Molecular Ecology Resources
November 2020, Volume 20, Issue 6, Pages 1610-1619
https://doi.org/10.1111/1755-0998.13225
https://archimer.ifremer.fr/doc/00640/75231/

Methods for identifying and interpreting sex- linked SNP markers and carrying out sex assignment: application to thornback ray (*Raja clavata*)

Trenkel Verena [1, *], Boudry Pierre [2], Verrez-Bagnis Veronique [3], Lorance Pascal [1]

[1] Ifremer, EMH Nantes, France
[2] Univ Brest, Ifremer, CNRS, IRD, LEMAR Plouzané ,France
[3] Ifremer, LEMMMB Nantes ,France

* Corresponding author : Verena Trenkel, email address : verena.trenkel@ifremer.fr

Abstract :

Sex determining modes remain unknown in numerous species, notably in fishes, in which a variety of modalities have been reported. Additionally, non-invasive individual sexing is problematic for species without external sex attributes or for early life stages, requiring cytogenetic or molecular analyses when sex chromosomes or sex-linked markers have been characterized. Genomics now provide a means to achieve this. Here, we review common sex-determination systems and corresponding statistical methods for identifying sex-linked genetic markers and their use for sex assignment, focusing on single nucleotide polymorphism (SNP) markers derived from reduced representation sequencing methods. We demonstrate the dependence of expected sex assignment error on the number of sex-linked SNPs and minor allele frequency. The application of three methods was made here: i) identification of heterozygote excess in one sex, ii) F ST outlier analysis between the two sexes and iii) neuronal net modelling. These methods were applied to a large SNP dataset (4604 SNPs) for 1680 thornback rays (Raja clavata). Using method i), nineteen putative sex-linked SNPs were identified. Comparison with the reference genome of a related species (Amblyraja radiata) indicated that all 19 SNPs are likely located on the same chromosome. These results suggest that thornback ray has a XX/XY sex-determination system. Method ii) identified eight SNPs probably located on different chromosomes. Method i) led to the lowest sex assignment error among the three methods (4.2% error for females and 3.7% for males).

Keywords : association-based approach, F ST outlier analysis, neural net, sex assignment error, sex determination system, sex-linked SNP, thornback ray

Introduction

Sex determination systems greatly vary between taxa and are generally classified as genetic, including monogenic or polygenic, or environmental. Genetic-based sex determination systems have been characterized for a wide range of taxa, including plants (Montgomery, Sadeque, Giacomini, Brown, & Tranel, 2019), algae (Coelho, Mignerot, & Cock, 2019), bivalves (Breton, Capt, Guerra, & Stewart, 2018) and vertebrates (Galindo, Loher, & Hauser, 2011; Heule, Salzburger, & Bohne, 2014; Mank, Promislow, & Avise, 2006). Individual sex assignment from genetic data is becoming increasingly important as researchers seek to reduce the impact of scientific monitoring of wildlife populations (V. M. Trenkel et al., 2019) or to replace indirect sexing methods (e.g. Loher, Woods, Jimenez-Hidalgo, & Hauser, 2016). For example, the sex in fishes is generally not identifiable from external criteria, making it necessary to kill fish to inspect internal organs. This can be problematic for species of conservation concern or in an aquaculture context (e.g. production of caviar) and more generally goes against the principle of minimising harm inflicted by scientific studies (Costello et al., 2016). Another situation where genetic-based sex determination can be useful is when only tissue samples are available but not whole individuals.

Methods for developing genetic-based sex identification depend on the sex-determination system, the possibility to analyse sex-ratio in progenies resulting from controlled crosses and whether a reference genome exists or not (Palmer, Rogers, Dean, & Wright, 2019). We first briefly review

common sex-determination systems and then summarize methods that can be used to identify sex-linked markers and carry out sex assignment focusing on single nucleotide polymorphism (SNP) markers derived from reduced representation sequencing methods (RRS). We illustrate some of the methods for thornback ray (*Raja clavata*), for which a large SNP dataset was available. The sex-determination system of this species was unknown. However, those sex-determination systems for rays which are known are primarily XX/XY (Devlin & Nagahama, 2002).

Sex-determination systems

Sex-determination systems vary across animal and plant clades, with some exhibiting a variety of systems (Bachtrog et al., 2014). In fishes, sex can be genetically or environmentally determined (Devlin & Nagahama, 2002), or result from the interaction of both types of factor. Temperature driven sex-determination, with more males developing in warmer waters, has notably been shown in southern flounder (Honeycutt et al., 2019) and European sea bass (see review in Baroiller, D'Cotta, & Saillant, 2009). In some species, polygenic sex determination combines genetic and environmental components that may vary between populations (e.g. in sea bass, Faggion, Vandeputte, Chatain, Gagnaire, & Allal, 2019). For many taxa, however, the sex-determination system is unknown. Therefore, to be most useful, methods for identifying sex-linked genetic markers need to work for different genetic sex-determination systems, in addition to being applicable for species for which genomes remain to be sequenced.

Some taxa have major sex-determining genes (see review in Bachtrog et al., 2014). One strategy is, therefore, to search for genes known to be involved in spermatogenesis (e.g. Rocco, Bencivenga, Archimandritis, & Stingo, 2009), testes development or thought to be in some way to be linked to sex determination (e.g. Bewick et al., 2013). The alternative strategy is to scan the genome using genetic markers for differences between sexes (Benestan et al., 2017; Gamble & Zarkower, 2014; Loher et al., 2016; Montgomery et al., 2019; Utsunomia et al., 2017). The differences between sexes will depend on the sex-determination system (Palmer et al., 2019). We summarise below relevant methods which are applicable independent of the degree of sex chromosome divergence, focussing on the use of SNP derived from RRS.

Sex-determining genes (major genes or genes with quantitative effects) can be located on the sex chromosomes or be spread throughout the genome over several chromosomes (Bachtrog et al., 2014). For SNP markers located on sex chromosomes, the expected pattern in the proportions of

male and female heterozygous individuals will be characteristic of the sex-determination system and the location of the genes (Table 1). For SNPs located on the sex chromosomes or linked to major sex-determining genes, females or males (according to the sex determination system) may show visibly greater heterozygosity. If, however, the sex-determining genes are spread over several chromosomes, heterozygosity will be similar between sexes for all linkage groups. For the proportion of heterozygous individuals to differ between sexes, allele frequencies need to differ. This could occur if sex differentiation was the result of genetic divergence. In contrast, the proportion of male and female heterozygous individuals will not differ for SNPs on loci unlinked to sex determination. Thus, the proportion of heterozygous individuals and the presence of SNP can be used to devise statistical methods for identifying sex-linked SNPs.

Identifying sex-linked SNPs

The appropriate method for identifying sex-linked SNPs depends on the sex-determination system (Palmer et al., 2019). If this system is unknown, then results obtained by applying different approaches will provide insights into the potential system.

For taxa with XX/XY sex-determination, only female individuals can be heterozygous for an SNP located on the X chromosome (type 1 in table 1), with the actual proportion depending on the allele frequency of the SNP. For an SNP located on the Y chromosome (type 2 in table 1) the locus will be missing for all females, and all males will be homozygous. The type 1 pattern was found for salmon louse (Carmichael et al., 2013), while the type 2 pattern was identified for rockfishes (Fowler & Buonaccorsi, 2016), a lizard (Gamble & Zarkower, 2014) and fur seal (Stovall et al., 2018). Drinan, Loher, and Hauser (2018) found SNPs of type 1 and type 2 in Pacific halibut.

The proportion of heterozygous individuals might differ between sexes for taxa with sex-determining genes spread throughout the genome (type 3 in table 1). SNPs with significantly different allele frequencies between the two sexes have been detected using the fixation index F_{ST} (Benestan et al., 2017; Drinan et al., 2018; Galindo et al., 2011). In this case, the F_{ST} is defined as the relative difference between the average expected heterozygosity of the two sexes given their respective allele frequencies and the expected heterozygosity ignoring sex. An F_{ST} value close to 1 means strong divergence between the two sexes and a value close to 0 means little divergence. Drinan et al. (2018) used a fixed cut off value of $F_{ST} \geq 0.3$ to identify sex-linked loci for Pacific

halibut, while Benestan et al. (2017) carried out an outlier analysis for F_{ST} values using the BayeScan program developed by Foll and Gaggiotti (2008). Alternatively, such sex-linked SNPs can be identified using SNP-wise likelihood ratio tests (G-test) for the null hypothesis of identical allele frequencies or logistic regression to select SNPs with significant explanatory power, as done by Utsunomia et al. (2017) using GENEPOP for the G-test (Rousset, 2008) and GenABEL for the logistic regression (Aulchenko, Ripke, Isaacs, & Van Duijn, 2007).

Most recent studies use either whole genome sequencing (e.g. Vicoso, Emerson, Zektser, Mahajan, & Bachtrog, 2013) or genotyping-by-sequencing (GBS) and other RRS protocols (e.g., RADseq, ddRAD), which are prone to genotyping errors and missing data (Marandel et al., 2020; Mastretta-Yanes et al., 2015). Another approach for reduced representation is transcriptome sequencing (Rovatsos, Rehak, Velensky, & Kratochvil, 2019; Rovatsos, Vukic, & Kratochvil, 2016). In a species for which segregation analysis in progenies resulting from controlled crosses is feasible (*Cannabis sativa)*, Prentout et al. (2020) recently proposed a RNA-seq based approach to identify the sex chromosomes. Note that, when using an RRS protocol for SNP development, the proportion of individuals required to have the SNPs needs to be set to 0 in order not to exclude potential sex-linked SNPs in the bioinformatics pipeline (Utsunomia et al., 2017). This concerns the *r* parameter in the commonly used Stacks program (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013). Further, the absence of non-polymorphic loci from one of the sexes should be checked for. These loci would correspond to sequences absent from the other sex because they occur only on the Y or W chromosome. Loci only present in females were identified for the neotropical fish *Characidium gomesi* (Utsunomia et al., 2017). To account for RADseq errors, Stovall et al. (2018) developed a statistical method to set sex-specific threshold values when identifying loci only present in one sex.

Assigning sex using SNPs

Sex assignment using SNPs can be done in the following ways. If the sex-linked SNPs have been found on the Y or W chromosome, individuals for which these SNPs are present are declared males in an XX/XY system (females if ZW/ZZ). Assignment error will depend on the reliability and completeness of the genotype data.

If sex-linked SNPs have been identified as those with heterozygous individuals predominantly occurring in females, the male sex is assigned to individuals who are homozygous for all sex-

linked SNPs. Individuals that are heterozygous for at least one sex-linked SNP are correspondingly assigned as females. By chance, depending on allele frequency, some females will be homozygous for some of the sex-linked SNPs. The expected assignment error for females therefore decreases when the available number of sex-linked SNPs increases (Figure 1a). For example, if all sex-linked SNPs had minor allele frequency (MAF) of 0.02, around 100 unlinked SNPs would be needed for an expected assignment error of 2%. For a MAF of 0.1, the number of SNPs needed for the same precision would only be 20. The assignment error depicted in figure 1a applies for unlinked SNPs; linkage (correlation) will increase the number of required SNPs to achieve the same assignment error.

For sex assignment using SNPs under sex-related selection or differing between sexes due to other mechanisms, the log-likelihood of observing the given pattern of genotypes (e.g., homozygous for the minor allele on SNP 1, heterozygous for SNP 2, etc.) can be used to assign the sex with the largest log-likelihood. This method is implemented as the frequency criterion in GENECLUST2 (Piry et al., 2004). The assignment error decreases rapidly with the number of SNPs under sex-related or correlated selection (Figure 1b). For example, if the F_{ST} between females and males is 0.25, then around 40 unlinked SNPs are needed to achieve an expected assignment error of 2%. Note that the number of SNPs is greater than the difference in allele frequency summarised by the F_{ST} value. GENECLUST2 implements additional criteria for group assignment, which could also be used. Loher et al. (2016) used GENECLUST2 for assigning sex to Pacific halibut.

Instead of searching for sex-linked SNPs or those under sex-related selection first, sex assignment can also be tackled directly. Supervised machine learning provides another way to create a sex assignment model (Palaiokostas et al., 2013).

Interpreting and validating sex-linked SNPs

Once putative sex-linked SNPs have been identified, some insights on their relative positions in the genome can be gained by studying the correlation between these SNPs. The expectation is that sex-linked SNPs found on the X or W chromosome will be more correlated than a random selection of SNPs spread across numerous autosomes. The linkage-disequilibrium based correlation coefficient is a suitable choice for calculating pairwise correlations (Russell & Fewster, 2009).

Validation of sex-linked loci can be achieved in several ways: i) cross-validation using part of the dataset used for identifying sex-linked loci, and/or ii) predicting sex for an independent sample with morphologically determined sex. Certain authors have also developed specific assays for DNA sequences including the putative sex-linked identified SNPs and then used them to identify the sex for an independent sample to be compared with morphologically determined sex (Carmichael et al., 2013; Drinan et al., 2018; Shi et al., 2018; Suda et al., 2019; Utsunomia et al., 2017).

To move from sex-correlated SNPs to gene function, comparison and alignment of the loci including the putative sex-linked SNPs with known interpreted genomes is needed (e.g. Benestan et al., 2017; Carmichael et al., 2013; Utsunomia et al., 2017). The most common tool for this is BLAST (basic local alignment search tool) first developed by Altschul, Gish, Miller, Myers, and Lipman (1990), and available from the National Centre for Biotechnology Information (NCBI).

Application

Data

The full set of SNPs for thornback ray was derived in two steps. First a RADseq protocol was applied to 225 individuals sampled in the Bay of Biscay and elsewhere (Northeast Atlantic and Mediterranean Sea) as described in Marandel et al. (2020) using DNA extracted from fin-clips. From this, the most polymorphic 9317 loci (allele frequency > 0.08 for individuals from the Bay of Biscay) were selected (Le Cam et al., 2019), which corresponds to ~ 3 markers/1 Mb distributed across the genome.

Next, 1680 samples (943 females, 737 males) from the Bay of Biscay were genotyped at these SNP loci using an Infinium® XT iSelect-96 SNP-array (chip). For this species, sex can be determined morphologically because males have two external organs called claspers. However, to avoid mistakes individuals should be turned on their backs. The individuals were sampled by different persons, including fishers and scientists.

Individual genotypes were scored using the clustering algorithm implemented in the Illumina® GenomeStudio Genotyping Analysis Module v2.0.3. The chip was created with DesignStudio Microarray Assay Designer. The high-throughput solution Infinium XT requires integrated systems that streamline sample preparation and analysis, so the Illumina Automation Control

software for the Tecan liquid handling robot was used. Genotyping reports were analysed with GenomeStudio, which normalizes the intensities of signals for each locus and assigns a cluster position to each sample. The GenCall score was then calculated for each genotype. A no-call threshold of 0.15 was used to not call individuals too far away from the cluster centre. Data quality controls included removing non-polymorphic and non-scoring SNPs. During chip development, 987 SNPs had been duplicated as there was a second SNP in the 50 nucleotide bases flanking sequence (Le Cam et al., 2019). When both versions were scored, one was removed from the final dataset, which contained 4604 valid SNPs (Trenkel & Lorance, 2020). The mean percentage of missing data was 0.3% (0.6–4%) per SNP and was, therefore, not an issue for this dataset.

Analyses

Two association-based methods outlined below were compared for identifying putative sex-linked SNPs and three methods were compared for assigning sex.

Method 1, referred to as the heterozygosity method, consisted of first searching for SNPs presenting at least five times higher heterozygosity in one sex than in the other. The identified putative sex-linked SNPs were then used for sex assignment, assuming an individual was female if it was heterozygous for at least one putative sex-linked SNP. Assignment error was determined by comparing the assigned sex with the sex recorded during sampling. Assignment error for a given number of sex-linked SNPs (1 to 19) was calculated by randomly drawing (without replacement) 100 SNP datasets.

Method 2 consisted of first carrying out an F_{ST} outlier analysis using BayeScan 2.1 (Foll & Gaggiotti, 2008) with default parameters. This method detects loci under selection by comparing populations, here sexes. BayeScan implements a Bayesian estimator of F_{ST} using the model proposed by Beaumont and Balding (2004) based on allele frequency counts. The identified K SNPs were then used for sex assignment. The log-likelihood for sex s was calculated as

$$logL_s = \sum_{k=1}^{K} \sum_{g=1}^{3} I_{k,g} \log\left(p_{s,k,g}\right)$$

where $p_{s,k,g}$ is the probability of observing genotype $g \in (AA, AB, BB)$ for SNP k if the individual is of sex s. $I_{k,g}$ is an indicator function that takes value 1 if the individual has genotype g for SNP k and 0 otherwise. Genotype probabilities were estimated from genotype counts of the full data.

Assigned and recorded sex was then compared to calculate assignment error. Only the full set of SNPs was considered because the number of SNPs identified to be potentially under selection was small.

Method 3, a neural net with one hidden layer, was used for sex assignment alone. The neural net was fitted to the full list of SNPs using the R package *neuralnet* (Günther & Fritsch, 2010). The number of neurons was varied between one and ten, leading to ten models. Individuals were randomly split into two datasets, one for fitting (3/4) and one for testing (1/4); ten pairs of such random sets were created. Assignment error was again calculated by comparing assigned and recorded sex.

To investigate linkage between SNPs the correlation coefficient $\hat{r}_{AB}$ was calculated using the linkage-disequilibrium approach (Russell & Fewster, 2009). Assuming SNP 1 with alleles *A* and *A*' is compared with SNP 2 with alleles *B* and *B*', the correlation is

$$\hat{r}_{AB} = \frac{\hat{\Delta}_{AB}}{\sqrt{\{\hat{p}_A(1-\hat{p}_A) + (\hat{h}_{AA} - \hat{p}_A^2)\}\{\hat{q}_B(1-\hat{p}_B) + (\hat{h}_{BB} - \hat{q}_B^2)\}}}$$

where $\hat{p}_A$ and $\hat{q}_B$ are the sample proportions of alleles *A* and *B* in the *n* sampled individuals. $\hat{h}_{AA}$ and $\hat{h}_{BB}$ are the proportions of homozygous individuals *AA* and *BB*, respectively. Δ_{AB} is called Burrow's composite linkage disequilibrium and is estimated from genotype data as

$$\hat{\Delta}_{AB} = \frac{n_{AB}}{n} - 2\hat{p}_A\hat{q}_B$$

where n_{AB} is calculated by summing the counts of different genotype combinations, e.g., if n_{AABB} is the number of individuals that are homozygous AA at SNP 1 and homozygous BB at SNP 2, then

$$n_{AB} = 2n_{AABB} + n_{AA'BB} + n_{AABB'} + \frac{n_{AA'BB'}}{2}.$$

The pairwise correlation coefficients $\hat{r}_{AB}$ estimated for putative sex-linked SNPs were compared to those obtained for other randomly selected pairs of SNPs. The expected correlation is zero for

unlinked SNPs. All statistical analyses were carried out using custom R code (version R3.4.4) (R Core Team, 2018).

A BLAST search using the executable BLAST+ 2.6.0 package (Altschul et al., 1990) available from NCBI optimized for short sequences was carried out for the identified putative sex-linked SNPs on 65–279 bp SNP sequences against a representative genome of *Amblyraja radiata* (starry ray, male adult, testis and liver tissues, GenBank accession number GCA_010909765.1 for whole genome assembly and GCF_010909765.1 for the annotated assembly), the closest related species with a publicly available genome. This assembly of the starry ray whole genome represents the principal haplotype of the diploid genome and includes the 49 chromosomes, which is the same number as thornback ray (Stingo & Rocco, 2001). .

Results

Applying the heterozygosity method, males were identified as being primarily homozygous at 19 SNPs (Figure 2ab); primer sequences of these SNPs are available in Trenkel and Lorance (2020). The 19 SNPs had a wide range of MAFs (mean 0.25, range 0.06 – 0.46). Assignment error decreased for females and increased for males with the number of the sex-linked SNPs used for assignment, reaching 4.2% for females and 3.7% for males when all 19 SNPs were used (Figure 2cd). Among the 19, ten putative sex-linked SNPs were significantly more correlated among each other than all other non-sex-linked SNPs (Figure 3); 16 sex-linked pairs had correlation coefficients $\hat{r}_{AB}$ outside the 99% central percentile range of non-sex-linked SNPs (-0.085 - 0.087). Thus, several of the 19 putative sex-linked SNPs are probably located in close proximity on the same chromosome.

The F_{ST} method identified eight outlier SNPs that might be under differential selection between males and females or show differences due to other mechanisms. The eight SNPs were all different from the 19 SNPs identified using the heterozygosity method (Trenkel & Lorance, 2020). These eight SNPs had larger F_{ST} values (0.008 to 0.017) compared with F_{ST} values for other SNPs (< 0.001) and were all uncorrelated ($\hat{r}_{AB}$: mean = 0.005; range = 0.048–0.060). Closer inspection revealed that there was a significant ($p < 0.01$) deficit of homozygous individuals for the minor allele in both sexes for one of these SNPs. Significance was determined by calculating the binomial probability of the observed number of minor-allele homozygous individuals given the sex-specific allele frequencies. It should be noted, however, that the minor allele frequencies for

these eight SNPs were small (0.04-0.08). Using all eight SNPs, assignment error was 59% for females and 39% for males.

The neural nets fitted to 3/4 of a random selection of the individuals and used for assigning sex to individuals in the withheld 1/4 had assignment errors around 10% for both males and females (Figure 4). Increasing the number of neurons in the hidden layer beyond two did not reduce misclassification any further.

The BLAST search for the 19 putative sex-linked SNPs identified with the heterozygosity method against the starry ray genome provided significant alignments for all of them. All 19 sequences were located on chromosome 46 of *Amblyraja radiata*. Hence, we can hypothesize that these putative sex-linked SNPs are located on the chromosome that is involved in sex determination (X chromosome). In contrast, among the eight SNPs identified with the F_{ST} outlier method, significant alignments were found for six SNP sequences that were all on different chromosomes but none of which were on chromosome 46. Only one SNP sequence could be related to an annotated sequence. It corresponds to the sequence of the *cluha* clustered mitochondria (cluA/CLU1) gene on chromosome 28. This gene is predicted to have mRNA binding activity and to be involved in intracellular distribution of mitochondria (Gao et al., 2014). The link with sex differentiation is, therefore, not obvious.

Discussion

The overview of sex determination systems and statistical methods for identifying putative sex-linked SNPs stresses the logical link between the two; this subject has been reviewed in greater detail by Palmer et al. (2019). The most powerful statistical methods depend on the sex determination system. If this is unknown, then applying several methods can provide insights into the potential sex determination system of the species of interest. However, no definite confirmation of the sex determination system can be obtained from SNP data alone. Furthermore, the identified putative sex-linked SNPs are not necessarily located on genes coding for processes involved in sex determination. As more genomes become well annotated, it should become easier in increasing numbers of species to identify the genes with SNPs associated with sex determination and their function.

Sex assignment is an important practical application of sex-linked SNP markers. There are many research and management related situations where sex assignment is essential. For example, for determining the sex of tissue samples without access to the individual or of individuals without external sex signs that are caught during scientific monitoring surveys and released alive to reduce monitoring impacts (Trenkel et al., 2019). Sex assignment is also useful in the food industry, e.g. for determining the sex-related quality category of derived meat products (Abdulmawjood, Krischek, Wicke, & Klein, 2012), broodstock management (Slembrouck et al., 2019) or to allow early culling of undesired individuals to maximize production (Falahatkar, Akhavan, Gilani, & Abbasalizadeh, 2013). Our study was motivated by a scientific project for which fishers collected a large number of thornback ray tails for later tissue sampling. Due to time constraints they were unable to record any individual sex information, therefore generating a need to find a sex assignment method.

Two methods for identifying putative sex-linked SNPs and assigning sex and one method for sex assignment only were applied to thornback ray using a large dataset of 4604 SNPs for 1681 individuals. Nineteen putative sex-linked SNPs were identified by comparing the proportion of heterozygous individuals in male and female individuals. In comparison, for a teleost fish species with ZW/ZZ sex determination, Utsunomia et al. (2017) identified 25 SNPs out of 9863 that exhibited extreme heterozygote deficiency in females.

BLAST comparisons with the starry ray reference genome provided satisfactory alignments with genes. All 19 putative sex-linked SNPs are located in the chromosome 46 of the starry ray genome. The observed pattern of lacking heterozygous males points to an XX/XY sex-determination system for this species, similar to other ray species (Devlin & Nagahama, 2002). Among these 19 putative sex-linked SNPs, ten were significantly correlated. This indicated that they are located in relatively close proximity on the X chromosome. The assignment error using all 19 putatively sex-linked SNPs was around 4%. This error level is comparable to Stovall et al. (2018) (1.1% females, 4.2% males). For females, it corresponds rather well to the expected value given the number of SNPs and their MAF values (Figure 1). Hence, reduction in assignment error for females could only be achieved by increasing the number of sex-linked SNPs. Unfortunately, no further such SNPs could be identified in the dataset. Inspection of the 67 misclassified individuals suggested that, for some of them, the misclassification was probably due to an error in the recorded sex data, as several consecutively sampled male and female individuals were

misclassified, i.e. they might have been shifted by one line when the information for sampled individuals was entered. For males this explanation is plausible for at least 14 out of 27 miss-assigned individuals and for females for 12 out of 40. Further undetected recording errors and some small level of genotyping errors might explain the existence of heterozygous males as none should occur if the identified SNPs were really located on the X chromosome. An alternative explanation is that the identified SNPs are on sex-determination regions, but that there is a small probability of recombination between them and the sex-determination gene occurring in males.

The F_{ST} outlier analysis identified eight SNPs potentially under sex-related selection, with one SNP having a deficit of homozygous individuals for the minor allele, which might indicate that it is deleterious for its carrier. Furthermore, even the largest F_{ST} value (0.06) was much smaller than the threshold value 0.3 used by Drinan et al. (2018) for identifying SNPs under selection. Comparison with the starry ray genome indicated the locations of six of these SNPs on different chromosomes. One SNP corresponded to the *cluha* clustered mitochondria (cluA/CLU1) gene whose role in sex determination if any is not obvious to us. The assignment error for this method based on assigning the sex with the largest log-likelihood was large. For females (59%) it was larger than the 45% expected error for eight SNPs (Figure 1b), while it was smaller for males (39%). Unless more SNPs under sex-related selection can be identified for thornback ray, this assignment method is unsuitable for practical applications in this species.

Compared to the heterozygosity method the neural net analysis had about twice as large assignment error (10%). Its performance might have been reduced by the potential data recording errors discussed above. However, given the large number of SNPs used in the analysis it is conceivable that the relatively small signal of a few sex-linked SNPs and SNPs under sex-related selection became swamped. This difficulty will be similar for other applications, making the method a second choice for sex assignment in species with sex chromosomes.

In conclusion, SNPs can provide insights into sex determination mechanisms and can be used for sex assignment. However, for practical applications, the number of SNPs that are sex-linked or subject to sex-related selection needs to be sufficiently large to achieve acceptable assignment errors.

Acknowledgements

This study received funding from the French national research agency *Agence Nationale de la Recherche* (project GenoPopTaille, contract ANR-14-CE02-0006-01) and the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773713 (PANDORA). We thank E. Blanc, APECS (*Association Pour l'Étude et Conservation des Sélaciens*) volunteers and scientists for collecting tissue samples and sex information on the surveys carried out by INRAE (*Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement*), in the Gironde (Sturat survey) and Ifremer on the Bay of Biscay shelf (Evhoe survey) and in the bay of Douarnenez (RaiesJuves survey). We thank F. Marandel, G. Charrier and A. Bidault for help with tissue preparation. Finally, we would like to thank the four anonymous referees whose constructive comments helped to improve the manuscript.

References

Abdulmawjood, A., Krischek, C., Wicke, M., & Klein, G. (2012). Determination of pig sex in meat and meat products using multiplex real time-PCR. *Meat Science, 91*(3), 272-276. doi:10.1016/j.meatsci.2012.02.001

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Aulchenko, Y. S., Ripke, S., Isaacs, A., & Van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics, 23*(10), 1294-1296. doi:10.1093/bioinformatics/btm108

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T.-L., . . . Tree Sex, C. (2014). Sex determination: Why so many ways of doing It? *Plos Biology, 12*(7). doi:10.1371/journal.pbio.1001899

Baroiller, J. F., D'Cotta, H., & Saillant, E. (2009). Environmental effects on fish sex determination and differentiation. *Sexual development, 3*(2-3), 118-135. doi:10.1159/000223077

Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology, 13*(4), 969-980. doi:10.1111/j.1365-294X.2004.02125.x

Benestan, L., Moore, J. S., Sutherland, B. J. G., Le Luyer, J., Maaroufi, H., Rougeux, E., . . . Bernatchez, L. (2017). Sex matters in massive parallel sequencing: Evidence for biases in genetic parameter estimation and investigation of sex determination systems. *Molecular Ecology, 26*(24), 6767-6783. doi:10.1111/mec.14217

Bewick, A. J., Chain, F. J. J., Zimmerman, L. B., Sesay, A., Gilchrist, M. J., Owens, N. D. L., . . . Evans, B. J. (2013). A large pseudoautosomal region on the sex chromosomes of the frog *Silurana tropicalis*. *Genome Biology and Evolution, 5*(6), 1087-1098. doi:10.1093/gbe/evt073

Breton, S., Capt, C., Guerra, D., & Stewart, D. (2018). Sex-determining mechanisms in bivalves. In J. Leonard (Ed.), *Transitions between sexual systems* (pp. 165-192): Springer, Cham.

Carmichael, S. N., Bekaert, M., Taggart, J. B., Christie, H. R. L., Bassett, D. I., Bron, J. E., . . . Sturm, A. (2013). Identification of a sex-linked SNP marker in the salmon louse (*Lepeophtheirus salmonis*) using RAD sequencing. *PLoS ONE, 8*(10). doi:10.1371/journal.pone.0077832

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology, 22*(11), 3124-3140. doi:10.1111/mec.12354

Coelho, S. M., Mignerot, L., & Cock, J. M. (2019). Origin and evolution of sex-determination systems in the brown algae. *New Phytologist, 222*(4), 1751-1756. doi:10.1111/nph.15694

Costello, M. J., Beard, K. H., Corlett, R. T., Cumming, G. S., Devictor, V., Loyola, R., . . . Primack, R. B. (2016). Field work ethics in biological research. *Biological Conservation, 203*, 268-271.

Devlin, R. H., & Nagahama, Y. (2002). Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture, 208*(3-4), 191-364. doi:10.1016/s0044-8486(02)00057-1

Drinan, D. P., Loher, T., & Hauser, L. (2018). Identification of genomic regions associated with sex in Pacific halibut. *Journal of Heredity, 109*(3), 326-332. doi:10.1093/jhered/esx102

Faggion, S., Vandeputte, M., Chatain, B., Gagnaire, P. A., & Allal, F. (2019). Population-specific variations of the genetic architecture of sex determination in wild European sea bass *Dicentrarchus labrax* L. *Heredity, 122*(5), 612-621. doi:10.1038/s41437-018-0157-z

Falahatkar, B., Akhavan, S. R., Gilani, M. H. T., & Abbasalizadeh, A. (2013). Sex identification and sexual maturity stages in farmed great sturgeon, *Huso huso* L. through biopsy. *Iranian Journal of Veterinary Research, 14*(2), 133-139.

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics, 180*(2), 977-993. doi:10.1534/genetics.108.092221

Fowler, B. L. S., & Buonaccorsi, V. P. (2016). Genomic characterization of sex-identification markers in *Sebastes carnatus* and *Sebastes chrysomelas* rockfishes. *Molecular Ecology, 25*(10), 2165-2175. doi:10.1111/mec.13594

Galindo, H. M., Loher, T., & Hauser, L. (2011). Genetic sex identification and the potential evolution of sex determination in Pacific halibut (*Hippoglossus stenolepis*). *Marine Biotechnology, 13*(5), 1027-1037. doi:10.1007/s10126-011-9366-7

Gamble, T., & Zarkower, D. (2014). Identification of sex-specific molecular markers using restriction site-associated DNA sequencing. *Molecular Ecology Resources, 14*(5), 902-913. doi:10.1111/1755-0998.12237

Gao, J., Schatton, D., Martinelli, P., Hansen, H., Pla-Martin, D., Barth, E., . . . Rugarli, E. I. (2014). CLUH regulates mitochondrial biogenesis by binding mRNAs of nuclear-encoded mitochondrial proteins. *Journal of Cell Biology, 207*(2), 213-223. doi:10.1083/jcb.201403129

Günther, F., & Fritsch, S. (2010). neuralnet: Training of neural networks. *The R Journal, 2*, 30-38.

Heule, C., Salzburger, W., & Bohne, A. (2014). Genetics of Sexual Development: An Evolutionary Playground for Fish. *Genetics, 196*(3), 579-591. doi:10.1534/genetics.114.161158

Honeycutt, J. L., Deck, C. A., Miller, S. C., Severance, M. E., Atkins, E. B., Luckenbach, J. A., . . . Godwin, J. (2019). Warmer waters masculinize wild populations of a fish with temperature-dependent sex determination. *Scientific Reports, 9*. doi:10.1038/s41598-019-42944-x

Le Cam, S., Bidault, A., Charrier, G., Cornette, F., Lamy, J.-B., Lapegue, S., . . . Trenkel, V. (2019). SNPs for thornback ray *Raja clavata*, validated from genotyped individuals. *SEANOE. https://doi.org/10.17882/70546*.

Loher, T., Woods, M. A., Jimenez-Hidalgo, I., & Hauser, L. (2016). Variance in age-specific sex composition of Pacific halibut catches, and comparison of statistical and genetic methods for reconstructing sex ratios. *Journal of Sea Research, 107*, 90-99. doi:10.1016/j.seares2015.06.004

Mank, J. E., Promislow, D. E. L., & Avise, J. C. (2006). Evolution of alternative sex-determining mechanisms in teleost fishes. *Biological Journal of the Linnean Society, 87*(1), 83-93. doi:10.1111/j.1095-8312.2006.00558.x

Marandel, F., Charrier, G., Lamy, J.-B., Le Cam, S., Lorance, P., & Trenkel, V. M. (2020). Estimating effective population size using RADseq: effects of SNP selection and sample size. *Ecology and Evolution, 10*(4), 1929-1937.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Pinero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources, 15*(1), 28-41. doi:10.1111/1755-0998.12291

Montgomery, J. S., Sadeque, A., Giacomini, D. A., Brown, P. J., & Tranel, P. J. (2019). Sex-specific markers for waterhemp (*Amaranthus tuberculatus*) and Palmer amaranth (*Amaranthus palmeri*). *Weed Science, 67*(4), 412-418. doi:10.1017/wsc.2019.27

Palaiokostas, C., Bekaert, M., Davie, A., Cowan, M. E., Oral, M., Taggart, J. B., . . . Migaud, H. (2013). Mapping the sex determination locus in the Atlantic halibut (*Hippoglossus hippoglossus*) using RAD sequencing. *BMC Genomics, 14*. doi:10.1186/1471-2164-14-566

Palmer, D. H., Rogers, T. F., Dean, R., & Wright, A. E. (2019). How to identify sex chromosomes and their turnover. *Molecular Ecology, 28*(21), 4709-4724. doi:10.1111/mec.15245

Piry, S., Alapetite, A., Cornuet, J. M., Paetkau, D., Baudouin, L., & Estoup, A. (2004). GENECLASS2: A software for genetic assignment and first-generation migrant detection. *Journal of Heredity, 95*(6), 536-539. doi:10.1093/jhered/esh074

Prentout, D., Razumova, O., Rhone, B., Badouin, H., Henri, H., Feng, C., . . . Marais, G. A. B. (2020). An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of *Cannabis sativa*. *Genome Research, 30*(2), 164-172. doi:10.1101/gr.251207.119

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rocco, L., Bencivenga, S., Archimandritis, A., & Stingo, V. (2009). Molecular characterization and chromosomal localization of spermatogenesis related sequences in *Torpedo torpedo* (Chondrichthyes, Torpediniformes). *Marine Genomics, 2*(2), 99-102. doi:10.1016/j.margen.2009.06.001

Rousset, F. (2008). GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources, 8*(1), 103-106. doi:10.1111/j.1471-8286.2007.01931.x

Rovatsos, M., Rehak, I., Velensky, P., & Kratochvil, L. (2019). Shared Ancient Sex Chromosomes in Varanids, Beaded Lizards, and Alligator Lizards. *Molecular Biology and Evolution, 36*(6), 1113-1120. doi:10.1093/molbev/msz024

Rovatsos, M., Vukic, J., & Kratochvil, L. (2016). Mammalian X homolog acts as sex chromosome in lacertid lizards. *Heredity, 117*(1), 8-13. doi:10.1038/hdy.2016.18

Russell, J. C., & Fewster, R. M. (2009). Evaluation of the linkage disequilibrium method for estimating effective population size. In D. L. Thomson, E. G. Cooch, & M. J. Conroy (Eds.), *Modeling Demographic Processes in Marked Populations* (pp. 291–320). New York, NY: Springer Verlag.

Shi, X., Waiho, K., Li, X. C., Ikhwanuddin, M., Miao, G. D., Lin, F., . . . Ma, H. Y. (2018). Female-specific SNP markers provide insights into a WZ/ZZ sex determination system for mud crabs *Scylla paramamosain*, *S. tranquebarica* and *S. serrata* with a rapid method for genetic sex identification. *BMC Genomics, 19*. doi:10.1186/s12864-018-5380-8

Slembrouck, J., Arifin, O. Z., Pouil, S., Subagja, J., Yani, A., Kristanto, A. H., & Legendre, M. (2019). Gender identification in farmed giant gourami (*Osphronemus goramy*): A methodology for better broodstock management. *Aquaculture, 498*, 388-395. doi:10.1016/j.aquaculture.2018.08.056

Stingo, V., & Rocco, L. (2001). Selachian cytogenetics: a review. *Genetica, 111*(1-3), 329-347. doi:10.1023/a:1013747215866

Stovall, W. R., Taylor, H. R., Black, M., Grosser, S., Rutherford, K., & Gemmell, N. J. (2018). Genetic sex assignment in wild populations using genotyping-by-sequencing data: A statistical threshold approach. *Molecular Ecology Resources, 18*(2), 179-190. doi:10.1111/1755-0998.12767

Suda, A., Nishiki, I., Iwasaki, Y., Matsuura, A., Akita, T., Suzuki, N., & Fujiwara, A. (2019). Improvement of the Pacific bluefin tuna (*Thunnus orientalis*) reference genome and development of male-specific DNA markers. *Scientific Reports, 9*. doi:10.1038/s41598-019-50978-4

Trenkel, V., & Lorance, P. (2020). SNPs for thornback ray *Raja clavata* used for identifying sex-linked SNPs and carrying out sex assignment. *SEANOE. https://doi.org/10.17882/74390*.

Trenkel, V. M., Vaz, S., Albouy, C., Brind'Amour, A., Laffargue, P., Romagnan, J.-B., & Lorance, P. (2019). We can reduce the impact of monitoring on marine living resources. *Marine Ecology Progress Series, 609*, 277–282. doi:https://doi.org/10.3354/meps12834

Utsunomia, R., Scacchetti, P. C., Hermida, M., Fernandez-Cebrian, R., Taboada, X., Fernandez, C., . . . Martinez, P. (2017). Evolution and conservation of Characidium sex chromosomes. *Heredity, 119*(4), 237-244. doi:10.1038/hdy.2017.43

Vicoso, B., Emerson, J. J., Zektser, Y., Mahajan, S., & Bachtrog, D. (2013). Comparative Sex Chromosome Genomics in Snakes: Differentiation, Evolutionary Strata, and Lack of Global Dosage Compensation. *Plos Biology, 11*(8). doi:10.1371/journal.pbio.1001643

Data Accessibility

The data and sequences for identified SNPs are available at www.seanoe.org https://doi.org/10.17882/74390.

Author Contributions

VT and PL designed the study, VT carried out the statistical analyses and VVB the BLAST search. All authors contributed to the manuscript and critically revised it.

Table 1. Expected heterozygosity pattern of a sex-linked SNP for different sex determination systems and locations of the SNP. ♂ ≠ ♀ means the proportion of heterozygous individuals differs between sexes.

SNP type	Location of the SNP	Genetic sex determination system		
		XX/XY	ZW/ZZ	Polygenic on different chromosomes
1	X or Z chromosome	♂ no heterozygotes ♀ heterozygotes	♂ heterozygotes ♀ no heterozygotes	-
2	Y or W chromosome	♂ no heterozygotes ♀ absent	♂ absent ♀ no heterozygotes	-
3	divergence between X-Y, Z-W, or other chromosome pairs	♂ ≠ ♀	♂ ≠ ♀	♂ ≠ ♀

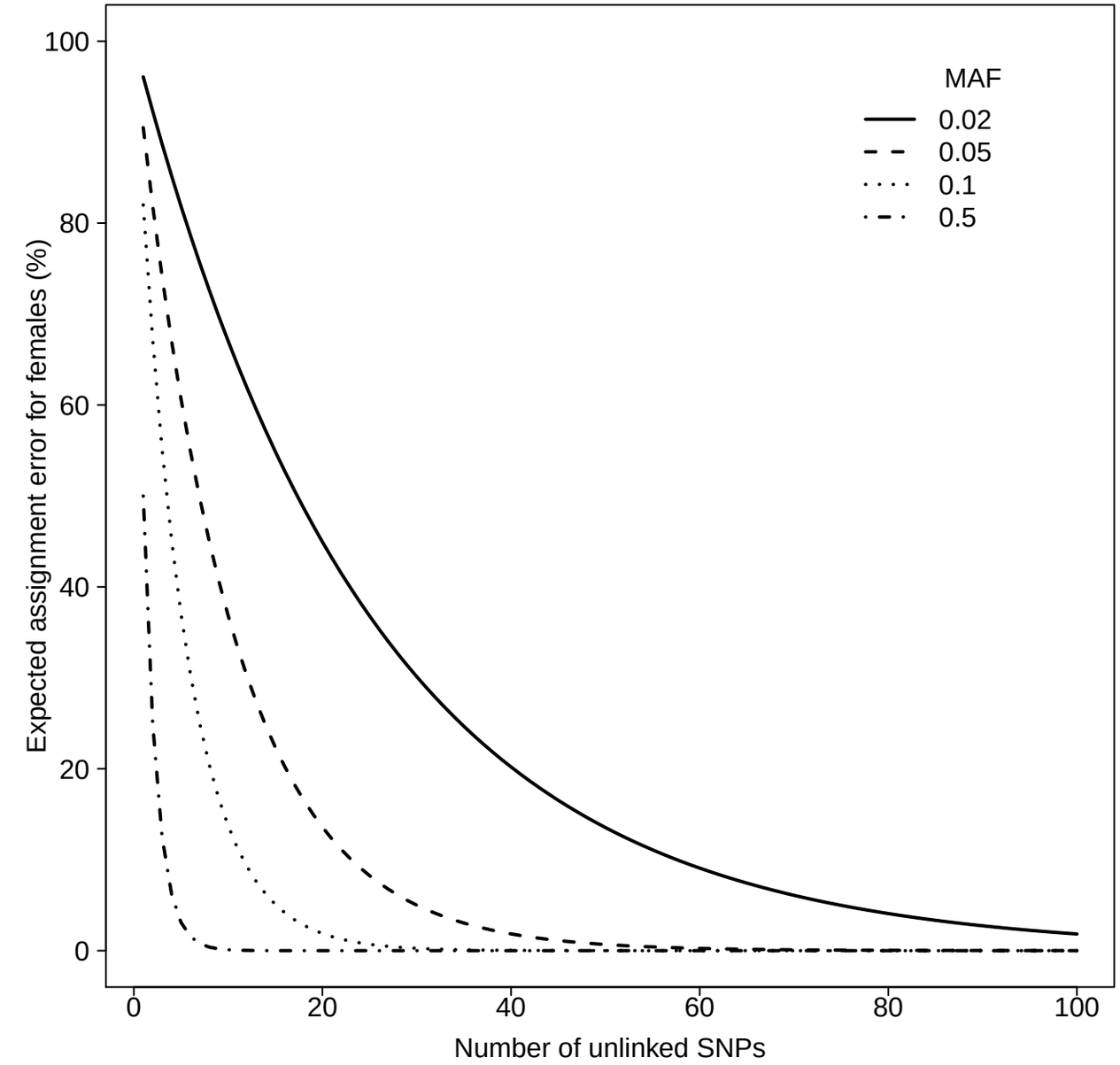
Figures

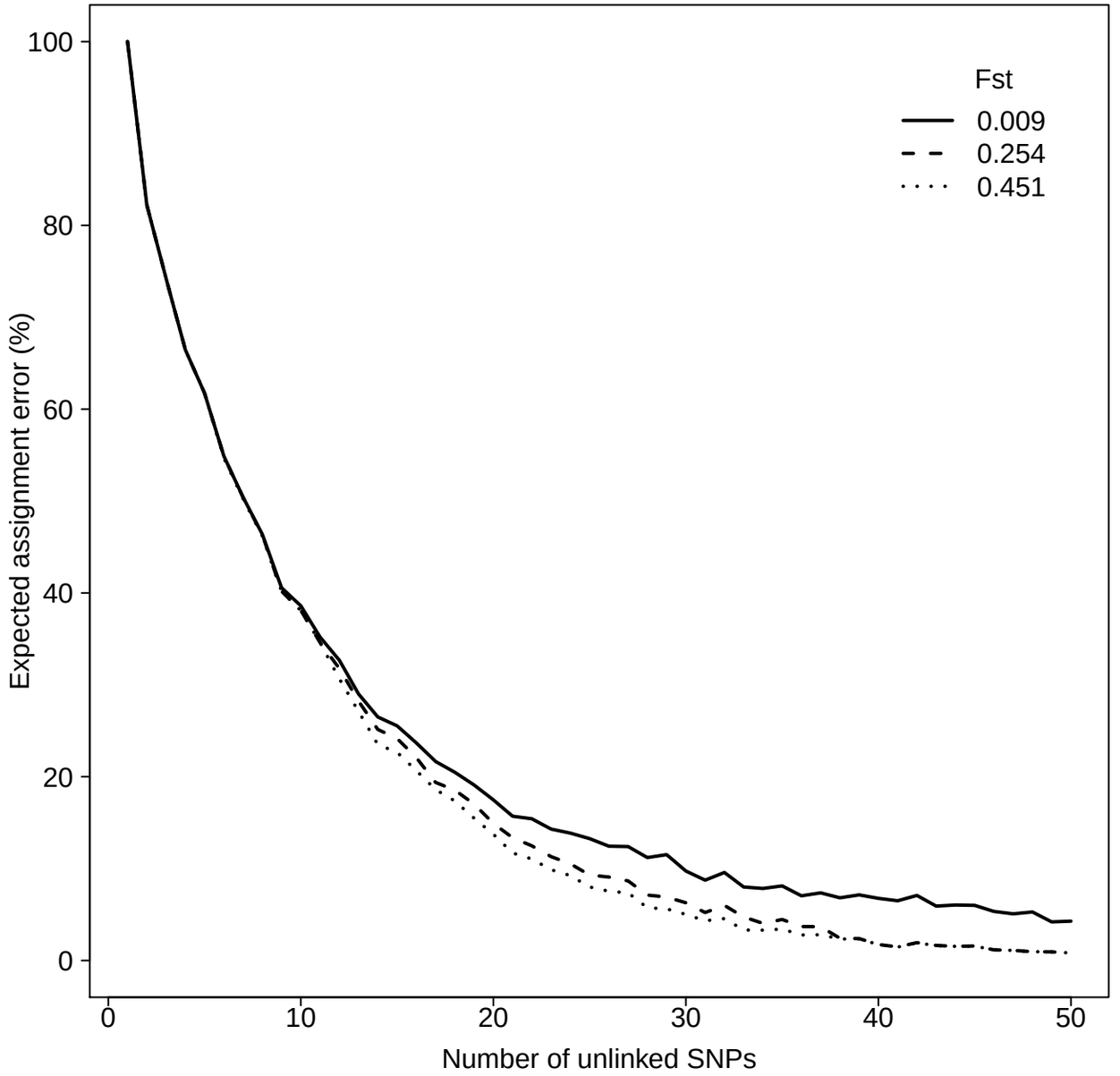
Figure 1. Expected sex assignment error. a) Assignment error for females for the heterozygosity method, for which all individuals with at least one heterozygous sex-linked SNP are assigned the female sex as a function of the number sex-linked SNPs and their minor allele frequency (MAF). b) Assignment error for the F_{ST} method, for which most likely sex is assigned given the likelihood derived from sex-specific minor allele frequencies as a function of the number of SNPs under sex selection and the F_{ST} value.

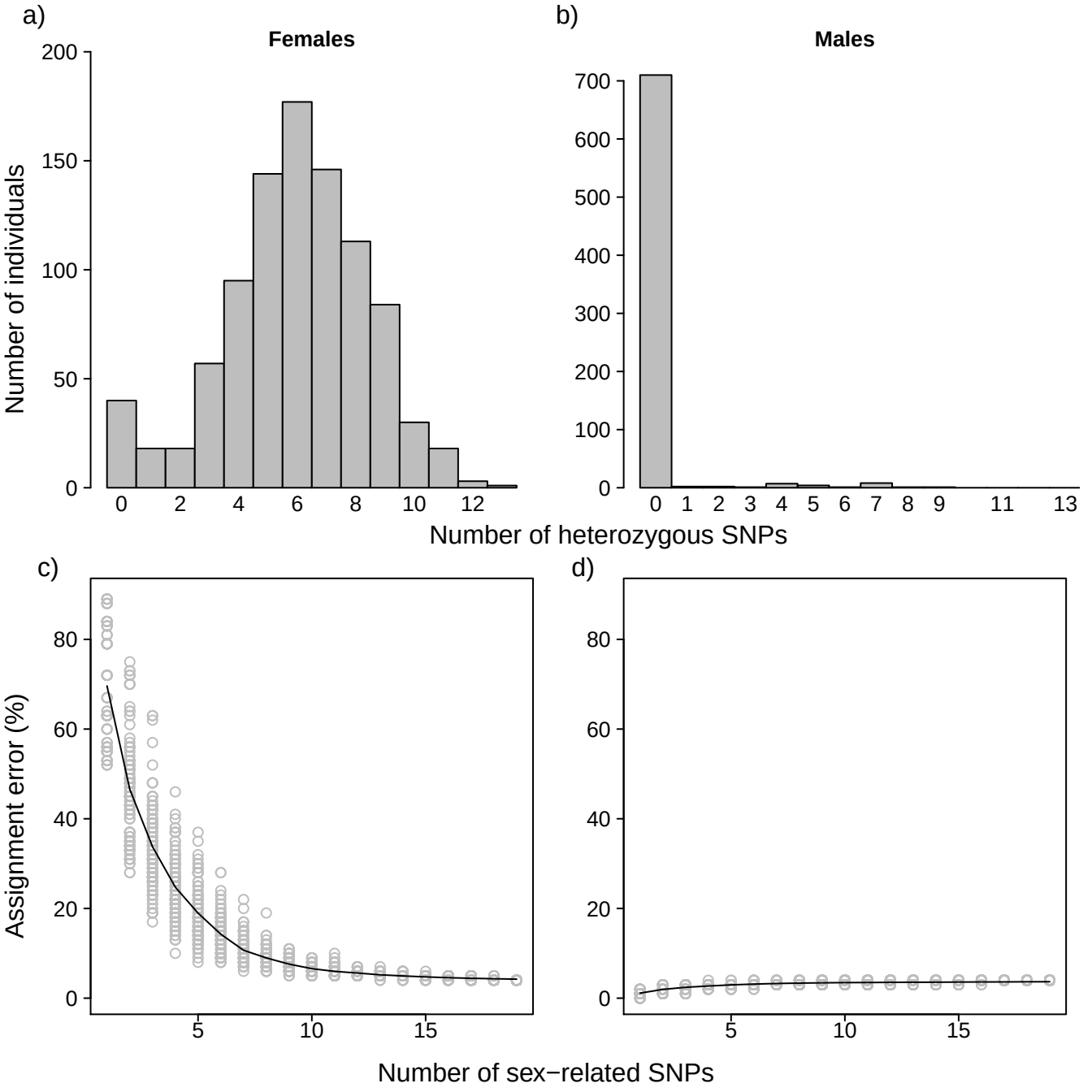
Figure 2. Histograms of the number of heterozygous individuals for a) females and b) males for 19 putative sex-linked SNPs and assignment error for c) females and d) males as a function of the number of SNPs (100 random SNPs combinations) for thornback ray in the Bay of Biscay.

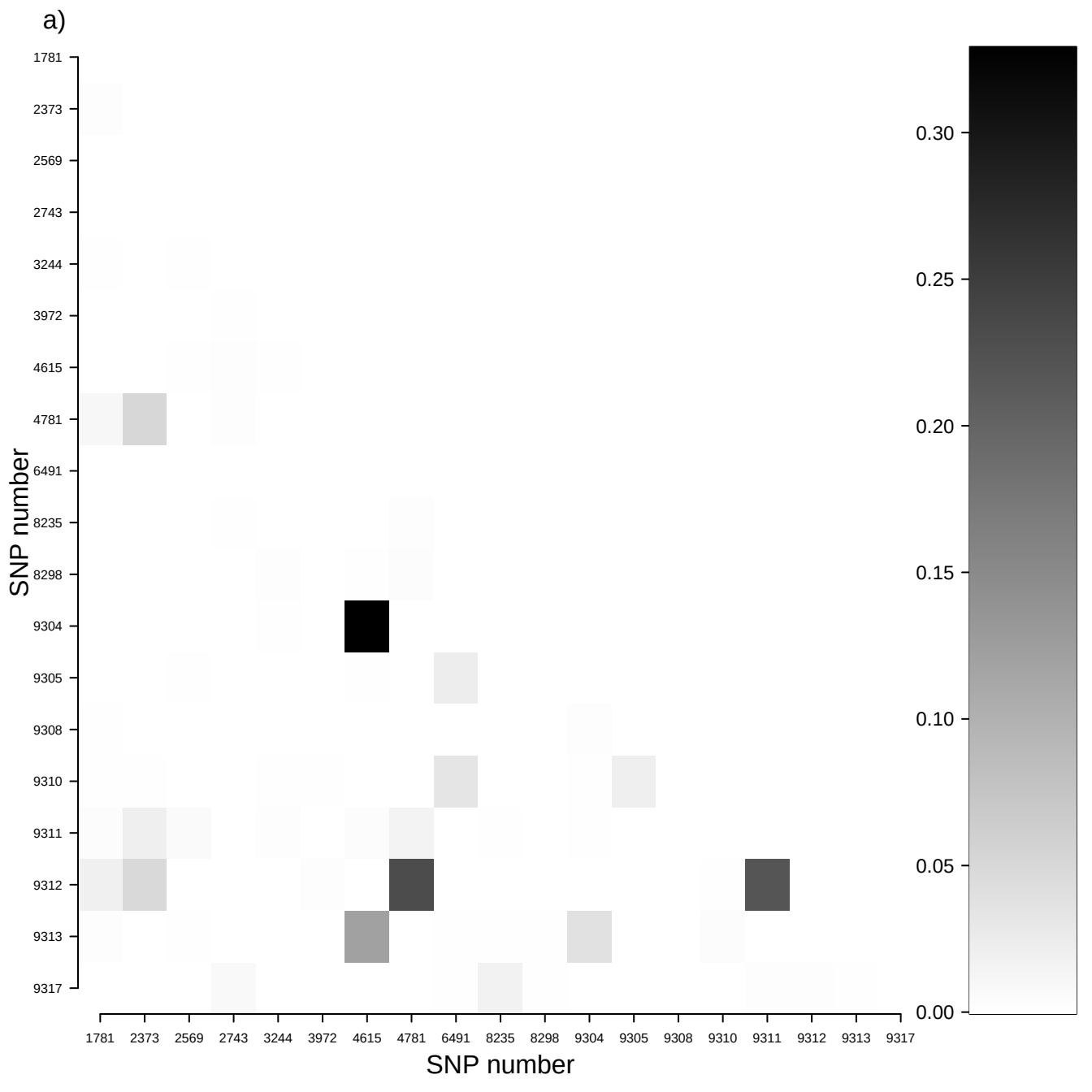
Figure 3. a) Linkage-disequilibrium based pairwise squared correlation coefficients between putative sex-linked SNPs. b) Histogram of pairwise correlations between putative sex-linked and between all other SNPs.

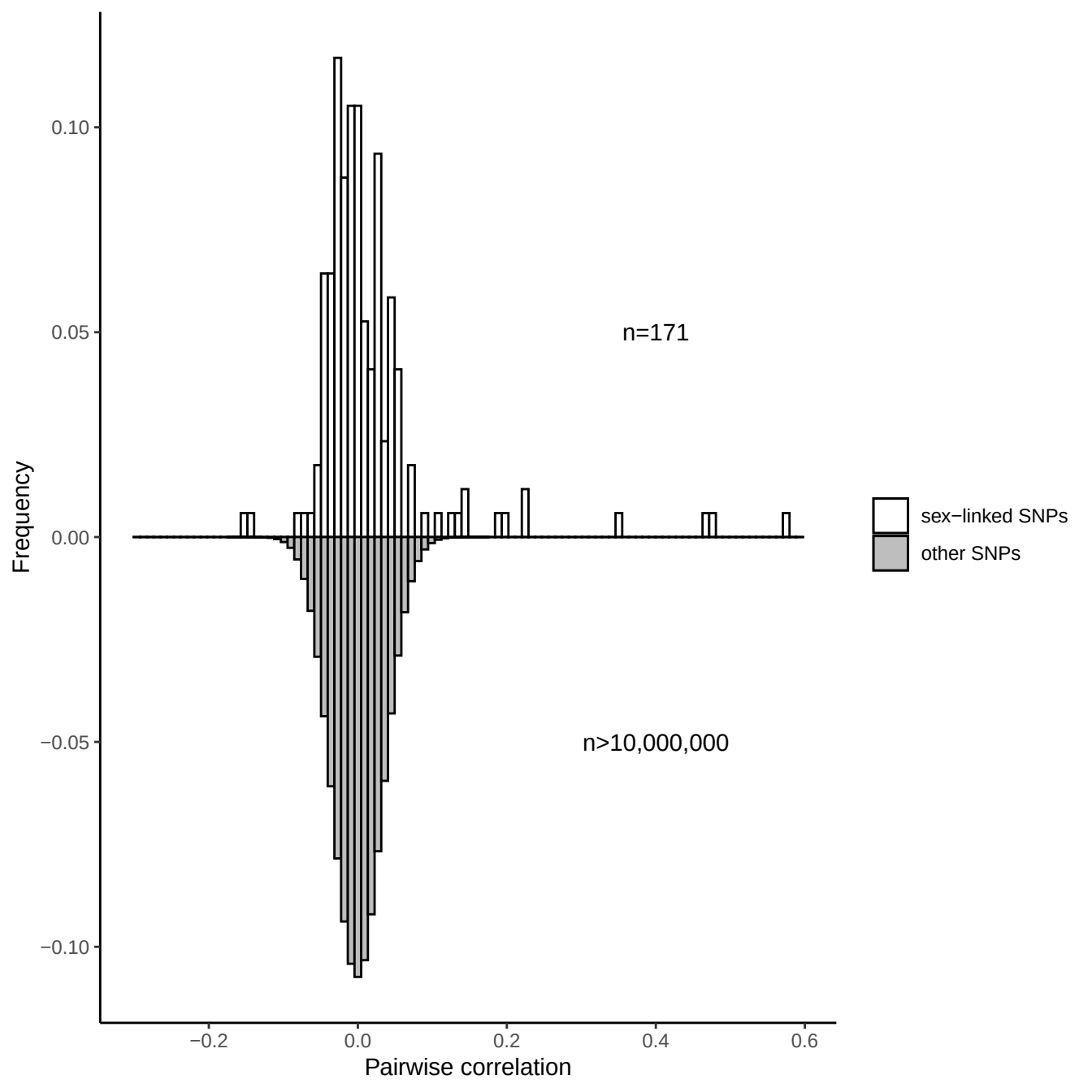
Figure 4. Sex assignment error for a neural net with one hidden layer and 1 to 10 neurons fitted to SNP data for male and female thornback ray in the Bay of Biscay. Results in grey are for ten random training and test datasets, black lines show mean values.

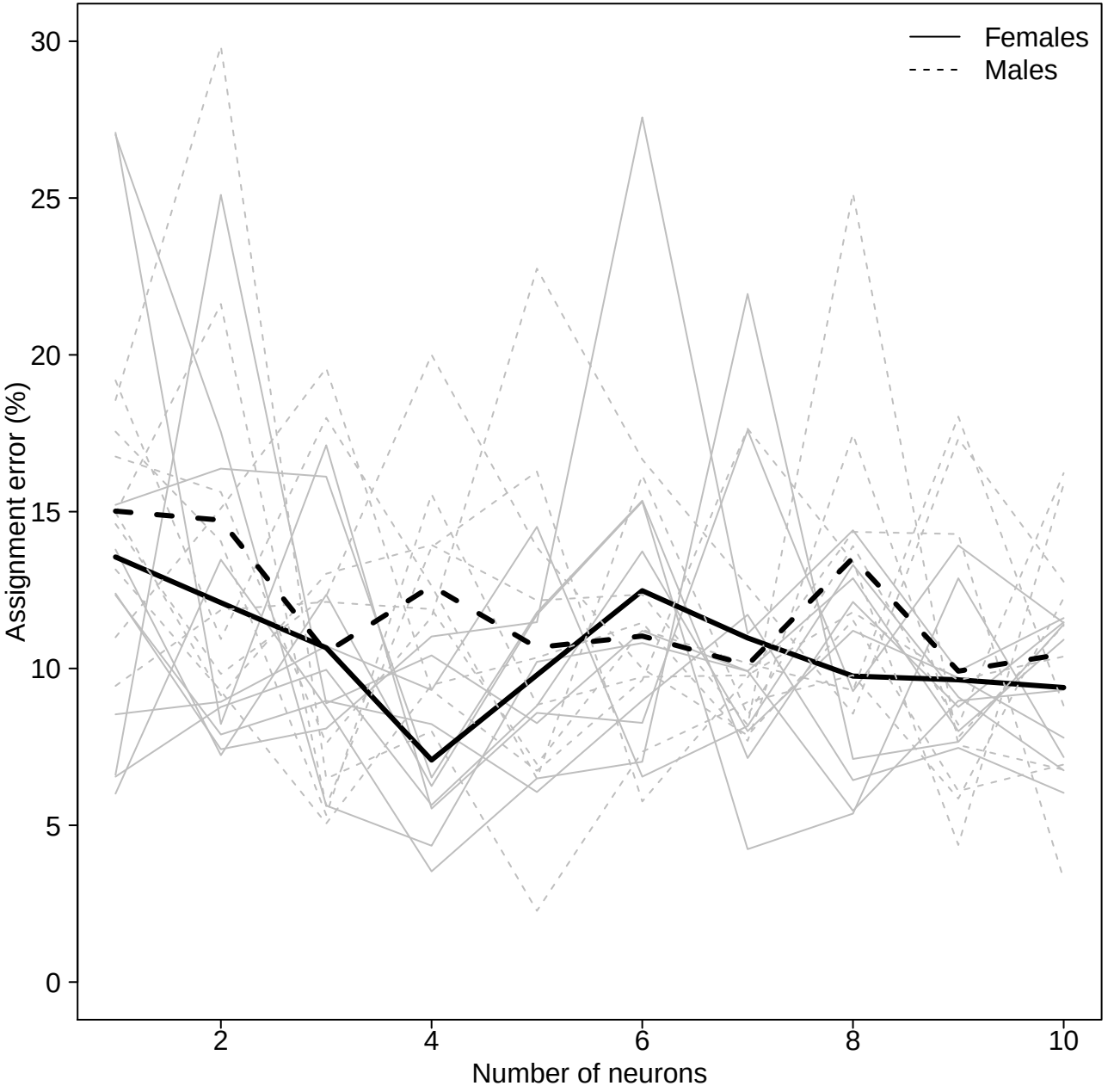

Expected assignment error for females (%)
Number of unlinked SNPs
MAF
0.02
0.05
0.1
0.5
0
20
40
60
80
100


Expected assignment error (%)
Number of unlinked SNPs
Fst
0.009
0.254
0.451
0
10
20
30
40
50
0
20
40
60
80
100


a)
Females
b)
Males
Number of individuals
Number of heterozygous SNPs
c)
d)
Assignment error (%)
Number of sex−related SNPs


a)
SNP number
1781
2373
2569
2743
3244
3972
4615
4781
6491
8235
8298
9304
9305
9308
9310
9311
9312
9313
9317
SNP number
1781 2373 2569 2743 3244 3972 4615 4781 6491 8235 8298 9304 9305 9308 9310 9311 9312 9313 9317
0.30
0.25
0.20
0.15
0.10
0.05
0.00


0.10
0.05
0.00
−0.05
−0.10
Frequency
n=171
n>10,000,000
sex−linked SNPs
other SNPs
−0.2
0.0
0.2
0.4
0.6
Pairwise correlation


Females
Males
Assignment error (%)
0
5
10
15
20
25
30
2
4
6
8
10
Number of neurons