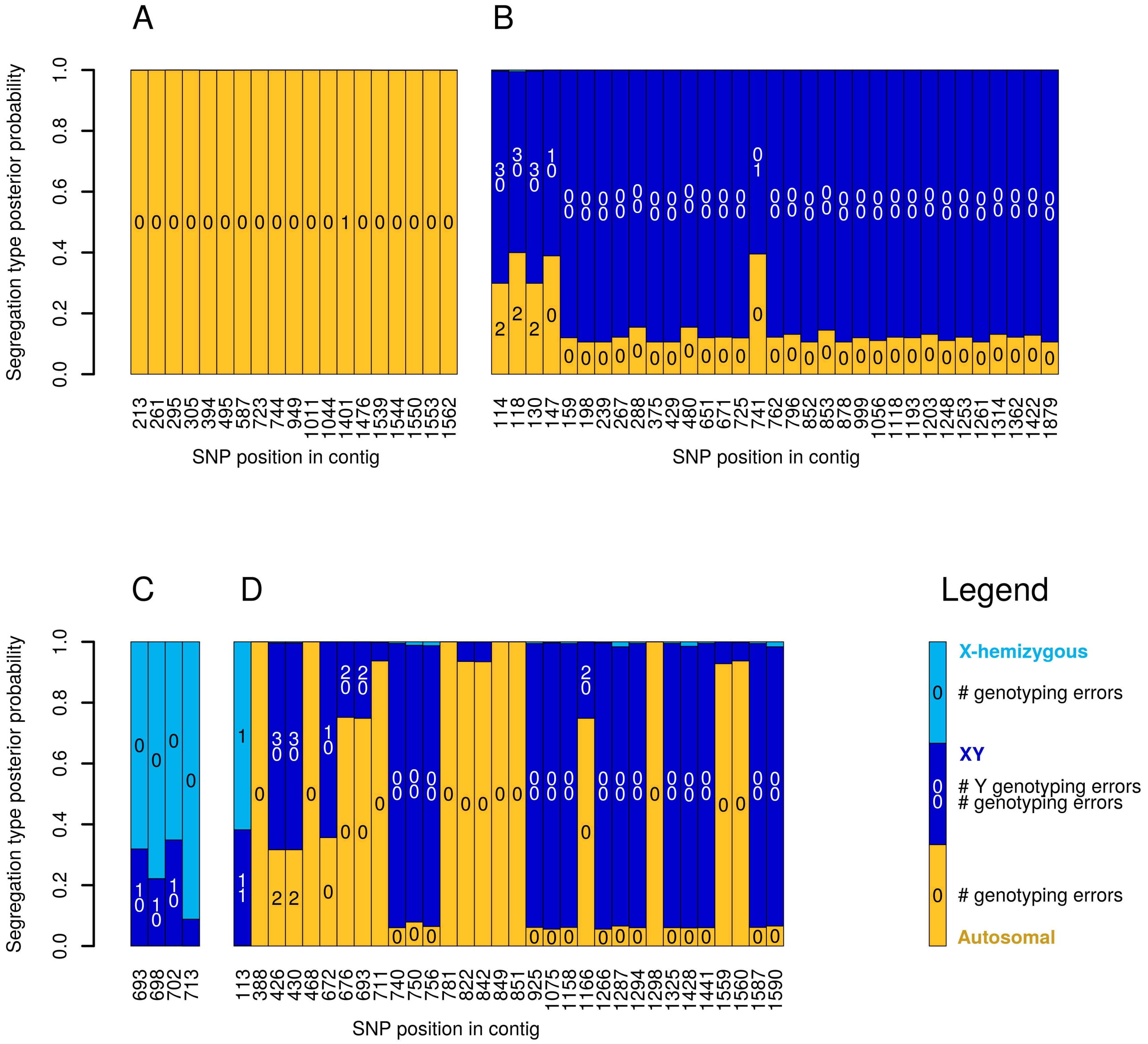

A
B
C
D
Legend
Segregation type posterior probability
0.0
0.2
0.4
0.6
0.8
1.0
SNP position in contig
213 261 295 305 394 495 587 723 744 949 1011 1044 1401 1476 1539 1544 1550 1553 1562
114 118 130 147 159 198 239 267 288 375 429 480 651 671 725 741 762 796 852 853 878 999 1056 1118 1193 1203 1248 1253 1261 1314 1362 1422 1879
693 698 702 713
113 388 426 430 468 672 676 693 711 740 750 756 781 822 842 849 851 925 1075 1158 1166 1266 1287 1294 1298 1325 1428 1441 1559 1560 1587 1590
X-hemizygous
0 # genotyping errors
XY
0 # Y genotyping errors
0 # genotyping errors
0 # genotyping errors
Autosomal

Figure 3.

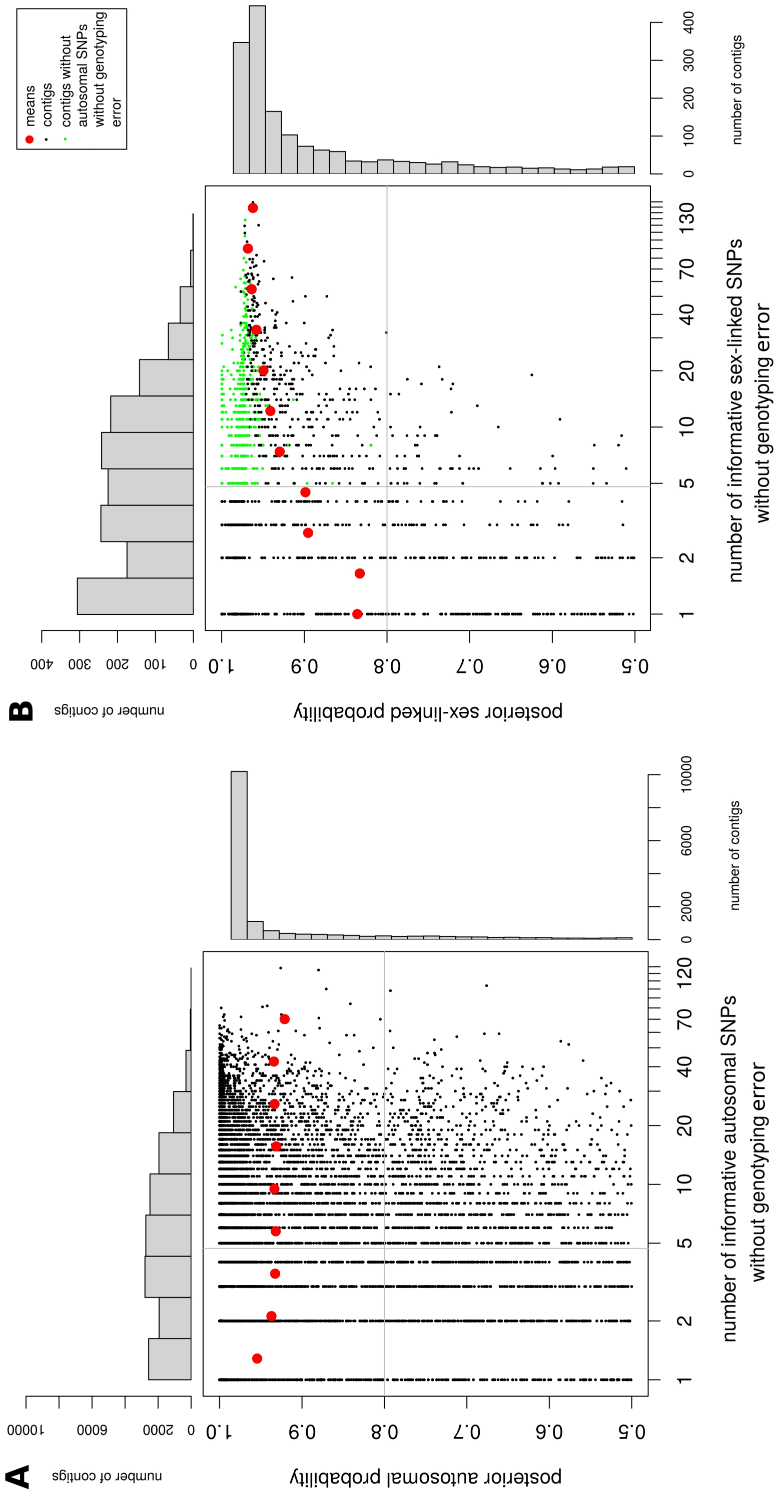
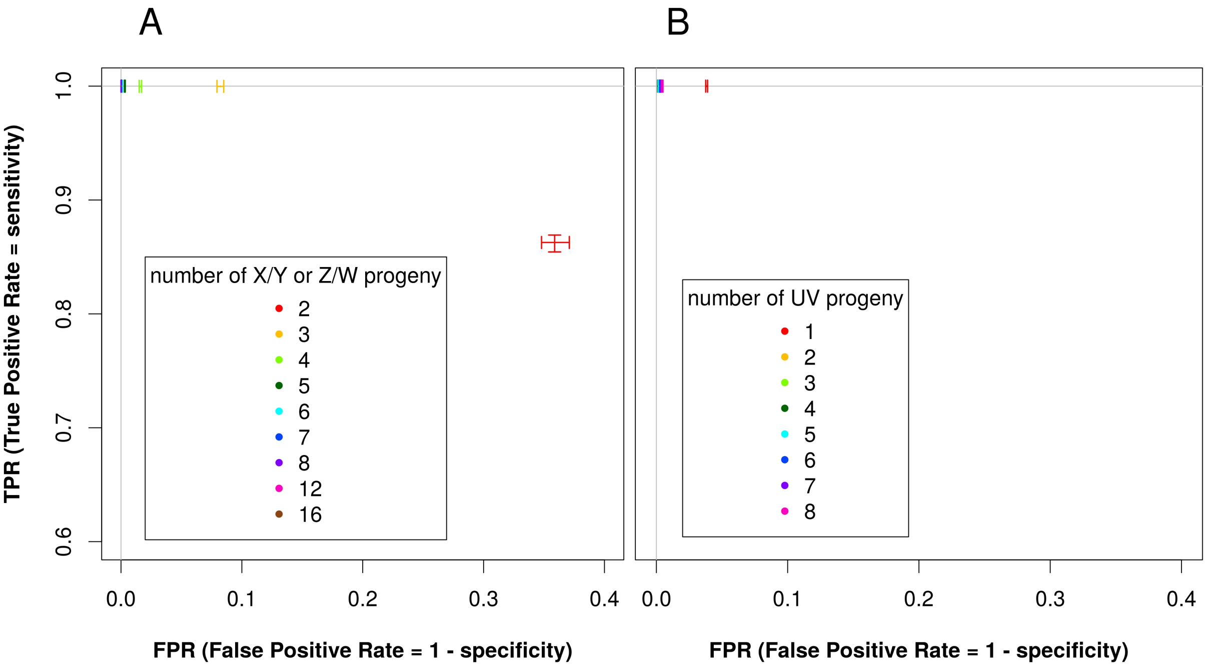

A
B
means
contigs
contigs without autosomal SNPs without genotyping error
posterior autosomal probability
posterior sex-linked probability
number of informative autosomal SNPs without genotyping error
number of informative sex-linked SNPs without genotyping error
number of contigs
1.0
0.9
0.8
0.7
0.6
0.5
1
2
5
10
20
40
70
120
130
0
2000
6000
10000
100
200
300
400

Figure 4.



A
B
TPR (True Positive Rate = sensitivity)
1.0
0.9
0.8
0.7
0.6
number of X/Y or Z/W progeny
● 2
● 3
● 4
● 5
● 6
● 7
● 8
● 12
● 16
number of UV progeny
● 1
● 2
● 3
● 4
● 5
● 6
● 7
● 8
0.0
0.1
0.2
0.3
0.4
FPR (False Positive Rate = 1 - specificity)
FPR (False Positive Rate = 1 - specificity)

Supplementary Material

Supplementary Text S1: Detailed explanation of the model for X/Y and Z/W systems.

Supplementary Text S2: Detailed explanation of the model for U/V systems.

Supplementary Figure S1: Method without a cross: male and female genotypes are studied, a SNP is considered sex-linked if all males are heterozygous (XY) and all females homozygous (XX), the pattern is reversed in the case of a ZW system. Genotypes are either inferred using the genotyper reads2snp or from read counts.

Supplementary Figure S2: simulations design for the program ms (Hudson 2002), for X/Y system (upper part) and U/V system (lower part).

Supplementary Figure S3: ROC curve showing the effect of different parameters on sensitivity (TPR, true positive rate) and specificity (1-FPR, false positive rate) in simulated data. A perfect classification of contigs would lead to a point having TPR equal to one and FPR equal to zero (top left corner of the graph). **A)** effect of X-Y divergence time. **B)** effect of the number of sex-linked contigs out of 10,000 simulated contigs. **C)** effect of theta (polymorphism).

Supplementary Figure S4: expected segregation result in a family when only one gene is sex-linked, closely linked genes will also seem sex-linked so that there is never a single sex-linked gene in a dataset that could not be detected by SEX-DETector (unlike what was observed in simulations where recombination events are not simulated).

Supplementary Figure S5: Schematic steps of the SEX-DETector pipeline with parameters of the model. Π_j are the segregation types probabilities in the dataset (j=1 for autosomal, j=2 for X/Y and j=3 for X-hemizygous). α_j are the proportions of real homogametic parent genotypes, in segregation type j, in the whole dataset. β_j are the proportions of real heterogametic parent genotypes, in segregation type j, in the whole dataset. ε is the probability of a genotyping error happening on any allele. p is the probability of a genotyping error happening on the Y allele, which is higher than ε due to low Y expression level in RNA-seq data. Observed genotypes can differ from real genotypes due to genotyping errors. The different parameters are estimated using an EM algorithm which then allows to compute a posterior segregation type probability for each SNP and then each contig.

Supplementary Table S1: Segregation tables, observed genotypes probabilities given true parent genotypes, segregation types and genotyping errors.

Supplementary Table S2: Known genes used to test SEX-DETector and compare it with other methods in *S. latifolia.*

Supplementary Table S3: Library sizes (in number of reads) and mapping statistics.

Supplementary Table S4: Details of model comparison using SEX-DETector on real datasets (*Silene latifolia* which has sex chromosomes and *Silene vulgaris* which does not have sex chromosomes) and simulated X/Y datasets with varying number of sex-linked contigs out of 10,000 simulated contigs. The best model is chosen as the one having the lowest BIC value (bold and stressed).