Received: 19 March 2024 | Revised: 27 May 2024 | Accepted: 28 May 2024

DOI: 10.1111/1755-0998.13985

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

SLRfinder: A method to detect candidate sex-linked regions with linkage disequilibrium clustering

Xueling Yi[1] | Petri Kemppainen[1] | Juha Merilä[1,2]

[1]Area of Ecology and Biodiversity, School of Biological Sciences, The University of Hong Kong, Hong Kong SAR, Hong Kong

[2]Ecological Genetics Research Unit, Organismal and Evolutionary Biology Programme, University of Helsinki, Helsinki, Finland

Correspondence
Xueling Yi and Petri Kemppainen, Area of Ecology and Biodiversity, School of Biological Sciences, The University of Hong Kong, Hong Kong SAR, Hong Kong.
Email: xuelingyi5@gmail.com and petrikemppainen2@gmail.com

Funding information
National Natural Science Foundation of China/Research Grants Council (RGC) Joint Research Scheme, Grant/Award Number: N_HKU763/21

Handling Editor: Nick Hamilton Barton

Abstract

Despite their critical roles in genetic sex determination, sex chromosomes remain unknown in many non-model organisms, especially those having recently evolved sex-linked regions (SLRs). These evolutionarily young and labile sex chromosomes are important for understanding early sex chromosome evolution but are difficult to identify due to the lack of Y/W degeneration and SLRs limited to small genomic regions. Here, we present SLRfinder, a method to identify candidate SLRs using linkage disequilibrium (LD) clustering, heterozygosity and genetic divergence. SLRfinder does not rely on specific sequencing methods or a specific type of reference genome (e.g., from the homomorphic sex). In addition, the input of SLRfinder does not require phenotypic sexes, which may be unknown from population sampling, but sex information can be incorporated and is necessary to validate candidate SLRs. We tested SLRfinder using various published datasets and compared it to the local principal component analysis (PCA) method and the depth-based method Sex Assignment Through Coverage (SATC). As expected, the local PCA method could not be used to identify unknown SLRs. SATC works better on conserved sex chromosomes, whereas SLRfinder outperforms SATC in analysing labile sex chromosomes, especially when SLRs harbour inversions. Power analyses showed that SLRfinder worked better when sampling more populations that share the same SLR. If analysing one population, a relatively larger sample size (around 50) is needed for sufficient statistical power to detect significant SLR candidates, although true SLRs are likely always top-ranked. SLRfinder provides a novel and complementary approach for identifying SLRs and uncovering additional sex chromosome diversity in nature.

KEYWORDS
heterozygosity, inversion, LD, sex chromosomes, sex-determining region, SLR

Xueling Yi and Petri Kemppainen contributed equally to this work.

1 | INTRODUCTION

Sex chromosomes play critical roles in genetic sex determination and yet remain unknown in many non-model organisms. Early studies in mammals and birds have demonstrated highly conserved and heteromorphic (i.e., having different morphologies) sex chromosomes with conserved sex-determining genes and degenerated Y or W chromosomes (Cortez et al., 2014). On the contrary, accumulating studies have found less conserved but much more labile sex chromosomes that may be different between closely related lineages in fish, amphibians, reptiles and some invertebrates (Dufresnes et al., 2015; Furman et al., 2020; Hearn et al., 2022; Jeffries et al., 2018; Myosho et al., 2012; Ogata et al., 2021; Tree of Sex Consortium, 2014; Vicoso, 2019; Yi et al., 2024). These labile sex chromosomes tend to be homomorphic (i.e., sex chromosomes having indistinguishable morphologies) and are featured by little or no degeneration, low inter-sex differentiation, variable sex-determining genes and sex-linked regions (SLRs) restricted to narrow genomic regions. These features make labile sex chromosomes and their SLRs difficult to identify using traditional methods such as karyotyping and PCR of conserved sex-determining genes (Palmer et al., 2019; Tree of Sex Consortium, 2014). However, labile sex chromosomes likely represent early evolutionary stages of sex chromosome evolution and their study is critical for our understanding of sex chromosome evolution (Blaser et al., 2014; Furman et al., 2020; Perrin, 2021; Vicoso, 2019). Therefore, additional work is needed to identify labile sex chromosomes and their SLRs in non-model species.

Recently, several methods have been developed to help identify sex chromosomes and their SLRs, but these methods mostly work for conserved sex chromosomes and are limited to certain types of sequencing data. For example, RADSex (Feron et al., 2021) was developed to identify sex determination systems (i.e., XX/XY or ZZ/ZW) and sex-linked markers of labile sex chromosomes specifically from restriction site-associated DNA sequencing (RADseq) data, and Pooled Sequencing Analysis for Sex Signal (PSASS ver. 3.1.0; https://github.com/SexGenomicsToolkit/PSASS) was developed to detect sex-linked signals by comparing pooled sequencing data from males and females (e.g., in Kitano et al., 2023). These methods are not applicable to individual-level whole-genome sequencing (WGS) data, which has been increasingly used in studies of non-model species. In addition, these methods require known phenotypic sexes which may not be available in non-invasive sampling or may be difficult to identify in individuals that are not sexually mature or have limited or no sexual dimorphism. FindZX (Sigeman et al., 2022) was developed to detect sex chromosomes using WGS data. This method has been applied to diverse systems including both conserved and labile sex chromosomes, and it can work on very small sample sizes (Sigeman et al., 2022). However, this method also relies on known phenotypic sexes, and it requires a reference genome of the homogametic sex (i.e., XX female or ZZ male), which may not be available or may be unknown when the sex determination system is unclear. Sex Assignment Through Coverage (SATC) (Nursyifa et al., 2022) was developed to jointly identify sex chromosomes and genetic sex using WGS data. This method does not require known phenotypic sexes, but it assumes that only X/Z scaffolds are assembled in the reference genome, which is practically the same as requiring a reference genome of the homogametic sex. In addition, these available methods are mostly based on sequencing depth (RADSex and SATC) or depth and heterozygosity (PSASS, FindZX), but many studies have shown that depth may not differ between sexes on labile sex chromosomes that are homomorphic and have narrow SLRs (Jeffries et al., 2022; Yi et al., 2024). Therefore, new methods are needed to help identify labile sex chromosomes in non-model species.

A previous study showed that sex chromosomes leave distinct patterns of linkage disequilibrium (LD) in population genomic data (McKinney et al., 2020). Linkage disequilibrium refers to the correlation between alleles at different loci: stronger correlation indicates higher LD and lower recombination rates (Barton, 2011; Kemppainen et al., 2015). LD can be caused by population demography (inbreeding, admixture and drift) and selection, and the decay of LD is modulated by evolutionary processes that affect recombination rates (e.g., inversions and sex chromosomes). Although some studies have demonstrated high LD in SLRs on labile sex chromosomes (Hearn et al., 2022; McKinney et al., 2020), LD has remained under-exploited in studies of SLR identification and sex chromosome evolution. Here, we present a method (SLRfinder) to identify candidate SLRs among LD clusters of highly correlated single-nucleotide polymorphisms (SNPs) based on the differentiation in heterozygosity and the genetic variation captured by principal component analysis (PCA). LD clusters from SLRs are expected to have the strongest LD due to recombination suppression between sex chromosomes, different individual heterozygosity between homogametic and heterogametic sexes, and the clearest genetic divergence between sexes captured by PCA. Unique patterns of regional PCA on conserved sex chromosomes were also detected in a recent study of cuckoos (Merondun et al., 2024) using the local PCA method (Li & Ralph, 2019). However, local PCA may not be able to identify unknown SLRs or labile sex chromosomes without the additional signals (LD and heterozygosity) used in SLRfinder. SLRfinder is also expected to outperform the depth-based methods (e.g., SATC) in identifying homomorphic sex chromosomes that tend to have similar depths of coverage in males and females, and it does not rely on specific types of sequencing methods or a reference genome of the homomorphic sex.

Below, we describe the workflow of SLRfinder and its application to published datasets of various taxa having identified labile sex chromosomes, including nine-spined sticklebacks (*Pungitius pungitius*), chum salmon (*Oncorhynchus keta*), guppies (*Poecilia reticulata*) and intertidal snails (*Littorina saxatilis*). We also tested the effectiveness of SLRfinder in conserved sex chromosomes using a dataset of African leopards (*Panthera pardus*). In addition, we compared the performance of SLRfinder to the local PCA method, which also relies on PCA patterns, and SATC, a depth-based method that also does not require known phenotypic sexes in the input. The results show that, as expected, the local PCA method could not be used to identify unknown SLRs. SATC only worked on conserved sex chromosomes and might yield wrong sex inferences when using a reference

genome of the heterogametic sex. On the contrary, SLRfinder does not rely on specific types of reference genomes and it outperforms SATC in analysing labile sex chromosomes, especially when the SLR is associated with genomic inversions. Since SLRfinder and SATC are based on independent signals (i.e., LD and heterozygosity versus depths of coverage), they are complementary to each other and should thus be considered jointly to maximize the ability to identify sex chromosomes in non-model species.

2 | MATERIALS AND METHODS

2.1 | Identify LD clusters from VCF inputs

The workflow of SLRfinder is summarized in Figure 1. The input data is a VCF file of filtered biallelic SNPs from populations genotyped using WGS or reduced-representation sequencing methods (e.g., RADseq). LD is estimated in VCFtools (Danecek et al., 2011) as squared coefficient of correlation (r^2) between pairs of loci within windows of 100 SNPs (--geno-r2 --ld-window 100). LD clusters are identified using a network analytical framework (Kemppainen et al., 2015) as illustrated in Figure S1. Briefly, pairs of highly correlated loci ($r^2 > min\_LD$, default 0.85) are extracted to generate a 'graph object' using the function *graph.edgelist* from the package igraph (Csardi and Nepusz, 2006) in R v4 (R Core Team, 2022). The graph object is further decomposed into separate LD clusters using the function *decompose.graph*. Loci belonging to the same LD cluster always have $r^2 > min\_LD$, whereas loci from different LD clusters must have $r^2 \leq min\_LD$ (Figure S1). Finally, only clusters with a minimum size (i.e., number of SNPs, default $min.cl.size \geq 20$) are retained for downstream analyses. Because *min_LD* and *min.cl.size* of LD clusters depend on the input SNP density, lower thresholds may be applied to non-WGS datasets that have fewer loci (e.g., from reduced-representation sequencing).

2.2 | Estimate heterozygosity and conduct PCA in each LD cluster

All SNPs from each identified LD cluster are used to conduct a PCA using the R package SNPRelate (Zheng et al., 2012) and estimate the observed heterozygosity as the proportion of heterozygous SNPs in the non-missing SNPs genotyped in each individual. A linear model is fitted to regress the estimated individual heterozygosity on scaled PC1 (also polarized if the original relationship is negative). The heterozygosity~PC1 plots are expected to show no grouping pattern in most LD clusters, three groups in a triangular shape representing three genotypes in autosomal inversions (Ma & Amos, 2012), and two groups corresponding to the homogametic sex (bottom-left corner) and the heterogametic sex (top-right corner) in the SLRs that are shared among individuals in the dataset (Figure 1a). Accordingly, candidate SLRs are expected to have the strongest association between heterozygosity and PC1, which is estimated by the adjusted R-squared value of the linear regression. Candidate SLRs are also expected to show stronger inter-sex than inter-population genetic divergence on PC1, which is estimated by the χ^2 goodness-of-fit tests on an equal separation (assuming equal sex ratio) of samples in each population on PC1. Smaller χ^2 statistics indicate that all populations include individuals from both bottom-left and top-right groups that represent two sexes on the heterozygosity~PC1 plot of SLRs, therefore indicating potentially stronger inter-sex differentiation than population structure in this region. If input datasets have skewed sex ratios, user-specified probabilities of sampling heterogametic and homogametic groups can be provided to get more accurate χ^2 estimates.

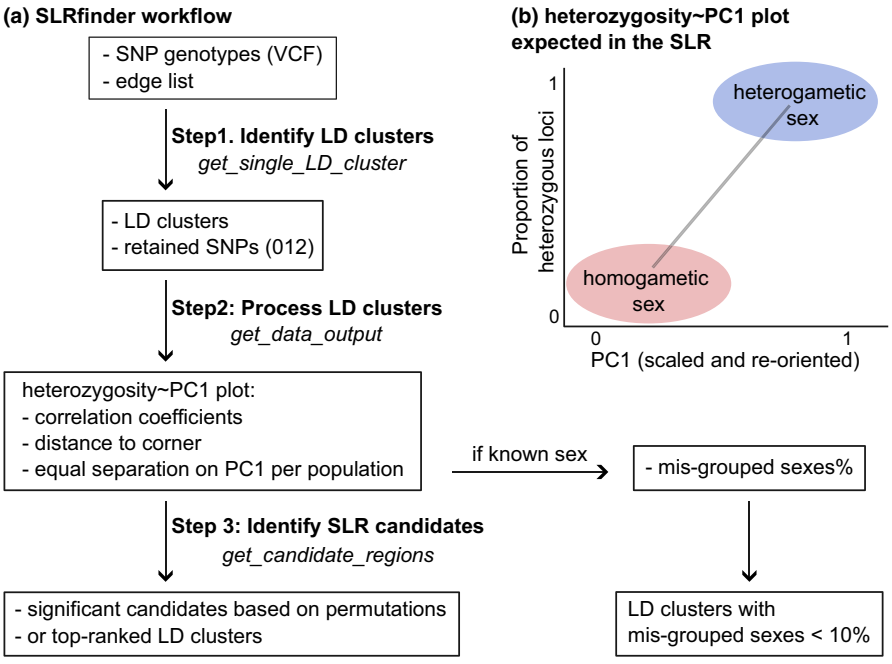


FIGURE 1 Workflow of SLRfinder. (a) The three major steps of SLRfinder. (b) Illustration of the expected heterozygosity~PC1 plot in the sex-linked region (SLR).

Candidate SLRs are also expected to show clear separation between and tight clustering within groups on the heterozygosity~PC1 plot, which is estimated by the scaled Euclidean distance between each individual and its nearest corner individual (i.e., the individuals having the highest or lowest heterozygosity, and if equal heterozygosity the highest or lowest scaled PC1 scores).

2.3 | Identify SLR candidates

Based on the above expectations, we identify candidate SLRs among LD clusters using ranks of the estimated parameters. An LD cluster is ranked higher (i.e., more likely to be a SLR) if it has a larger size (i.e., more SNPs), stronger heterozygosity~PC1 regression, smaller variation of the Euclidean distance (i.e., better grouping on the heterozygosity~PC1 plot) and smaller χ^2 statistic (i.e., stronger inter-sex divergence than population structure on PC1). The summed ranks of these parameters are permuted (default 10,000 times) among LD clusters to generate a null distribution of the summed ranks and estimate how often the permuted values are lower than the observed value (i.e., the *p*-value) of each LD cluster. In addition, we correct potential *p*-value inflation using genomic control (Devlin et al., 2001; Devlin & Roeder, 1999). The $-\log_{10}(p)$ values are divided by the inflation factor (λ) estimated as the linear slope in a quantile-quantile plot between the observed $-\log_{10}(p)$ and those expected under the null hypothesis of a uniform distribution of *p*-values. Significant candidates (adjusted *p*-value < .05) are reported with their heterozygosity~PC1 plots. If no statistical significance is detected, the five top-ranked LD clusters and their plots are reported.

Although SLRfinder does not require phenotypic sexes in the input, known sex information can be incorporated to filter LD clusters where the two sexes are fully separated on the heterozygosity~PC1 plot, which can provide additional inference on candidate sex chromosomes. To do this, we estimate the percentage of sexed individuals that are likely placed in the wrong group (i.e., the minority sex in a group is regarded as misplaced). The LD clusters that have less than 10% misplaced individuals (allowing for rare phenotypic misidentifications; the 10% threshold can be modified using the sex_filter parameter) are reported as candidate SLRs.

2.4 | Test of SLRfinder using published datasets

To test the efficiency and accuracy of SLRfinder, we applied it to published empirical datasets of various species (Table S1). First, we applied SLRfinder to nine-spined sticklebacks (*Pungitius pungitius*) that have heteromorphic sex chromosomes (SLR identified as LG12:1–16900000; Dixon et al., 2019; Kivikoski et al., 2021; Natri et al., 2019) in the non-European and Eastern European (EL) lineages, whereas homomorphic sex chromosomes (SLR identified as LG3:17260000–17340000; Yi et al., 2024) in the Western European lineage (WL). Both lineages have the XX/XY sex determination. In addition, two UK populations have unidentified sex chromosomes, and a hybrid Polish population was identified with both types of sex chromosomes (10 LG3-determined males and 1 LG12-determined male; Yi et al., 2024). The WGS data of nine-spined sticklebacks were published in a previous study (Feng et al., 2022) and available on ENA (project PRJEB39599). Raw sequencing data were re-mapped to the version 7 reference genome of *Pungitius pungitius* (GCA_902500615.3; Kivikoski et al., 2021) using bwa-mem in BWA v0.7.17 (Li, 2013), sorted and indexed using SAMtools version 1.16.1 (Danecek et al., 2021), and genotyped by Genome Analysis ToolKit (GATK) following the best practice protocol (Depristo et al., 2011; Van der Auwera et al., 2013). Biallelic SNPs were extracted using commands -m2 -M2 -v snps –min-ac=1 in BCFtools (Li, 2011) and data mapped to unassembled contigs were removed. The SNP genotypes were split into five datasets representing the WL, EL, non-European, UK and Polish populations. Each dataset was further filtered in VCFtools by quality (--minGQ 20 –minQ 30), missing data (--max-missing 0.75) and minor allele frequency (--maf 0.15) before being analysed by SLRfinder using default settings. The same SNP filtering was used below in the other test datasets using WGS. Phenotypic sexes are known in one EL and one WL population and were provided to SLRfinder. The previously identified genetic sexes (Yi et al., 2024) were used when analysing the Polish population.

Next, we applied SLRfinder to chum salmon (*Oncorhynchus keta*) whose sex chromosomes (XX/XY) have been identified as LG15 both using RADseq (SLR unspecified; McKinney et al., 2020) and using WGS data (SLR identified as LG15:40010001–46610001; Rondeau et al., 2023). We re-analysed both datasets using SLRfinder. The WGS data were mapped to the newly assembled male reference genome of *Oncorhynchus keta* (GCF_023373465.1), and the VCF file of genotyped biallelic SNPs was downloaded from the corresponding publication (Rondeau et al., 2023) and filtered before being analysed by SLRfinder. In addition, to test the potential influence of different reference genomes, we downloaded the raw WGS data published in Rondeau et al. (2023) from NCBI (BioProject PRJNA556729), mapped them to a female reference genome (GCF_012931545.1), and genotyped and filtered SNPs in the same way described above. The demultiplexed RADseq data published in McKinney et al. (2020) were downloaded from NCBI (BioProject PRJNA611968) and mapped to the male reference genome (GCF_023373465.1) using bwa-mem. The mapped reads were sorted, indexed and marked with duplicates using SAMtools and genotyped using the program *ref_map.pl* with default settings in Stacks 2.65 (Rochette et al., 2019). The genotyped data were further filtered using the program *populations* by minor allele frequency (--min-maf 0.15) and missing data (-R 0.75), and the ordered genotypes were output in the VCF format. We did not output a single SNP per stack locus as the following analyses are based on the information of LD. The output VCF file was analysed by SLRfinder using a lower threshold for detecting LD clusters (*min_LD* = 0.2, *min.cl.size* = 5) due to lower SNP density in RADseq data. Phenotypic sexes are known for the WGS dataset (Rondeau et al., 2023) but not the RADseq dataset (McKinney et al., 2020).

We also applied SLRfinder to datasets of guppies (*Poecilia reticulata*) whose sex chromosomes (XX/XY) have been identified as the LG12

with two SLR candidates (4,800,000–5,200,000 bp, 24,500,000–25,400,000 bp) in the newly assembled male reference genome (Fraser et al., 2020). Raw WGS data of the previously studied populations (Fraser et al., 2020; Kü Nstner et al., 2016) were downloaded from NCBI (BioProject PRJEB10680 and PRJNA238429) and mapped separately to the male reference genome (GCA_904066995.1) and a female reference genome (GCA_000633615.2) to test potential impacts of different references. Data mapping, genotyping and SNP filtering were done in the same way as in nine-spined sticklebacks. Phenotypic sexes are known for these individuals (Fraser et al., 2020) and were provided to SLRfinder.

In addition, we applied SLRfinder to a dataset of the intertidal snail, *Littorina saxatilis*, in a Swedish hybrid zone between two ecotypes. The crab ecotype found in shores sheltered from waves was identified with the ZZ/ZW system and the LG12 sex chromosomes (SLR unspecified), whereas the wave ecotype had unidentified sex chromosomes that were not LG12 (Hearn et al., 2022). Raw WGS data of these individuals have been published (Westram et al., 2018) and were downloaded from NCBI (BioProject PRJNA483347) and mapped to the male reference genome of *Littorina saxatilis* (GCA_037325665.1). Individuals were split into two ecotype-specific datasets based on their relative position on the transect: individuals at <68 m to the main transition were considered as crab ecotypes, whereas individuals at >88 m to the main transition were considered as wave ecotypes (Hearn et al., 2022; Westram et al., 2018). Individuals of the hybrid ecotype were excluded from analyses for clarity. Data mapping, genotyping and SNP filtering were the same as above and done for each dataset independently. Each dataset was treated as one population and phenotypic sexes (Westram et al., 2018) were provided to SLRfinder.

Lastly, we applied SLRfinder to African leopards (*Panthera pardus*), which have conserved sex chromosomes (XX/XY). Due to computational constraints, we only analysed the WGS data of 26 individuals published in a previous study (Pečnerová et al., 2021). The raw data were downloaded from NCBI (BioProject PRJEB41230) and mapped to a scaffold-level female reference genome of *Panthera pardus* (GCF_001857705.1). Scaffolds from sex chromosomes were indicated using SATC in previous studies (Nursyifa et al., 2022; Pečnerová et al., 2021). Data mapping, genotyping and SNP filtering were done in the same way as in nine-spined sticklebacks. Because sample information was not provided for the raw sequencing data, we assigned these individuals into genetic populations based on PCA using separately filtered biallelic SNPs (--minGQ 20 –minQ 30 –maf 0.05 –max-missing 0.8). No phenotypic sexes were provided and the genetic sexes inferred by SATC (see below) were used as the sex information in SLRfinder.

2.5 | Comparing SLRfinder with the local PCA method

The local PCA method was developed to detect local variation of population structure (Li & Ralph, 2019). Here, we test whether the local PCA method can also identify unknown SLRs or sex chromosomes using the datasets of nine-spined sticklebacks, the crab ecotype of intertidal snails and African leopards. Local PCA was conducted by chromosome (or on the whole dataset of leopards) using the lostruct R package (Li & Ralph, 2019) and the same VCF inputs as in SLRfinder. The VCF files were first re-formatted using the function *read_vcf* and eigenvectors and eigenvalues were estimated using the function *eigen_windows*. We set window size at 100 SNPs in the snail_crab dataset, 500 SNPs in the stickleback_WL and stickleback_UK datasets, and 1000 SNPs in the stickleback_EL, stickleback_nonEU and leopard datasets. Distances between eigenvector/eigenvalue matrices were estimated using *pc_dist* and analysed by multidimensional scaling (MDS) using *cmdscale*. MDS plots of the first two dimensions are expected to show different patterns between autosomes and sex chromosomes, and windows from SLRs are expected to be outliers.

2.6 | Comparing SLRfinder with SATC

We also compared the effectiveness of SLRfinder with SATC (Nursyifa et al., 2022) using the above datasets, excluding the salmon WGS data mapped to the male reference genome because this dataset was a VCF file downloaded from the previous publication (Rondeau et al., 2023), and the bam files were not available. To run SATC, the depth of coverage was calculated by SAMtools-idxstats using the mapped and duplicates-marked individual bam files. Then, the idx files were processed by SATC with default settings which filter scaffolds by minimum 100 kb, normalize length by the five longest scaffolds and identify sex scaffolds by the Gaussian model.

2.7 | Test the power of SLRfinder using different sample sizes, sex ratios and SLR components

To assess the statistical power of SLRfinder, we first applied it to subsets of the WL and EL nine-spined stickleback datasets where we varied the number of individuals or populations. To test the effects of total sample size, we kept all populations and randomly selected three to five individuals per population in the WL or EL dataset. To test the effects of population size, we randomly selected 1–5 WL or EL populations and included all individuals from the selected populations. Because the stickleback datasets had around 20 samples per population, we used the snail_crab dataset to test the effects of sample size if only one population is collected. The outlier and mismatched samples (snail_ID ANG275, ANG147, ANG237 and ANG179, see results) were excluded, leaving 88 females and 60 males which were analysed again by SLRfinder. Then, we randomly selected 60, 30, 25 or 20 samples per sex to generate datasets with the equal an sex ratio but different sample sizes.

To test the effects of sex ratios, we used the previously identified genetic sexes of nine-spined sticklebacks (Yi et al., 2024) and only included the seven WL populations and the 24 EL populations

that have at least four individuals per sex. We kept the same total number of individuals ($n = 28$ in WL, $n = 96$ in EL) and modified sex ratios by randomly selecting two individuals per sex per population (even sex ratio), or one individual from one sex and three from the other in each population (sex ratios 1:3 or 3:1). For sex ratios 1:2 or 2:1, we randomly selected nine individuals from one sex and 19 from the other across WL populations, and 32 individuals from one sex and 64 from the other across EL populations. For sex ratios 1:10 or 10:1, we randomly selected three individuals from one sex and 25 from the other across WL populations, and nine individuals from one sex and 87 from the other across EL populations. We also tested extreme scenarios where only one sex was sampled in the dataset.

The subset VCF files of the selected individuals were filtered and processed by SLRfinder as described above. We first used the default expectation of an equal sex ratio in all tests. When the true SLR was not included in top-ranked candidates, we re-analysed the data with SLRfinder using the true sex ratios as the expectation in χ^2 tests to see whether SLRfinder results could be improved. If not, we further modified the rank parameters to see whether the results could be improved.

3 | RESULTS

3.1 | SLRfinder analyses of nine-spined sticklebacks, chum salmon and intertidal snails

SLRfinder successfully identified the sex chromosomes and SLRs of nine-spined sticklebacks (Table 1; Figure 2). In the WL dataset, SLRfinder identified a single significant candidate on LG3 that highly overlaps with but is narrower than the previously described WL SLR (LG3:17260000–17340000; Yi et al., 2024). Similarly, in the EL and non-European datasets, SLRfinder identified a single significant candidate on LG12 that highly overlaps with the previously reported EL SLR (Kivikoski et al., 2021). The SLRfinder-inferred genetic sexes are also consistent with known phenotypic sexes and the previous identifications of genetic sex in these populations (Yi et al., 2024). When analysing the Polish population where both types of sex chromosomes coexist (Yi et al., 2024), SLRfinder detected no statistical significance but the two top-ranked regions ($p = .16$) included the prevalent LG3 SLR which separated two sexes on the heterozygosity~PC1 plot (Table 1, Figure S2A). Filtering the percentage of misplaced sexes retained both the LG3 SLRs and a few autosomal regions (Table 1). The LG12 SLR carried by one individual in this dataset was not detected by SLRfinder (Table 1). The UK dataset did not generate significant candidates, possibly due to mixed SLRs and/or small sample sizes (see below). However, none of the top-ranked candidates in the UK dataset were located on LG12 or LG3 (Table S1), consistent with the previous findings that sex chromosomes of the UK populations are likely neither LG12 nor LG3 (Yi et al., 2024). Instead, the LD clusters having the lowest adjusted p-values ($p = .2179$) included a 225-bp region on LG7 and a 203-bp region on LG16 (Table 1, Figure S2B). Additional sampling of individuals with known sexes is required to validate whether these regions can separate the two sexes and to identify the yet unknown sex chromosomes of the UK populations.

SLRfinder also identified the sex chromosomes and SLRs of chum salmon (Table 1; Figure 2). When using the WGS data mapped to the male reference, SLRfinder identified LG15 and LG3 as significant candidates, both highly overlapping with the previously reported sex-associated regions (LG3:750001–1950001, LG15:40010001–46610001 and LG26:1–280001) in genome-wide association studies (GWAS; Rondeau et al., 2023). In addition, our identified LG15 SLRs are located in a much narrower region (1,770,035 bp including both candidates detected by ranks and sex filtering; Table S1) than the region identified by GWAS (6,600,000 bp). While LG15 was inferred as sex chromosomes by independent studies using different datasets and analyses (McKinney et al., 2020; Rondeau et al., 2023), the LG3 cluster most likely represents a true sex-linked autosomal region. Interestingly, despite the complete separation between two sexes on the heterozygosity~PC1 plot of this LG3 cluster, the few individuals of unknown sex were not grouped with either sex in the LG3 cluster but were clearly grouped with females in the significant LG15 cluster, which is the true SLR (Figure 2d). Another LG15 cluster located within the previously identified SLR (Rondeau et al., 2023) was also detected by filtering the percentage of misplaced sexes. Therefore, this cluster was a false negative ($p = .08$) with a marginal rank probably due to an outlier male individual on the heterozygosity~PC1 plot (Figure 2e). Similarly, when using the WGS data mapped to the female reference, the autosomal LG3 cluster was identified as significant and three LG15 clusters were detected by filtering the misplaced sexes but were ranked as false negatives ($p > .2$), probably due to an outlier male that had relatively low heterozygosity (Figure 2f,g, Table 1, Table S1). On the contrary, SLRfinder identified a false positive ($p = .03$) LG24 cluster, which showed a similar pattern but did not separate the two sexes on the heterozygosity~PC1 plot (Figure 2f). When using the RADseq data, no significant candidate was identified but the true LG15 SLR was the top-ranked cluster having 319 SNPs and a marginal p-value of .06 (Table 1). This false negative result was possibly due to the sparse RADseq SNPs and loose LD filtering ($min\_LD = 0.2$) of this dataset, which generated a weak grouping on the heterozygosity~PC1 plot (Figure S2C).

When analysing the crab ecotype of intertidal snails, SLRfinder correctly indicated the ZZ/ZW system where females are the heterogametic sex, and detected false negative ($p > .2$) top-ranked clusters on the sex chromosome LG12 (Figure 2h, Table 1, Table S1). Interestingly, one female (ANG179) was consistently identified as an outlier in the top-ranked SLRs (Figure 2h). Because samples were collected along a transect in the hybrid zone (Westram et al., 2018), it is possible that this female carries the different unknown sex chromosomes of the wave ecotype. This outlier female might have resulted in the lack of statistical significance in this dataset, which was further explored in the power tests below. SLRfinder did not detect any significant or sex-separated clusters in the wave ecotype of snails (Table 1, Table S1, Figure S2D).

TABLE 1 Summary of the SLRfinder results using test datasets.

Dataset	# Ind	# Pop	# LD cluster	Sex_filter	Rank_candidates
stickleback_WL	162	8	2737	**LG3 (1 cluster)**	**LG3: 17269450–17332740**
stickleback_EL	598	29	1149	**LG12 (6 clusters)**	**LG12: 335099–17815098**
stickleback_nonEU	78	5	1329	Sex unknown	**LG12: 11477–17786235**
stickleback_POL	20	1	1862	**LG3 (2 clusters)** LG11 (1 cluster) LG19 (2 clusters)	**LG3: 17259548–17352126**, LG11: 16831493–17118358 ($p = .16$)
stickleback_UK	29	2	5331	Sex unknown	LG7: 3628961–3664806, LG16: 12008573–12103926 ($p = .22$)
salmon_male	59	11	25,646	LG3 (2 clusters), **LG15 (2 clusters)**	LG3: 1206464–1520135, **LG15: 44853640–45359574**
salmon_female	59	11	28,294	LG3 (2 clusters), **LG15 (3 clusters)**, LG26 (2 clusters), LG32 (1 cluster)	LG3: 1105000–1335501, LG24: 13905805–14138084
salmon_RAD	288	6	1498	Sex unknown	LG14: 53640831–53640941, **LG15: 22646022–46527777** ($p = 0.06$)
snail_crab	152	1	65	**LG12 (2 clusters)**	**LG12: 27879164–69715730** ($p = .21$)
snail_wave	100	1	74	No cluster retained.	LG13: 43354999–43374258 ($p = .13$)
guppy_female	170	10	103	No cluster retained.	All clusters had $p > .5$
guppy_male	170	10	78	No cluster retained.	All clusters had $p > .5$
leopard	26	3	90	NW_017619865.1 NW_017619916.1 NW_017619950.1 NW_017619951.1 NW_017619964.1 NW_017620089.1	All clusters had $p > .5$

Note: Each dataset is presented with the total number of individuals, the total number of populations and the total number of linkage disequilibrium (LD) clusters detected in the first step of SLRfinder. Sex-filtered results are the LD clusters having less than 10% misplaced sexed individuals. Ranked candidates are the LD clusters tested significant (adjusted $p < .05$) or, if non-significant, the clusters having the lowest adjusted p-value (only those with $p < .5$ are listed). Clusters on the known sex chromosomes are indicated in bold.

3.2 | SLRfinder analyses of guppies and leopard

SLRfinder did not identify significant candidates using the guppy datasets (Table 1), and none of the top-ranked clusters were located on the previously identified sex chromosomes LG12 (Fraser et al., 2020; Table S1, Figure 3a, Figure S3A). In fact, despite a relatively large sample size (170 individuals and 10 populations), the guppy datasets were identified with very few LD clusters (Table 1), including only two LG12 clusters using the female reference genome (Figure S3B,C) and one LG12 cluster using the male reference genome (Figure 3b), none of which showed a separation between sexes. To further investigate the signal of SLRs in guppies, we extracted SNPs located in the previously reported candidate SLRs (LG12:4800000–5200000, LG12:24500000–25400000; Fraser et al., 2020) using the filtered VCF mapped to the male reference genome and generated the heterozygosity~PC1 plot for each SLR. We found similar heterozygosity in males and females and stronger population structure than sex differentiation on PC1 in both candidate SLRs (Figure 3c,d). Therefore, these results indicate that the guppy datasets do not have the expected signal for SLRs (i.e., inter-sex differentiation in heterozygosity and stronger inter-sex divergence than population structure on PC1), which explains why SLRfinder was not able to identify these SLRs.

SLRfinder also did not find significant candidates in the African leopard dataset, using the Sex Assignment Through Coverage (SATC)-inferred genetic sex and the principal component analysis (PCA)-inferred genetic populations (Figure S4A). However, six LD clusters were detected after filtering the misplaced sexes (Figure S4B) and two of them were located on the scaffolds that were also identified with abnormal depth ratios in SATC (see below), indicating that these clusters are likely truly sex-linked.

3.3 | Local PCA on test datasets

The local PCA analyses of the EL and non-European sticklebacks showed diverged clusters of SLRs and pseudoautosomal regions on MDS plots of the heteromorphic LG12 sex chromosomes but no clustering on MDS plots of autosomes (Figure S5A). However, the MDS plot of the homomorphic LG3 sex chromosomes in the WL sticklebacks did not show clear outlier regions, and similar MDS plots were found across chromosomes in the UK sticklebacks (Figure S5A). On the contrary, the local PCA method of the crab ecotype of snails showed diverged clustering patterns in both autosomes and the homomorphic LG12 sex chromosomes (Figure S5B). Several outlier regions were indicated in the local PCA analyses of the whole leopard

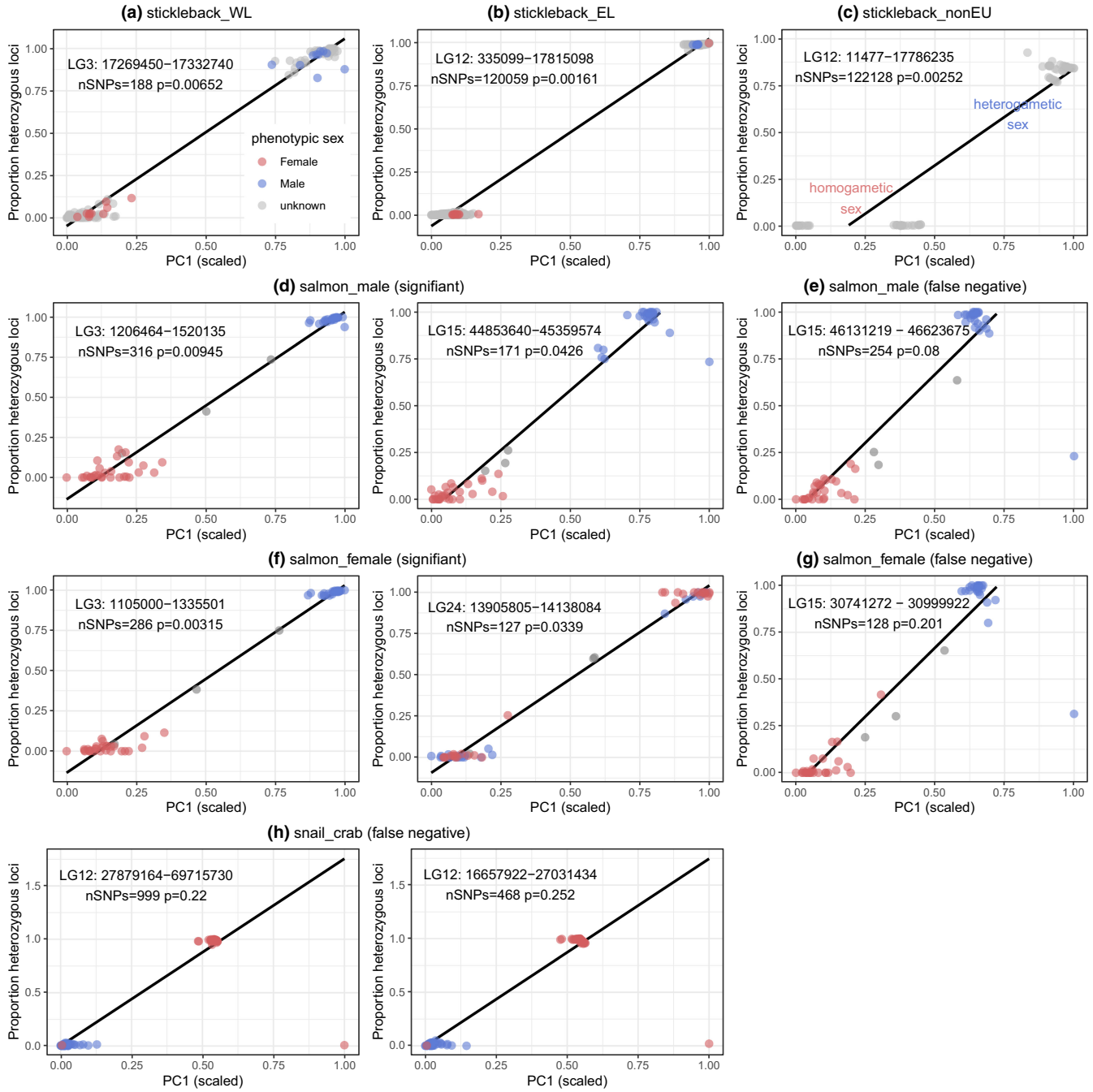
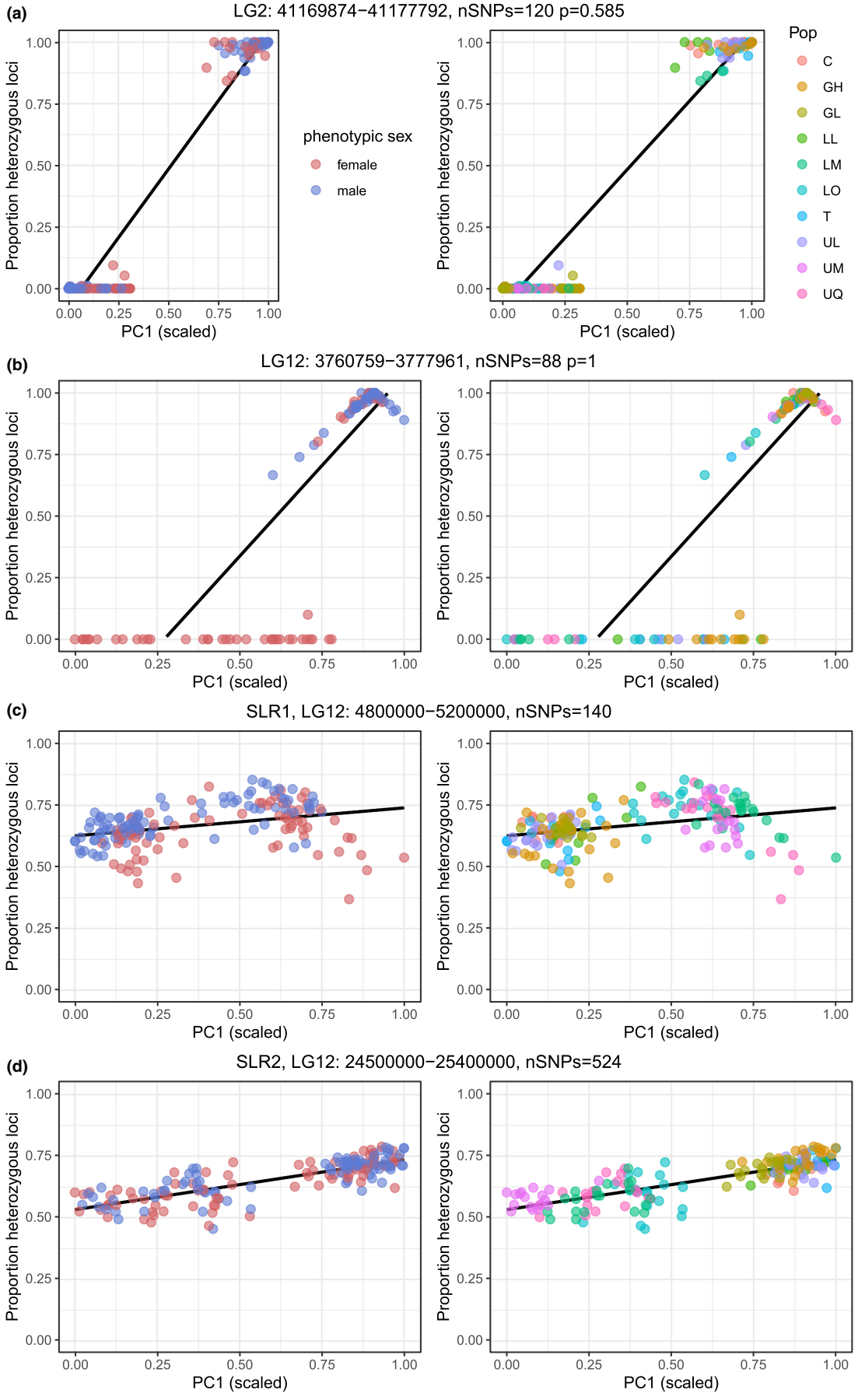


FIGURE 2 Heterozygosity-PC1 plots of the SLRfinder-identified candidates using datasets of nine-spined sticklebacks (a–c), chum salmon (d–g) and intertidal snails (h). Dots represent individuals coloured by the phenotypic sex. The black line represents the fitted linear regression. (b) The single phenotypic female in the top-right group is the individual 16-f that was also found to be a genetic male in previous studies (Feng et al., 2022; Yi et al., 2024). (e, g) The false negative clusters on LG15 detected by filtering the percentage of misplaced sexed individuals. Two more false negatives were detected in the salmon_female dataset but had much fewer single-nucleotide polymorphisms (SNPs; 26 and 96) and thus were not plotted. (h) The top-ranked two false negative sex-linked regions (SLRs). In both regions, the cluster of genetic males (ZZ) included one phenotypic female (sample ID ANG275), whereas the cluster of genetic females (ZW) included two phenotypic males (ANG147, ANG237), and one phenotypic female (ANG179) was identified as the bottom-right outlier.

FIGURE 3 Heterozygosity~PC1 plots of the guppy dataset mapped to the male reference genome. Each dot is one individual coloured by phenotypic sex (left) or population (right). (a) The top candidate identified by SLRfinder. (b) The single Linkage disequilibrium (LD) cluster identified on the sex chromosomes LG12. (c, d) Plots using single-nucleotide polymorphisms (SNPs) from the two previously reported sex-linked region (SLR) candidates (Fraser et al., 2020). The two sexes did not differ in heterozygosity, and the PC1 divergence mostly reflects population structure.



dataset (Figure S5C), but these regions were distributed on 1021 scaffolds that include both sex-linked and autosomal scaffolds.

3.4 | SATC analyses of test datasets

SATC could not analyse the datasets of WL sticklebacks, UK sticklebacks, chum salmon mapped to the female reference, guppies mapped to the male reference genome, or intertidal snails. In these datasets, the *sexDetermine* command found no good candidate based on the depth of coverage, which is consistent with the similar depth between sexes shown in previous studies on most of these populations (Fraser et al., 2020; Yi et al., 2024). Although SATC was able to process the EL nine-spined sticklebacks, the SATC-inferred sexes (i.e., heterogametic XY or homogametic XX) were opposite to the known phenotypic sexes and previously identified genetic sexes (Yi et al., 2024; Figure 4a). This is likely because SATC assumes only X/Z-linked scaffolds in the reference genome and therefore always identifies the homogametic sex as those having a higher depth of coverage (Nursyifa et al., 2022). However, when putatively Y-linked scaffolds/contigs are included in the reference genome, these contigs may show the largest depth differences and higher depths in the heterogametic sex, opposite to the SATC expectation. As a result, the single SATC-identified X/Z-linked scaffold in the EL sticklebacks was a putatively Y-linked unassembled contig (ctg7180000006428, Kivikoski et al., 2021) and XY males which had higher depths on this contig were misidentified as homogametic in SATC (Figure 4a). When analysing non-European sticklebacks, SATC did not detect X/Z-linked regions and only indicated several regions with abnormal depth ratios (Figure 4b). Interestingly, the SATC-inferred sexes were consistent with genetic sexes identified in our previous study (Yi et al., 2024), except for a Canadian population (CAN-FLO) whose individuals were indicated as genetic males in SLRfinder and our previous study (Figure 2c; Yi et al. 2024) but homogametic in SATC (Figure 4b). Additional sampling with known phenotypic sexes is required to validate the sex identification of these non-European populations. Similarly, SATC was able to process the guppy dataset mapped to the female reference, but all phenotypic females were indicated as heterogametic (XY), and all indicated homogametic individuals (XX) were phenotypic males (Figure 4d). Only one X/Z linked contig was identified in the guppy dataset (Figure 4c) and only one contig having the abnormal depth ratio was identified in the chum salmon dataset (Figure S6C,D). The known SLRs or sex chromosomes were not identified probably because chromosome-level depth differences were small, and SATC could not break down assembled chromosomes into smaller regions that would include SLRs. Overall, these results showed limited application of SATC to the identification of labile sex chromosomes.

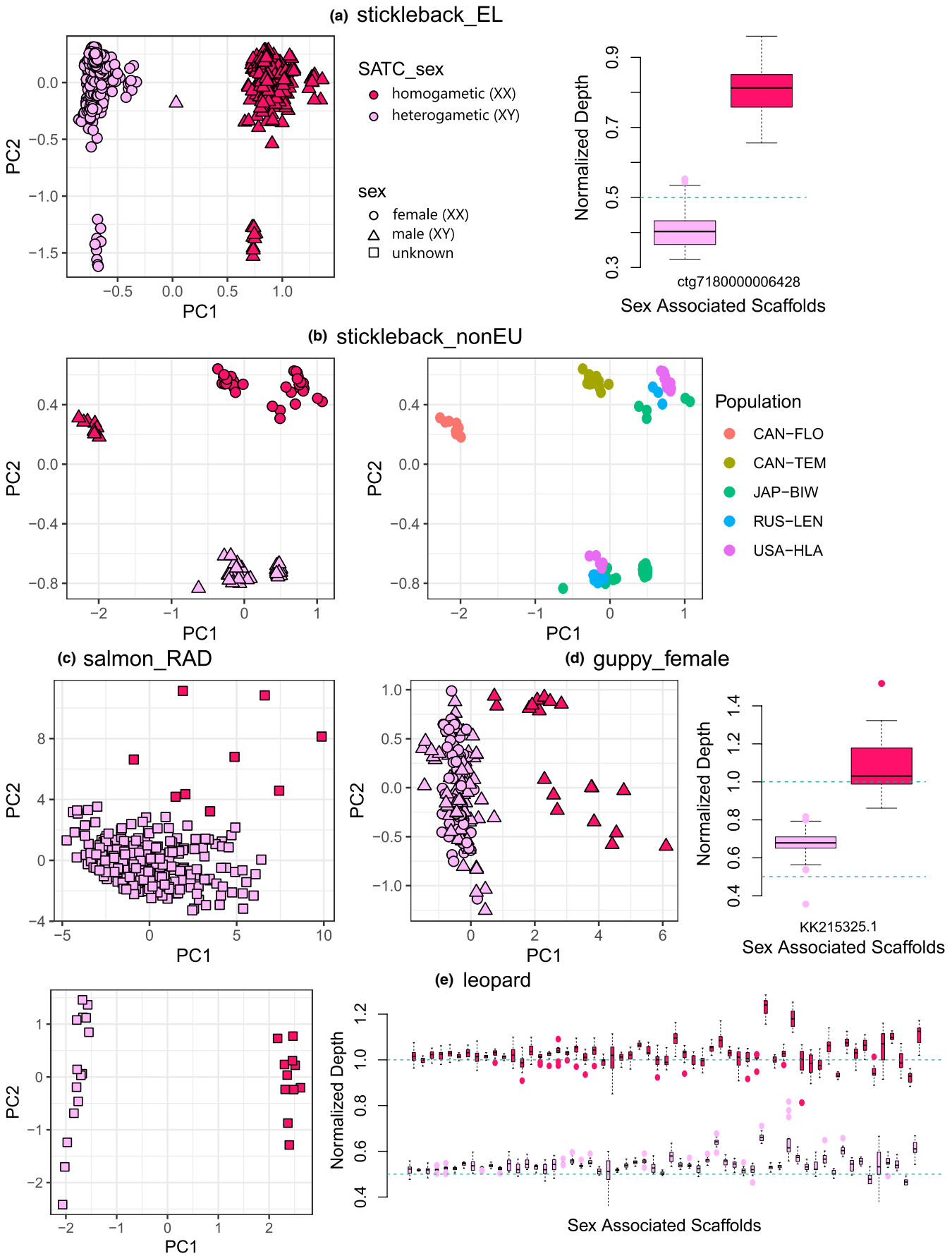
On the contrary, SATC was successfully applied to the dataset of African leopards which have conserved sex chromosomes and were mapped to a scaffold-level female reference genome. Using only 29 individuals, we identified 58 scaffolds as X/Z-linked and eight scaffolds having abnormal depth ratios (Figure 4e; Figure S6E), including all of the reported sex-linked scaffolds in previous studies using the same dataset (Nursyifa et al., 2022; Pečnerová et al., 2021).

3.5 | Power tests of SLRfinder

Results of the power tests using the WL sticklebacks, EL sticklebacks and crab ecotype of snails are summarized in Tables S2–S4. In the WL sticklebacks, SLRfinder accurately detected the LG3 SLR as the only significant candidate when using all eight populations with three to five randomly selected individuals per population (minimum 24 individuals in total). The LG3 SLR was always identified with the lowest *p*-value when using one to five populations in the dataset, although only the test using five populations showed statistical significance. SLRfinder identified the LG3 SLR as the significant candidate when testing the sex ratios (male:female) of 1:1 or 3:1 and with the lowest *p*-value when testing the sex ratios of 1:2, 2:1, or 10:1. No statistical significance was found and the LG3 SLR was not identified among top-ranked candidates when testing the sex ratios 1:3 and 1:10 with default settings that expect even sex ratios (i.e., .5 probability of sampling each sex). We then re-analysed the dataset having sex ratio 1:3 using the expectation of uneven sex ratio (e.g., .25 probability of sampling males and .75 of females) and without the rank of cluster size. The LG3 SLR had the lowest *p*-value (Table S2) but was still not significant ($p = .2$) probably because few SNPs from the SLR were genotyped when few individuals of the heterogametic sex were included in the dataset. However, even if using the expectation of uneven sex ratio and no rank of cluster size, the LG3 SLR was not included in the top-ranked candidates when the sex ratio was extremely skewed (1:10 or 10:1). The LG3 SLR was detected by filtering the percentage of misplaced sexes in most of these datasets. When one sex was completely missing, neither sex filtering nor the candidate ranking could work and no false positives were found.

When applied to the EL sticklebacks, SLRfinder accurately detected the LG12 SLR as the only significant candidate when using all 29 populations with three to five randomly selected individuals per population, and when using three to five randomly selected populations (Table S3). When only one population (equal sex ratio)

FIGURE 4 Sex Assignment Through Coverage (SATC) results of test datasets. All test datasets are known to have the XX/XY sex determination. Principal component analysis (PCA) plots show variation in the normalized depth of coverage across all samples. Shapes represent genetic sex (a, b), phenotypic sex (d) or unknown sex (c, e). Colours represent the SATC-inferred homogametic or heterogametic sex, except for (b) where individuals are also coloured by population (right). Boxplots show the SATC-identified X/Z-linked scaffolds. Each scaffold has two boxes showing the normalized depth of coverage in the SATC-indicated homogametic sex (expected depth 1.0) and heterogametic sex (expected depth 0.5).



was included, the LG12 SLR was identified with a marginal *p*-value (.08) as the top-ranked candidate. However, when using two populations, the LG12 SLR was not included in the top-ranked candidates, indicating some uncertainty when the population size is small and the sex ratio is uneven (around 1:3 in this test, Table S3). Using the expectation of uneven sex ratio allowed SLRfinder to add a LG12 cluster to the top-ranked candidate (Table S3). When testing uneven sex ratios up to 1:3 and 3:1 using larger sample sizes (24 populations, 96 individuals), SLRfinder identified the LG12 SLR as the significant candidates even using the default setting of equal sex ratio (Table S3). However, when using the most skewed sex ratios (1:10, 10:1), no statistical significance was found and the LG12 SLR was not identified among the top-ranked candidates. Three significant candidates were detected after using the expectation of uneven sex ratios in the dataset having the sex ratio 10:1, including the LG12 SLR and two false positives ($p = .0491$, Table S3). The LG12 SLR was detected by filtering based on the percentage of misplaced sexes in all datasets except for those having the most skewed sex ratios (1:10, 10:1). Again, neither sex filtering nor the candidate ranking could work when one sex was completely missing.

When analysing the crab ecotype of snails without the four outlier and mismatched individuals (Figure 2h), the LG12 SLRs were top-ranked but still not significant ($p = .11$; Table S4). Because this dataset had a skewed sex ratio (male:female = 0.68), we re-analysed it using the expectation of uneven sex ratio (.6 probability of sampling males and .4 of females) and the largest LG12 SLR was identified as the only significant candidate ($p = .022$; Table S4). Next, we kept the equal sex ratio and tested SLRfinder on a total of 120, 60 and 50 samples, all of which showed significant LG12 candidates (Table S4). However, the LG12 SLRs were still top-ranked but not statistically significant using a total of 40 samples ($p = .07$; Table S4). Therefore, our tests indicated that statistical significance in SLRfinder would require a minimum of around 50 individuals at an equal sex ratio in this dataset of one population. It should be noted that this population of the crab ecotype was collected along the transect in a hybrid zone, and therefore, these samples may be genetically more diverse than those from the populations at the core of the species distribution range. Therefore, the minimum sample size if only one population is analysed may differ depending on the genetic diversity of the sampled individuals and the biological features of their sex chromosomes.

4 | DISCUSSION

Linkage disequilibrium has been shown to be highly informative with respect to chromosomal evolution, adaptation and population structure (Fang et al., 2020, 2021; Faria et al., 2019; Guzmán et al., 2022; Kemppainen et al., 2015), and has been also suggested to be potentially useful in identifying SLRs (Hearn et al., 2022; McKinney et al., 2020). However, signals of LD have remained under-exploited in studies of population genomics and sex chromosomes. Here, we present a method, SLRfinder, which utilizes LD to identify candidate SLRs and the sex chromosomes in which they are located. Results show that SLRfinder successfully identified known SLRs as significant candidates when analysing the published population data of nine-spined sticklebacks, the chum salmon dataset mapped to the male reference genome and the crab ecotype of intertidal snails after using the expectation of uneven sex ratio. In addition, using LD clustering, the SLRfinder-identified SLRs were narrower than those identified using GWAS (Rondeau et al., 2023) or sliding windows (Yi et al., 2024), which indicates that SLRfinder can be beneficial by further narrowing down the highly linked SLR even when the pair of sex chromosomes is already known. Interestingly, the SLRs of nine-spined sticklebacks, chum salmon and intertidal snails have been indicated to involve genomic inversions that might facilitate the recombination suppression in SLRs and the early sex chromosome evolution (Hearn et al., 2022; McKinney et al., 2020; Natri et al., 2019; Yi et al., 2024). The sex-linked inversions might have strengthened the signals of LD and heterozygosity and thus made these SLRs easier to detect in SLRfinder. However, despite these accumulating studies, it remains unclear how often inversions (and other structural variants) might be associated with labile SLRs in natural populations. We propose that SLRfinder might be helpful to answer this question as it is likely most sensitive to the SLRs having structural variants and can be easily applied to genomic data of populations in non-model species.

Our comparison between SLRfinder and local PCA further showed the power of LD clustering (Table 2). When running local PCA by chromosome, the relatively more diverged sex chromosomes (e.g., LG12 in the EL and non-European sticklebacks) can be distinguished from autosomes based on their clearly diverged outlier windows on the MDS plot, similar to findings in the previous study of birds (Merondun et al., 2024). However, local PCA could not differentiate autosomes and homomorphic sex chromosomes (e.g., LG3 in the WL sticklebacks and LG12 in the crab ecotype of snails). When running local PCA on all scaffolds, such as in the leopard dataset, even windows from conserved sex chromosomes are difficult to identify due to the extra noise introduced by combining all unsorted scaffolds. In addition, local PCA works on SNP windows in fixed sizes that could not be too small to avoid high proportions of missing data. On the contrary, SLRfinder works on LD clusters that can have various sizes and are more biologically meaningful. The first step of LD clustering also largely reduces the data size of downstream analyses because SNPs that are not correlated with each other are discarded. Accordingly, although both methods share the idea of regional PCA, we show that SLRfinder is specialized for identifying unknown SLRs and labile sex chromosomes, whereas the local PCA method is better for indicating adaptive genomic regions across the whole genome (Li & Ralph, 2019) but not necessarily sex-related regions.

We also compared SLRfinder to the previously developed depth-based method SATC (Nursyifa et al., 2022). As expected, SLRfinder outperformed SATC in analysing labile sex chromosomes that tend to have similar depths between sexes, such as in the WL sticklebacks, chum salmon and the crab ecotype of snails.

TABLE 2 Comparison between SLRfinder, local principal component analysis (PCA) and Sex Assignment Through Coverage (SATC).

	SLRfinder	Local PCA	SATC
Recommended type of sex chromosomes	Labile	Conserved	Conserved
Considered signals	Linkage disequilibrium, number of SNPs in the LD cluster, heterozygosity, genetic differentiation	Genetic differentiation	Depth of coverage
Sequencing data	Whole-genome resequencing (preferred) or restriction site-associated DNA sequencing	Whole-genome resequencing	Whole-genome resequencing
Reference genome	Best if chromosome-level; homogametic or heterogametic sex	Chromosome-level; homogametic or heterogametic sex	Best if scaffold-level; homogametic sex (no Y/W-linked scaffolds)
Phenotypic sex	Not required in the input but can be incorporated to provide extra supports	Not used	Not used
Computational burden	Medium: memory and speed depend on data size	Small	Small

Both SATC and SLRfinder can process data without the input of phenotypic sexes, which can be difficult to obtain from non-invasive sampling or difficult to identify without clear phenotypic sexual dimorphism. However, it should be noted that phenotypic sexes are needed to validate whether a detected candidate region is the true SLR. This validation may be easier in SLRfinder, which can readily incorporate phenotypic sexes and report patterns of inter-sex separation. SATC assumes no Y/W-linked scaffolds in the reference genome, which is usually true in the taxa having conserved sex chromosomes because the highly degenerated Y/W chromosomes are difficult to assemble and often excluded when the reference genome comes from the heterogametic sex. However, reference genomes of the taxa having labile sex chromosomes are more likely a mosaic combination of scaffolds from both sex chromosomes if the sequences were from a heterogametic individual (e.g., the version 7 reference of the nine-spined stickleback; Kivikoski et al., 2021), making SATC less applicable and even misleading (such as in the case of EL sticklebacks). In addition, SATC was designed for data mapped to scaffold-level reference genomes (Nursyifa et al., 2022) and could not break down long assembled chromosomes, which may prevent the identification of narrow SLRs of labile sex chromosomes when using chromosome-level reference genomes. On the contrary, SATC worked better than SLRfinder on conserved sex chromosomes that have clear inter-sex differences in the depth of coverage. Accordingly, our study suggests that SLRfinder and SATC are complementary methods that specialize on different types of sex chromosomes and datasets (Table 2). Therefore, we recommend testing both methods (and potentially other methods as well) when trying to identify SLRs in new populations or species to get complementary results.

SLRfinder has several advantages as illustrated in our analyses. First, SLRfinder does not require known sex determination systems or a specific type of reference genome (i.e., from the heterogametic or homogametic sex). It is worth noting that SLRfinder may work better using the chromosome-level than the scaffold-level reference genome because the former generates more and larger LD clusters. Second, SLRfinder does not require a specific sequencing method (e.g., WGS or RADseq) and can be easily applied to any SNP genotypes in the VCF format. The highly flexible R scripts allow manual parameter settings (e.g., *min_LD*, expected sex ratio and rank parameters) and can be easily extended to include additional ranking or filtering parameters (e.g., F_{ST} between sexes). Third, SLRfinder is a conservative method. Our test found very few false positives, which could be identified by the separation between phenotypic sexes and the usually higher *p*-values than the top-ranked true SLRs. On the contrary, SLRfinder did not have enough power to detect significant SLRs in several cases but the false negatives can be identified by filtering the misplaced sexes. In addition, false negatives tend to be the top-ranked clusters with the lowest non-significant *p*-values and the largest numbers of SNPs. Thus, even in the absence of statistical significance, SLRfinder can suggest top-ranked LD clusters that may be worth analysing further.

Like all the other methods, SLRfinder also has its limitations which were explored using the test datasets. First, SLRfinder may have limited power when sample sizes are small, especially with a limited number of populations. For example, SLRfinder successfully identified the true SLR using as few as 24 individuals from eight WL populations of nine-spined sticklebacks, but not when using as many as 79 individuals from four WL populations (Table S2). This is likely because more diverse populations generate more and larger LD clusters, which increases the power of SLRfinder. In the dataset of the crab ecotype of snails, only one population was included and minimum around 50 samples at the equal sex ratio are needed to get statistical significance, although the true SLRs were always top-ranked (Table S4). The required minimum sample size also depends on the state of sex chromosomes. For example, the homomorphic LG3 SLR had high $p = .44$ using one WL population (BEL-MAL, $n = 20$), whereas the heteromorphic LG12 SLR had much lower $p = .082$ using one EL population (FIN-HEL, $n = 22$) despite similar sample sizes and the equal sex ratio

(Table S3). Therefore, more samples may be needed to detect significant SLRs on homogametic sex chromosomes that had less differentiation. Second, SLRfinder assumes sampling of both sexes in relatively equal proportions. Although slightly skewed sex ratios (max 1:3 or 3:1) could work in most cases and can be accounted for in χ^2 tests, SLRfinder appears not to work when sex ratios are highly skewed (e.g., 1:10 or 10:1). Third, SLRfinder assumes that all individuals in the input dataset share the same SLRs. Tests of the Polish sticklebacks showed that when this assumption is slightly violated (i.e., one heterogametic individual carrying a different SLR), SLRfinder could still detect the prevalent SLR as the top-ranked region, although the results seemed to have more noise and were not significant. SLRfinder might not work very well if the individuals included in the analyses have more diverse SLRs. In this case, each population may need to be analysed separately. However, although not tested here, these limitations from sample size, sex ratio and shared SLRs likely also apply to most of the other methods for SLR identification with few exceptions (e.g., FindZX may work on a single individual; Sigeman et al., 2022). Lastly, SLRfinder may be biased to reporting SLRs having inversions, which are easier to detect due to stronger signals of LD and difference in heterozygosity. However, because inversions may be important for the early formation of SLRs and the evolution of sex chromosomes, we expect more empirical cases that can apply SLRfinder than those that cannot.

Unsurprisingly, SLRfinder only works when the expected signals (differential heterozygosity and genetic differentiation between sexes in SLRs) are present in the data. However, these signals may not be clear in every dataset. When sample sizes are small and include low signal-to-noise ratios, these expected signals can occur by chance rather than driven by linkage to sex. In addition, some biological systems may exhibit complicated signals in their SLRs. For example, guppies showed no difference in male and female heterozygosity and stronger population structure than inter-sex divergence in the previously identified candidate SLRs (Figure 3c,d). The high heterozygosity in both sexes and strong population signal might be explained by the maintenance of many different Y haplotypes among these populations via balancing selection (Fraser et al., 2020). It is also possible that these populations actually have different SLRs which would generate noisy signals in SLRfinder. Similarly, a previously developed depth-based method, RADSex, was applied to 15 teleost fishes having labile sex chromosomes but only six were successfully identified with sex markers (Feron et al., 2021). Taken together, these results show that no single method is universally applicable to all taxa having diverse sex chromosome systems.

In summary, SLRfinder provides a novel approach for the identification of labile sex chromosomes in populations of non-model species using LD and heterozygosity. Given the lack of a universal method for identifying SLRs across diverse sex chromosome systems, SLRfinder complements the previously developed depth-based methods (e.g., SATC) by serving the same purpose in different contexts. SLRfinder seems to work best when applied to a large number of divergent populations and when sex ratios are relatively equal. Although phenotypic sexes are not required to run SLRfinder, they can be incorporated for additional filtering and are needed to validate whether the identified candidates are true SLRs. In addition, SLRfinder is most sensitive to SLRs that involve inversions and can detect autosomal regions that may have become sex-linked (e.g., the LG3 region in chum salmon), which can be interesting in the contexts of sexual selection, sexual antagonism and sex chromosome evolution.

AUTHOR CONTRIBUTIONS

P.K. and X.Y. conceptualized the study. P.K. designed the method and wrote the raw scripts. X.Y. polished the method and analysed empirical datasets. J.M. supervised the study and provided resources. X.Y. and P.K. drafted the manuscript. All authors edited the manuscript.

ACKNOWLEDGEMENTS

We thank Bonnie Fraser for the help to get access to the raw data of guppies. Thanks to Katherine Hearn and Anja Westram for the help with access to phenotypic sexes of the snail datasets.

FUNDING INFORMATION

This study was supported by the National Natural Science Foundation of China/Research Grants Council (RGC) Joint Research Scheme 2021/2022 ('N_HKU763/21' to JM). We are grateful to the IT Centre for Scientific Computing (CSC), Finland, for access to computing resources.

CONFLICT OF INTEREST STATEMENT

The authors claim no conflict of interest.

DATA AVAILABILITY STATEMENT

SLRfinder scripts are publicly available on GitHub (https://github.com/xuelingyi/SLRfinder) with a step-by-step tutorial. The sample information of our tested datasets is also provided on GitHub.

ORCID

Xueling Yi https://orcid.org/0000-0003-4860-7429
Petri Kemppainen https://orcid.org/0000-0002-7228-8133
Juha Merilä https://orcid.org/0000-0001-9614-0072

REFERENCES

Barton, N. H. (2011). Estimating linkage disequilibria. *Heredity*, *106*(2), 205–206.

Blaser, O., Neuenschwander, S., & Perrin, N. (2014). Sex-chromosome turnovers: The hot-potato model. *American Naturalist*, *183*(1), 140–146. https://doi.org/10.1086/674026

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

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, *1695*(5), 1–9.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean,

G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x

Devlin, B., Roeder, K., & Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*, *60*(3), 155–166. https://doi.org/10.1006/tpbi.2001.1542

Dixon, G., Kitano, J., & Kirkpatrick, M. (2019). The origin of a new sex chromosome by introgression between two stickleback fishes. *Molecular Biology and Evolution*, *36*(1), 28–38. https://doi.org/10.1093/MOLBEV/MSY181

Dufresnes, C., Borzee, A., Horn, A., Stock, M., Ostini, M., Sermier, R., Wassef, J., Litvinchuck, S. N., Kosch, T. A., Waldman, B., Jang, Y., Brelsford, A., & Perrin, N. (2015). Sex-chromosome homomorphy in Palearctic tree frogs results from both turnovers and X–Y recombination. *Molecular Biology and Evolution*, *32*(9), 2328–2337. https://doi.org/10.1093/MOLBEV/MSV113

Fang, B., Kemppainen, P., Momigliano, P., Feng, X., & Merilä, J. (2020). On the causes of geographically heterogeneous parallel evolution in sticklebacks. *Nature Ecology & Evolution*, *4*(8), 1105–1115. https://doi.org/10.1038/s41559-020-1222-6

Fang, B., Kemppainen, P., Momigliano, P., & Merilä, J. (2021). Population structure limits parallel evolution in sticklebacks. *Molecular Biology and Evolution*, *38*(10), 4205–4221. https://doi.org/10.1093/MOLBEV/MSAB144

Faria, R., Chaube, P., Morales, H. E., Larsson, T., Lemmon, A. R., Lemmon, E. M., Rafajlović, M., Panova, M., Ravinet, M., Johannesson, K., Westram, A. M., & Butlin, R. K. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Molecular Ecology*, *28*(6), 1375–1393. https://doi.org/10.1111/mec.14972

Feng, X., Merilä, J., & Löytynoja, A. (2022). Complex population history affects admixture analyses in nine-spined sticklebacks. *Molecular Ecology*, *31*(20), 5386–5401. https://doi.org/10.1111/MEC.16651

Feron, R., Pan, Q., Wen, M., Imarazene, B., Jouanno, E., Anderson, J., Herpin, A., Journot, L., Parrinello, H., Klopp, C., Kottler, V. A., Roco, A. S., Du, K., Kneitz, S., Adolfi, M., Wilson, C. A., McCluskey, B., Amores, A., Desvignes, T., … Guiguen, Y. (2021). RADSex: A computational workflow to study sex determination using restriction site-associated DNA sequencing data. *Molecular Ecology Resources*, *21*(5), 1715–1731. https://doi.org/10.1111/1755-0998.13360

Fraser, B. A., Whiting, J. R., Paris, J. R., Weadick, C. J., Parsons, P. J., Charlesworth, D., Bergero, R., Bemm, F., Hoffmann, M., Kottler, V. A., Liu, C., Dreyer, C., & Weigel, D. (2020). Improved reference genome uncovers novel sex-linked regions in the guppy (*Poecilia reticulata*). *Genome Biology and Evolution*, *12*(10), 1789–1805. https://doi.org/10.1093/GBE/EVAA187

Furman, B. L. S., Metzger, D. C. H., Darolti, I., Wright, A. E., Sandkam, B. A., Almeida, P., Shu, J. J., Mank, J. E., & Fraser, B. (2020). Sex chromosome evolution: So many exceptions to the rules. *Genome Biology and Evolution*, *12*(6), 750–763. https://doi.org/10.1093/gbe/evaa081

Guzmán, N. V., Kemppainen, P., Monti, D., Castillo, E. R. D., Rodriguero, M. S., Sánchez-Restrepo, A. F., Cigliano, M. M., & Confalonieri, V. A. (2022). Stable inversion clines in a grasshopper species group despite complex geographical history. *Molecular Ecology*, *31*(4), 1196–1215. https://doi.org/10.1111/mec.16305

Hearn, K. E., Koch, E. L., Stankowski, S., Butlin, R. K., Faria, R., Johannesson, K., & Westram, A. M. (2022). Differing associations between sex determination and sex-linked inversions in two ecotypes of *Littorina saxatilis*. *Evolution Letters*, *6*(5), 358–374.

Jeffries, D. L., Lavanchy, G., Sermier, R., Sredl, M. J., Miura, I., Borzée, A., Barrow, L. N., Canestrelli, D., Crochet, P. A., Dufresnes, C., Fu, J., Ma, W. J., Garcia, C. M., Ghali, K., Nicieza, A. G., O'Donnell, R. P., Rodrigues, N., Romano, A., Martínez-Solano, Í., … Perrin, N. (2018). A rapid rate of sex-chromosome turnover and non-random transitions in true frogs. *Nature Communications*, *9*(1), 1–11. https://doi.org/10.1038/s41467-018-06517-2

Jeffries, D. L., Mee, J. A., & Peichel, C. L. (2022). Identification of a candidate sex determination gene in *Culaea inconstans* suggests convergent recruitment of an *Amh* duplicate in two lineages of stickleback. *Journal of Evolutionary Biology*, *35*(12), 1683–1695. https://doi.org/10.1111/JEB.14034

Kemppainen, P., Knight, C. G., Sarma, D. K., Hlaing, T., Prakash, A., Maung Maung, Y. N., Somboon, P., Mahanta, J., & Walton, C. (2015). Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Molecular Ecology Resources*, *15*(5), 1031–1045. https://doi.org/10.1111/1755-0998.12369

Kitano, J., Ansai, S., Fujimoto, S., Kakioka, R., Sato, M., Mandagi, I. F., Sumarto, B. K. A., & Yamahira, K. (2023). A cryptic sex-linked locus revealed by the elimination of a master sex-determining locus in medaka fish. *The American Naturalist*, *202*(2), 231–240. https://doi.org/10.1086/724840

Kivikoski, M., Rastas, P., Löytynoja, A., & Merilä, J. (2021). Automated improvement of stickleback reference genome assemblies with LepAnchor software. *Molecular Ecology Resources*, *21*(6), 2166–2176. https://doi.org/10.1111/1755-0998.13404

Kü Nstner, A., Hoffmann, M., Fraser, B. A., Kottler, V. A., Sharma, E., Weigel, D., & Dreyer, C. (2016). The genome of the Trinidadian guppy, *Poecilia reticulata*, and variation in the Guanapo population. *PLoS One*, *11*, e0169087. https://doi.org/10.1371/journal.pone.0169087

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/BIOINFORMATICS/BTR509

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. (arXiv:1303.3997). https://doi.org/10.48550/arXiv.1303.3997

Li, H., & Ralph, P. (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics*, *211*(1), 289–304.

Ma, J., & Amos, C. I. (2012). Investigation of inversion polymorphisms in the human genome using principal components analysis. *PLoS One*, *7*(7), 40224. https://doi.org/10.1371/journal.pone.0040224

McKinney, G., McPhee, M. V., Pascal, C., Seeb, J. E., & Seeb, L. W. (2020). Network analysis of linkage disequilibrium reveals genome architecture in chum salmon. *G3: Genes, Genomes, Genetics*, *10*(5), 1553–1561. https://doi.org/10.1534/G3.119.400972

Merondun, J., Marques, C. I., Andrade, P., Meshcheryagina, S., Galván, I., Afonso, S., Alves, J. M., Araújo, P. M., Bachurin, G., Balacco, J., Bán, M., Fedrigo, O., Formenti, G., Fossøy, F., Fülöp, A., Golovatin, M., Granja, S., Hewson, C., Honza, M., … Wolf, J. B. (2024). Evolution and genetic architecture of sex-limited polymorphism in cuckoos. *Science Advances*, *10*(17), eadl5255.

Myosho, T., Otake, H., Masuyama, H., Matsuda, M., Kuroki, Y., Fujiyama, A., Naruse, K., Hamaguchi, S., & Sakaizumi, M. (2012). Tracing the emergence of a novel sex-determining gene in medaka *Oryzias luzonensis*. *Genetics*, *191*(1), 163–170.

Natri, H. M., Merilä, J., & Shikano, T. (2019). The evolution of sex determination associated with a chromosomal inversion. *Nature Communications*, *10*(1), 1–13. https://doi.org/10.1038/s41467-018-08014-y

Nursyifa, C., Brüniche-Olsen, A., Garcia-Erill, G., Heller, R., & Albrechtsen, A. (2022). Joint identification of sex and sex-linked scaffolds in non-model organisms using low depth sequencing data. *Molecular Ecology Resources*, *22*(2), 458–467. https://doi.org/10.1111/1755-0998.13491

Ogata, M., Suzuki, K., Yuasa, Y., & Miura, I. (2021). Sex chromosome evolution from a heteromorphic to a homomorphic system by interpopulation hybridization in a frog. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1833), 20200105. https://doi.org/10.1098/rstb.2020.0105

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. https://doi.org/10.1111/MEC.15245

Pečnerová, P., Garcia-Erill, G., Liu, X., Nursyifa, C., Waples, R. K., Santander, C. G., Quinn, L., Frandsen, P., Meisner, J., Stæger, F. F., Rasmussen, M. S., Brüniche-Olsen, A., Hviid Friis Jørgensen, C., da Fonseca, R. R., Siegismund, H. R., Albrechtsen, A., Heller, R., Moltke, I., & Hanghøj, K. (2021). High genetic diversity and low differentiation reflect the ecological versatility of the African leopard. *Current Biology*, *31*(9), 1862–1871.e5. https://doi.org/10.1016/J.CUB.2021.01.064

Perrin, N. (2021). Sex-chromosome evolution in frogs: What role for sex-antagonistic genes? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1832), 20200094. https://doi.org/10.1098/RSTB.2020.0094

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. https://doi.org/10.1111/MEC.15253

Rondeau, E. B., Christensen, K. A., Johnson, H. A., Sakhrani, D., Biagi, C. A., Wetklo, M., Despins, C. A., Leggatt, R. A., Minkley, D. R., Withler, R. E., Beacham, T. D., Koop, B. F., & Devlin, R. H. (2023). Insights from a chum salmon (*Oncorhynchus keta*) genome assembly regarding whole-genome duplication and nucleotide variation influencing gene function. *G3: Genes, Genomes, Genetics*, *13*(8), jkad127. https://doi.org/10.1093/G3JOURNAL/JKAD127

Sigeman, H., Sinclair, B., & Hansson, B. (2022). Findzx: An automated pipeline for detecting and visualising sex chromosomes using whole-genome sequencing data. *BMC Genomics*, *23*(1), 1–14. https://doi.org/10.1186/S12864-022-08432-9/FIGURES/5

Tree of Sex Consortium. (2014). *Tree of sex: A database of sexual systems* [dataset]. Nature Publishing Group. https://doi.org/10.5061/DRYAD.V1908

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.BI1110S43

Vicoso, B. (2019). Molecular and evolutionary dynamics of animal sex-chromosome turnover. *Nature Ecology and Evolution*, *3*(12), 1632–1641. https://doi.org/10.1038/s41559-019-1050-8

Westram, A. M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., Ravinet, M., Blomberg, A., Mehlig, B., Johannesson, K., & Butlin, R. (2018). Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evolution Letters*, *2*(4), 297–309.

Yi, X., Wang, D., Reid, K., Feng, X., Löytynoja, A., & Merilä, J. (2024). Sex chromosome turnover in hybridizing stickleback lineages. *Evolution Letters*, qrae019. https://doi.org/10.1093/evlett/qrae019

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/BIOINFORMATICS/BTS606

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Yi, X., Kemppainen, P., & Merilä, J. (2024). SLRfinder: A method to detect candidate sex-linked regions with linkage disequilibrium clustering. *Molecular Ecology Resources*, *24*, e13985. https://doi.org/10.1111/1755-0998.13985