Springer

Conservation Genetics Resources

Supplementary Information for:

An R-based tool for identifying sex-linked markers from Restriction site Associated DNA sequencing with applications to elasmobranch conservation

Floriaan Devloo-Delva*, Thierry Gosselin, Paul A. Butcher, Peter M. Grewe, Charlie Huveneers, Robin B. Thomson, Jonathan M. Werry, and Pierre Feutry

Author affiliations
CSIRO Environment
CSIRO Australian National Fish Collection
School of Natural Sciences – Quantitative Marine Science, University of Tasmania
NSW Department of Primary Industries, National Marine Science Centre
Southern Shark Ecology Group, College of Science and Engineering, Flinders University
Griffith Centre for Coastal Management, Griffith University
* *Corresponding author: Floriaan.devloo.delva@csiro.au*

October 2023

Contents

1 Running sexy_markers

THe `sexy_markers` function is developed within the `radiator` R package. More information on the instalation and input requirements can be found here: **https://thierrygosselin.github.io/radiator/index.html**

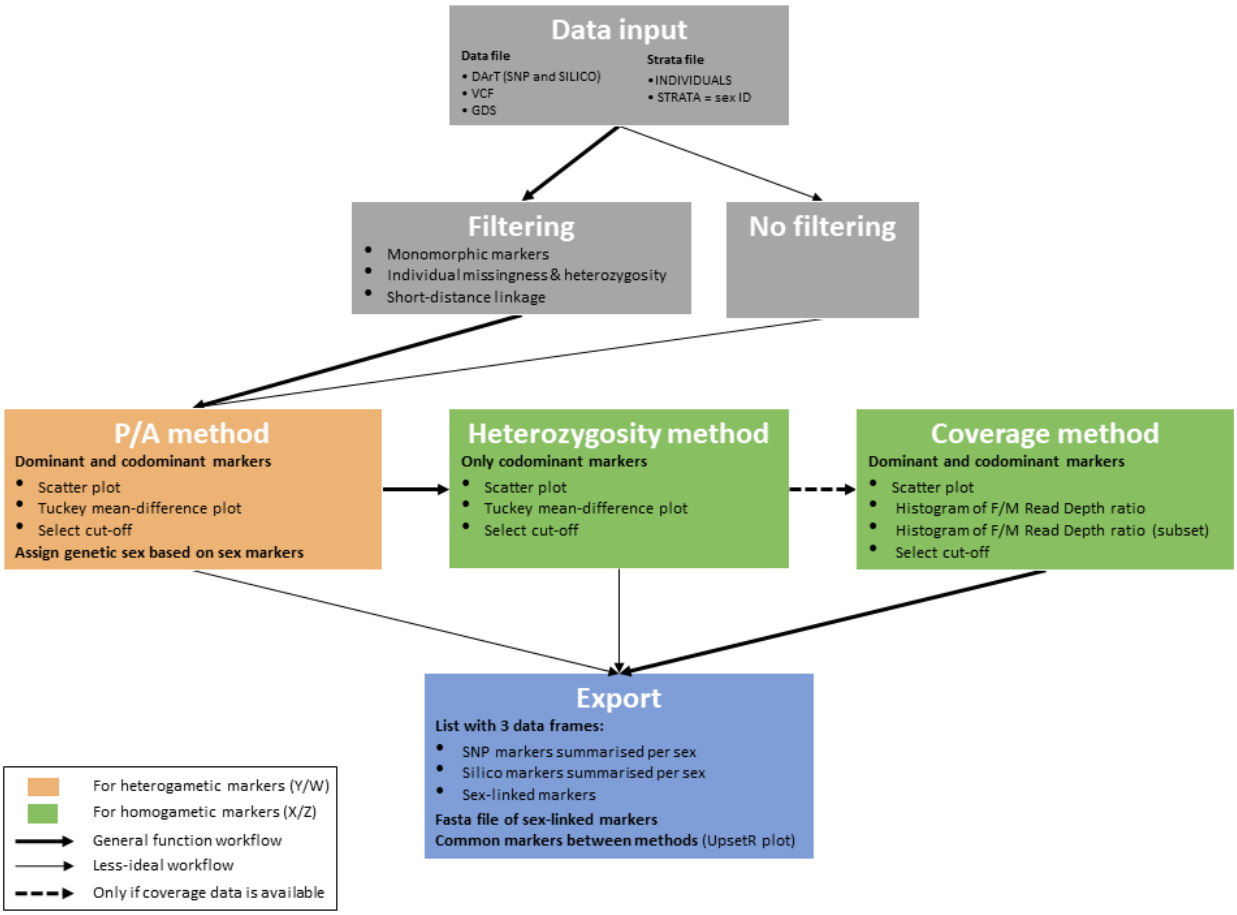
The `sexy_markers` function is designed to identify single nucleotide polymorphism (SNP) markers that have a sex-linked pattern. These patterns can be decided into three methods to detect sex-linked markers:

1. **Presence/absence (P/A) of markers**: To identify markers on Y or W chromosomes, we look at the presence or absence of a marker between females and males. More specifically, if a marker is always present in males but never in females, they are putatively located in the Y-chromosome; and vice versa for the W-linked markers.
2. **Heterozygosity of markers**: By looking at the heterozygosity of a marker between sexes, we can identify markers that are always homozygous in one sex (e.g. males for an XY system), while exhibiting an intermediate range of heterozygosity in the other sex (0.1 - 0.5).
3. **Average coverage of markers**: If the data includes read depth information, this function will look for markers that have double the coverage for either of the sexes. For example if an XY/XX system is present, females are expected to have double the number of counts for markers on the X chromosome.

In addition, if Y- or W-linked markers are identified (i.e. P/A method), the `sexy_markers` function will compare the phenotypic and genetic sex IDs, and allow you to reassign the gender to all individuals.

1.1 Workflow

1. Read data
2. Filter data
3. Summarise the data by sex
4. Identify heterogamatic (Y or W) markers: P/A method
5. Identify homogametic (X or Z) markers: Heterozygosity method
6. Identify homogametic (X or Z) markers: Coverage method
7. Export the summarized data and sex-linked markers



1.2 Making strata file

The strata file is a tab-separated file with a minimum of 2 columns for VCF files (INDIVIDUALS, STRATA) and 3 columns for DArT files (TARGET_ID, INDIVIDUALS, STRATA).

Here the STRATA should be the phenotypic (or genetic) sex. **M for male, F for female and U for unknown.**

```
strata <- radiator::read_strata("whiteshark_strata.tsv", keep.two = FALSE)$strata
knitr::kable(strata[1:9,])
```

TARGET_ID	INDIVIDUALS	STRATA	Species
737306	WS-m-EC-A-AU05-0254	M	White Shark
818272	WS-M-EC-J-AU05-0235	M	White Shark
818213	WS-M-EC-J-AU05-0249	M	White Shark
737317	WS-m-EC-J-AU05-1016	M	White Shark
721589	WS-m-EC-J-AU05-1018	M	White Shark
721601	WS-m-EC-J-AU05-1019	M	White Shark
737321	WS-m-EC-J-AU05-1020	M	White Shark
737325	WS-m-EC-J-AU05-1024	M	White Shark
737329	WS-m-EC-J-AU05-1035	M	White Shark

1.3 Setting data arguments

This function accepts a range of data types, including co-dominant and dominant markers:

- DArT data: Input data as .csv file.
  - Co-dominant markers:
    - DArT counts: Has the read depth per allele
    - DArT genotypes: Either 1-row (coded 0, 1 or 2 for each marker) or 2-rows (coded 0 or 1 for each allele)
  - Dominant markers:
    - DArT silico (counts): has the read depth per marker
    - DArT silico (genotypes): has the P/A per marker (i.e. 0 or 1)
- ddRAD data: input data as .gds or .vcf file
  - Radiator can import vcf files from the following genotype callers: STACKS, GATK, platypus, samtools and ipyrad
- RAD data (*untested*)

Power to detect sex-linked markers (*from personal experience*): DArT silico (counts) > DArT counts or RAD-type data with allele read depth > DArT silico (genotypes) > RAD-type data (genotypes) and DArT (1-row, 2-rows genotypes).

```
data <- "WhiteShark_SNP_counts.csv"
silicodata <- "WhiteShark_silico_count.csv"
```

1.4 Run sexy_markers - Basic input

This is the minimum requirement if `interactive.filter = TRUE`.

```
sex.results <- radiator::sexy_markers(
  data = data,
  silicodata = silicodata, #only for DArT
  strata = strata,
  interactive.filter = TRUE)
```

1.5 sexy_markers output

1.5.1 Read data

First the data is read with the `radiator::read_dart`, `radiator::read_vcf` or `radiator::read_rad` functions.

```
################################################################################
######################### radiator::sexy_markers################################
################################################################################
Execution date@time: 20220422@1502
Folder created: sexy_markers_20220422@1502
```

```
File written: radiator_sexy_markers_args_2022042290@1502.tsv
Reading DArT file...
Number of blacklisted samples: 276
DArT SNP format: alleles coverage in 2 Rows counts
Generating genotypes and calibrating REF/ALT alleles...
Number of markers recalibrated based on counts of allele read depth: 2448
Generating GDS...
File written: radiator_2022042290@1502.gds.rad

Number of chrom: 1
Number of locus: 21502
Number of SNPs: 23393
Number of populations: 3
Number of individuals: 558

Number of ind/pop:
M = 250
F = 251
U = 57

Number of duplicate id: 0

Computation time, overall: 30 sec
```

1.5.1.1 Filter data

It is important that the data is not filtered too stringently, since this would remove potential sex-linked markers. Yet some filters are included to reduce false positive results.

1.5.1.2 Monomorphic loci

Monomorphic markers are filtered out because they are not informative.

```
DArT conversion timing: 100 sec
There is currently an issue with the cluster allocation in WINDOWS systems.
Consequently, we set the 'parallel.core' to 1. This will only affect the
data-filtering time.

Filter monomorphic markers
Number of individuals / strata / chrom / locus / SNP:
    Blacklisted: 0 / 0 / 0 / 1144 / 1322
```

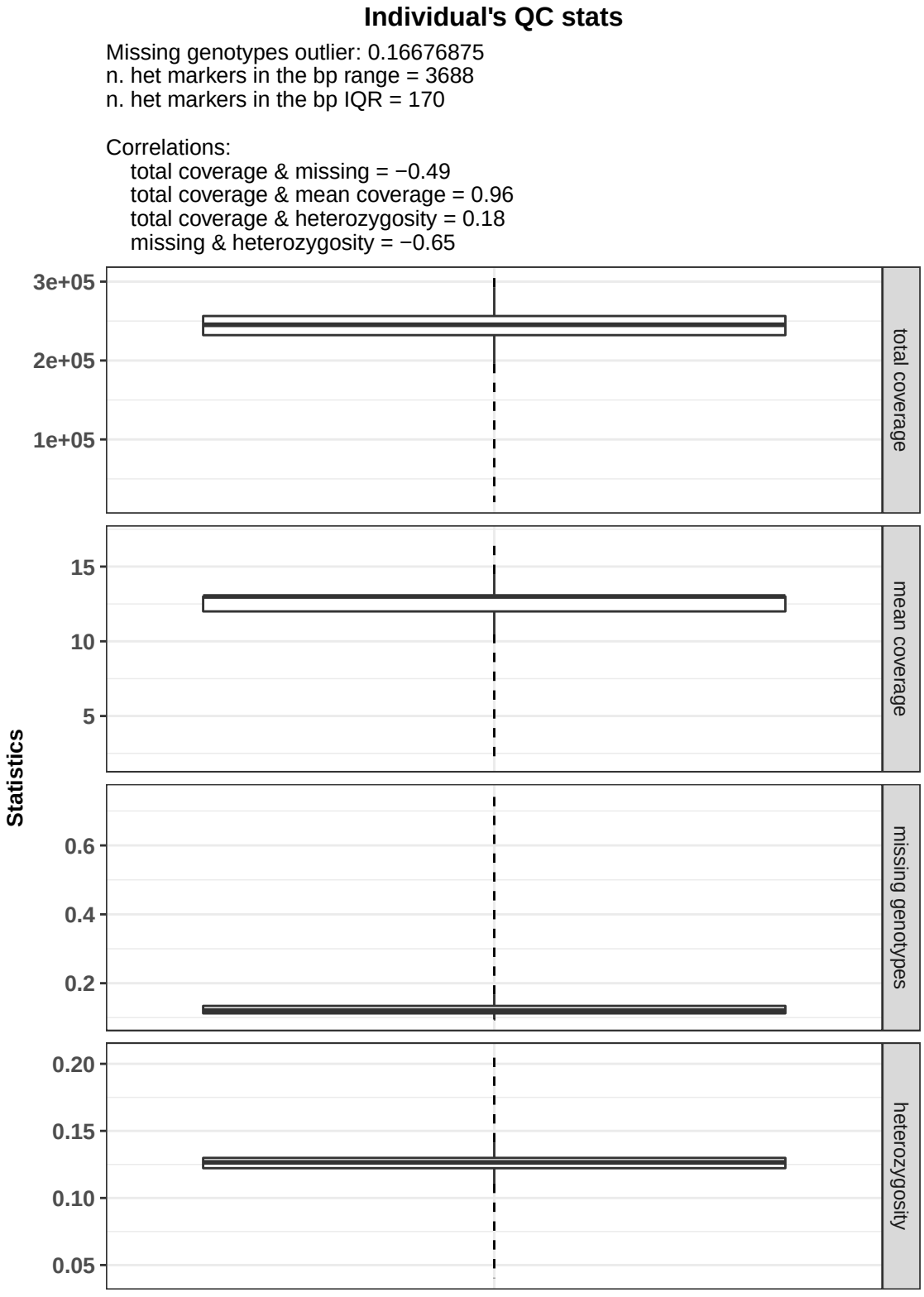
1.5.1.3 Individual missingness and heterozygosity & monomorphic loci

Individuals with high missingness and high heterozygosity could resemble bad DNA, cross-contamination or different population/species. This would affect the number of sex-linked markers that are identified with the heterozygosity method.

```
################################################################################
######################### radiator::filter_individuals #########################
################################################################################
Execution date@time: 2022042290@1504
Function call and arguments stored in: radiator_filter_individuals_args_2022042290@1504.tsv
Interactive mode: on

Step 1. Visualization
Step 2. Missingness
```

```
Step 3. Heterozygosity
Step 4. Total Coverage (if available)


Filters parameters file generated: filters_parameters_20220422@1504.tsv
Filters parameters file: initiated

Step 1. Visualization of samples QC

[================================================] 100%, completed in 0s
File written: individuals qc info and stats summary
File written: individuals qc plot

Step 2. Filtering markers based individual missingness/genotyping

Do you want to blacklist samples based on missingness ? (y/n):
y

2 options to blacklist samples:
1. based on the outlier statistics
2. enter your own threshold
1

Removing outliers individuals based on genotyping statistics: 0.16676875
Filters parameters file: updated
################################### RESULTS ###################################

Filter individuals based on missingness: 0.16676875
Number of individuals / strata / chrom / locus / SNP:
    Before: 558 / 3 / 1 / 20358 / 22071
    Blacklisted: 70 / 0 / 0 / 0 / 0
    After: 488 / 3 / 1 / 20358 / 22071

Step 3. Filtering markers based on individual heterozygosity

Do you want to blacklist samples based on heterozygosity ? (y/n):
y

2 options to blacklist samples:
1. based on the outlier statistics
2. enter your own threshold
1

Removing outliers individuals based on heterozygosity statistics: 0.11068875 / 0.14141075
Filters parameters file: updated
################################### RESULTS ###################################

Filter individuals based on heterozygosity: 0.11068875 0.14141075
Number of individuals / strata / chrom / locus / SNP:
    Before: 488 / 3 / 1 / 20358 / 22071
    Blacklisted: 8 / 0 / 0 / 0 / 0
    After: 480 / 3 / 1 / 20358 / 22071

Step 4. Filtering markers based on individual's total coverage
```

```
Do you want to blacklist samples based on total coverage ? (y/n):
y

2 options to blacklist samples:
1. based on the outlier statistics
2. enter your own threshold
1

Removing outliers individuals based on total coverage statistics: 195778.625 / 292781.625
Filters parameters file: updated
################################### RESULTS ###################################

Filter individuals based on total coverage: 195778.625 292781.625
Number of individuals / strata / chrom / locus / SNP:
    Before: 480 / 3 / 1 / 20358 / 22071
    Blacklisted: 35 / 0 / 0 / 0 / 0
    After: 445 / 3 / 1 / 20358 / 22071

Filter monomorphic markers
Number of individuals / strata / chrom / locus / SNP:
    Blacklisted: 0 / 0 / 0 / 1834 / 2212

Computation time, overall: 184 sec
######################### completed filter_individuals ########################
```



1.5.1.4 Short-distance linkage Markers with more than 1 SNP per fragment are removed, since they do not add extra information and add computational time.

While it is possible to filter on long-distance linkage (R^2), we believe this will filter out many sex-linked markers.

```
################################################################################
############################## radiator::filter_ld ##############################
################################################################################
Execution date@time: 2022042229@1507
Function call and arguments stored in: radiator_filter_ld_args_202204229@1507.tsv

Interactive mode: on

Step 1. Short distance LD threshold selection
Step 2. Filtering markers based on short distance LD
Step 3. Long distance LD pruning selection
Step 4. Threshold selection
Step 5. Filtering markers based on long distance LD


Filters parameters file generated: filters_parameters_202204229@1507.tsv
Filters parameters file: initiated
Minimizing short distance LD...
    The range in the number of SNP/locus is: 1-7

Step 1. Short distance LD threshold selection
the goal is to keep only 1 SNP per read/locus
Choose the filter.short.ld threshold
Options include:
1: mac (Not sure ? use mac...)
2: random
3: first
4: middle
5: last
1

Step 2. Filtering markers based on short distance LD
filter.short.ld = mac
[==================================================] 100%, completed in 0s
File written: whitelist.short.ld.tsv
File written: blacklist.short.ld.tsv
Filters parameters file: updated
################################### RESULTS ###################################

Filter short ld threshold: mac
Number of individuals / strata / chrom / locus / SNP:
    Before: 445 / 3 / 1 / 18524 / 19859
    Blacklisted: 0 / 0 / 0 / 0 / 1335
    After: 445 / 3 / 1 / 18524 / 18524

Do you want to continue filtering using long distance ld  ? (y/n):
n

Computation time, overall: 13 sec
############################## completed filter_ld ##############################
```

```
Sex-ratio (F/M): 0.98
```

1.5.2 Read SILICO data

If DArT silico data is available, this will be read with the `radiator::read_dart` function.

Any individuals that were filtered out for the SNP data, will also be filtered for the SILICO data

```
################################################################################
############################## radiator::read_dart #############################
################################################################################
Execution date@time: 20220422 9@1508
Reading DArT file...
    Number of individuals: 834
Analyzing strata file
    Number of strata: 3
    Number of individuals: 445
Using individuals in strata file to filter individuals in DArT file

Note: Careful if using DArT statistics generated for all samples...
Number of blacklisted samples: 389

DArT characteristics:
DArT SNP format: silico DArT
Synchronizing data and strata...
    Number of strata: 3
    Number of individuals: 445
File written: radiator.silico.dart_20220422 9@1509.rad
################################### SUMMARY ####################################

Number of clones: 116206
Number of populations: 3
Number of individuals: 445

Number of ind/pop:
M = 203
F = 199
U = 43

Number of duplicate id: 0

Computation time, overall: 45 sec
############################## completed read_dart #############################
```

1.5.3 Summarise the data by sex

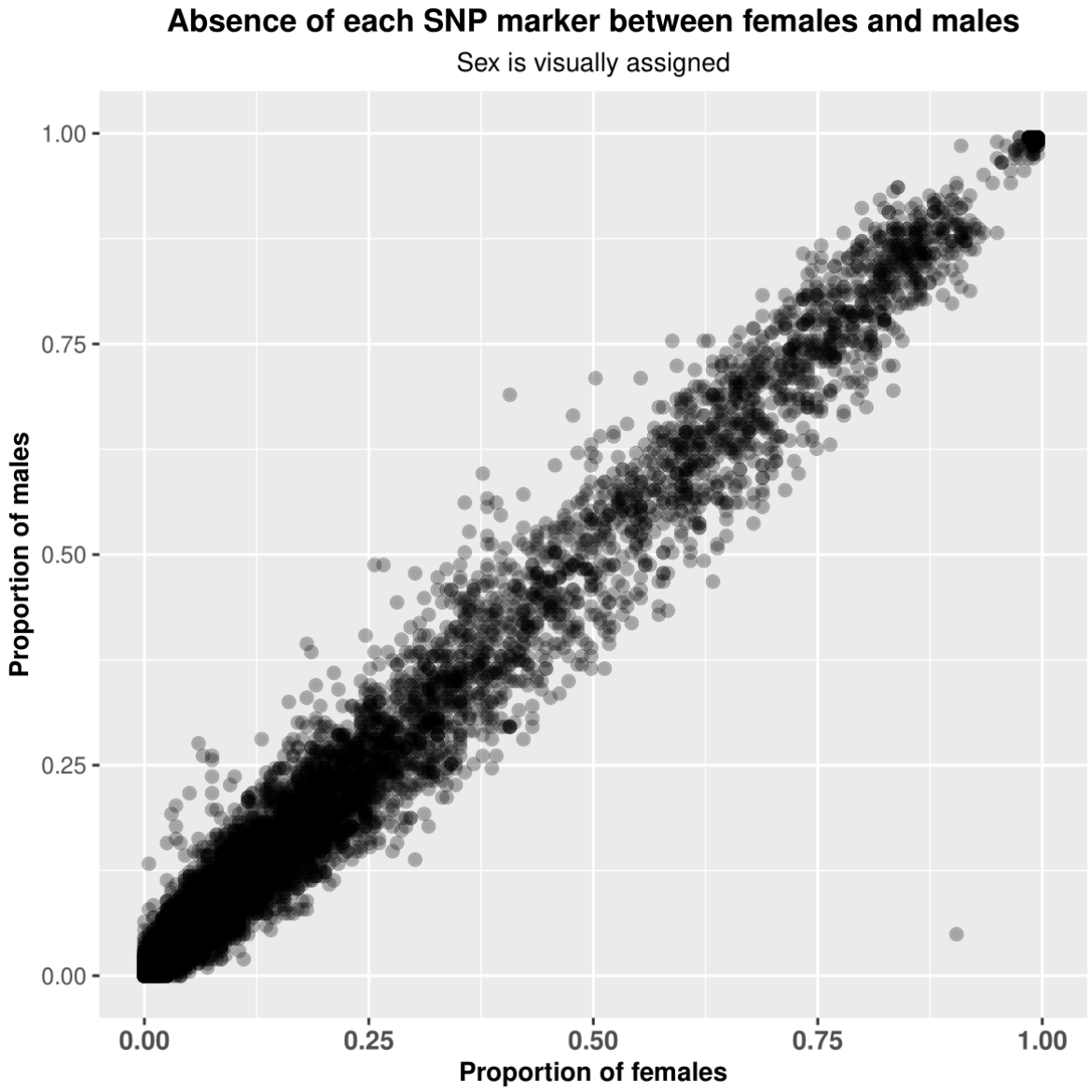
Per sex, the *mean presence-absence*, *mean heterozygosity* and *mean read depth* (if present) is calculated.

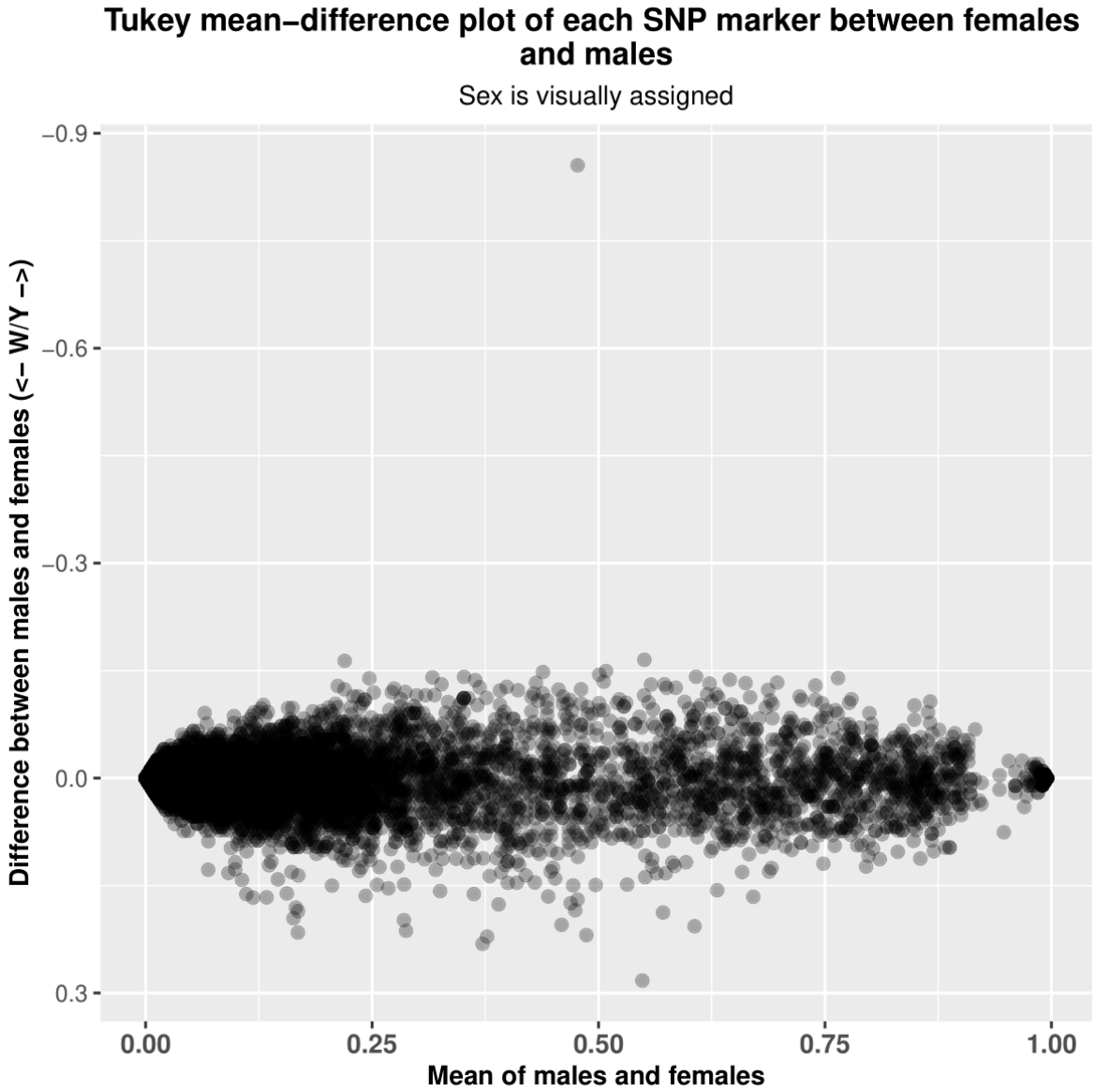
1.5.4 Find heterogametic sex-linked markers

As described above and in function documentation.

1.5.4.1 Presence-absence method for SNP data

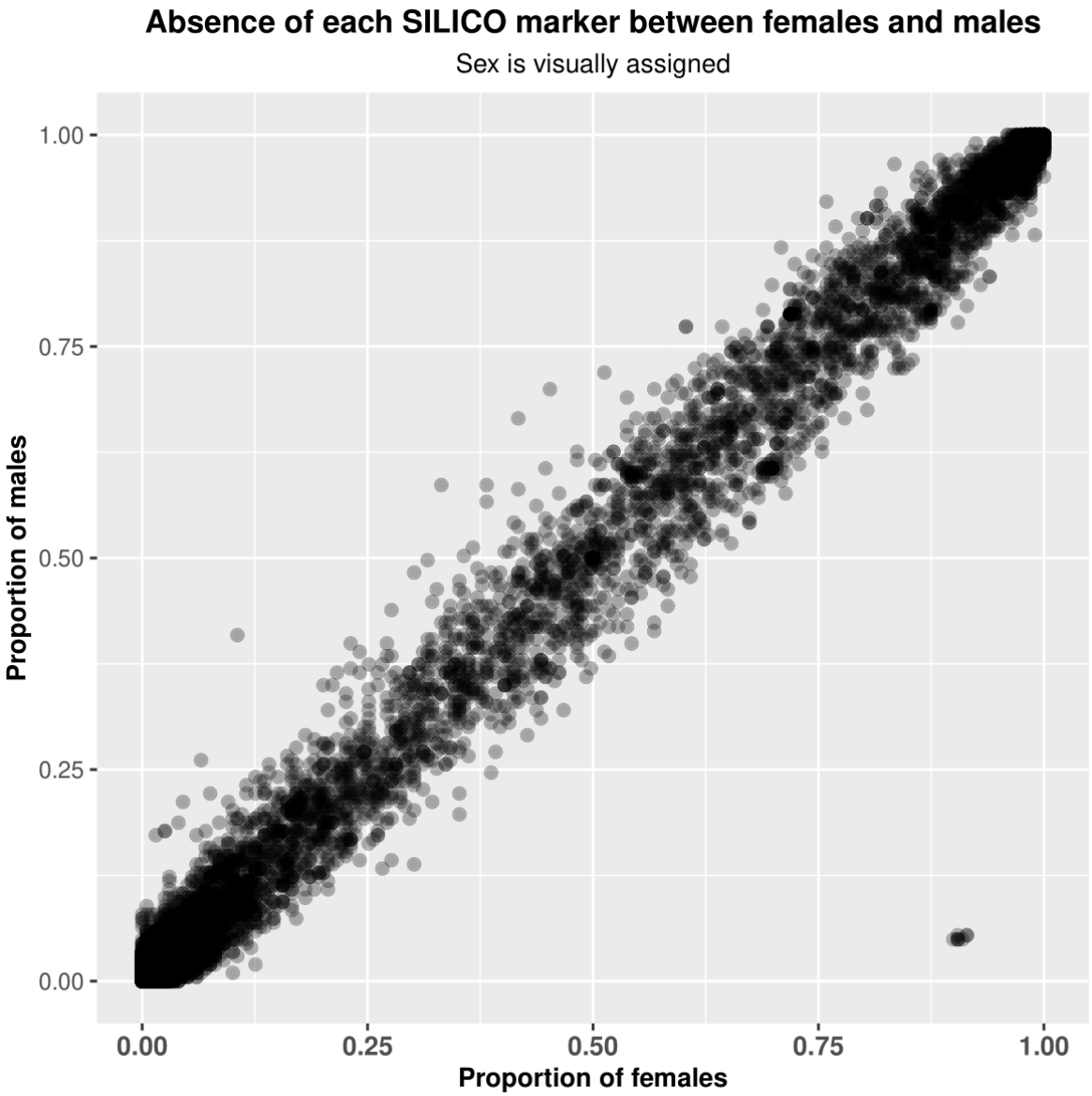
```
################################################################################
######################## Start finding sex-linked markers ######################
################################################################################
Files written: '1A.sexy_markers_PA_scatter_plot.pdf' &
'1B.sexy_markers_PA_tuckey_plot.pdf'

P/A method of SNPs:
Look at the figures: Do you want to select Y/W-linked markers (y/n):
y
Choose the threshold for Y/W-linked markers, note that the Y-axis is inverted (-1 to 1):
-0.6
Files written: '2A.sexy_markers_SILICO_PA_scatter_plot.pdf' &
'2B.sexy_markers_SILICO_PA_tuckey_plot.pdf'
```

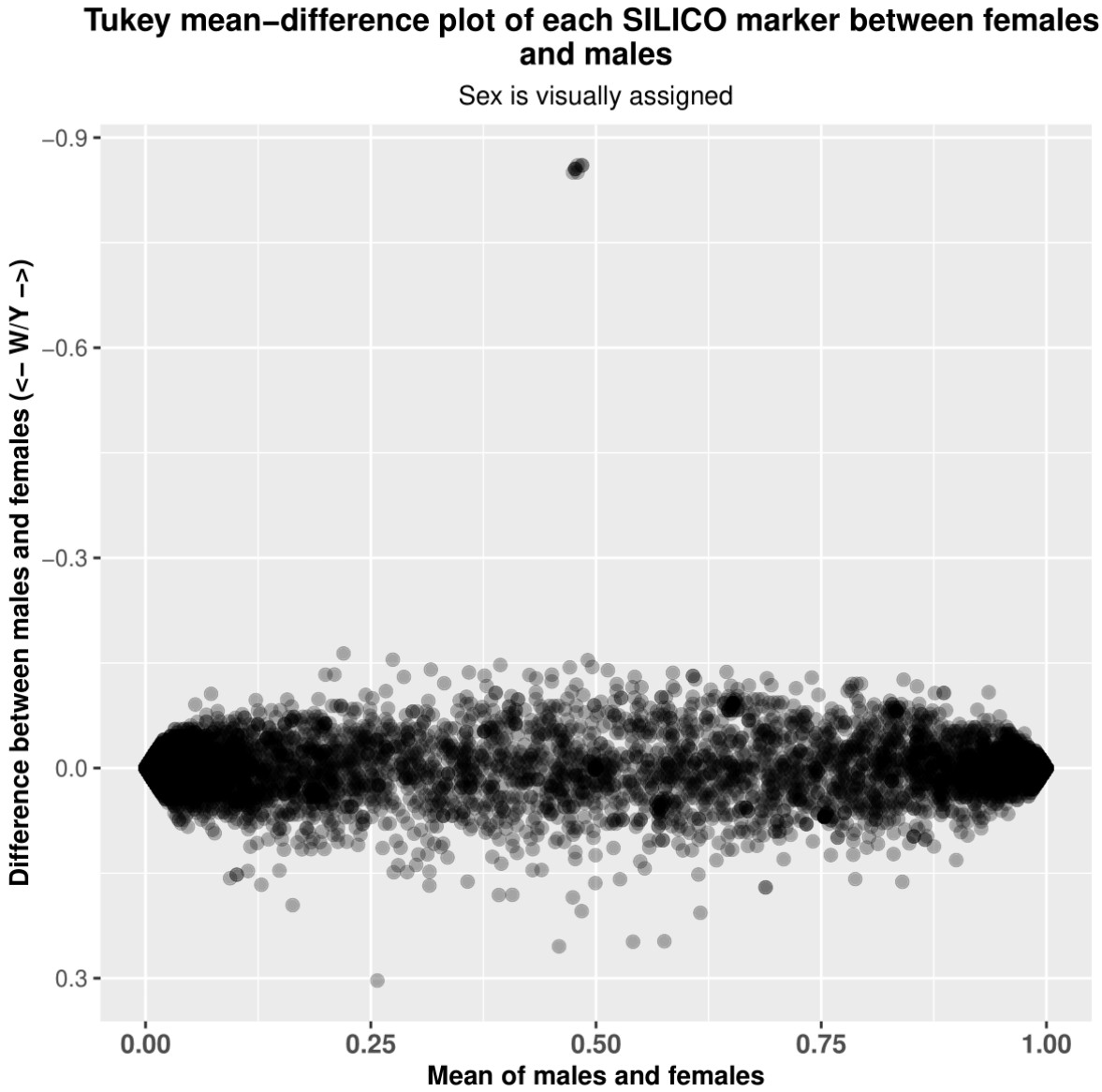




1.5.4.2 Presence-absence method for SILICO data

```
P/A method of SILICOs:
Look at the figures: Do you want to select Y/W-linked markers (y/n):
y
Choose the threshold for Y/W-linked SILICO markers,
note that the Y-axis is inverted (-1 to 1):

-0.6
```


Absence of each SILICO marker between females and males
Sex is visually assigned
Proportion of males
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Proportion of females



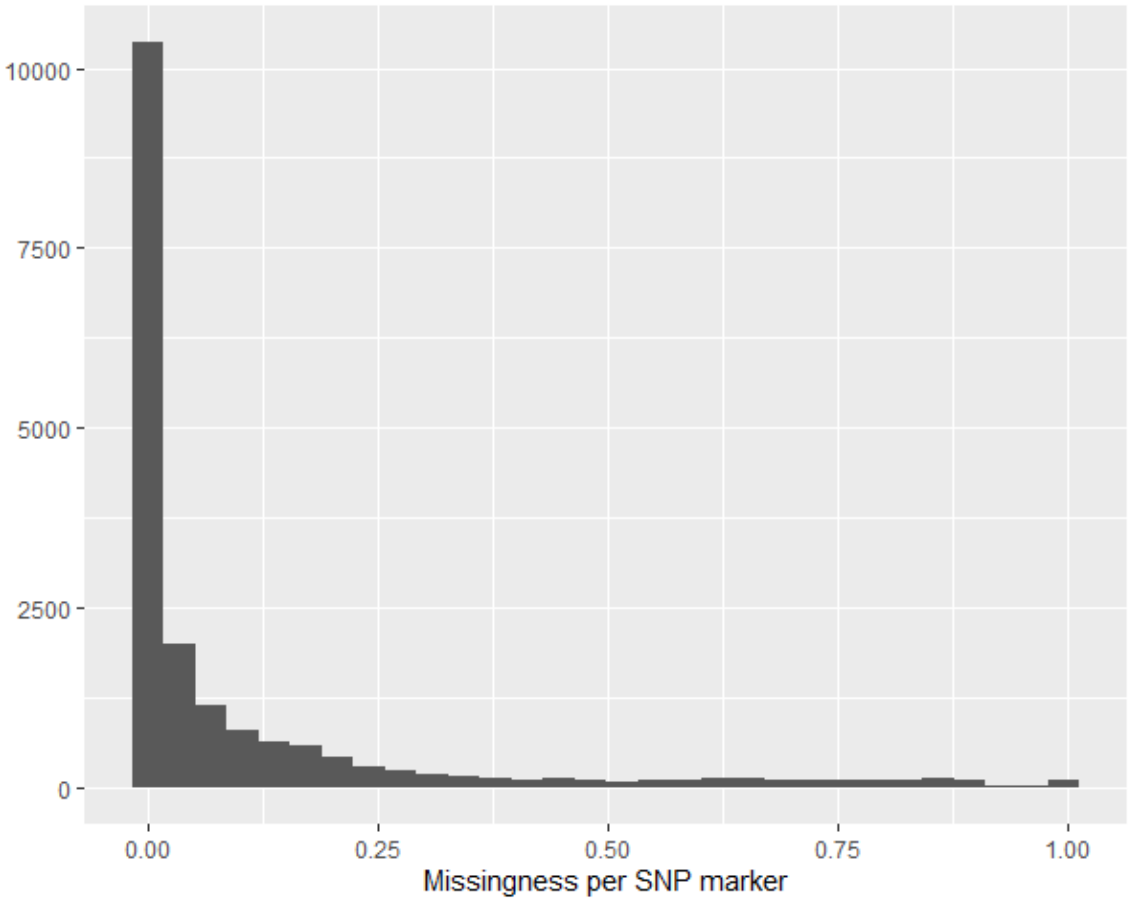
1.5.5 Recalculate the sex ID based on genetic sex assignment

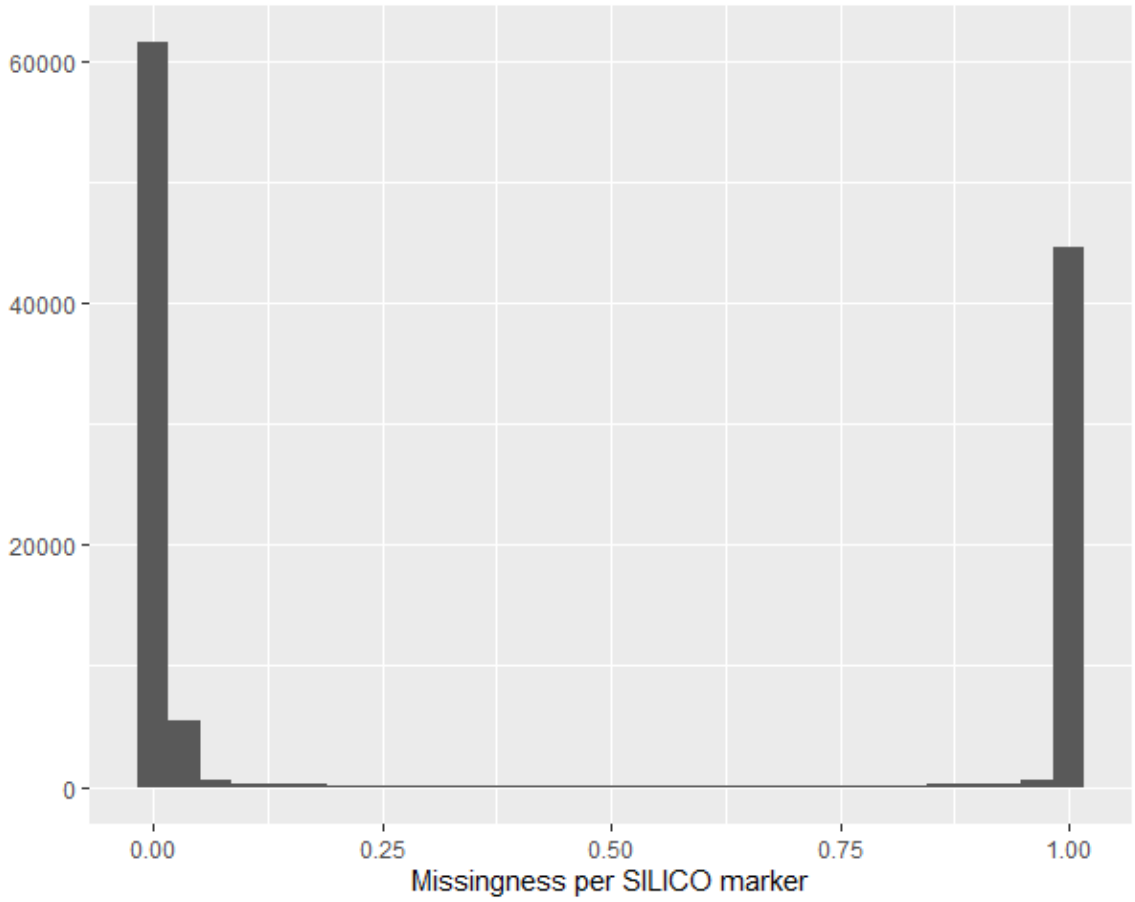
```
# A tibble: 6 x 3
  Visual_Sex Genetic_Sex_SNP     n
  <chr>      <chr>           <int>
1 F          F                 182
2 F          M                  17
3 M          F                  10
4 M          M                 193
5 U          F                  20
6 U          M                  23
# A tibble: 6 x 3
  Visual_Sex Genetic_Sex_SILICO     n
  <chr>      <chr>              <int>
1 F          F                    182
```

```
2 F          M                     17
3 M          F                     10
4 M          M                    193
5 U          F                     20
6 U          M                     23
# A tibble: 2 x 3
  Genetic_Sex_SNP Genetic_Sex_SILICO     n
  <chr>           <chr>              <int>
1 F               F                    212
2 M               M                    233
For further analysis, do you want to continue based on
(1) visual, (2) genetic SNP or (3) genetic SILICO sex?
We advise (3) for better results
3

Sex and summary statistics will be calculated accoriding to: 3
New strata file with genetic sex written.


Individuals with unknown sex ID are now included in the analysis.
        The new sex-ratio (F/M) is: 0.91
# A tibble: 2 x 2
  GENETIC_SEX     N
  <chr>       <int>
1 F             212
2 M             233
```

1.5.6 Filter SNP and SILICO data for markers with high missingness

```
Have a look at the plot: Choose the upper threshold for
missingness per SNP marker (e.g. 0.2):
0.25
For detecting heterogametic markers, the SILICO data is filtered on a
missingness >: 0.25

Have a look at the plot: Choose the upper threshold for missingness per
SILICO marker (e.g. 0.2):
0.25
For detecting homogametic markers, the SILICO data is filtered on a
missingness > 0.25
```


10000
7500
5000
2500
0
0.00
0.25
0.50
0.75
1.00
Missingness per SNP marker

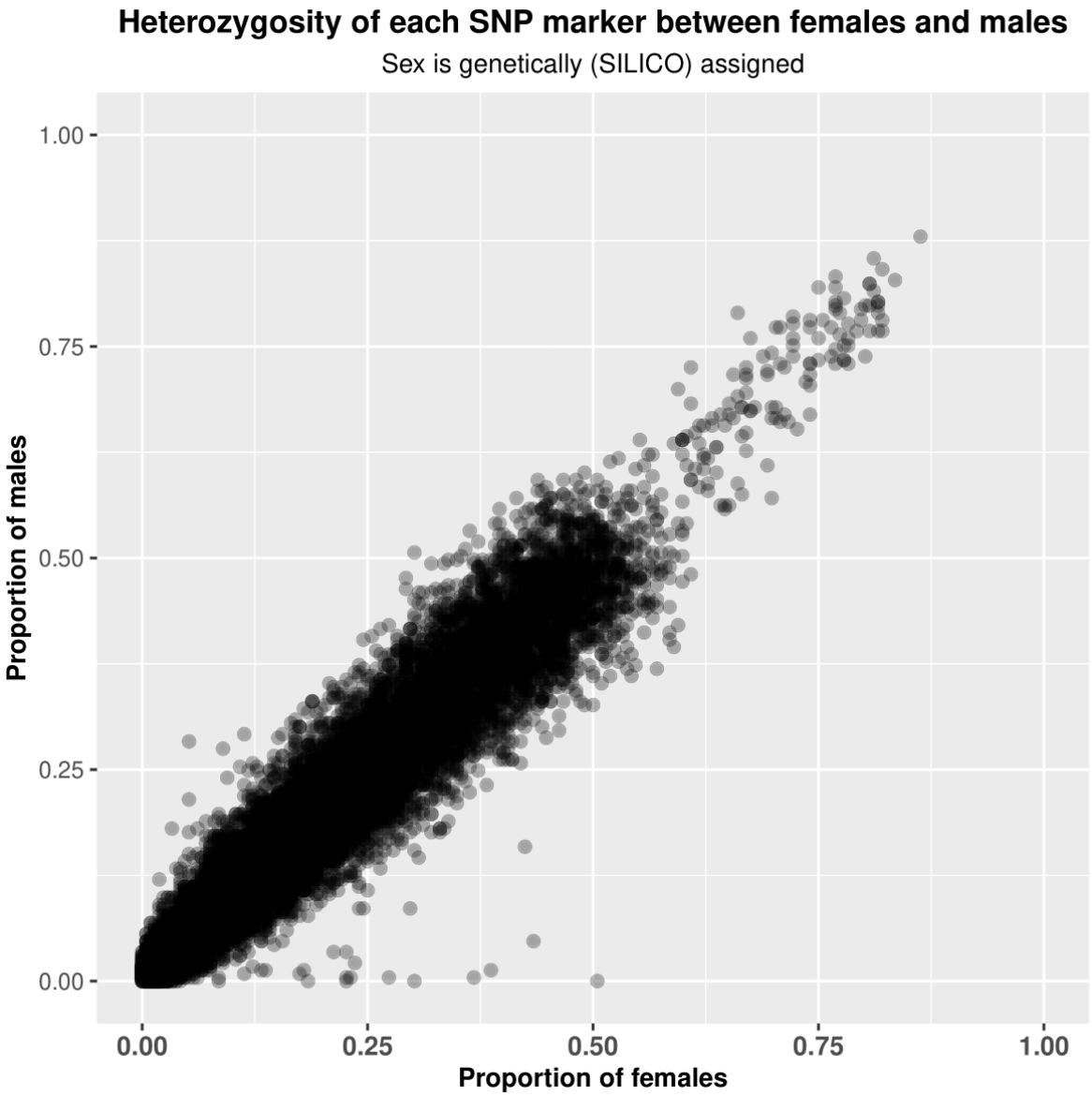


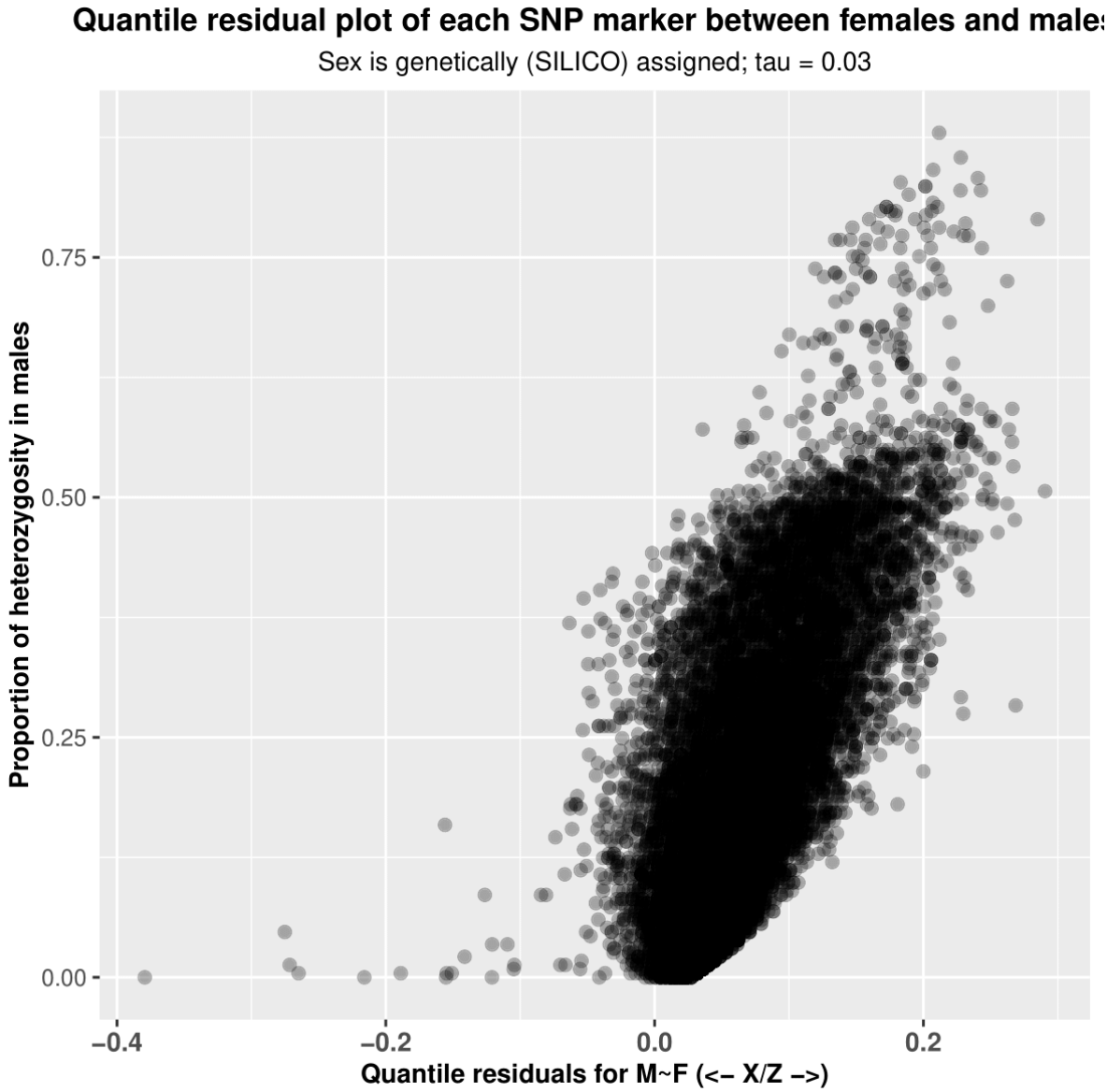
1.5.7 Find homogametic sex-linked markers

As described above and in function documentation.

1.5.7.1 Heterozygosity method for SNP data The threshold to identify sex markers is based on the residuals of the quantile regression; specifically, this depends on the `tau` value. This value is the quantile to calculate the residuals to. In our case, the value of tau to distinguish heterozygosity outliers depends on the total number of markers and the number of homogametic markers. Since you generally expect to have very few X-or Z-linked markers compared to all the total number of markers, we set `tau = 0.03`.

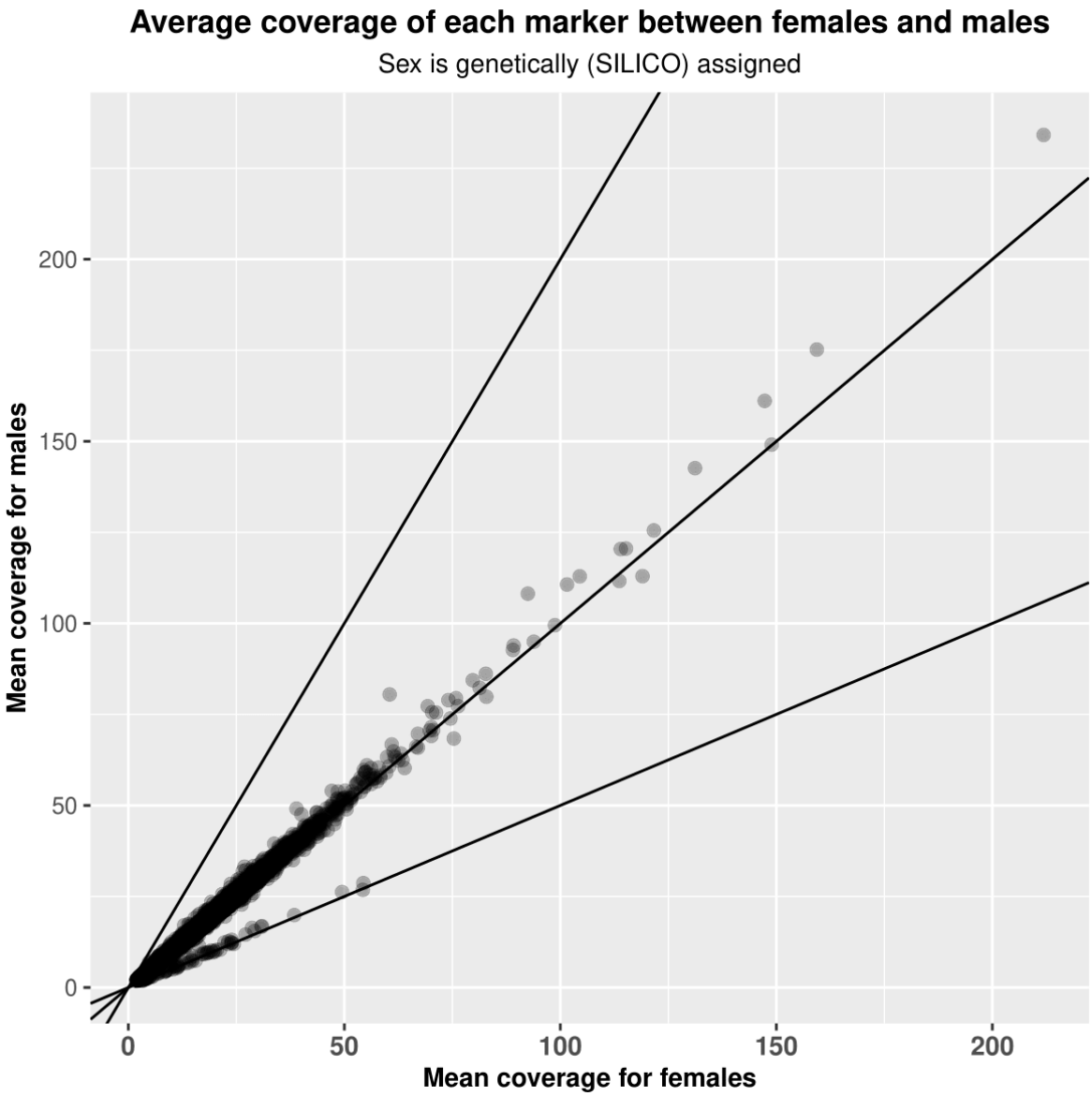
```
Start finding X- or Z-linked markers.
CAUTION: this step is largely affected by structure in your data
(e.g. families, populations or species).

Heterozygosity method of SNPs:
Do you want to plot the residuals for
(1) X-linked markers: i.e. heterozygous for females
(2) Z-linked markers: i.e. heterozygous for males
1
Look at the figures: Do you want to select X/Z-linked markers (y/n):
y
Choose the threshold for X/Z-linked markers (-1 to 1):
-0.1
```

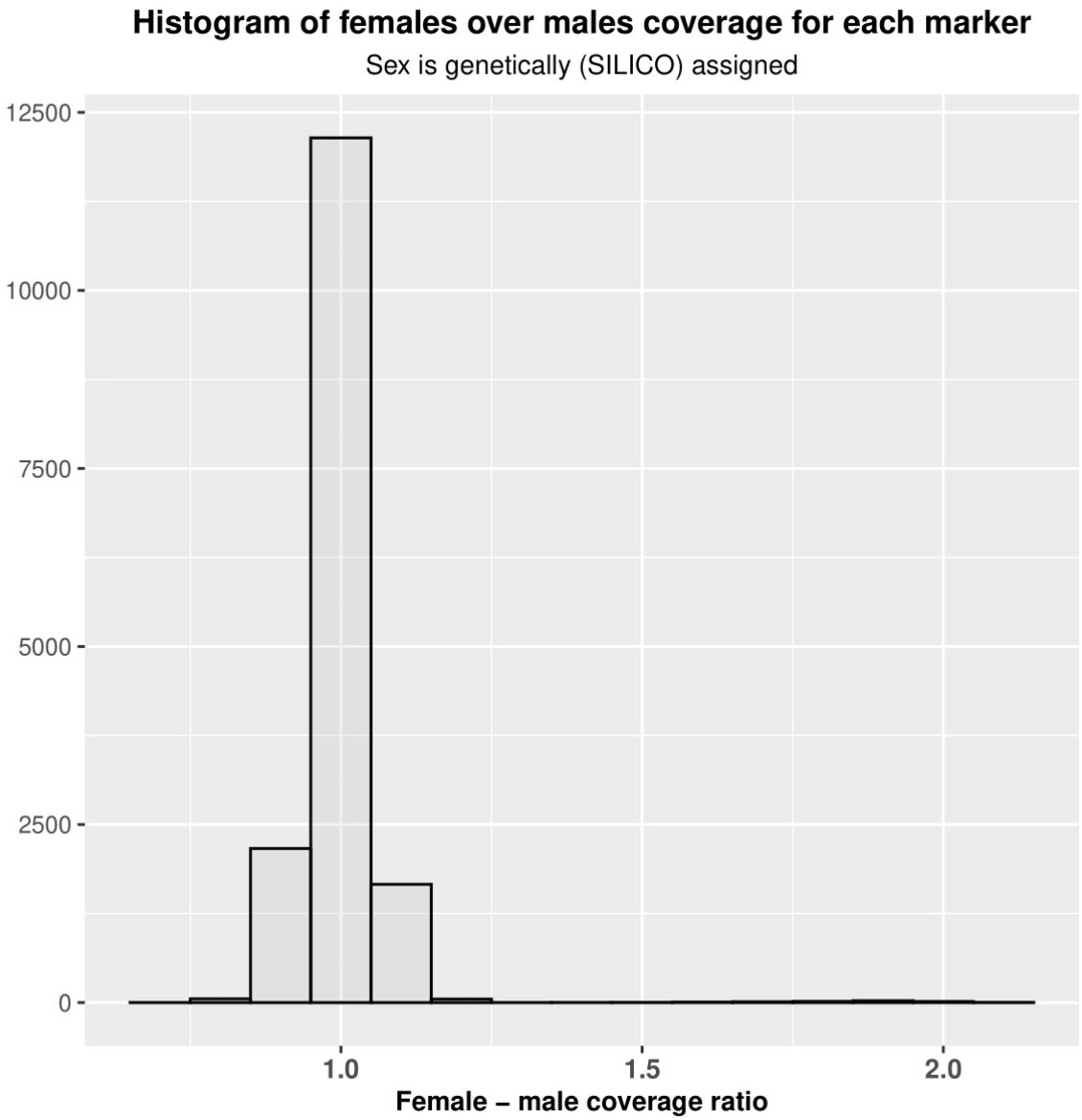

Heterozygosity of each SNP marker between females and males
Sex is genetically (SILICO) assigned
Proportion of males
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Proportion of females

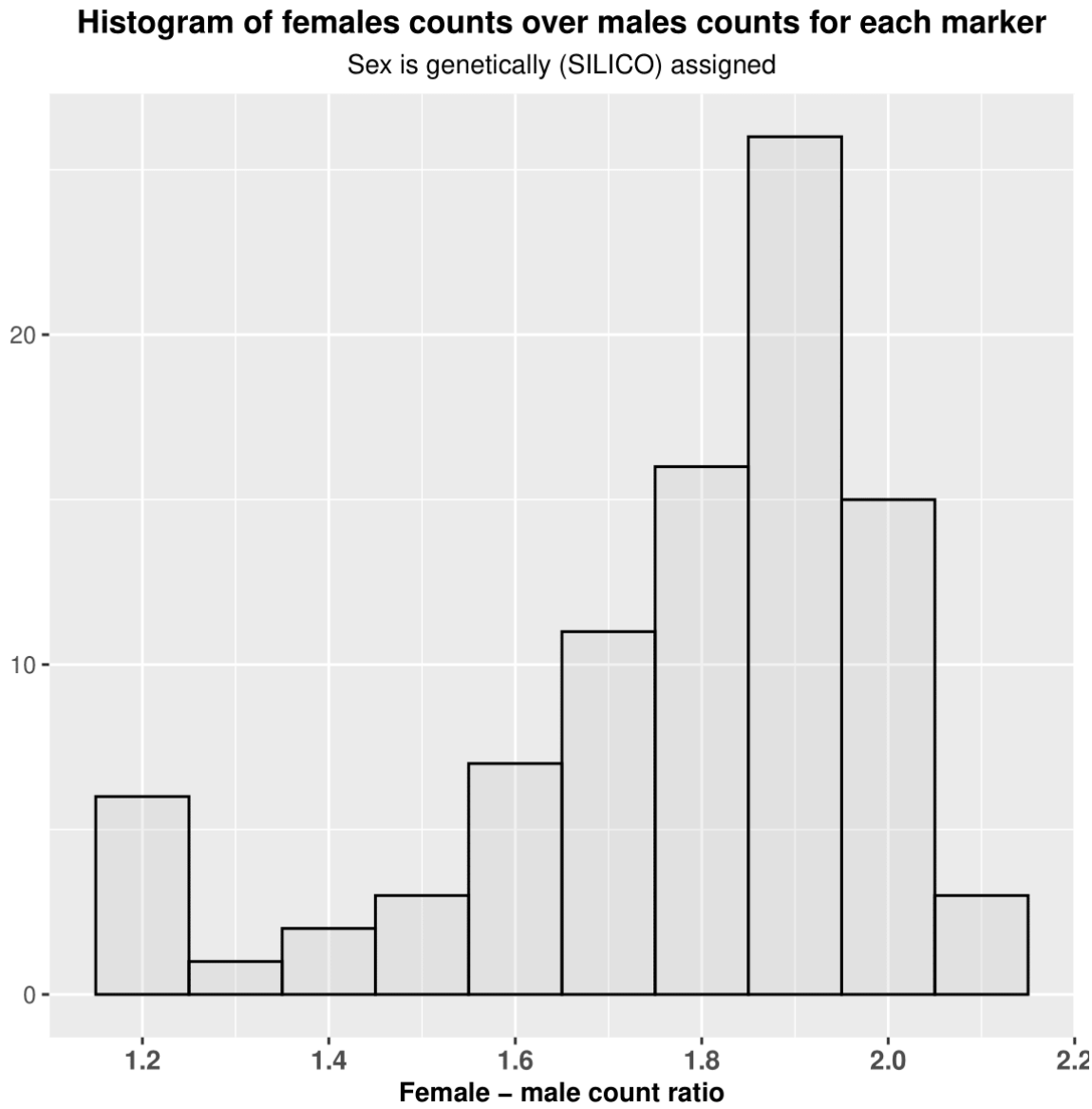


1.5.7.2 Coverage method for SNP data

```
Choose the lower OR upper threshold to subset the histogram
(e.g. scaling the plot to a ratio < 0.8 or ratio > 1.2)
1.2
Files written: '4A.sexy_markers_RD_scat_plot.pdf' & '4B.sexy_markers_RD_hist_plot.pdf' &
'4C.sexy_markers_RD_hist_subsetted_plot.pdf'
Coverage method of SNPs:
Look at the figures: Do you want to select X/Z-linked markers (y/n):
y
Choose the RATIO threshold for X/Z-linked markers:
1.5
```

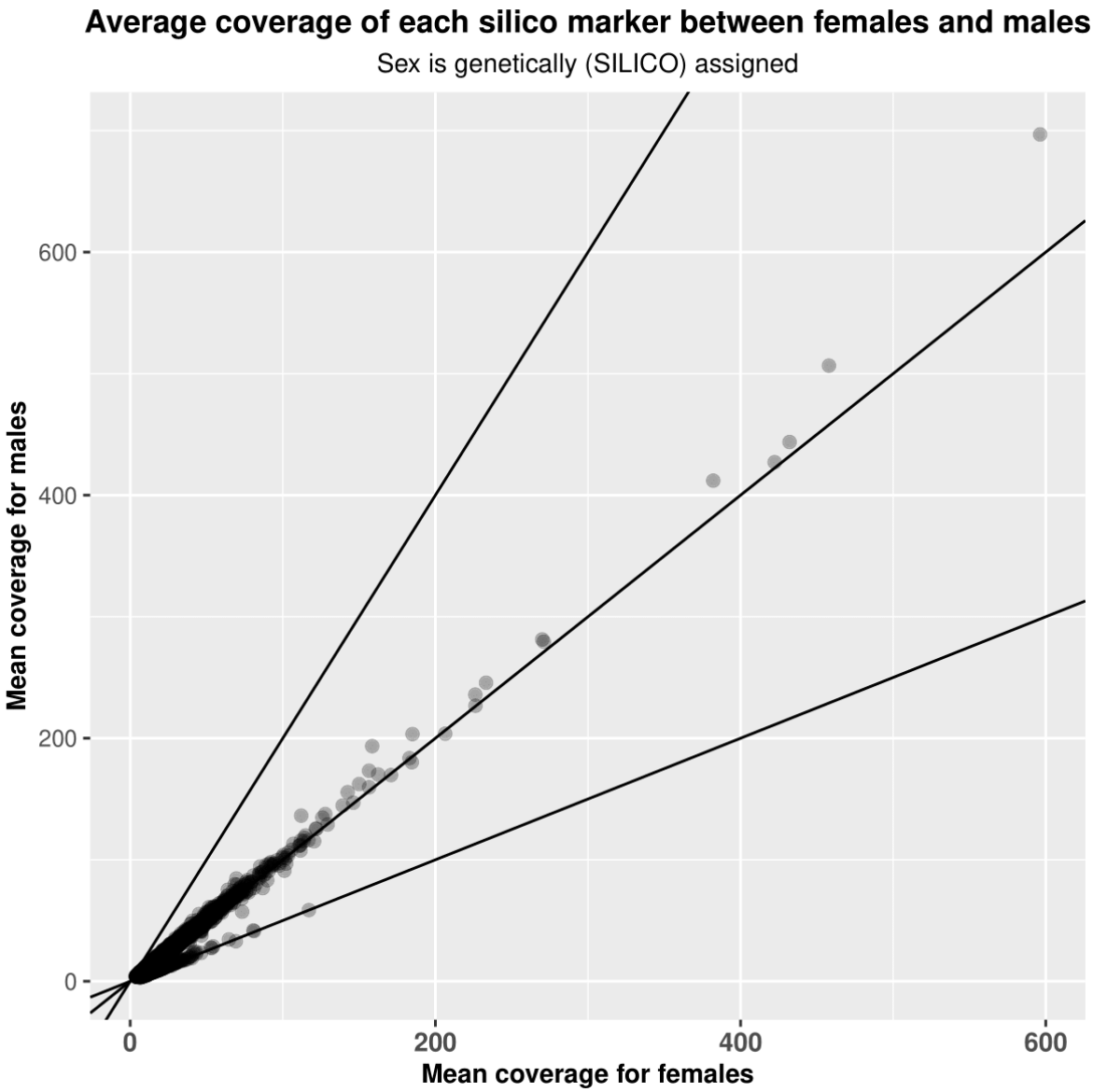

Average coverage of each marker between females and males
Sex is genetically (SILICO) assigned
Mean coverage for males
Mean coverage for females
0
50
100
150
200

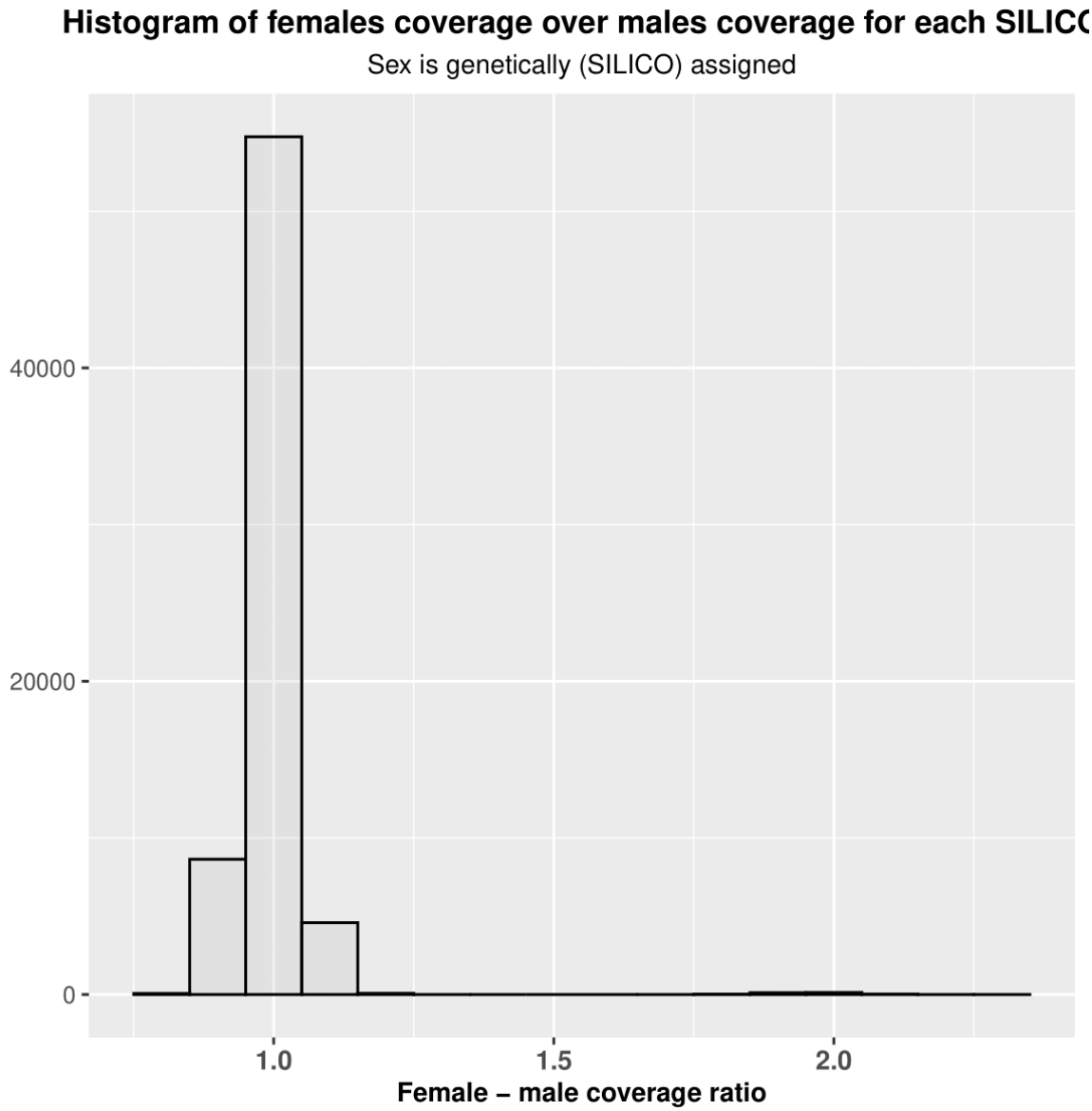

Histogram of females over males coverage for each marker
Sex is genetically (SILICO) assigned
12500
10000
7500
5000
2500
0
1.0
1.5
2.0
Female – male coverage ratio

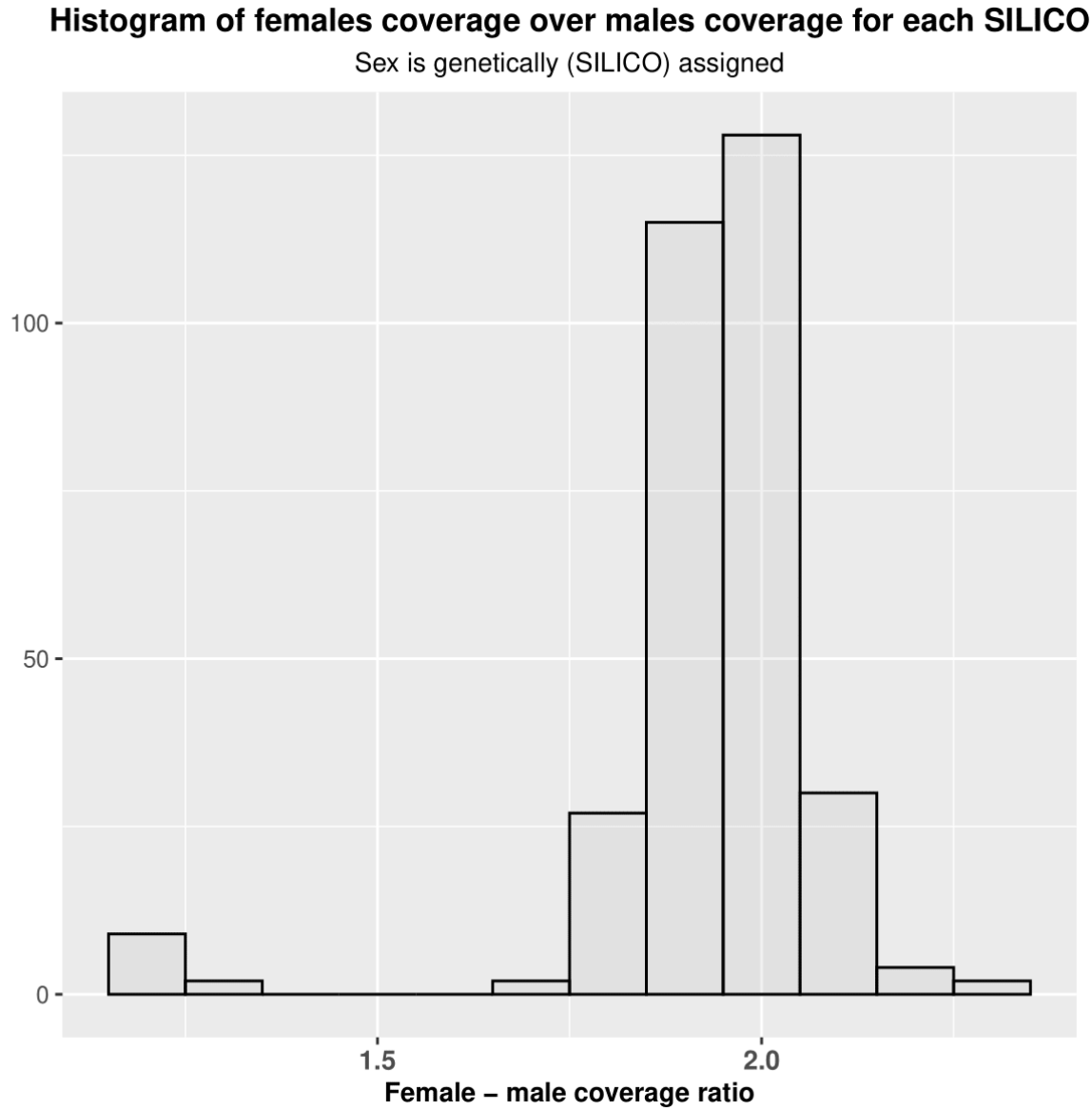


1.5.7.3 Coverage method for SILICO data

```
Choose the lower OR upper threshold to subset the histogram
(e.g. scaling the plot to a ratio < 0.8 or ratio > 1.2)
1.2
Files written: '5A.sexy_markers_SILICO_RD_scat_plot.pdf' &
'5B.sexy_markers_SILICO_RD_hist_plot.pdf' &
'5C.sexy_markers_SILICO_RD_hist_subsetted_plot.pdf'
Coverage method of SILICOs:
Look at the figures: Do you want to select X/Z-linked markers (y/n):
y
Choose the RATIO threshold for X/Z-linked SILICO markers:
1.5
```

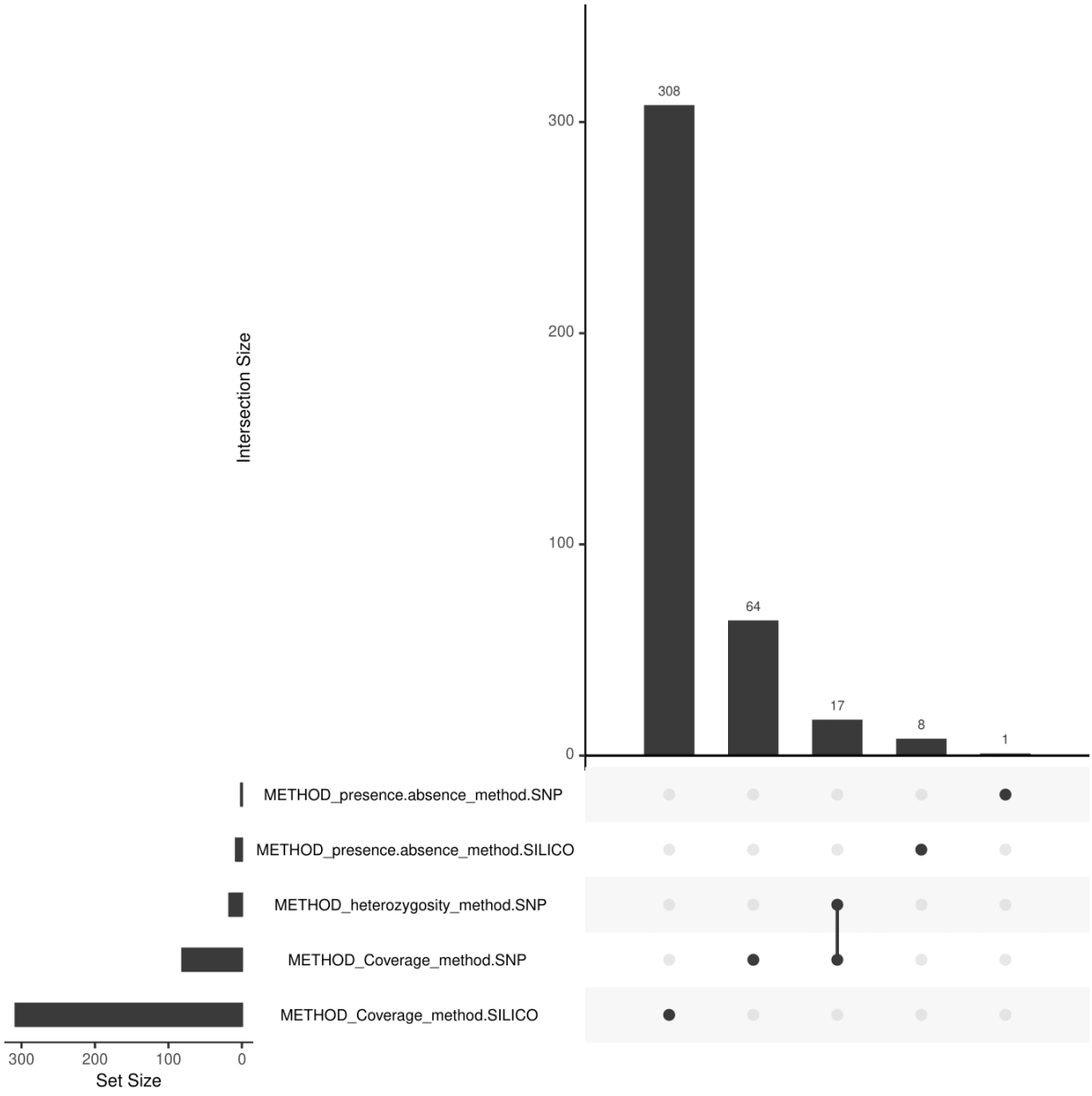

Average coverage of each silico marker between females and males
Sex is genetically (SILICO) assigned
Mean coverage for males
Mean coverage for females
0
200
400
600


Histogram of females coverage over males coverage for each SILICO
Sex is genetically (SILICO) assigned
40000
20000
0
1.0
1.5
2.0
Female – male coverage ratio



1.5.8 Summary

```
Summary table of sex-linked markers by method of discovery:
        Coverage_method-SILICO            Coverage_method-SNP
                           308                             81
     heterozygosity_method-SNP presence/absence_method-SILICO
                            17                              8
   presence/absence_method-SNP
                             1
The 'upset' plot shows any overlapping sex-linked markers between methods
File written: '6.sexy.markers_upsetrplot.pdf'
File written:'7.sexy_markers_sequences.fasta'

Computation time, overall: 282 sec
############################ sexy markers completed ##############################
```



1.5.9 Advanced input

To run the function without the need to provide input, we set `interactive.filter = FALSE`. This allows for the knitting of an R markdown document. However, this requires additional arguments to run the function. **Note that we advice to run this function with `interactive.filter = TRUE`**

```
sex.results <- radiator::sexy_markers(
  data = data,
  silicodata = silicodata,
  strata = strata,
  coverage.thresholds = 1,
  filters = TRUE,
```

```
  parallel.core = 1,
  folder.name = NULL,
  tau = 0.03,
  species = "White_shark", #optional
  population = "Australia & New Zealand", #optional
  interactive.filter = FALSE,

  # The arguments below are only if interactive.filter = FALSE
  mis.threshold.data = 0.2,
  mis.threshold.silicodata = 0.2,
  threshold.y.markers = (-0.6),
  threshold.y.silico.markers = (-0.6),
  sex.id.input = 3,
  het.qr.input = 1,
  threshold.x.markers.qr = (-0.2),
  zoom.data = 1.2,
  threshold.x.markers.RD = 1.6,
  zoom.silicodata = 1.2,
  threshold.x.markers.RD.silico = 1.6,
  )
save(sex.results, file = "sexy-markers_results.Rdata")
```

---

2 Other approaches to identify SLMs

2.1 Read and filter the data

We applied the same data-quality filters as in the 'sexy_markers' function.

```
strata <- strata[!strata$STRATA=="U",]
WS.data <- radiator::read_dart(
  data,
  strata,
  tidy.dart = FALSE,
  verbose = FALSE,
  parallel.core = parallel::detectCores() - 1
)
```

```
Reading DArT file..
Number of blacklisted samples: 333

DArT SNP format: alleles coverage in 2 Rows counts
Generating genotypes and calibrating REF/ALT alleles...
Number of markers recalibrated based on counts of allele read depth: 2445
Generating GDS...
File written: radiator_20220422@1050.gds.rad

Number of chrom: 1
Number of locus: 21502
Number of SNPs: 23393
Number of strata: 2
Number of individuals: 501

Number of ind/strata:
M = 250
F = 251

Number of duplicate id: 0

Computation time, overall: 33 sec
```

```
WS.data <- radiator::filter_monomorphic(
  WS.data,
  verbose = TRUE,
  parallel.core = parallel::detectCores() - 1
)
```

```
################################################################################
######################### radiator::filter_monomorphic #########################
################################################################################
Execution date@time: 20220422@1051
Folder created: 18_filter_monomorphic_20220422@1051
Function call and arguments stored in: radiator_filter_monomorphic_args_20220422@1051.tsv
Filters parameters file generated: filters_parameters_20220422@1051.tsv
File written: blacklist.monomorphic.markers_20220422@1051.tsv
```

```
Synchronizing markers.meta
File written: whitelist.polymorphic.markers_20220422@1051.tsv
################################### RESULTS ###################################

Filter monomorphic markers
Number of individuals / strata / chrom / locus / SNP:
    Before: 501 / 2 / 1 / 21502 / 23393
    Blacklisted: 0 / 0 / 0 / 1814 / 2157
    After: 501 / 2 / 1 / 19688 / 21236

Computation time, overall: 1 sec
######################### completed filter_monomorphic ########################
```

```
WS.data <- radiator::filter_individuals(
      data = WS.data,
      interactive.filter = FALSE,
      filter.individuals.missing = "outliers",
      filter.individuals.heterozygosity = "outliers",
      filter.individuals.coverage.total = "outliers",
      parallel.core = parallel::detectCores() - 1,
      verbose = TRUE,
      internal = FALSE
    ) %>%
      radiator::filter_ld(
        data = .,
        interactive.filter = FALSE,
        filter.short.ld = "mac",
        long.ld.missing = FALSE,
        parallel.core = 1,
        verbose = TRUE,
        internal = FALSE

      )
```

```
################################################################################
############################## radiator::filter_ld ##############################
################################################################################
Execution date@time: 20220422@1051
################################################################################
######################### radiator::filter_individuals #########################
################################################################################
Execution date@time: 20220422@1051
Folder created: 19_filter_individuals_20220422@1051
Function call and arguments stored in: radiator_filter_individuals_args_20220422@1051.tsv
Filters parameters file generated: filters_parameters_20220422@1051.tsv
File written: individuals qc info and stats summary
File written: individuals qc plot

Removing outliers individuals based on genotyping statistics: 0.1636125
################################### RESULTS ###################################

Filter individuals based on missingness: 0.1636125
Number of individuals / strata / chrom / locus / SNP:
    Before: 501 / 2 / 1 / 19688 / 21236
```

```
    Blacklisted: 58 / 0 / 0 / 0 / 0
    After: 443 / 2 / 1 / 19688 / 21236

Removing outliers individuals based on heterozygosity statistics: 0.1155665 / 0.1462585
################################### RESULTS ###################################

Filter individuals based on heterozygosity: 0.1155665 0.1462585
Number of individuals / strata / chrom / locus / SNP:
    Before: 443 / 2 / 1 / 19688 / 21236
    Blacklisted: 8 / 0 / 0 / 0 / 0
    After: 435 / 2 / 1 / 19688 / 21236

Removing outliers individuals based on total coverage statistics: 193273 / 278361
################################### RESULTS ###################################

Filter individuals based on total coverage: 193273 278361
Number of individuals / strata / chrom / locus / SNP:
    Before: 435 / 2 / 1 / 19688 / 21236
    Blacklisted: 9 / 0 / 0 / 0 / 0
    After: 426 / 2 / 1 / 19688 / 21236

Filter monomorphic markers
Number of individuals / strata / chrom / locus / SNP:
    Blacklisted: 0 / 0 / 0 / 1486 / 1791

Computation time, overall: 13 sec
######################### completed filter_individuals ########################
Folder created: 20_filter_ld_20220422@1051
Function call and arguments stored in: radiator_filter_ld_args_20220422@1051.tsv
Filters parameters file generated: filters_parameters_20220422@1051.tsv
Minimizing short distance LD...
    The range in the number of SNP/locus is: 1-7

Step 2. Filtering markers based on short distance LD
filter.short.ld = mac
[==================================================] 100%, completed, 0s
File written: whitelist.short.ld.tsv
File written: blacklist.short.ld.tsv
################################### RESULTS ###################################

Filter short ld threshold: mac
Number of individuals / strata / chrom / locus / SNP:
    Before: 426 / 2 / 1 / 18202 / 19445
    Blacklisted: 0 / 0 / 0 / 0 / 1243
    After: 426 / 2 / 1 / 18202 / 18202

Computation time, overall: 15 sec
############################## completed filter_ld ##############################
```

```
convert <- radiator::genomic_converter(
  data = WS.data,
  strata = NULL,
  output = c("genlight", "pcadapt"),
  filename = NULL,
```

```
  parallel.core = parallel::detectCores() - 1
)
WS.genlight <- convert$genlight
WS.pcadapt <- convert$pcadapt
save(WS.genlight, WS.pcadapt, file = "WS_SLM_objects.Rdata")
```

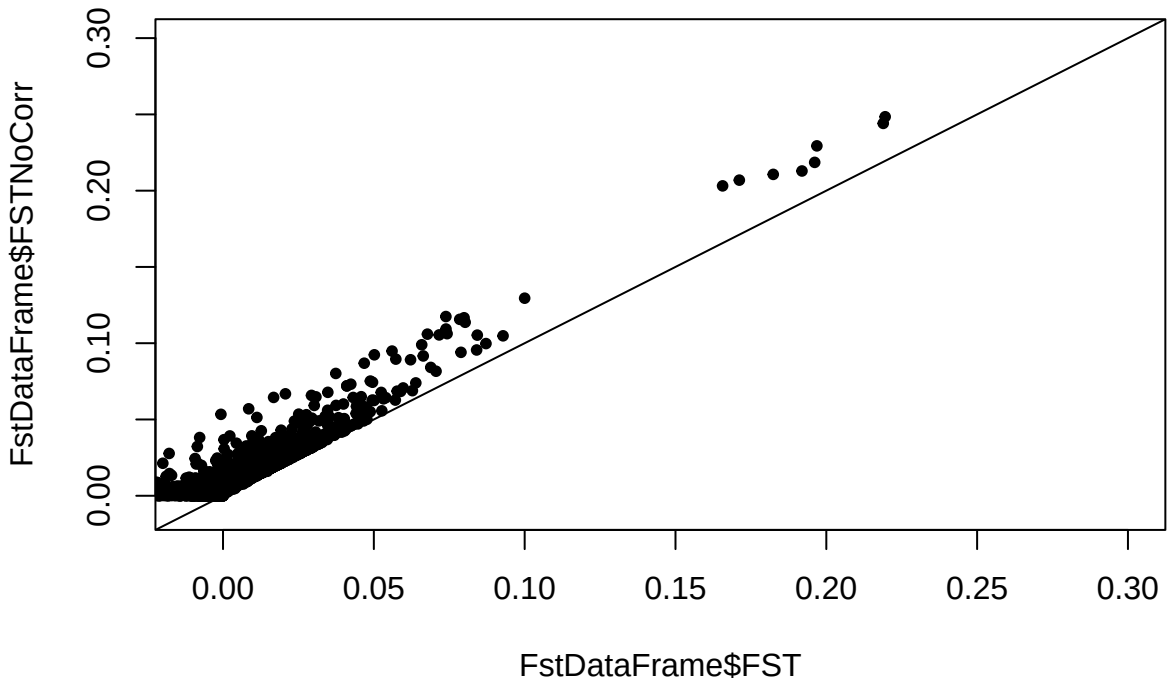
```
################################################################################
########################## radiator::genomic_converter ##########################
################################################################################
Execution date@time: 20220422@1052
Folder created: 21_radiator_genomic_converter_20220422@1052
Function call and arguments stored in: radiator_genomic_converter_args_20220422@1052.tsv
Filters parameters file generated: filters_parameters_20220422@1052.tsv

Importing data: SeqVarGDSClass
Calibrating REF/ALT alleles...

Writing tidy data set:
radiator_data_20220422@1052.rad

Preparing data for output

Data is bi-allelic
Generating adegenet genlight object
Loading required package: parallel
Generating pcadapt file and object
Generating pcadapt file...
################################################################################
######################### radiator::filter_common_markers #######################
################################################################################
Execution date@time: 20220422@1054
Scanning for common markers...

Computation time, overall: 4 sec
######################### completed filter_common_markers #######################
################################################################################
########################## radiator::filter_monomorphic #########################
################################################################################
Execution date@time: 20220422@1054
Scanning for monomorphic markers...

Computation time, overall: 3 sec
########################## completed filter_monomorphic #########################
```

2.2 Outlier detection

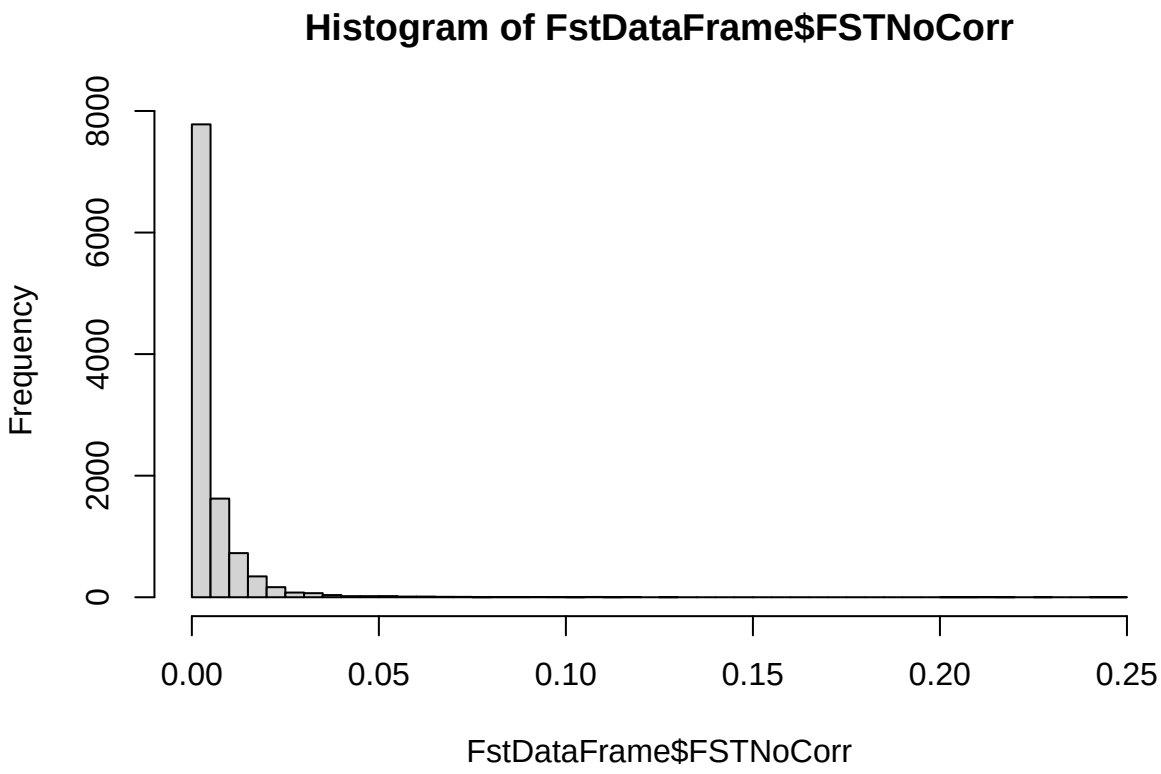
2.2.1 OutFLANK

```
library(OutFLANK)
library(qvalue)
load("WS_SLM_objects.Rdata")
```

```
SNPmat <- as.matrix(WS.genlight)
colnames(SNPmat) <- NULL
row.names(SNPmat) <- NULL
SNPmat[is.na(SNPmat)] <- 9
WS.popNames <- WS.genlight$pop
WS.locnames <- WS.genlight$loc.names

FstDataFrame <- OutFLANK::MakeDiploidFSTMat(SNPmat = SNPmat,
                                            locusNames = WS.locnames,
                                            popNames = WS.popNames)
plot(FstDataFrame$FST, FstDataFrame$FSTNoCorr, xlim = c(-0.01, 0.3),
     ylim = c(-0.01, 0.3), pch = 20) + abline(0, 1)
```




```
hist(FstDataFrame$FSTNoCorr, breaks = 50)
```

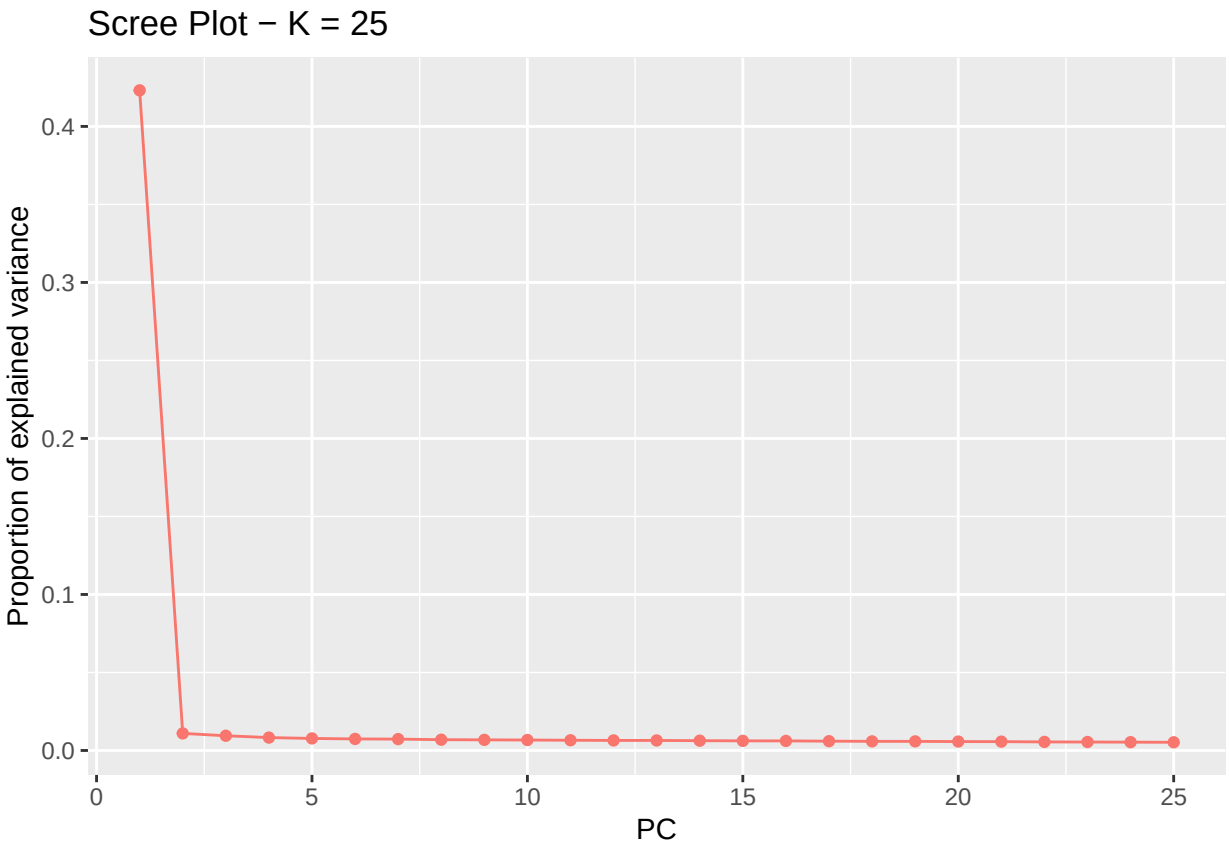



```
Outliers <- OutFLANK::OutFLANK(FstDataFrame,
                               NumberOfSamples = length(levels(WS.genlight@pop)),
                               LeftTrimFraction = 0.012, RightTrimFraction = 0.012,
                               Hmin = 0.001, qthreshold = 0.01)
Outliers$numberLowFstOutliers #0
Outliers$numberHighFstOutliers #0
OutflankNames <- Outliers$results$LocusName[Outliers$results$OutlierFlag == TRUE]
```

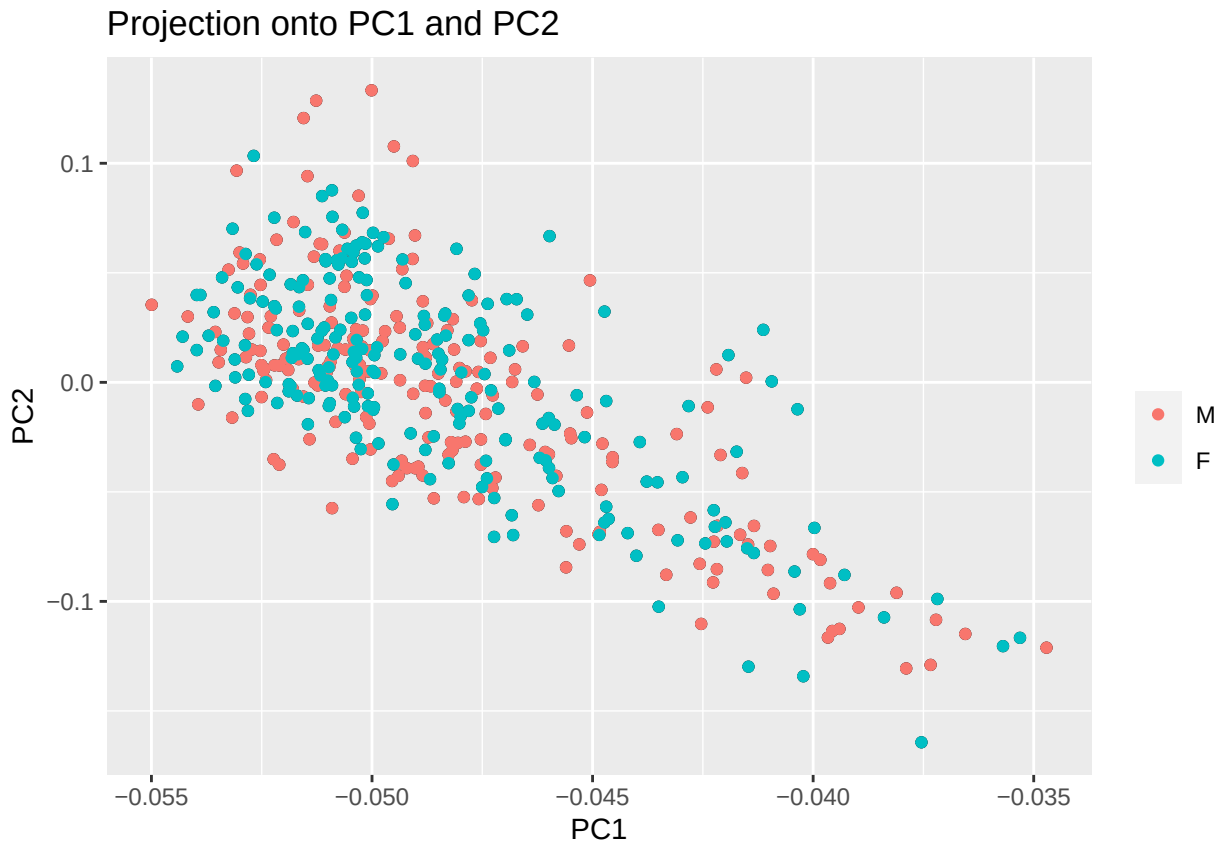
```
## Calculating FSTs, may take a few minutes...
## [1] "10000 done of 10941"
## integer(0)
## [1] 0
## [1] 131
```

2.2.2 PCadapt

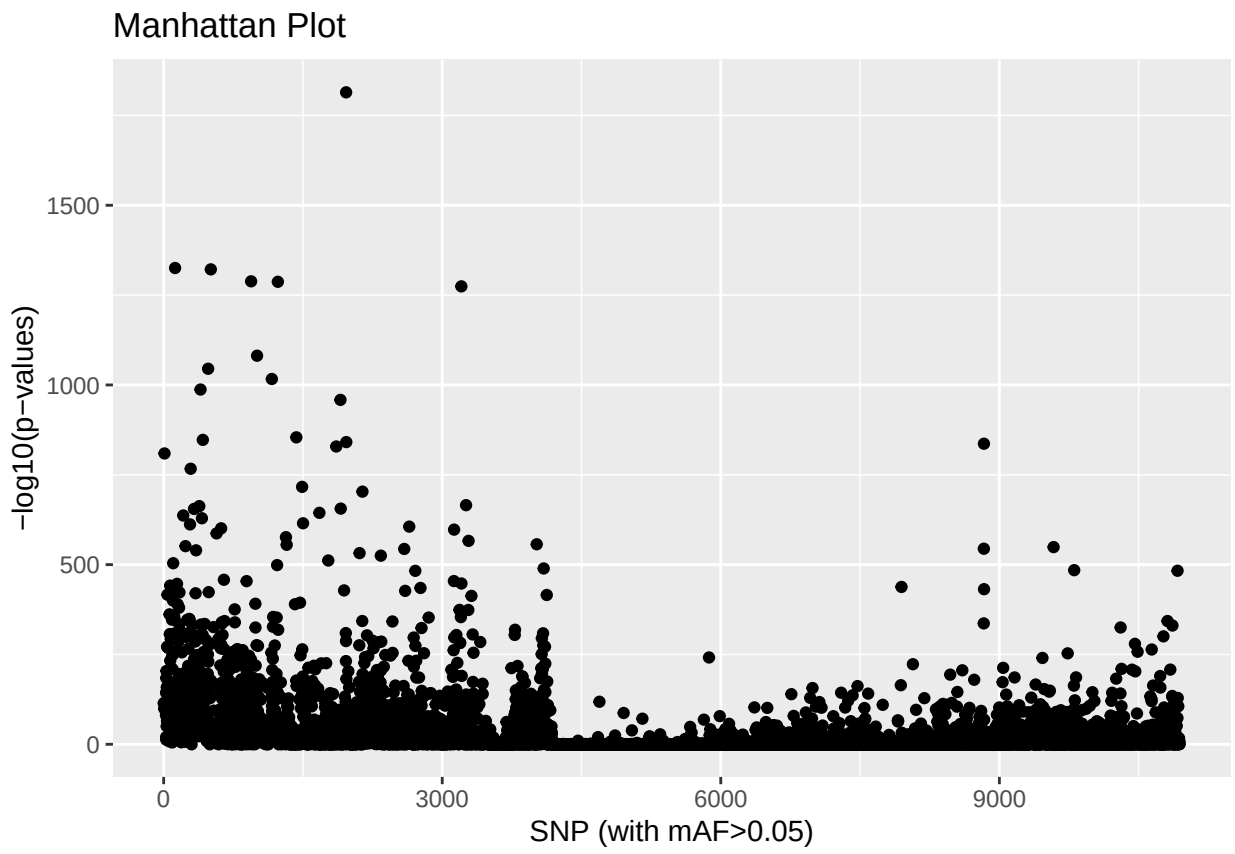
```
library(pcadapt)
poplist <- WS.pcadapt$pop.string
data <- WS.pcadapt$genotype.matrix
data <- pcadapt::read.pcadapt(data)
K <- 25
x <- pcadapt::pcadapt(data, K = K, min.maf = 0.05)

plot(x, option = "screeplot") ## K = 2
```



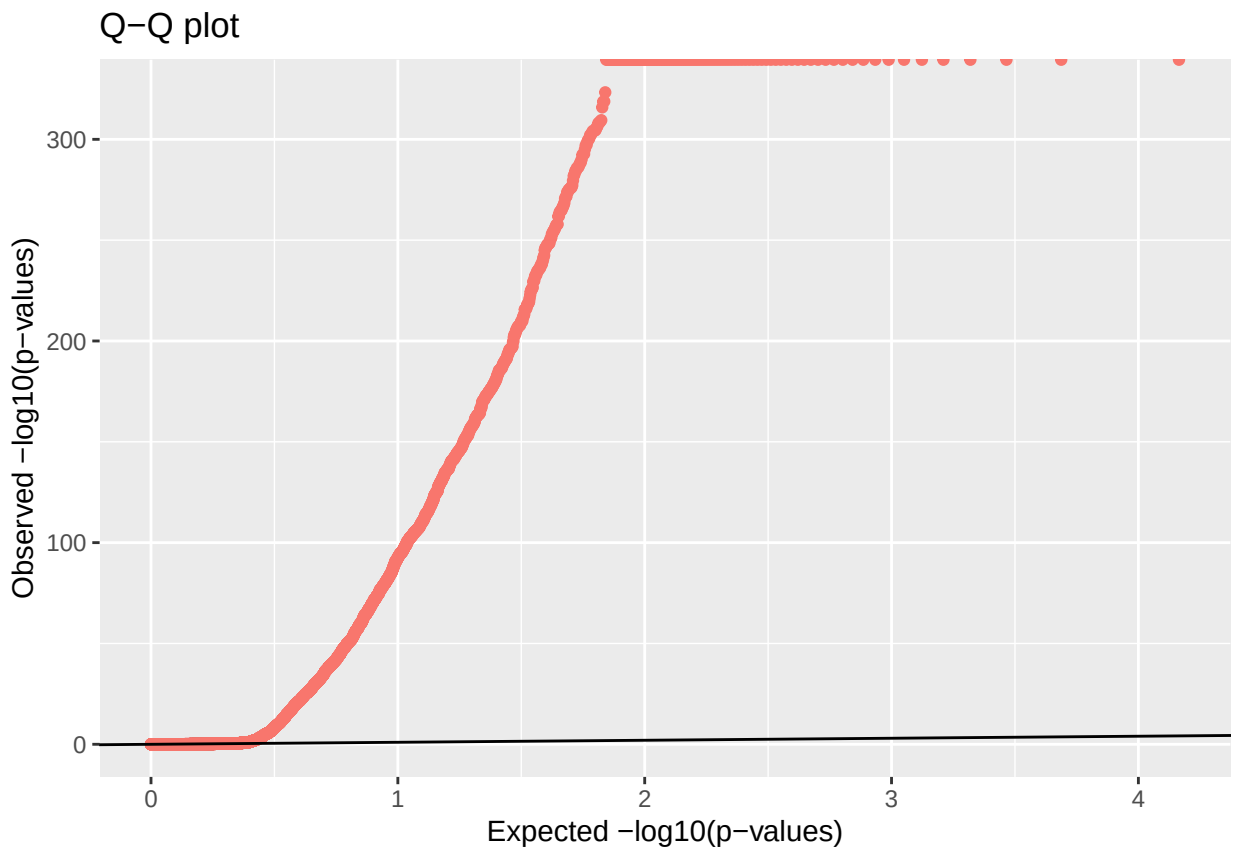

```
plot(x, option = "scores", i = 1, j = 2, pop = poplist)
```



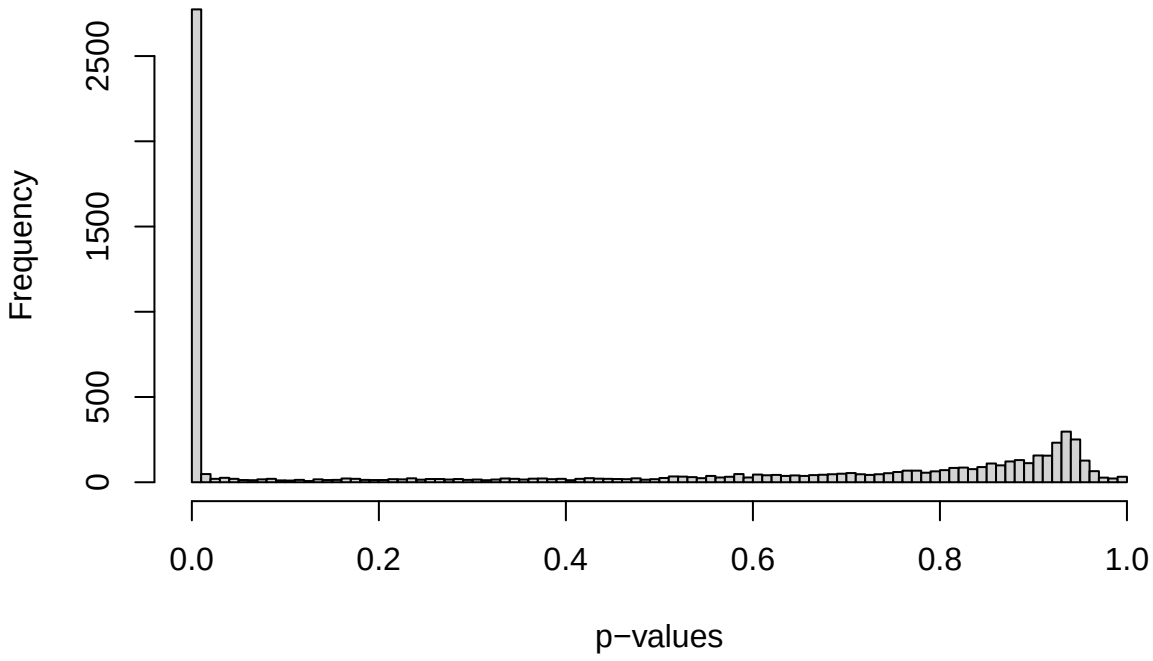

```
K <- 2
x <- pcadapt::pcadapt(data, K = K, min.maf = 0.05)
x$singular.values
plot(x, option = "manhattan")
```




```
plot(x, option = "qqplot", threshold = 0.05)
```




```
hist(x$pvalues, xlab = "p-values", main = NULL, breaks = 100)
```




```
qval <- qvalue::qvalue(x$pvalues)$qvalues
alpha <- 0.01
pcadapt.outliers <- which(qval < alpha)
paste0("number of PCadapt_Outliers = ", length(pcadapt.outliers))
PCadaptkNames <- WS.genlight$loc.names[pcadapt.outliers]

# save(OutflankNames, PCadaptkNames, file = "Sex-linked_outliers.Rdata")
```

```
## [1] 0.6504852 0.1046095
## [1] "number of PCadapt_Outliers = 2720"
```

2.3 Test how 'sexy_markers' compares to other approaches

```
## [1] "number of OuFLANK markers = 131"
## [1] "number of PCadapt markers = 2720"
## [1] "number of sexy_markers markers = 403"
## [1] "sexy_markers vs OutFLANK outliers = 0"
## [1] "sexy_markers vs PCadapt outliers = 16"
## [1] "OutFLANK vs PCadapt outliers = 10"
## [1] "sexy_markers vs combined outliers = 0"
```

3 BLAST SLMs to reference genome

The SLMs that were identified from the full dataset with 'sexy_markers' were blasted against the reference genomes (Marra et al. 2019 and VGP).

The 415 SLMs were blasted against the reference genomes, using the Megablast algorithm as implemented in Geneious v.11.1.5. Word size was set to 20 bp and only up to 10 hits with a max E-value of 10^-5 were returned.

3.1 Marra et al 2019 - reference genome

Only 304 unique SLMs had a BLAST hit.

```
f.in <- "WS_SLMs_mapped_to_Marra_genome.csv"
data <- readr::read_csv(file = f.in) %>%
  tidyr::separate(col = Query,
                  into = c("Marker_type", "Method", "LOCUS", "SPECIES"),
                  sep = "\\|", remove = FALSE)
```

```
## Rows: 1495 Columns: 45
```

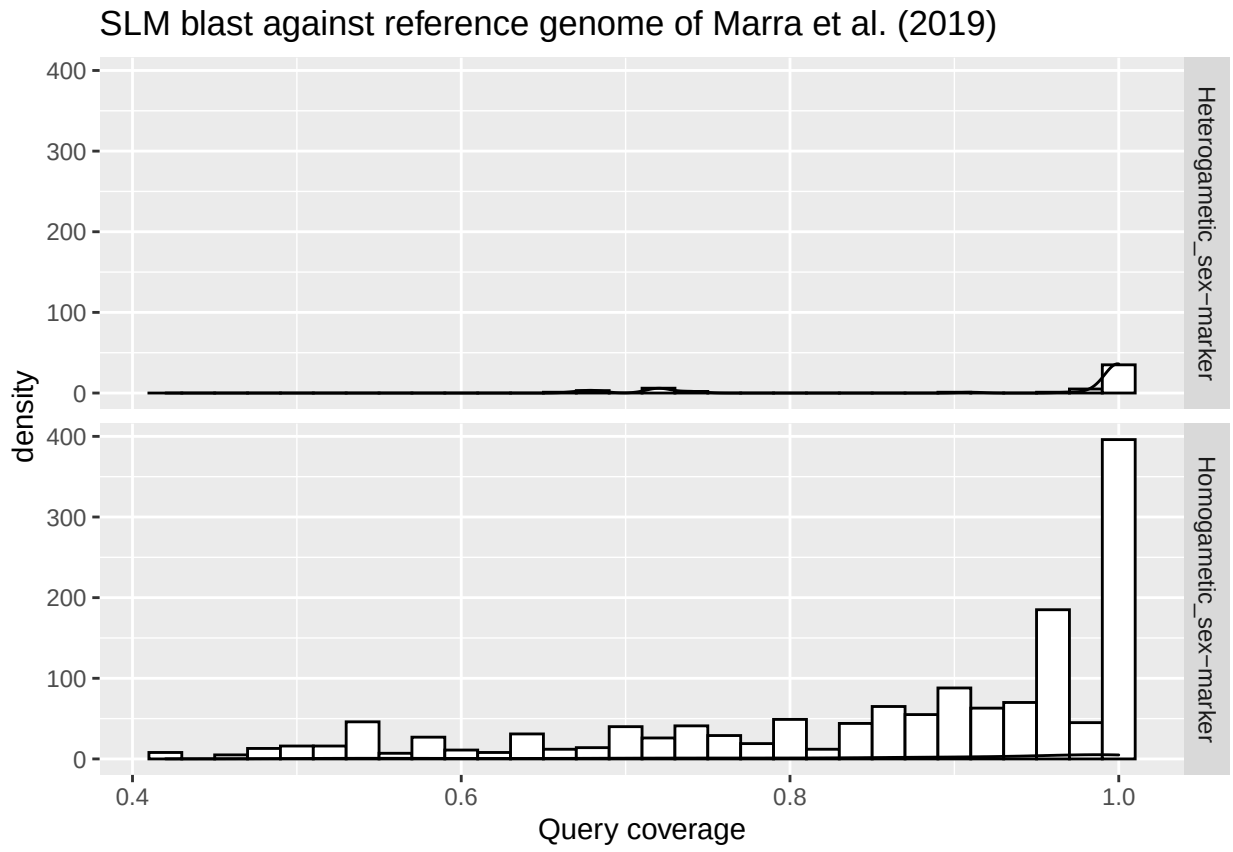
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (20): Name, Accession, Common Name, Created Date, Database, Description,...
## dbl (22): # Nucleotide Sequences With Mates, # Nucleotides, # Sequences, % I...
## lgl  (3): Free end gaps, Ref Seq Name, Sequence List Name
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

3.1.1 Quality filtering of the data

```
p <- ggplot2::ggplot(data, ggplot2::aes(x = `Query coverage`)) +
  ggplot2::geom_histogram(colour = "black", fill = "white") +
  ggplot2::geom_density(alpha = .2,size = 0.5) +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("SLM blast against reference genome of Marra et al. (2019)")
print(p)
```

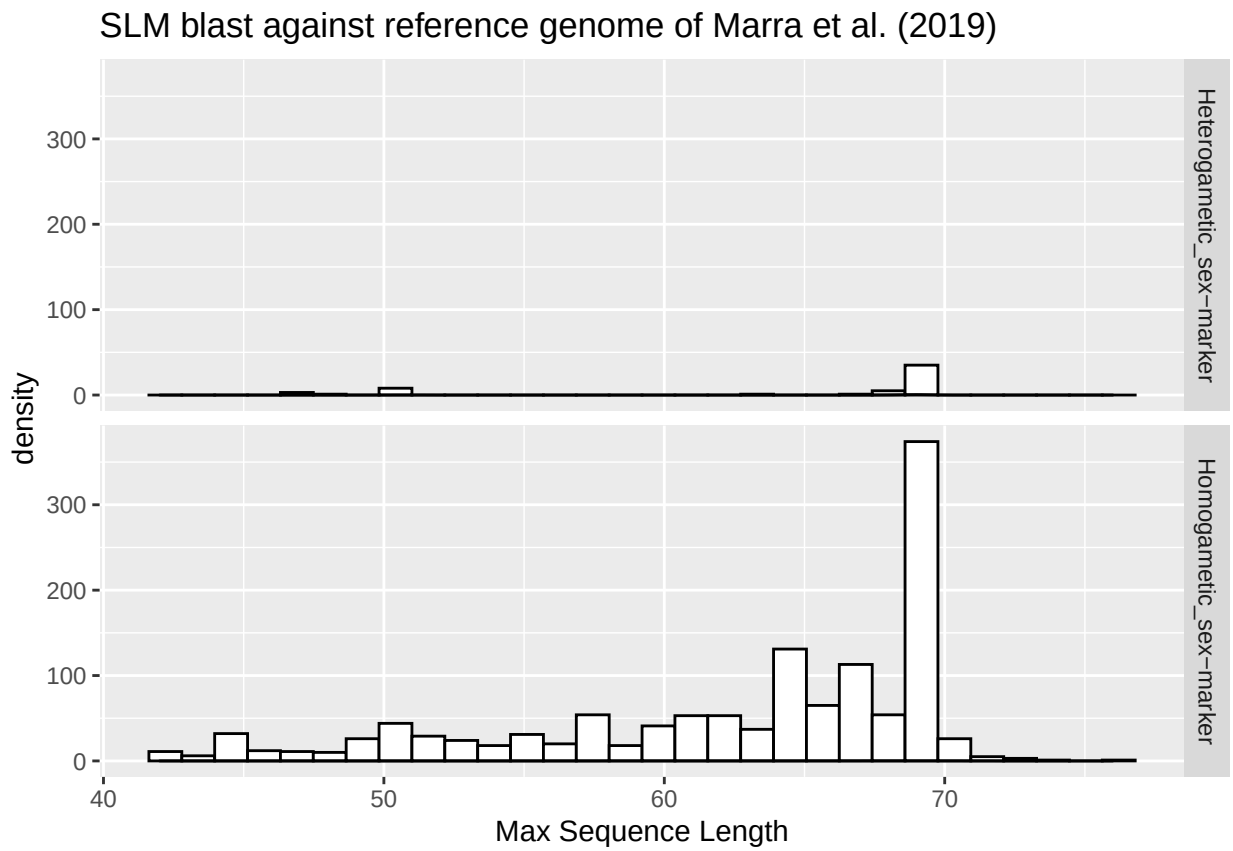
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
query.threshold <- 0.6
subdata <- data[data$`Query coverage` > query.threshold,]
```

```
p <- ggplot2::ggplot(subdata, ggplot2::aes(x = `Max Sequence Length`)) +
  ggplot2::geom_histogram(colour = "black", fill = "white") +
  ggplot2::geom_density(alpha = .2,size = 0.5) +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("SLM blast against reference genome of Marra et al. (2019)")
print(p)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

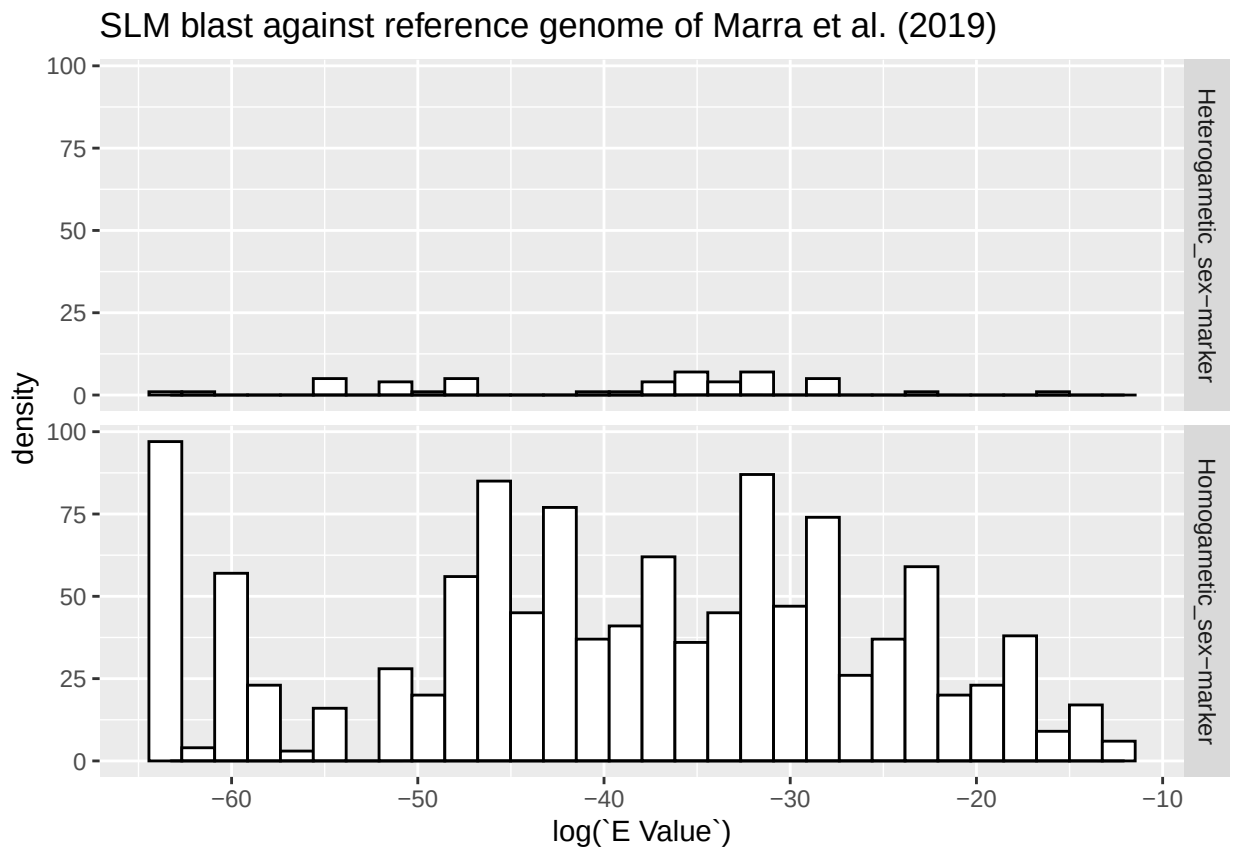



```
seq.length <- 50
subdata <- subdata[subdata$`Max Sequence Length` > seq.length,]
```

Because the Y-chromosome is expected to mutate at a faster rate, we decreased the E-value threshold for heterogametic markers.

```
p <- ggplot2::ggplot(subdata, ggplot2::aes(x = log(`E Value`))) +
  ggplot2::geom_histogram(colour = "black", fill = "white") +
  ggplot2::geom_density(alpha = .2,size = 0.5) +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("SLM blast against reference genome of Marra et al. (2019)")
print(p)
```

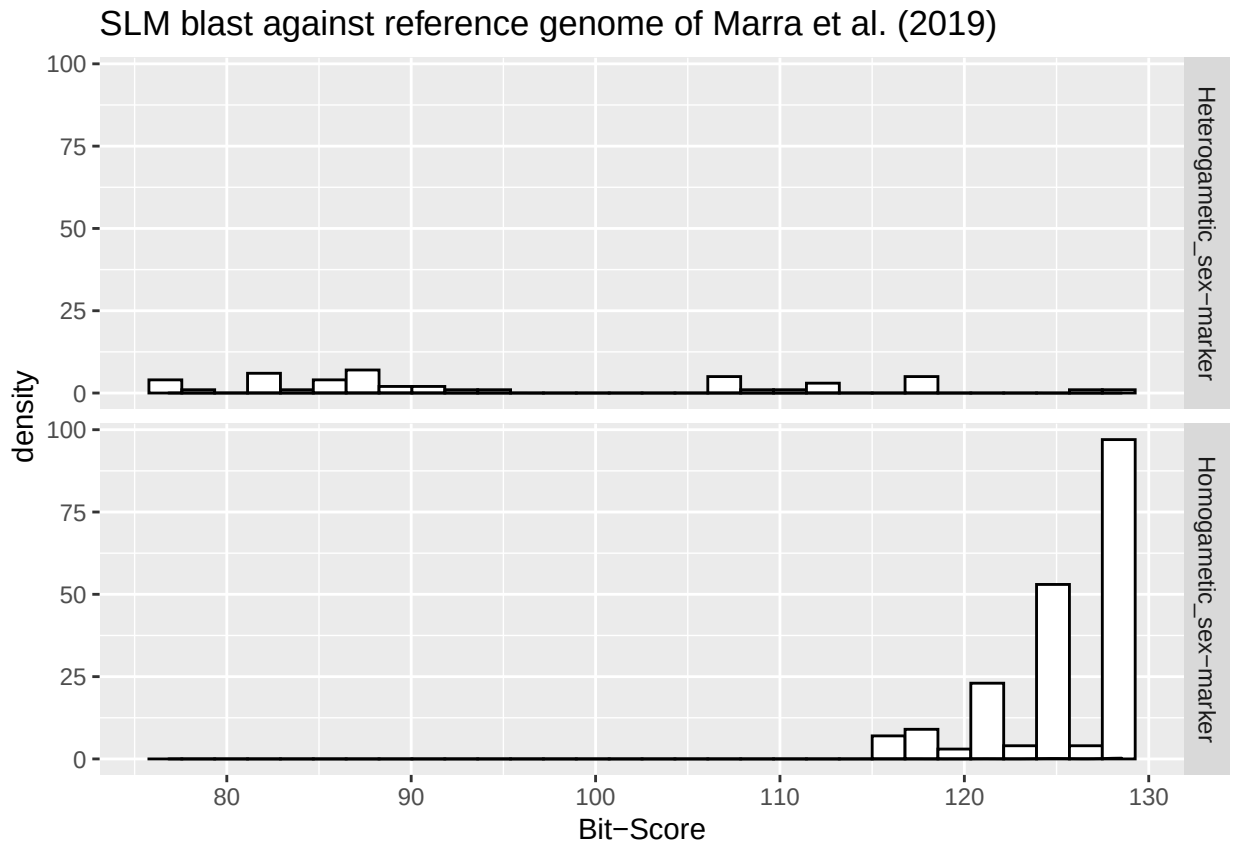
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
eval.threshold.het <- exp(-25)
eval.threshold.hom <- exp(-52)
subdata <- subdata[(subdata$Marker_type == "Heterogametic_sex-marker" &
                      subdata$`E Value` < eval.threshold.het) |
                     (subdata$Marker_type == "Homogametic_sex-marker" &
                        subdata$`E Value` < eval.threshold.hom),]
```

```
p <- ggplot2::ggplot(subdata, ggplot2::aes(x = `Bit-Score`)) +
  ggplot2::geom_histogram(colour = "black", fill = "white") +
  ggplot2::geom_density(alpha = .2,size = 0.5) +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("SLM blast against reference genome of Marra et al. (2019)")
print(p)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
data %<>%
  dplyr::filter(`Query coverage` >  query.threshold) %>%
  dplyr::filter(`Max Sequence Length` > seq.length) %>%
  dplyr::filter((Marker_type == "Heterogametic_sex-marker" &
                   `E Value` < eval.threshold.het) |
                (Marker_type == "Homogametic_sex-marker" &
                   `E Value` < eval.threshold.hom))
```

3.1.2 Make a table

```
tab <- table(data$SPECIES[!duplicated(data$Query)],
             data$Marker_type[!duplicated(data$Query)])
tab
```

```
##
##                          Heterogametic_sex-marker Homogametic_sex-marker
##   Carcharodon_carcharias                        7                     172
```

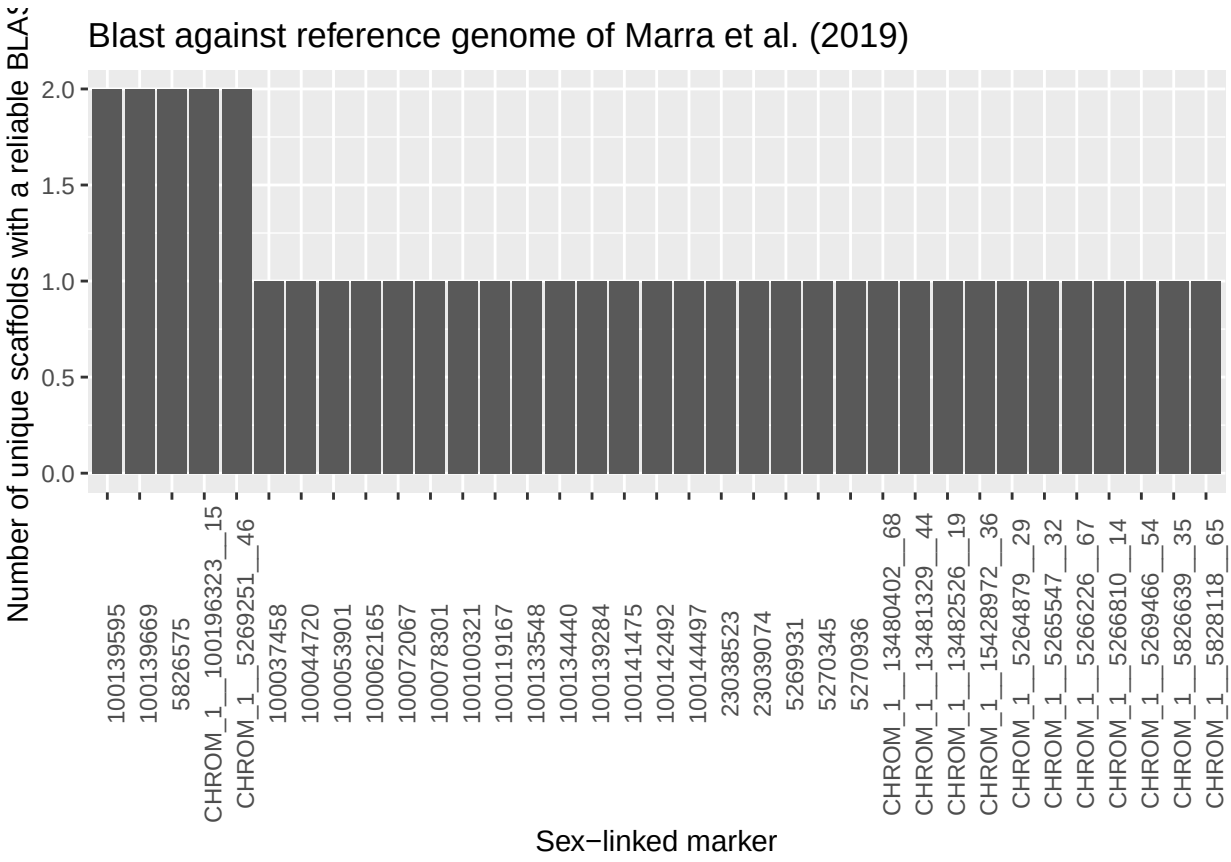
3.1.3 Remove putative paralogous sequences

Paralogs are SLMs that have too many high quality hits on different fragments

```
dups <- unique(data$Query[duplicated(data$Query)])
length(dups) ## 23 SLMs with multiple hits
```

```
## [1] 164
```

```
data %<>% dplyr::group_by(LOCUS) %>%
  dplyr::top_n(1, wt = `Bit-Score`)

dup.data <- data[data$Query %in% dups,] #51

plot.data <- dplyr::group_by(dup.data, LOCUS) %>%
  dplyr::count(DUP = duplicated(Name, fromLast = TRUE)) %>%
  dplyr::rename(unique.scaffolds = n) %>%
  dplyr::filter(DUP == FALSE)

p <- ggplot2::ggplot(plot.data,
                     ggplot2::aes(x = reorder(LOCUS, -unique.scaffolds),
                                  y = unique.scaffolds)) +
  ggplot2::geom_bar(stat = "identity") +
  ggplot2::ggtitle("Blast against reference genome of Marra et al. (2019)") +
  ggplot2::ylab("Number of unique scaffolds with a reliable BLAST hit") +
  ggplot2::xlab("Sex-linked marker") +
  ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 90))
print(p)
```



```
rm.names <- plot.data$LOCUS[plot.data$unique.scaffold > 1]
```

```
## Warning: Unknown or uninitialised column: 'unique.scaffold'.
```

```
data <- data[!data$LOCUS %in% rm.names,]
```

3.1.4 Make a table

```
tab <- table(data$SPECIES[!duplicated(data$Query)],
             data$Marker_type[!duplicated(data$Query)])
tab
```

```
##
##                           Heterogametic_sex-marker Homogametic_sex-marker
##   Carcharodon_carcharias                         7                     297
```

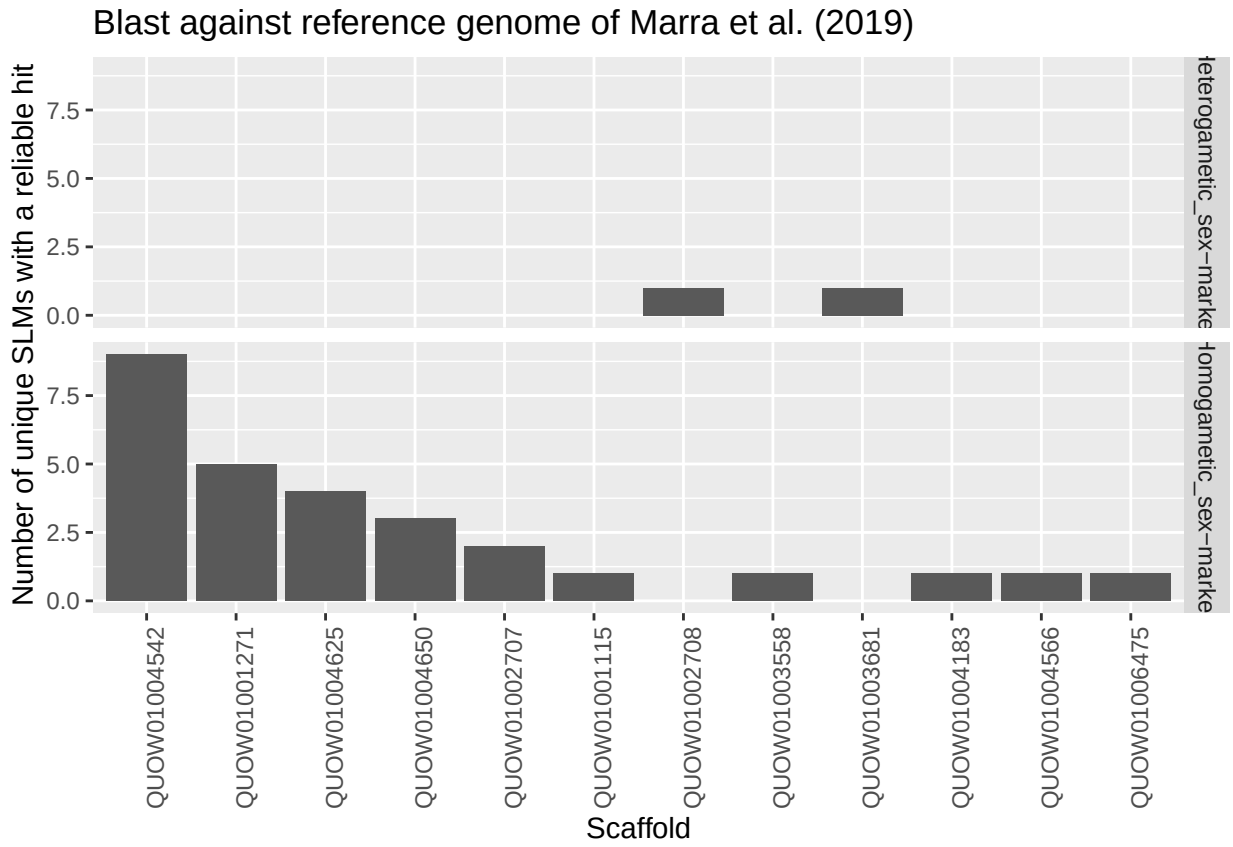
```
dups <- unique(data$Query[duplicated(data$Query)])
dup.data <- data[data$Query %in% dups,]
# several SLMs are repeated in the scaffold
tab <- table(dup.data$Marker_type, dup.data$Name)
tab
```

```
##
##                            QUOW01001115 QUOW01001271 QUOW01002707 QUOW01002708
##   Heterogametic_sex-marker            0            0            0            1
##   Homogametic_sex-marker              1            6            4            0
##
##                            QUOW01002857 QUOW01003388 QUOW01003558 QUOW01003681
##   Heterogametic_sex-marker            0            0            0            5
##   Homogametic_sex-marker              1            1            2            0
##
##                            QUOW01004183 QUOW01004201 QUOW01004276 QUOW01004289
##   Heterogametic_sex-marker            0            0            0            0
##   Homogametic_sex-marker              1            5            1            2
##
##                            QUOW01004542 QUOW01004566 QUOW01004625 QUOW01004650
##   Heterogametic_sex-marker            0            0            0            0
##   Homogametic_sex-marker             18            2            8            6
##
##                            QUOW01006475 QUOW01006560 QUOW01006806
##   Heterogametic_sex-marker            0            0            0
##   Homogametic_sex-marker              2            1            4
```

3.1.5 Make histogram of number of SLMs per scaffold

```
plot.data <- dplyr::group_by(dup.data, Name, Marker_type) %>%
  dplyr::count(duplicated(Query)) %>%
  dplyr::rename(unique.SLMs = n) %>%
```

```
  dplyr::filter(`duplicated(Query)` == FALSE)

p <- ggplot2::ggplot(plot.data, ggplot2::aes(x = reorder(Name, -unique.SLMs),
                                             y = unique.SLMs)) +
  ggplot2::geom_bar(stat = "identity") +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("Blast against reference genome of Marra et al. (2019)") +
  ggplot2::ylab("Number of unique SLMs with a reliable hit") +
  ggplot2::xlab("Scaffold") +
  ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 90))
print(p)
```




```
Marra.sex.scaffolds <- unique(c(plot.data$Name[plot.data$unique.SLMs > 10],
                                data$Name[data$Marker_type ==
                                            "Heterogametic_sex-marker"]))
Marra.sex.markers <- unique(data$LOCUS[data$Name %in% Marra.sex.scaffolds])
Marra.data <- data
```

3.2 Vertebrate Genome Project - reference genome

only 332 unique SLMs had a BLAST hit.

```
f.in <- "WS_SLMs_mapped_to_VPG_genome.csv"
data <- readr::read_csv(file = f.in) %>%
  tidyr::separate(col = Query, into = c("Marker_type", "Method", "LOCUS", "SPECIES"),
                  sep = "\\|", remove = FALSE)
```

```
## Rows: 1855 Columns: 35
```

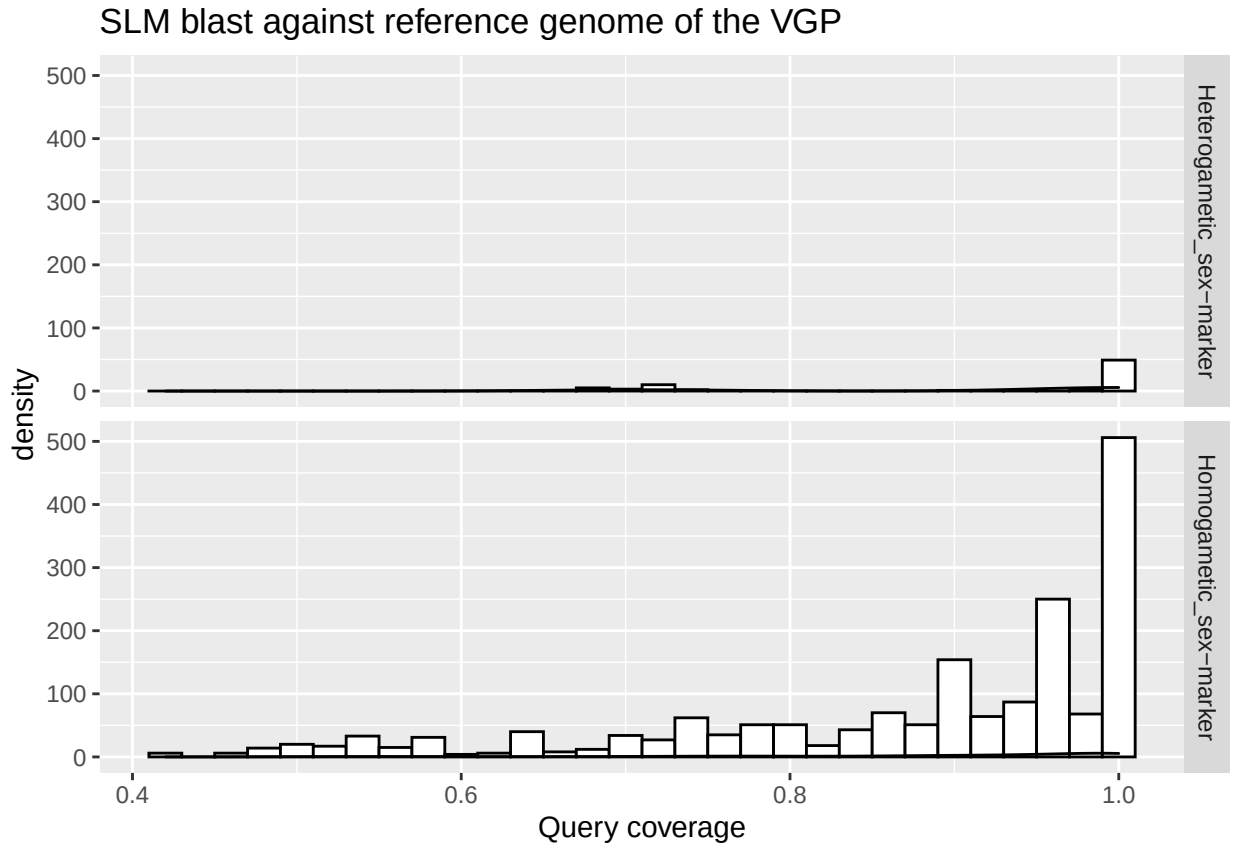
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): Name, Created Date, Database, Modified, Molecule Type, Query, Sequ...
## dbl (21): # Nucleotide Sequences With Mates, # Nucleotides, # Sequences, % I...
## lgl  (5): Accession, Description, Free end gaps, Ref Seq Name, Sequence List...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

3.2.1 Quality filtering of the data

```
p <- ggplot2::ggplot(data, ggplot2::aes(x = `Query coverage`)) +
  ggplot2::geom_histogram(colour = "black", fill = "white") +
  ggplot2::geom_density(alpha = .2,size = 0.5) +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("SLM blast against reference genome of the VGP")
print(p)
```

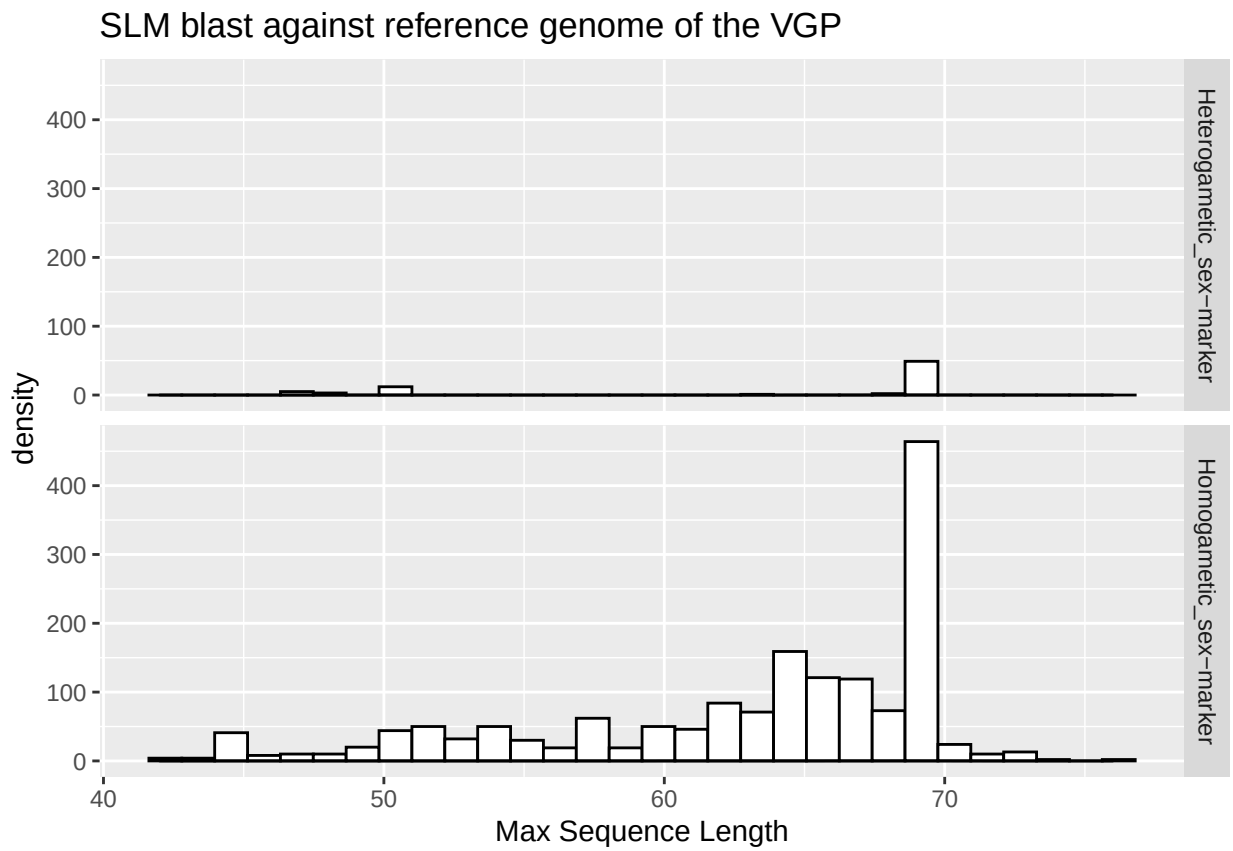
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
query.threshold <- 0.6
subdata <- data[data$`Query coverage` > query.threshold,]
```

```
p <- ggplot2::ggplot(subdata, ggplot2::aes(x = `Max Sequence Length`)) +
  ggplot2::geom_histogram(colour = "black", fill = "white") +
  ggplot2::geom_density(alpha = .2,size = 0.5) +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("SLM blast against reference genome of the VGP")
print(p)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

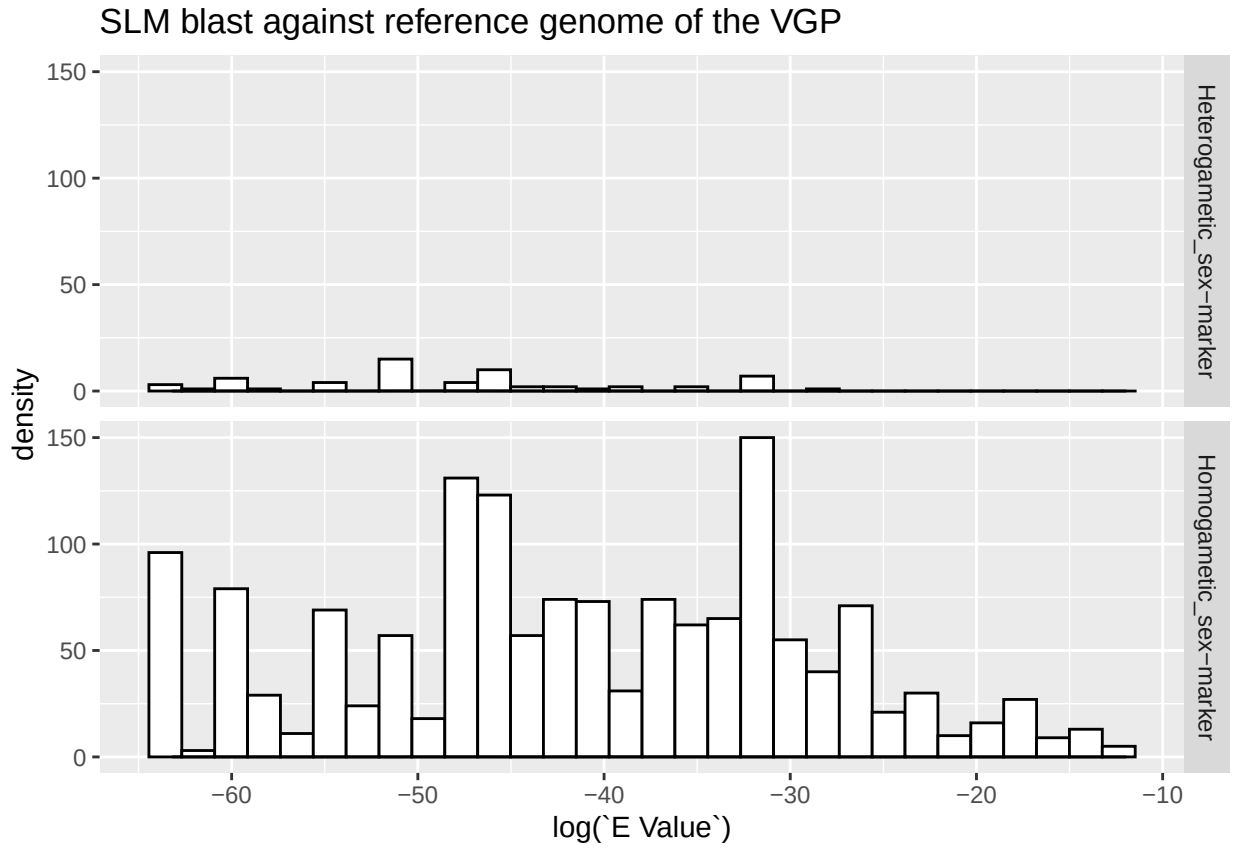



```
seq.length <- 50
subdata <- subdata[subdata$`Max Sequence Length` > seq.length,]
```

Because the Y-chromosome is expected to mutate at a faster rate, we decreased the E-value threshold for heterogametic markers.

```
p <- ggplot2::ggplot(subdata, ggplot2::aes(x = log(`E Value`))) +
  ggplot2::geom_histogram(colour = "black", fill = "white") +
  ggplot2::geom_density(alpha = .2,size = 0.5) +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("SLM blast against reference genome of the VGP")
print(p)
```

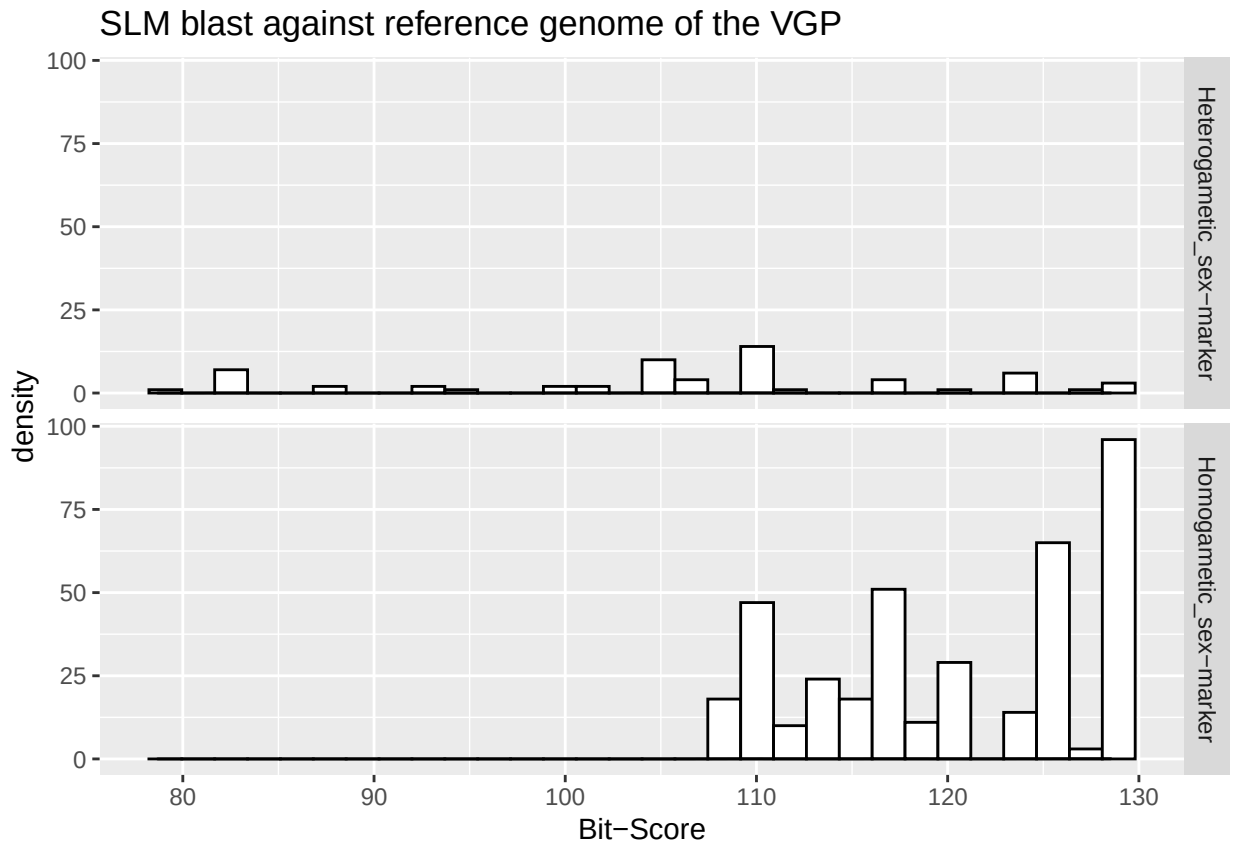
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
eval.threshold.het <- exp(-25)
eval.threshold.hom <- exp(-48)
subdata <- subdata[(subdata$Marker_type == "Heterogametic_sex-marker" &
                      subdata$`E Value` < eval.threshold.het) |
                     (subdata$Marker_type == "Homogametic_sex-marker" &
                        subdata$`E Value` < eval.threshold.hom),]
```

```
p <- ggplot2::ggplot(subdata, ggplot2::aes(x = `Bit-Score`)) +
  ggplot2::geom_histogram(colour = "black", fill = "white") +
  ggplot2::geom_density(alpha = .2,size = 0.5) +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("SLM blast against reference genome of the VGP")
print(p)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
data %<>%
  dplyr::filter(`Query coverage` >  query.threshold) %>%
  dplyr::filter(`Max Sequence Length` > seq.length) %>%
  dplyr::filter((Marker_type == "Heterogametic_sex-marker" &
                   `E Value` < eval.threshold.het) |
                (Marker_type == "Homogametic_sex-marker" &
                   `E Value` < eval.threshold.hom))
```

3.2.2 Make a table

```
tab <- table(data$SPECIES[!duplicated(data$Query)],
             data$Marker_type[!duplicated(data$Query)])
tab
```

```
##
##                          Heterogametic_sex-marker Homogametic_sex-marker
##   Carcharodon_carcharias                        8                     219
```

3.2.3 Remove putative paralogous sequences

Here paralogs are SLMs that have too many high-confidence hits on different fragments

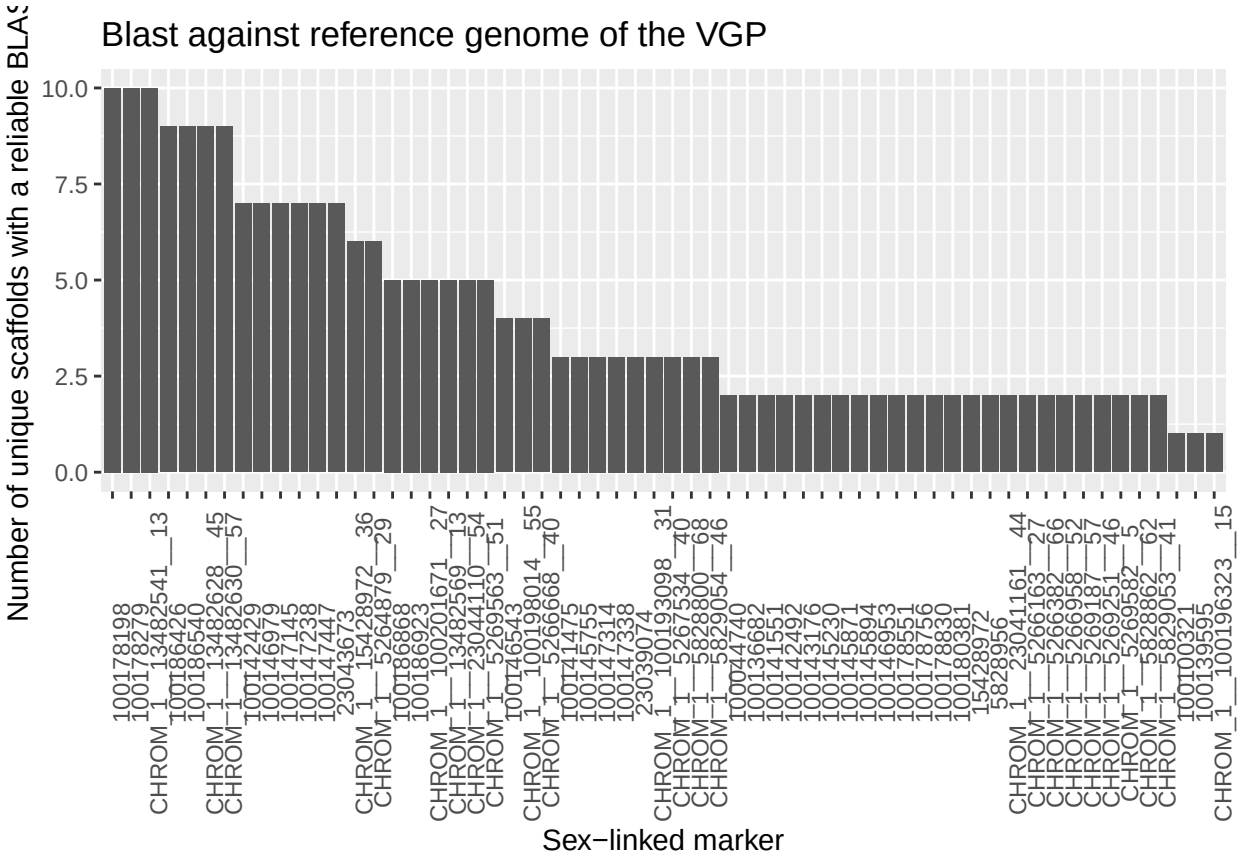
```
dups <- unique(data$Query[duplicated(data$Query)])
length(dups) ## 66 SLMs with multiple hits
```

```
## [1] 191
```

```
data %<>% dplyr::group_by(LOCUS) %>%
  dplyr::top_n(1, wt = `Bit-Score`)

dups <- unique(data$Query[duplicated(data$Query)])
length(dups) ## 47 SLMs with multiple hits
```

```
## [1] 64
```

```
dup.data <- data[data$Query %in% dups,]


plot.data <- dplyr::group_by(dup.data, LOCUS) %>%
  dplyr::count(DUP =duplicated(Name, fromLast = TRUE)) %>%
  dplyr::rename(unique.scaffolds = n) %>%
  dplyr::filter(DUP == TRUE)

p <- ggplot2::ggplot(plot.data, ggplot2::aes(x = reorder(LOCUS, -unique.scaffolds),
                                             y = unique.scaffolds)) +
  ggplot2::geom_bar(stat = "identity") +
  ggplot2::ggtitle("Blast against reference genome of the VGP") +
  ggplot2::ylab("Number of unique scaffolds with a reliable BLAST hit") +
  ggplot2::xlab("Sex-linked marker") +
  ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 90))
print(p)
```




```
rm.names <- plot.data$LOCUS[plot.data$unique.scaffold > 1]
```

```
## Warning: Unknown or uninitialised column: 'unique.scaffold'.
```

```
data <- data[!data$LOCUS %in% rm.names,]
```

3.2.4 Make a table

```
tab <- table(data$SPECIES[!duplicated(data$Query)],
             data$Marker_type[!duplicated(data$Query)])
tab
```

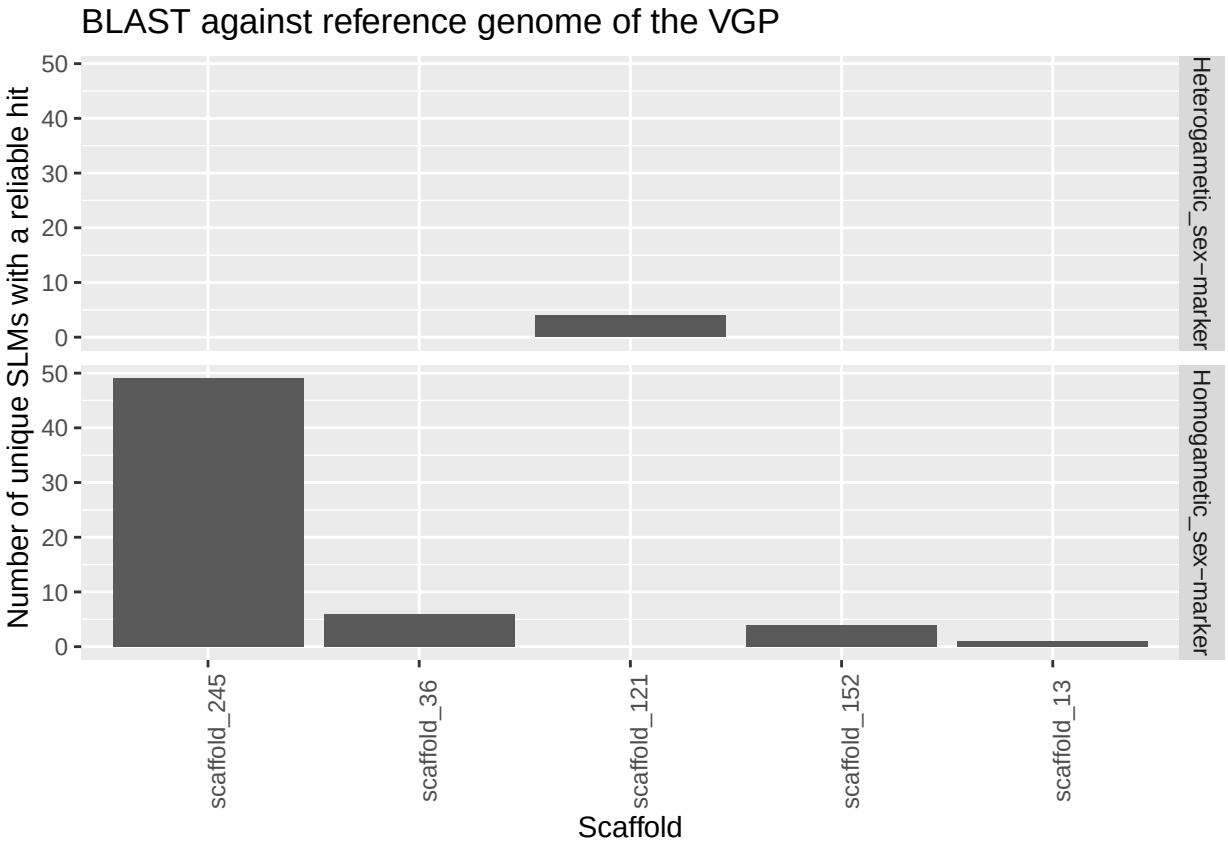
```
##
##                          Heterogametic_sex-marker Homogametic_sex-marker
##   Carcharodon_carcharias                        8                     324
```

```
dups <- unique(data$Query[duplicated(data$Query)])
dup.data <- data[data$Query %in% dups,]
# several SLMs are repeated in the scaffold
tab <- table(dup.data$Marker_type, dup.data$Name)
tab
```

```
##
##                            scaffold_121 scaffold_13 scaffold_152 scaffold_245
##   Heterogametic_sex-marker           19           0            0            0
##   Homogametic_sex-marker              0           1            4          204
##
##                            scaffold_36 scaffold_384 scaffold_580
##   Heterogametic_sex-marker           0            0            0
##   Homogametic_sex-marker            14            4            1
```

3.2.5 Make histogram of number of unique SLMs per scaffold

```
plot.data <- dplyr::group_by(dup.data, Name, Marker_type) %>%
  dplyr::count(duplicated(Query)) %>%
  dplyr::rename(unique.SLMs = n) %>%
  dplyr::filter(`duplicated(Query)` == FALSE)

p <- ggplot2::ggplot(plot.data, ggplot2::aes(x = reorder(Name, -unique.SLMs),
                                             y = unique.SLMs)) +
  ggplot2::geom_bar(stat = "identity") +
  ggplot2::facet_grid(Marker_type ~ .) +
  ggplot2::ggtitle("BLAST against reference genome of the VGP") +
  ggplot2::ylab("Number of unique SLMs with a reliable hit") +
  ggplot2::xlab("Scaffold") +
  ggplot2::theme(axis.text.x = ggplot2::element_text(angle = 90))
print(p)
```




```
VGP.sex.scaffolds <- unique(c(plot.data$Name[plot.data$unique.SLMs > 10],
                              data$Name[data$Marker_type == "Heterogametic_sex-marker"]))
VGP.sex.markers <- unique(data$LOCUS[data$Name %in% VGP.sex.scaffolds])
VGP.data <- data
```

---

4 PCR primers and conditions based on genome sequence data

Table 2: Primer sequences used for sex identification of White Shark samples.

Fragment	Primer_name	Primer_sequence	Amplicon_length
Y-chrom	WSMx(6363_for	5'-TGGTGGAAGTTGGCTATGGCTGTGA	655
Y-chrom	WSMx(6980_rev	5'-GTCTCACTTTATCTTTCCTGCTAACTTAAGGCCATG	655
Beta-actin	WSBact303_for	5'-GAGGCCCTCTTCCAACCATCTT	287
Beta-actin	WSBact572_rev	5'-ATCTCCTTCTGCATACGGTCT	287
X-chrom	WS-X3_222_F	5'-TGCACCACGCATATCCTGAG	164
X-chrom	WS-X3_366_R	5'-GCTGTTGTTCCACACTCCCT	164

Table 3: PCR reagents and concentrations

Final.volume	X20.µL
2x QIAGEN Multiplex PCR Master Mix -> 1x	10 µL
Primer MIX 1.666 µM -> 0.2 µM	2.4 µL
5x Q-solution -> 0.5x	2 µL
ddH20	3.6 µL
Template (<20 ng/uL)	2 µL

Table 4: PCR cycling parameters

Step	Temp	Time	Cycles
Initial denaturation	95C	15 min	1
Denaturation	94C	30 sec	35
Annealing	57	30 sec	35
Extension	72C	1 min	35
Final extension	72C	5min	1

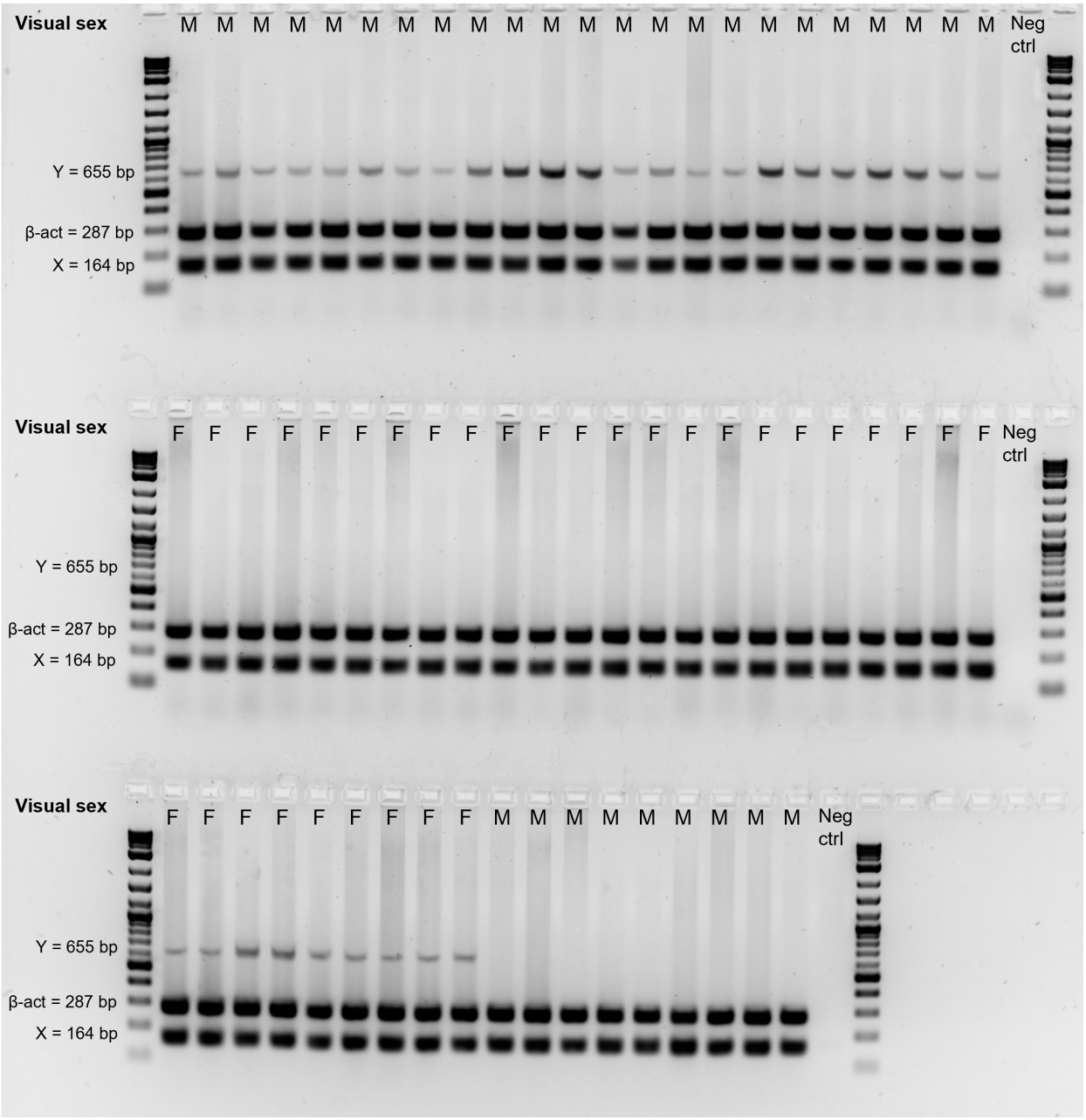

Visual sex
M M M M M M M M M M M M M M M M M M M M M M M
Neg ctrl
Y = 655 bp
β-act = 287 bp
X = 164 bp
Visual sex
F F F F F F F F F F F F F F F F F F F F F F F
Neg ctrl
Y = 655 bp
β-act = 287 bp
X = 164 bp
Visual sex
F F F F F F F F F M M M M M M M M M M
Neg ctrl
Y = 655 bp
β-act = 287 bp
X = 164 bp

Figure 1: PCR product of Y-chrom fragment

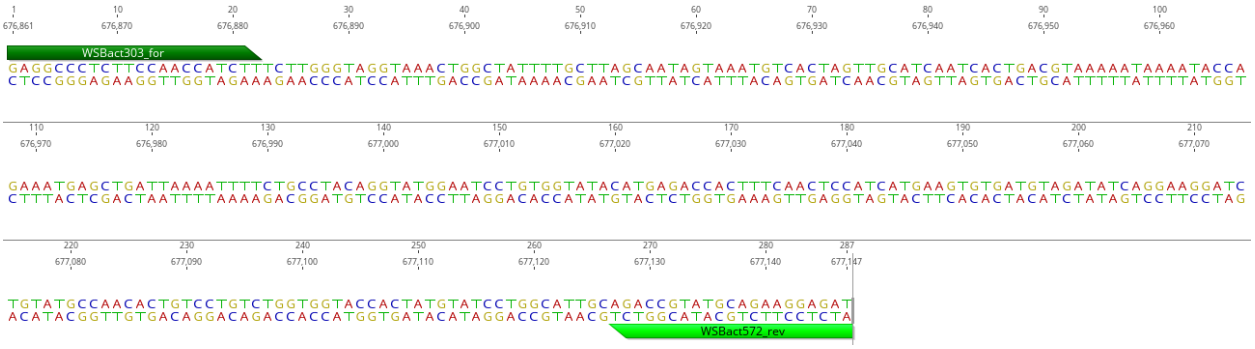


Figure 2: PCR product of Beta-actin fragment

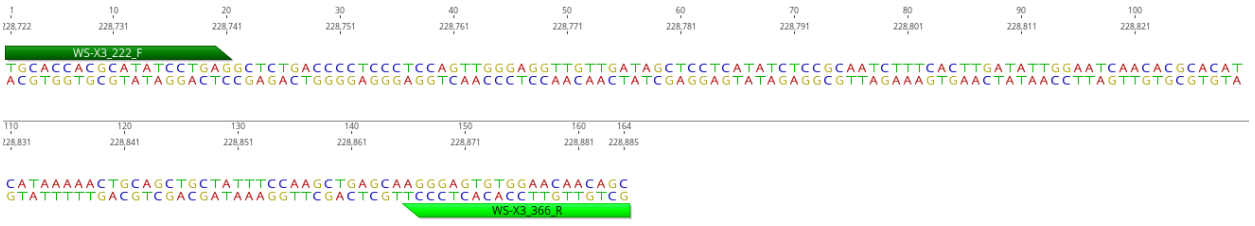


Figure 3: PCR product of X-chrom fragment

5 Test the accuracy of 'sexy_markers'

We test how low sample size, unequal sex ratio and low marker coverage will affect the power to detect SLMs. The number of SLMs from the full dataset that mapped to putative sex chromosomes were considered as reference markers

5.1 Locus names that map to sex chromosomes

```
std.sex.names <- unique((c(Marra.sex.markers, VGP.sex.markers)))
save(std.sex.names, file = "std_sex_names.Rdata")
```

5.2 Prepare subset data files

Data and strata sub-sampling Sample the data with variable numbers of individuals and SNPs

```
strata <- readr::read_tsv("Strata_according_to_genetic_sex.tsv")
data.file <- "WhiteShark_SNP_counts.csv"
whitelist.markers <- "whitelist_snp.tsv"
silicodata.file <- "WhiteShark_silico_count.csv"
source("Subset_function.R")


data <- radiator::read_dart(data = data.file, strata = strata, whitelist.markers)
silicodata <- radiator::read_dart(data = silicodata.file, strata = strata) %>%
  dplyr::rename(MARKERS = CLONE_ID)


shark.subset <- c(6, 12, 24, 48, 72, 96, 120, 144, 252, 348)
                  #not 480 (sex ratio is not perfectly 1:1)
marker.subset <- c(200, 1000, 2000, 5000, 7500, 10000, 12500, 15000, 17500, 20000)

shark.id <- SeqArray::seqGetData(data, "sample.id")
sex.id <- radiator::extract_individuals_metadata(data,
                                                 ind.field.select = "STRATA",
                                                 radiator.node = TRUE,
                                                 whitelist = FALSE,
                                                 blacklist = FALSE,
                                                 verbose = FALSE)$STRATA
# print(sum(sex.id == "F"))
# print(sum(sex.id == "M"))


parentfile <- "parentfile.gds"
gdsfmt.data <- createfn.gds(filename = parentfile, allow.duplicate = FALSE)
assign.gdsn(index.gdsn(gdsfmt.data, ""), src.node =  index.gdsn(data, ""))
gdsfmt.data <- new("SeqVarGDSClass", gdsfmt.data)
seqClose(gdsfmt.data)

k <- 0
pop <- c()
```

```
for (i in 1:10) {
  for (j in 1:10) {
    k <- k + 1
    print(k) # follow progress

    ind <- shark.subset[i]
    var <- marker.subset[j]
    samp.id.F <- sample(shark.id[sex.id == "F"],
                        ind / 2, replace = FALSE) # Sample for equal sex ratio
    samp.id.M <- sample(shark.id[sex.id == "M"],
                        ind / 2, replace = FALSE) # Sample for equal sex ratio
    # samp.id.F <- sample(shark.id[sex.id == "F"],
    #                     ind / 1.5, replace = FALSE) # Sample for 2:1 sex ratio
    # samp.id.M <- sample(shark.id[sex.id == "M"],
    #                     ind / 3, replace = FALSE) # Sample for 2:1 sex ratio
    # samp.id.F <- sample(shark.id[sex.id == "F"],
    #                     ind / 3, replace = FALSE) # Sample for 1:2 sex ratio
    # samp.id.M <- sample(shark.id[sex.id == "M"],
    #                     ind / 1.5, replace = FALSE) # Sample for 1:2 sex ratio
    samp.id <- c(samp.id.F, samp.id.M)
    var.id <- sample(seqGetData(data, "variant.id"), var, replace = FALSE)

    newfile <- paste0(k,".C.carcharias_subset.gds")

    radiator_gds_subset(parentfile, newfile, sample.include = samp.id,
                        snp.include = var.id)

    varsil <- 116200 * (var/20000) #percentage-wise

    silico <- dplyr::filter(silicodata, INDIVIDUALS %in% samp.id) %>%
      dplyr::filter(MARKERS %in% sample(unique(silicodata$MARKERS), varsil,
                                        replace = FALSE))
    save(silico, file =  paste0(k,".C.carcharias_subset_Silico.Rdata"))
    rm(silico)
    pop <- c(pop, paste0(ind, " sharks & ", var, " SNPs"))
  }
}
save(silicodata, pop, file = "0.C.carcharias_subset_Silicodata.Rdata")
```

5.3 Run for subsampled data

```
# setwd("D:/Radiator/Test_robustness/C.carcharias/1-1_FM_sexratio")
# setwd("D:/Radiator/Test_robustness/C.carcharias/2-1_FM_sexratio")
# setwd("D:/Radiator/Test_robustness/C.carcharias/1-2_FM_sexratio")

gdsfmt::showfile.gds(closeall = TRUE)
Nind <- c()
Nvarsnp <- c()
Nvarsil <- c()
Nsexmark <- c()
Nsexmark.intersect <- c()
```

```
Nsexmark.not.intersect <- c()
sex <- list()
k <- 0
for (k in 1:100) {
  print(k) # for progress
  data <- openfn.gds(paste0(k,".C.carcharias_subset.gds"), readonly = FALSE)
  data <- new("SeqVarGDSClass", data)
  load(paste0(k,".C.carcharias_subset_Silico.Rdata"))

  strata <- radiator::extract_individuals_metadata(
    gds = data,
    ind.field.select = c("TARGET_ID", "INDIVIDUALS", "STRATA"),
    whitelist = TRUE
  )
  Nind <- c(Nind, length(strata$INDIVIDUALS))
  Nvarsnp <- c(Nvarsnp,length(radiator::extract_markers_metadata(
    gds = data,markers.meta.select = "VARIANT_ID", whitelist = TRUE)$VARIANT_ID))
  Nvarsil <- c(Nvarsil, length(unique(silico$MARKERS)))

  sex[[k]] <- radiator::sexy_markers(
    data = data,
    silicodata = silico,
    strata = strata,
    coverage.thresholds = coverage.thresholds,
    filters = filters,
    interactive.filter = interactive.filter,
    parallel.core = parallel.core
    ,
    species = species,
    population = pop[k],
    folder.name = paste0(k,"._", pop[k]),
    sex.id.input = sex.id.input,
    tau = tau,
    mis.threshold.data = mis.threshold.data,
    mis.threshold.silicodata = mis.threshold.silicodata,
    zoom.data = zoom.data,
    zoom.silicodata = zoom.silicodata,
    threshold.y.markers = threshold.y.markers,
    threshold.y.silico.markers = threshold.y.silico.markers,
    het.qr.input = het.qr.input,
    threshold.x.markers.qr = threshold.x.markers.qr,
    threshold.x.markers.RD = threshold.x.markers.RD,
    threshold.x.markers.RD.silico = threshold.x.markers.RD.silico)

  Nsexmark <- c(Nsexmark,length(sex[[k]]$sexy.summary$SEX_MARKERS))
  Nsexmark.intersect <- c(Nsexmark.intersect,
                          sum(sex[[k]]$sexy.summary$SEX_MARKERS %in%
                                std.sex.names))
  Nsexmark.not.intersect <- c(Nsexmark.not.intersect,
                              sum(!(sex[[k]]$sexy.summary$SEX_MARKERS %in%
                                      std.sex.names)))
}
```

```
save(std.sex,Nstd.ind, Nstd.varsnp, Nstd.varsil, Nstd.sexmark,std.sex.names, sex,
     Nsexmark,Nsexmark.intersect,Nsexmark.not.intersect, Nind, Nvarsnp, Nvarsil,
     file = "0.C.carcharias_Sexy_markers.Rdata")
```

5.4 Heatmap of results

```
# save(ccar.data.11,ccar.data.21,ccar.data.12,ggal.data.11,ggal.data.21,
#      ggal.data.12, file = "data_for_heatmap.Rdata")
# load("data_for_heatmap.Rdata")
load("data_for_heatmap2.Rdata")
pal <- wesanderson::wes_palette("Zissou1", 100, type = "continuous")


title <- "Robustness of sexy_markers function for 1:1 F-M sex ratio"
subtitle <- "Total number of sex-linked markers\n(markers not in common with full data set)"
legend <- "Number of sex-linked markers\nin common with the full dataset"

heatmap.ccar.11a <- ggplot2::ggplot(ccar.data.11,
                                    ggplot2::aes(x = as.factor(Nind),
                                                 y = as.factor(Nvarsnp.prcnt))) +
  ggplot2::geom_tile(ggplot2::aes(fill = Nsexmark.intersect)) +
  ggplot2::labs(x = "Number of individuals subsampled",
                y = "Percentage of markers subsampled",
                title = title, subtitle = subtitle) +
  ggplot2::scale_fill_gradientn(colours = pal, name = legend) +
  ggplot2::geom_text(ggplot2::aes(label = paste0(Nsexmark,"\n(",
                                                 Nsexmark.not.intersect,")")),
                     colour = "grey30",size = 2) +
  ggplot2::scale_x_discrete(expand = c(0, 0)) +
  ggplot2::scale_y_discrete(expand = c(0, 0)) +
  ggplot2::coord_equal() +
  ggplot2::theme(plot.title = ggplot2::element_text( face = "bold", size = 10),
                 plot.subtitle = ggplot2::element_text(size = 8,
                                                       color = "grey30",
                                                       hjust = 0.5),
                 legend.title = ggplot2::element_text(size = 8,
                                                      color = "grey30"),
                 legend.direction = "horizontal", legend.position = "bottom") +
  ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth = 10,
                                                  title.position = "bottom"))
# print(heatmap.ccar.11a)

heatmap.ccar.11b <- ggplot2::ggplot(ccar.data.11,
                                    ggplot2::aes(x = as.factor(Nind),
                                                 y = as.factor(Nvarsnp))) +
  ggplot2::geom_tile(ggplot2::aes(fill = Nsexmark.intersect)) +
  ggplot2::labs(x = "Number of individuals subsampled",
                y = "Number of markers subsampled",
                title = title, subtitle = subtitle) +
  ggplot2::scale_fill_gradientn(colours = pal, name = legend) +
  ggplot2::geom_text(ggplot2::aes(label = paste0(Nsexmark,"\n(",
```

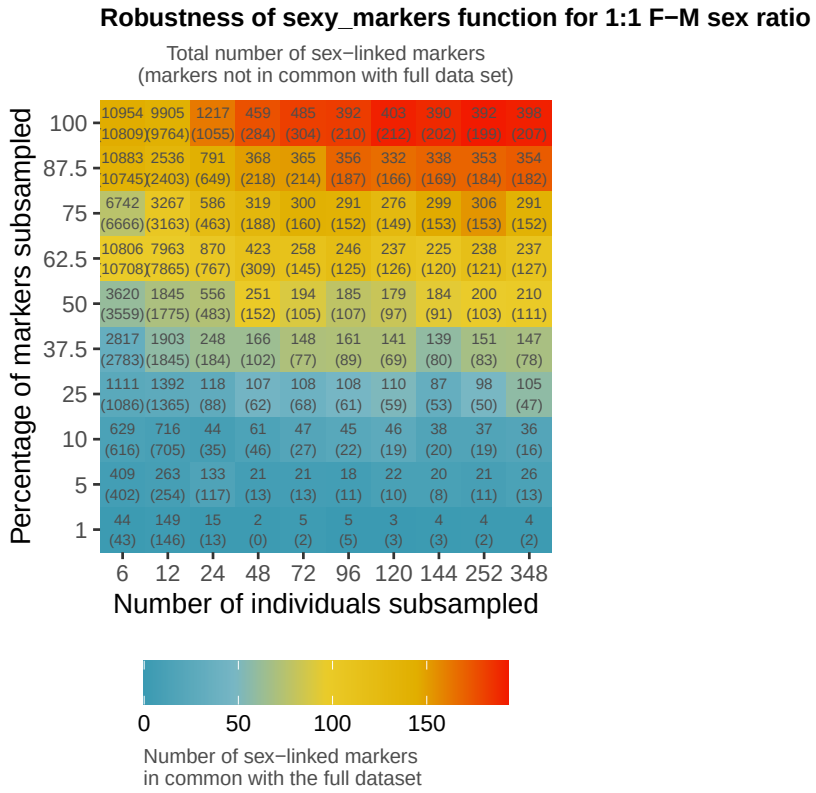
```
                                                Nsexmark.not.intersect,")")),
                    colour = "grey30",size = 2) +
  ggplot2::scale_x_discrete(expand = c(0, 0)) +
  ggplot2::scale_y_discrete(expand = c(0, 0)) +
  ggplot2::coord_equal() +
  ggplot2::theme(plot.title = ggplot2::element_text( face = "bold",
                                                     size = 10),
                 plot.subtitle = ggplot2::element_text(size = 8,
                                                       color = "grey30",
                                                       hjust = 0.5),
                 legend.title = ggplot2::element_text(size = 8,
                                                      color = "grey30"),
                 legend.direction = "horizontal", legend.position = "bottom") +
  ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth = 10,
                                                  title.position = "bottom"))
# print(heatmap.ccar.11b)

subtitle <- "Total number of sex-linked markers\n(markers in common with full data set)"
legend <- "Number of sex-linked markers\nnot in common with the full dataset"
heatmap.ccar.11c <- ggplot2::ggplot(ccar.data.11,
                                    ggplot2::aes(x = as.factor(Nind),
                                                 y = as.factor(Nvarsnp))) +
  ggplot2::geom_tile(ggplot2::aes(fill = Nsexmark.not.intersect)) +
  ggplot2::labs(x = "Number of individuals subsampled",
                y = "Number of markers subsampled",
                title = title,
                subtitle = subtitle) +
  ggplot2::scale_fill_gradientn(colours = pal,
                                name = legend,
                                values = scales::rescale(c(0,1,3,10,100,1000,5000, 10000))) +
  ggplot2::geom_text(ggplot2::aes(label = paste0(Nsexmark,"\n(",
                                                 Nsexmark.intersect,")")),
                     colour = "grey30",
                     size = 2) +
  ggplot2::scale_x_discrete(expand = c(0, 0)) +
  ggplot2::scale_y_discrete(expand = c(0, 0)) +
  ggplot2::coord_equal() +
  ggplot2::theme(plot.title = ggplot2::element_text( face = "bold",
                                                     size = 10),
                 plot.subtitle = ggplot2::element_text(size = 8,
                                                       color = "grey30",
                                                       hjust = 0.5),
                 legend.title = ggplot2::element_text(size = 8,
                                                      color = "grey30"),
                 legend.direction = "horizontal",
                 legend.position = "bottom") +
  ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth = 11,
                                                  title.position = "bottom"))
# print(heatmap.ccar.11c)


title <- "Robustness of sexy_markers function for 2:1 F-M sex ratio"
subtitle <- "Total number of sex-linked markers\n(markers not in common with full data set)"
```

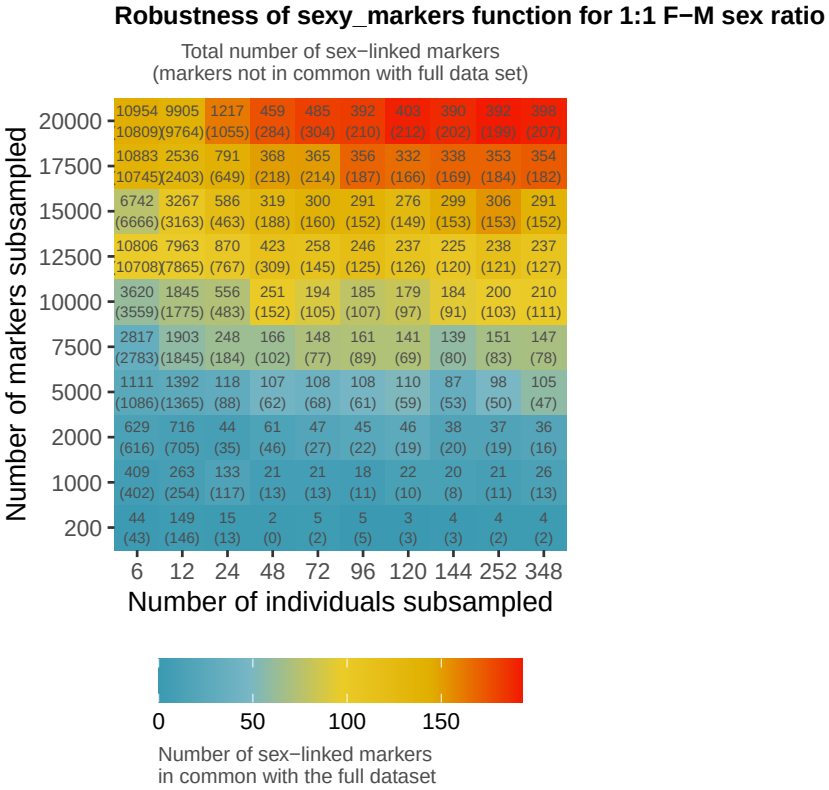
```
legend <- "Number of sex-linked markers\nin common with the full dataset"

heatmap.ccar.21a <- ggplot2::ggplot(ccar.data.21,
                                    ggplot2::aes(x = as.factor(Nind),
                                                 y = as.factor(Nvarsnp.prcnt))) +
  ggplot2::geom_tile(ggplot2::aes(fill = Nsexmark.intersect)) +
  ggplot2::labs(x = "Number of individuals subsampled",
                y = "Percentage of markers subsampled",
                title = title, subtitle = subtitle) +
  ggplot2::scale_fill_gradientn(colours = pal,
                                name = legend) +
  ggplot2::geom_text(ggplot2::aes(label = paste0(Nsexmark,"\n(",
                                                 Nsexmark.not.intersect,")")),
                     colour = "grey30",size = 2) +
  ggplot2::scale_x_discrete(expand = c(0, 0)) +
  ggplot2::scale_y_discrete(expand = c(0, 0)) +
  ggplot2::coord_equal() +
  ggplot2::theme(plot.title = ggplot2::element_text( face = "bold",
                                                     size = 10),
                 plot.subtitle = ggplot2::element_text(size = 8,
                                                       color = "grey30",
                                                       hjust = 0.5),
                 legend.title = ggplot2::element_text(size = 8,
                                                      color = "grey30"),
                 legend.direction = "horizontal", legend.position = "bottom") +
  ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth = 10,
                                                  title.position = "bottom"))
# print(heatmap.ccar.21a)

heatmap.ccar.21b <- ggplot2::ggplot(ccar.data.21,
                                    ggplot2::aes(x = as.factor(Nind),
                                                 y = as.factor(Nvarsnp))) +
  ggplot2::geom_tile(ggplot2::aes(fill = Nsexmark.intersect)) +
  ggplot2::labs(x = "Number of individuals subsampled",
                y = "Number of markers subsampled",
                title = title, subtitle = subtitle) +
  ggplot2::scale_fill_gradientn(colours = pal, name = legend) +
  ggplot2::geom_text(ggplot2::aes(label = paste0(Nsexmark,"\n(",
                                                 Nsexmark.not.intersect,")")),
                     colour = "grey30",size = 2) +
  ggplot2::scale_x_discrete(expand = c(0, 0)) +
  ggplot2::scale_y_discrete(expand = c(0, 0)) +
  ggplot2::coord_equal() +
  ggplot2::theme(plot.title = ggplot2::element_text( face = "bold",
                                                     size = 10),
                 plot.subtitle = ggplot2::element_text(size = 8,
                                                       color = "grey30",
                                                       hjust = 0.5),
                 legend.title = ggplot2::element_text(size = 8,
                                                      color = "grey30"),
                 legend.direction = "horizontal", legend.position = "bottom") +
  ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth = 10,
                                                  title.position = "bottom"))
```

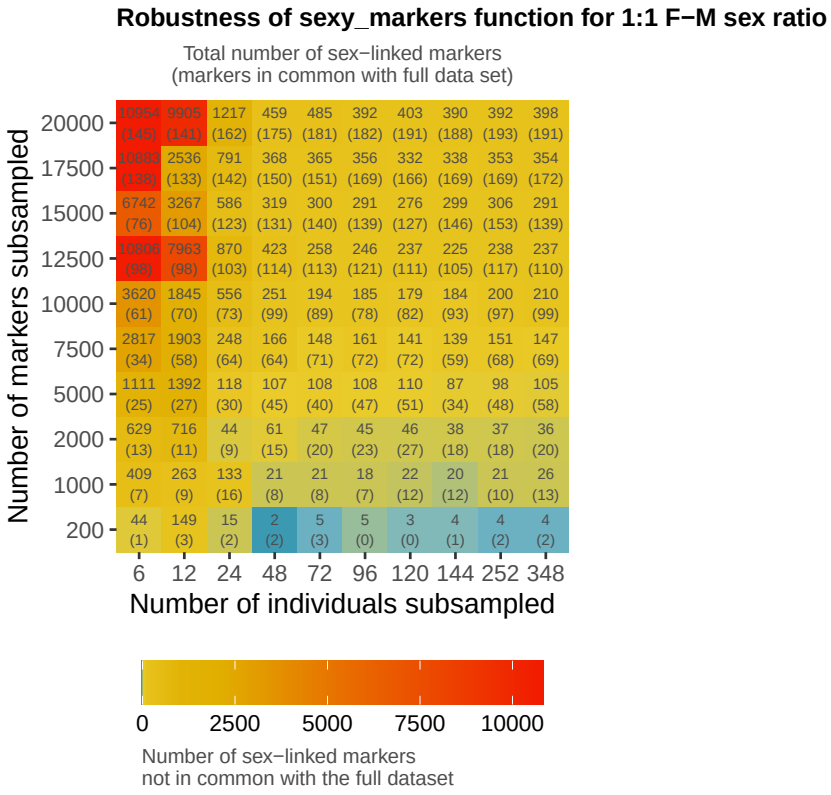
```
# print(heatmap.ccar.21b)

subtitle <- "Total number of sex-linked markers\n(markers in common with full data set)"
legend <- "Number of sex-linked markers\nnot in common with the full dataset"
heatmap.ccar.21c <- ggplot2::ggplot(ccar.data.21,
                                    ggplot2::aes(x = as.factor(Nind),
                                                 y = as.factor(Nvarsnp))) +
  ggplot2::geom_tile(ggplot2::aes(fill = Nsexmark.not.intersect)) +
  ggplot2::labs(x = "Number of individuals subsampled",
                y = "Number of markers subsampled",
                title = title, subtitle = subtitle) +
  ggplot2::scale_fill_gradientn(colours = pal, name = legend,
                                values = scales::rescale(c(0,1,3,10,100,1000,5000, 10000))) +
  ggplot2::geom_text(ggplot2::aes(label = paste0(Nsexmark,"\n(",
                                                 Nsexmark.intersect,")")),
                     colour = "grey30",size = 2) +
  ggplot2::scale_x_discrete(expand = c(0, 0)) +
  ggplot2::scale_y_discrete(expand = c(0, 0)) +
  ggplot2::coord_equal() +
  ggplot2::theme(plot.title = ggplot2::element_text( face = "bold",
                                                     size = 10),
                 plot.subtitle = ggplot2::element_text(size = 8,
                                                       color = "grey30",
                                                       hjust = 0.5),
                 legend.title = ggplot2::element_text(size = 8,
                                                      color = "grey30"),
                 legend.direction = "horizontal",
                 legend.position = "bottom") +
  ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth = 10,
                                                  title.position = "bottom"))
# print(heatmap.ccar.21c)


title <- "Robustness of sexy_markers function for 1:2 F-M sex ratio"
subtitle <- "Total number of sex-linked markers\n(markers not in common with full data set)"
legend <- "Number of sex-linked markers\nin common with the full dataset"

heatmap.ccar.12a <- ggplot2::ggplot(ccar.data.12,
                                    ggplot2::aes(x = as.factor(Nind),
                                                 y = as.factor(Nvarsnp.prcnt))) +
  ggplot2::geom_tile(ggplot2::aes(fill = Nsexmark.intersect)) +
  ggplot2::labs(x = "Number of individuals subsampled",
                y = "Percentage of markers subsampled",
                title = title,
                subtitle = subtitle) +
  ggplot2::scale_fill_gradientn(colours = pal, name = legend) +
  ggplot2::geom_text(ggplot2::aes(label = paste0(Nsexmark,"\n(",
                                                 Nsexmark.not.intersect,")")),
                     colour = "grey30",size = 2) +
  ggplot2::scale_x_discrete(expand = c(0, 0)) +
  ggplot2::scale_y_discrete(expand = c(0, 0)) +
  ggplot2::coord_equal() +
  ggplot2::theme(plot.title = ggplot2::element_text( face = "bold",
```

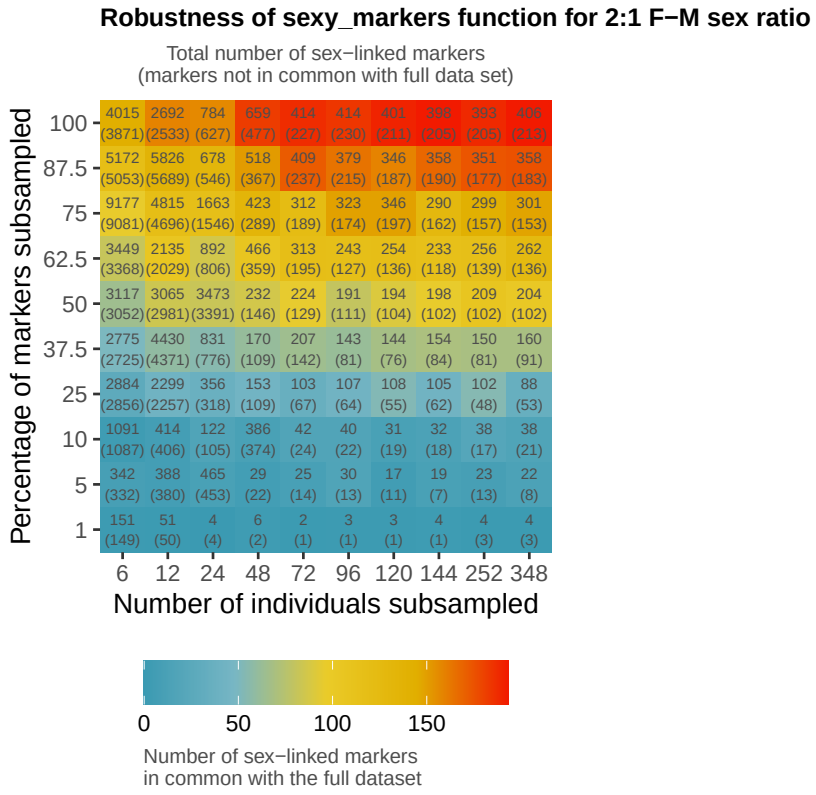
```
                                                  size = 10),
                plot.subtitle = ggplot2::element_text(size = 8,
                                                      color = "grey30",
                                                      hjust = 0.5),
                legend.title = ggplot2::element_text(size = 8,
                                                  color = "grey30"),
                legend.direction = "horizontal",
                legend.position = "bottom") +
  ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth = 10,
                                                  title.position = "bottom"))
# print(heatmap.ccar.12a)

heatmap.ccar.12b <- ggplot2::ggplot(ccar.data.12,
                                    ggplot2::aes(x = as.factor(Nind),
                                                 y = as.factor(Nvarsnp))) +
  ggplot2::geom_tile(ggplot2::aes(fill = Nsexmark.intersect)) +
  ggplot2::labs(x = "Number of individuals subsampled",
                y = "Number of markers subsampled",
                title = title,
                subtitle = subtitle) +
  ggplot2::scale_fill_gradientn(colours = pal, name = legend) +
  ggplot2::geom_text(ggplot2::aes(label = paste0(Nsexmark,"\n(",
                                                 Nsexmark.not.intersect,")")),
                     colour = "grey30",size = 2) +
  ggplot2::scale_x_discrete(expand = c(0, 0)) +
  ggplot2::scale_y_discrete(expand = c(0, 0)) +
  ggplot2::coord_equal() +
  ggplot2::theme(plot.title = ggplot2::element_text( face = "bold",
                                                  size = 10),
                plot.subtitle = ggplot2::element_text(size = 8,
                                                      color = "grey30",
                                                      hjust = 0.5),
                legend.title = ggplot2::element_text(size = 8,
                                                  color = "grey30"),
                legend.direction = "horizontal",
                legend.position = "bottom") +
  ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth = 10,
                                                  title.position = "bottom"))
# print(heatmap.ccar.12b)


subtitle <- "Total number of sex-linked markers\n(markers in common with full data set)"
legend <- "Number of sex-linked markers\nnot in common with the full dataset"
heatmap.ccar.12c <- ggplot2::ggplot(ccar.data.12,
                                    ggplot2::aes(x = as.factor(Nind),
                                                 y = as.factor(Nvarsnp))) +
  ggplot2::geom_tile(ggplot2::aes(fill = Nsexmark.not.intersect)) +
  ggplot2::labs(x = "Number of individuals subsampled",
                y = "Number of markers subsampled",
                title = title,
                subtitle = subtitle) +
  ggplot2::scale_fill_gradientn(colours = pal,
                                name = legend,
```

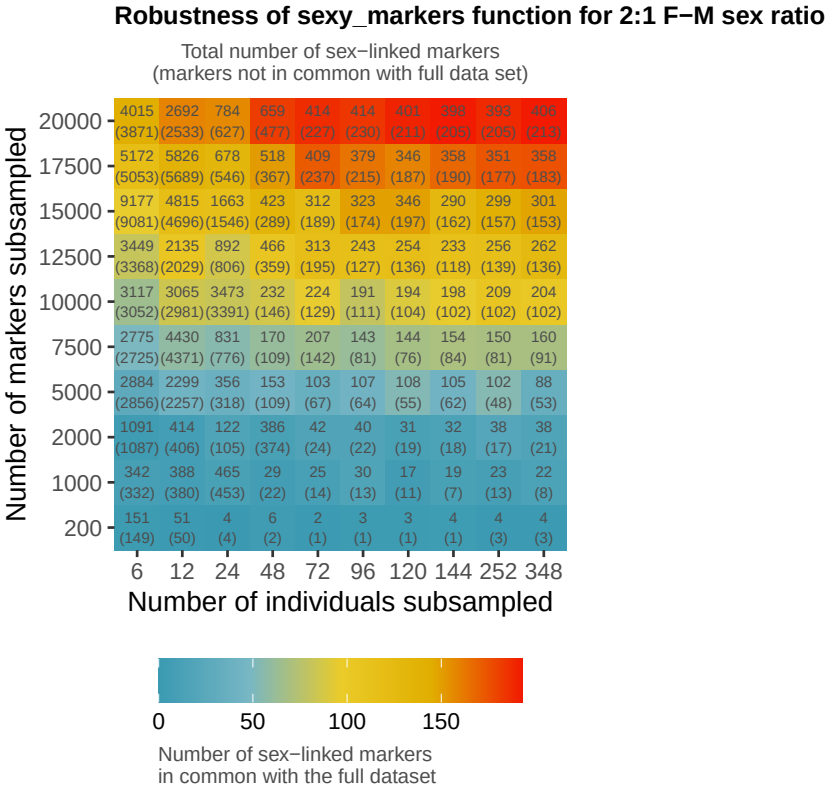
```
                                values = scales::rescale(c(0,1,3,10,100,1000,5000, 10000))) +
  ggplot2::geom_text(ggplot2::aes(label = paste0(Nsexmark,"\n(",
                                                 Nsexmark.intersect,")")),
                     colour = "grey30",size = 2) +
  ggplot2::scale_x_discrete(expand = c(0, 0)) +
  ggplot2::scale_y_discrete(expand = c(0, 0)) +
  ggplot2::coord_equal() +
  ggplot2::theme(plot.title = ggplot2::element_text( face = "bold",
                                                     size = 10),
                 plot.subtitle = ggplot2::element_text(size = 8,
                                                       color = "grey30",
                                                       hjust = 0.5),
                 legend.title = ggplot2::element_text(size = 8,
                                                      color = "grey30"),
                 legend.direction = "horizontal", legend.position = "bottom") +
  ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth = 10,
                                                  title.position = "bottom"))
# print(heatmap.ccar.12c)


#
# plot.a <- ggpubr::ggarrange(heatmap.ccar.11a, heatmap.ccar.21a,
#                             heatmap.ccar.12a, ncol = 1, nrow = 3,
#                             common.legend = TRUE, legend = "bottom")
# print(plot.a)
# plot.b <- ggpubr::ggarrange(heatmap.ccar.11b, heatmap.ccar.21b,
#                             heatmap.ccar.12b, ncol = 1, nrow = 3,
#                             common.legend = TRUE, legend = "bottom")
# print(plot.b)
# plot.c <- ggpubr::ggarrange(heatmap.ccar.11c, heatmap.ccar.21c,
#                             heatmap.ccar.12c, ncol = 1, nrow = 3,
#                             common.legend = TRUE, legend = "bottom")
# print(plot.c)


# filename <- "Ccarcharias_Heatmap_combined_a.pdf"
# width <- 45; height <- 15
#
# ggplot2::ggsave(
#   filename = filename,
#   plot = plot,
#   width = 15,
#   height = 15,
#   dpi = 300,
#   units = "cm"
# )
```

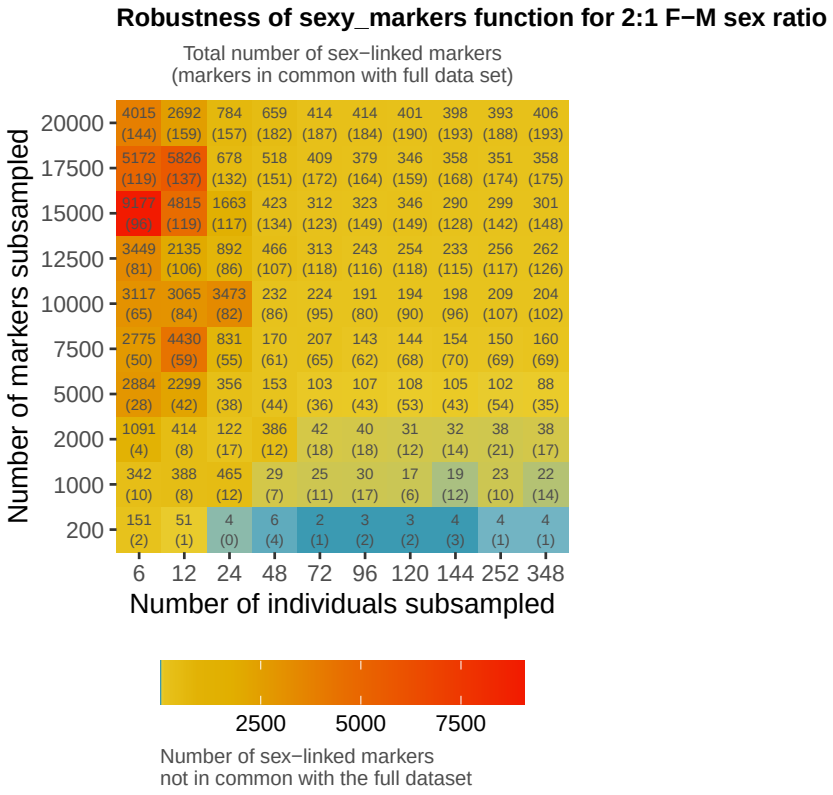

Robustness of sexy_markers function for 1:1 F−M sex ratio
Total number of sex−linked markers
(markers not in common with full data set)
Percentage of markers subsampled
100
87.5
75
62.5
50
37.5
25
10
5
1
10954 9905 1217 459 485 392 403 390 392 398
10809 (9764) (1055) (284) (304) (210) (212) (202) (199) (207)
10883 2536 791 368 365 356 332 338 353 354
10745 (2403) (649) (218) (214) (187) (166) (169) (184) (182)
6742 3267 586 319 300 291 276 299 306 291
(6666) (3163) (463) (188) (160) (152) (149) (153) (153) (152)
10806 7963 870 423 258 246 237 225 238 237
10708 (7865) (767) (309) (145) (125) (126) (120) (121) (127)
3620 1845 556 251 194 185 179 184 200 210
(3559) (1775) (483) (152) (105) (107) (97) (91) (103) (111)
2817 1903 248 166 148 161 141 139 151 147
(2783) (1845) (184) (102) (77) (89) (69) (80) (83) (78)
1111 1392 118 107 108 108 110 87 98 105
(1086) (1365) (88) (62) (68) (61) (59) (53) (50) (47)
629 716 44 61 47 45 46 38 37 36
(616) (705) (35) (46) (27) (22) (19) (20) (19) (16)
409 263 133 21 21 18 22 20 21 26
(402) (254) (117) (13) (13) (11) (10) (8) (11) (13)
44 149 15 2 5 5 3 4 4 4
(43) (146) (13) (0) (2) (5) (3) (3) (2) (2)
6 12 24 48 72 96 120 144 252 348
Number of individuals subsampled
0 50 100 150
Number of sex−linked markers
in common with the full dataset

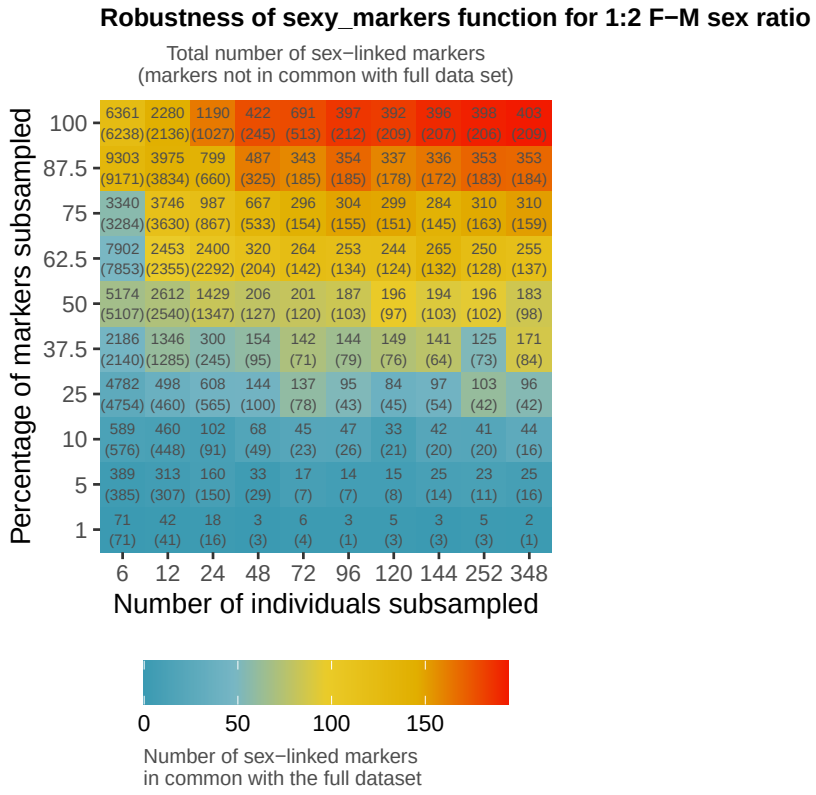

Robustness of sexy_markers function for 1:1 F−M sex ratio
Total number of sex−linked markers
(markers not in common with full data set)
Number of markers subsampled
20000
17500
15000
12500
10000
7500
5000
2000
1000
200
10954 9905 1217 459 485 392 403 390 392 398
10809 (9764) (1055) (284) (304) (210) (212) (202) (199) (207)
10883 2536 791 368 365 356 332 338 353 354
10745 (2403) (649) (218) (214) (187) (166) (169) (184) (182)
6742 3267 586 319 300 291 276 299 306 291
(6666) (3163) (463) (188) (160) (152) (149) (153) (153) (152)
10806 7963 870 423 258 246 237 225 238 237
10708 (7865) (767) (309) (145) (125) (126) (120) (121) (127)
3620 1845 556 251 194 185 179 184 200 210
(3559) (1775) (483) (152) (105) (107) (97) (91) (103) (111)
2817 1903 248 166 148 161 141 139 151 147
(2783) (1845) (184) (102) (77) (89) (69) (80) (83) (78)
1111 1392 118 107 108 108 110 87 98 105
(1086) (1365) (88) (62) (68) (61) (59) (53) (50) (47)
629 716 44 61 47 45 46 38 37 36
(616) (705) (35) (46) (27) (22) (19) (20) (19) (16)
409 263 133 21 21 18 22 20 21 26
(402) (254) (117) (13) (13) (11) (10) (8) (11) (13)
44 149 15 2 5 5 3 4 4 4
(43) (146) (13) (0) (2) (5) (3) (3) (2) (2)
6 12 24 48 72 96 120 144 252 348
Number of individuals subsampled
0 50 100 150
Number of sex−linked markers
in common with the full dataset

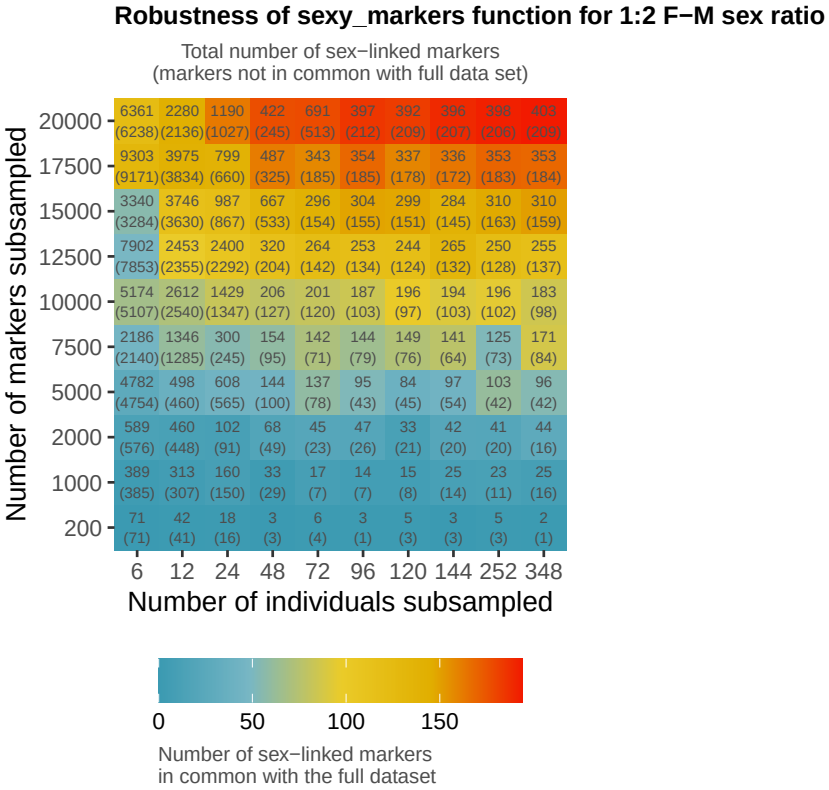

Robustness of sexy_markers function for 1:1 F−M sex ratio
Total number of sex−linked markers
(markers in common with full data set)
Number of markers subsampled
20000
17500
15000
12500
10000
7500
5000
2000
1000
200
10954 (145) 9905 (141) 1217 (162) 459 (175) 485 (181) 392 (182) 403 (191) 390 (188) 392 (193) 398 (191)
10883 (138) 2536 (133) 791 (142) 368 (150) 365 (151) 356 (169) 332 (166) 338 (169) 353 (169) 354 (172)
6742 (76) 3267 (104) 586 (123) 319 (131) 300 (140) 291 (139) 276 (127) 299 (146) 306 (153) 291 (139)
10806 (98) 7963 (98) 870 (103) 423 (114) 258 (113) 246 (121) 237 (111) 225 (105) 238 (117) 237 (110)
3620 (61) 1845 (70) 556 (73) 251 (99) 194 (89) 185 (78) 179 (82) 184 (93) 200 (97) 210 (99)
2817 (34) 1903 (58) 248 (64) 166 (64) 148 (71) 161 (72) 141 (72) 139 (59) 151 (68) 147 (69)
1111 (25) 1392 (27) 118 (30) 107 (45) 108 (40) 108 (47) 110 (51) 87 (34) 98 (48) 105 (58)
629 (13) 716 (11) 44 (9) 61 (15) 47 (20) 45 (23) 46 (27) 38 (18) 37 (18) 36 (20)
409 (7) 263 (9) 133 (16) 21 (8) 21 (8) 18 (7) 22 (12) 20 (12) 21 (10) 26 (13)
44 (1) 149 (3) 15 (2) 2 (2) 5 (3) 5 (0) 3 (0) 4 (1) 4 (2) 4 (2)
6 12 24 48 72 96 120 144 252 348
Number of individuals subsampled
0 2500 5000 7500 10000
Number of sex−linked markers
not in common with the full dataset

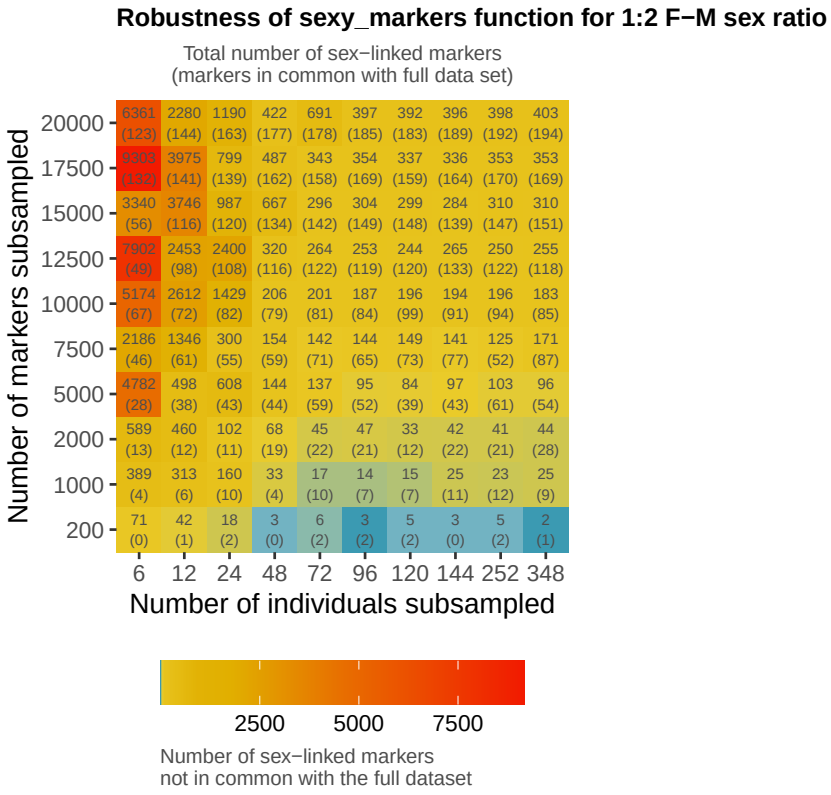

Robustness of sexy_markers function for 2:1 F−M sex ratio
Total number of sex−linked markers
(markers not in common with full data set)
Percentage of markers subsampled
100
87.5
75
62.5
50
37.5
25
10
5
1
4015 (3871) 2692 (2533) 784 (627) 659 (477) 414 (227) 414 (230) 401 (211) 398 (205) 393 (205) 406 (213)
5172 (5053) 5826 (5689) 678 (546) 518 (367) 409 (237) 379 (215) 346 (187) 358 (190) 351 (177) 358 (183)
9177 (9081) 4815 (4696) 1663 (1546) 423 (289) 312 (189) 323 (174) 346 (197) 290 (162) 299 (157) 301 (153)
3449 (3368) 2135 (2029) 892 (806) 466 (359) 313 (195) 243 (127) 254 (136) 233 (118) 256 (139) 262 (136)
3117 (3052) 3065 (2981) 3473 (3391) 232 (146) 224 (129) 191 (111) 194 (104) 198 (102) 209 (102) 204 (102)
2775 (2725) 4430 (4371) 831 (776) 170 (109) 207 (142) 143 (81) 144 (76) 154 (84) 150 (81) 160 (91)
2884 (2856) 2299 (2257) 356 (318) 153 (109) 103 (67) 107 (64) 108 (55) 105 (62) 102 (48) 88 (53)
1091 (1087) 414 (406) 122 (105) 386 (374) 42 (24) 40 (22) 31 (19) 32 (18) 38 (17) 38 (21)
342 (332) 388 (380) 465 (453) 29 (22) 25 (14) 30 (13) 17 (11) 19 (7) 23 (13) 22 (8)
151 (149) 51 (50) 4 (4) 6 (2) 2 (1) 3 (1) 3 (1) 4 (1) 4 (3) 4 (3)
6 12 24 48 72 96 120 144 252 348
Number of individuals subsampled
0 50 100 150
Number of sex−linked markers
in common with the full dataset


Robustness of sexy_markers function for 2:1 F−M sex ratio
Total number of sex−linked markers
(markers not in common with full data set)
Number of markers subsampled
20000
17500
15000
12500
10000
7500
5000
2000
1000
200
4015 (3871) 2692 (2533) 784 (627) 659 (477) 414 (227) 414 (230) 401 (211) 398 (205) 393 (205) 406 (213)
5172 (5053) 5826 (5689) 678 (546) 518 (367) 409 (237) 379 (215) 346 (187) 358 (190) 351 (177) 358 (183)
9177 (9081) 4815 (4696) 1663 (1546) 423 (289) 312 (189) 323 (174) 346 (197) 290 (162) 299 (157) 301 (153)
3449 (3368) 2135 (2029) 892 (806) 466 (359) 313 (195) 243 (127) 254 (136) 233 (118) 256 (139) 262 (136)
3117 (3052) 3065 (2981) 3473 (3391) 232 (146) 224 (129) 191 (111) 194 (104) 198 (102) 209 (102) 204 (102)
2775 (2725) 4430 (4371) 831 (776) 170 (109) 207 (142) 143 (81) 144 (76) 154 (84) 150 (81) 160 (91)
2884 (2856) 2299 (2257) 356 (318) 153 (109) 103 (67) 107 (64) 108 (55) 105 (62) 102 (48) 88 (53)
1091 (1087) 414 (406) 122 (105) 386 (374) 42 (24) 40 (22) 31 (19) 32 (18) 38 (17) 38 (21)
342 (332) 388 (380) 465 (453) 29 (22) 25 (14) 30 (13) 17 (11) 19 (7) 23 (13) 22 (8)
151 (149) 51 (50) 4 (4) 6 (2) 2 (1) 3 (1) 3 (1) 4 (1) 4 (3) 4 (3)
6 12 24 48 72 96 120 144 252 348
Number of individuals subsampled
0 50 100 150
Number of sex−linked markers
in common with the full dataset


Robustness of sexy_markers function for 2:1 F−M sex ratio
Total number of sex−linked markers
(markers in common with full data set)
Number of markers subsampled
20000
17500
15000
12500
10000
7500
5000
2000
1000
200
4015 (144) 2692 (159) 784 (157) 659 (182) 414 (187) 414 (184) 401 (190) 398 (193) 393 (188) 406 (193)
5172 (119) 5826 (137) 678 (132) 518 (151) 409 (172) 379 (164) 346 (159) 358 (168) 351 (174) 358 (175)
9177 (96) 4815 (119) 1663 (117) 423 (134) 312 (123) 323 (149) 346 (149) 290 (128) 299 (142) 301 (148)
3449 (81) 2135 (106) 892 (86) 466 (107) 313 (118) 243 (116) 254 (118) 233 (115) 256 (117) 262 (126)
3117 (65) 3065 (84) 3473 (82) 232 (86) 224 (95) 191 (80) 194 (90) 198 (96) 209 (107) 204 (102)
2775 (50) 4430 (59) 831 (55) 170 (61) 207 (65) 143 (62) 144 (68) 154 (70) 150 (69) 160 (69)
2884 (28) 2299 (42) 356 (38) 153 (44) 103 (36) 107 (43) 108 (53) 105 (43) 102 (54) 88 (35)
1091 (4) 414 (8) 122 (17) 386 (12) 42 (18) 40 (18) 31 (12) 32 (14) 38 (21) 38 (17)
342 (10) 388 (8) 465 (12) 29 (7) 25 (11) 30 (17) 17 (6) 19 (12) 23 (10) 22 (14)
151 (2) 51 (1) 4 (0) 6 (4) 2 (1) 3 (2) 3 (2) 4 (3) 4 (1) 4 (1)
6 12 24 48 72 96 120 144 252 348
Number of individuals subsampled
2500 5000 7500
Number of sex−linked markers
not in common with the full dataset


Robustness of sexy_markers function for 1:2 F−M sex ratio
Total number of sex−linked markers
(markers not in common with full data set)
Percentage of markers subsampled
100
87.5
75
62.5
50
37.5
25
10
5
1
6361 (6238) 2280 (2136) 1190 (1027) 422 (245) 691 (513) 397 (212) 392 (209) 396 (207) 398 (206) 403 (209)
9303 (9171) 3975 (3834) 799 (660) 487 (325) 343 (185) 354 (185) 337 (178) 336 (172) 353 (183) 353 (184)
3340 (3284) 3746 (3630) 987 (867) 667 (533) 296 (154) 304 (155) 299 (151) 284 (145) 310 (163) 310 (159)
7902 (7853) 2453 (2355) 2400 (2292) 320 (204) 264 (142) 253 (134) 244 (124) 265 (132) 250 (128) 255 (137)
5174 (5107) 2612 (2540) 1429 (1347) 206 (127) 201 (120) 187 (103) 196 (97) 194 (103) 196 (102) 183 (98)
2186 (2140) 1346 (1285) 300 (245) 154 (95) 142 (71) 144 (79) 149 (76) 141 (64) 125 (73) 171 (84)
4782 (4754) 498 (460) 608 (565) 144 (100) 137 (78) 95 (43) 84 (45) 97 (54) 103 (42) 96 (42)
589 (576) 460 (448) 102 (91) 68 (49) 45 (23) 47 (26) 33 (21) 42 (20) 41 (20) 44 (16)
389 (385) 313 (307) 160 (150) 33 (29) 17 (7) 14 (7) 15 (8) 25 (14) 23 (11) 25 (16)
71 (71) 42 (41) 18 (16) 3 (3) 6 (4) 3 (1) 5 (3) 3 (3) 5 (3) 2 (1)
6 12 24 48 72 96 120 144 252 348
Number of individuals subsampled
0 50 100 150
Number of sex−linked markers
in common with the full dataset


Robustness of sexy_markers function for 1:2 F−M sex ratio
Total number of sex−linked markers
(markers not in common with full data set)
Number of markers subsampled
20000
17500
15000
12500
10000
7500
5000
2000
1000
200
6361 (6238) 2280 (2136) 1190 (1027) 422 (245) 691 (513) 397 (212) 392 (209) 396 (207) 398 (206) 403 (209)
9303 (9171) 3975 (3834) 799 (660) 487 (325) 343 (185) 354 (185) 337 (178) 336 (172) 353 (183) 353 (184)
3340 (3284) 3746 (3630) 987 (867) 667 (533) 296 (154) 304 (155) 299 (151) 284 (145) 310 (163) 310 (159)
7902 (7853) 2453 (2355) 2400 (2292) 320 (204) 264 (142) 253 (134) 244 (124) 265 (132) 250 (128) 255 (137)
5174 (5107) 2612 (2540) 1429 (1347) 206 (127) 201 (120) 187 (103) 196 (97) 194 (103) 196 (102) 183 (98)
2186 (2140) 1346 (1285) 300 (245) 154 (95) 142 (71) 144 (79) 149 (76) 141 (64) 125 (73) 171 (84)
4782 (4754) 498 (460) 608 (565) 144 (100) 137 (78) 95 (43) 84 (45) 97 (54) 103 (42) 96 (42)
589 (576) 460 (448) 102 (91) 68 (49) 45 (23) 47 (26) 33 (21) 42 (20) 41 (20) 44 (16)
389 (385) 313 (307) 160 (150) 33 (29) 17 (7) 14 (7) 15 (8) 25 (14) 23 (11) 25 (16)
71 (71) 42 (41) 18 (16) 3 (3) 6 (4) 3 (1) 5 (3) 3 (3) 5 (3) 2 (1)
6 12 24 48 72 96 120 144 252 348
Number of individuals subsampled
0 50 100 150
Number of sex−linked markers
in common with the full dataset


Robustness of sexy_markers function for 1:2 F−M sex ratio
Total number of sex−linked markers
(markers in common with full data set)
Number of markers subsampled
20000
17500
15000
12500
10000
7500
5000
2000
1000
200
6361 (123) 2280 (144) 1190 (163) 422 (177) 691 (178) 397 (185) 392 (183) 396 (189) 398 (192) 403 (194)
9303 (132) 3975 (141) 799 (139) 487 (162) 343 (158) 354 (169) 337 (159) 336 (164) 353 (170) 353 (169)
3340 (56) 3746 (116) 987 (120) 667 (134) 296 (142) 304 (149) 299 (148) 284 (139) 310 (147) 310 (151)
7902 (49) 2453 (98) 2400 (108) 320 (116) 264 (122) 253 (119) 244 (120) 265 (133) 250 (122) 255 (118)
5174 (67) 2612 (72) 1429 (82) 206 (79) 201 (81) 187 (84) 196 (99) 194 (91) 196 (94) 183 (85)
2186 (46) 1346 (61) 300 (55) 154 (59) 142 (71) 144 (65) 149 (73) 141 (77) 125 (52) 171 (87)
4782 (28) 498 (38) 608 (43) 144 (44) 137 (59) 95 (52) 84 (39) 97 (43) 103 (61) 96 (54)
589 (13) 460 (12) 102 (11) 68 (19) 45 (22) 47 (21) 33 (12) 42 (22) 41 (21) 44 (28)
389 (4) 313 (6) 160 (10) 33 (4) 17 (10) 14 (7) 15 (7) 25 (11) 23 (12) 25 (9)
71 (0) 42 (1) 18 (2) 3 (0) 6 (2) 3 (2) 5 (2) 3 (0) 5 (2) 2 (1)
6 12 24 48 72 96 120 144 252 348
Number of individuals subsampled
2500 5000 7500
Number of sex−linked markers
not in common with the full dataset

6 References

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C. 2012. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28(12):1647-1649

Marra, N.J., Stanhope, M.J., Jue, N.K., Wang, M., Sun, Q., Pavinski Bitar, P., Richards, V.P., Komissarov, A., Rayko, M., Kliver, S., Stanhope, B.J., Winkler, C., O'Brien, S.J., Antunes, A., Jorgensen, S. , and Shivji, M.S. (2019) White shark genome reveals ancient elasmobranch adaptations associated with wound healing and the maintenance of genome stability. Proceedings of the National Academy of Sciences of the United States of America, 116, 4446-4455.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., ... & Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature, 592(7856), 737-746.

Vertebrate Genome Project; https://vgp.github.io/genomeark/

Genome 10K Community of Scientists. "Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species." Journal of Heredity 100.6 (2009): 659-674.

7 Session info

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_Australia.1252  LC_CTYPE=English_Australia.1252
## [3] LC_MONETARY=English_Australia.1252 LC_NUMERIC=C
## [5] LC_TIME=English_Australia.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] pcadapt_4.3.3  adegenet_2.1.5 ade4_1.7-18    OutFLANK_0.2   qvalue_2.26.0
## [6] formatR_1.11   knitr_1.37     magrittr_2.0.1 radiator_1.2.0
##
## loaded via a namespace (and not attached):
##  [1] tidyr_1.1.4       bit64_4.0.5       vroom_1.5.7       splines_4.1.0
##  [5] shiny_1.7.1       assertthat_0.2.1  highr_0.9         yaml_2.2.1
##  [9] pillar_1.6.4      lattice_0.20-44   glue_1.6.0        digest_0.6.29
## [13] promises_1.2.0.1  colorspace_2.0-2  htmltools_0.5.2   httpuv_1.6.4
## [17] Matrix_1.4-0      plyr_1.8.6        pkgconfig_2.0.3   wesanderson_0.3.6
## [21] purrr_0.3.4       xtable_1.8-4      scales_1.1.1      RSpectra_0.16-0
## [25] later_1.3.0       tzdb_0.2.0        tibble_3.1.6      mgcv_1.8-38
## [29] generics_0.1.1    farver_2.1.0      ggplot2_3.3.5     ellipsis_0.3.2
## [33] cli_3.1.0         crayon_1.4.2      mime_0.12         evaluate_0.14
## [37] bigassertr_0.1.5  fansi_0.5.0       nlme_3.1-153      MASS_7.3-54
## [41] vegan_2.5-7       tools_4.1.0       hms_1.1.1         bigutilsr_0.3.4
## [45] lifecycle_1.0.1   stringr_1.4.0     munsell_0.5.0     cluster_2.1.2
## [49] compiler_4.1.0    rlang_0.4.12      grid_4.1.0        rstudioapi_0.13
```

```
## [53] igraph_1.2.10     labeling_0.4.2    rmarkdown_2.11    gtable_0.3.0
## [57] codetools_0.2-18  DBI_1.1.2         reshape2_1.4.4    R6_2.5.1
## [61] dplyr_1.0.7       fastmap_1.1.0     seqinr_4.2-8      bit_4.0.4
## [65] utf8_1.2.2        permute_0.9-5     readr_2.1.1       ape_5.6
## [69] stringi_1.7.6     parallel_4.1.0    Rcpp_1.0.7        vctrs_0.3.8
## [73] tidyselect_1.1.1  xfun_0.29
```

8 Citations

```
## [1] "Gosselin, T. (2020). radiator: RADseq Data Exploration, Manipulation and"
## [2] "   Visualization using R. R package version 1.1.9"
## [3] "   https://thierrygosselin.github.io/radiator/. doi : 10.5281/zenodo.3687060"
##
##
## [1] "Gosselin, T, Anderson, E. C., Bradbury, I. (2020). assigner: Assignment"
## [2] "   Analysis with GBS/RAD Data using R. R package version 0.5.8."
## [3] "   http://thierrygosselin.github.io/assigner/. doi : 10.5281/zenodo.592677"
##
##
## [1] "H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New"
## [2] "   York, 2016."
##
##
## [1] "Michael C. Whitlock and Katie Lotterhos (2014). OutFLANK: Fst outliers with"
## [2] "   trimming. R package version 0.2."
##
##
## [1] "Privé, Florian, et al. \"Performing highly efficient genome scans for local"
## [2] "   adaptation with R package pcadapt version 4.\" Molecular Biology and"
## [3] "   Evolution (2020)."
##
##
## [1] "Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source"
## [2] "   Software, 4(43), 1686, https://doi.org/10.21105/joss.01686"
##
##
## [1] "Xiuwen Zheng, David Levine, Jess Shen, Stephanie M. Gogarten, Cathy Laurie,"
## [2] "   Bruce S. Weir. A High-performance Computing Toolset for Relatedness and"
## [3] "   Principal Component Analysis of SNP Data. Bioinformatics 2012; doi:"
## [4] "   10.1093/bioinformatics/bts606"
##
##
## [1] "Xiuwen Zheng, Stephanie M. Gogarten, Michael Lawrence, Adrienne Stilp, Matthew"
## [2] "   P. Conomos, Bruce S. Weir, Cathy Laurie, David Levine. SeqArray -- A"
## [3] "   storage-efficient high-performance data format for WGS variant calls."
## [4] "   Bioinformatics 2017; doi: 10.1093/bioinformatics/btx145"
```