Early/Late fortis/scandens project

Fan Han

November 20, 2018

prepare dataset

Extract the 6 pools and rename the samples

Sample information

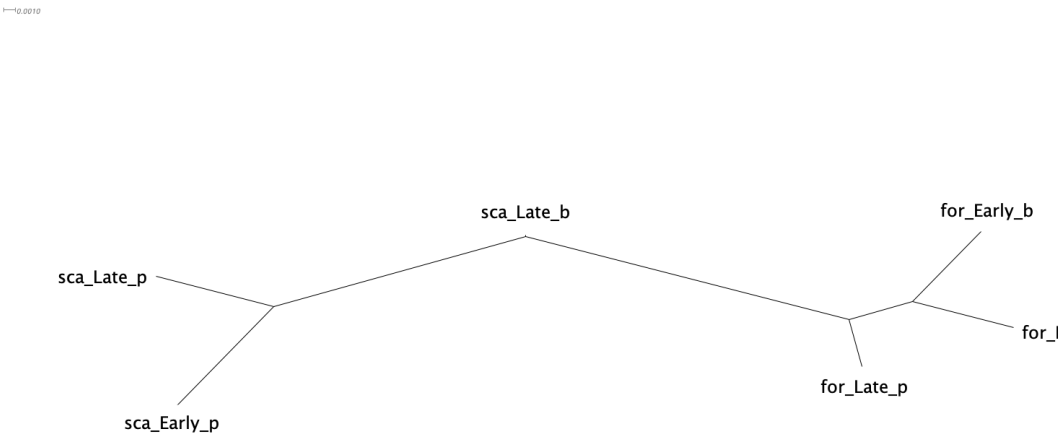
Original ID	Sample ID
150731_D00457_0112_BC71DHANXX/Sample_Pool1	sca_Early_p
150731_D00457_0112_BC71DHANXX/Sample_Pool2	sca_Late_b
150813_D00118_0225_BC79PNANXX/Sample_Pool3	sca_Late_p
150819_D00118_0226_AC79LFANXX/Sample_Pool4	for_Early_b
150819_D00118_0226_AC79LFANXX/Sample_Pool5	for_Late_b
150819_D00118_0226_AC79LFANXX/Sample_Pool6	for_Late_p

pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools

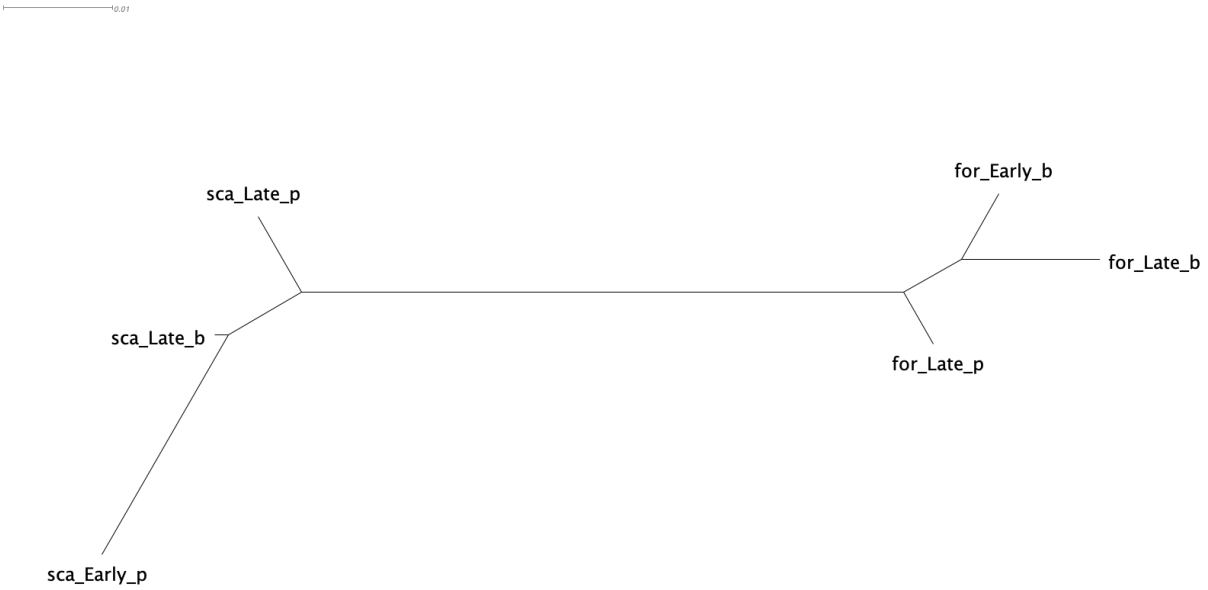
Number of SNPs: 10,814,262

Whole-genome NJ tree

- pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/WGTree
- No missing GT
- Number of SNPs: 10,795,687



- Plot of whole-genome trees



* It shows that the introgression primarily happened on autosomes

TreeScan with 50kb window

- Because the data is pooled sequencing, I can only build NJ tree for each window

- Calculate the allele frequency for each population
- Remember to midpoint-rooting the trees with retree program from PHYLIP otherwise the visualization in Densitree will be not pretty
- pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/TreeScan
- Run time 01:02:51

Number of trees: 20,632

DensiTree to visualize all the midpoint-rooted trees

- It is noticeable that within-species introgression happens more on Z and between-species introgression happens more on autosomes

Summary of the tree topologies across the genome

- The output of TreeScan is a list of newick trees. The tricky thing is to compare them in pairwise and decide which topologies are the same.
  It seems ETE Toolkit could help with this: http://etetoolkit.org/documentation/ete-compare/. It compares the Robinson-Foulds' distance of multiple trees with reference tree and report mismatches. But it needs reference trees as input. So first I need to produce all the possible topologies
- Treedist from PHYLIP package could also report pairwise distance and much faster than ETE. Change the Distance type to symmetric Difference. http://evolution.genetics.washington.edu/phylip/doc/treedist.html
- I cannot really determine whether two populations are clustered in unrooted trees if they are present on the internal branch. So I need to root every tree and compare if they are identical
- pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/TreeScan/Rooted
- Runtime: ~4 days
- Simplify tree categories. The trees that have two distinct clusters of fortis and scandens should be considered as one type no matter how the within-species divergence is. But how...

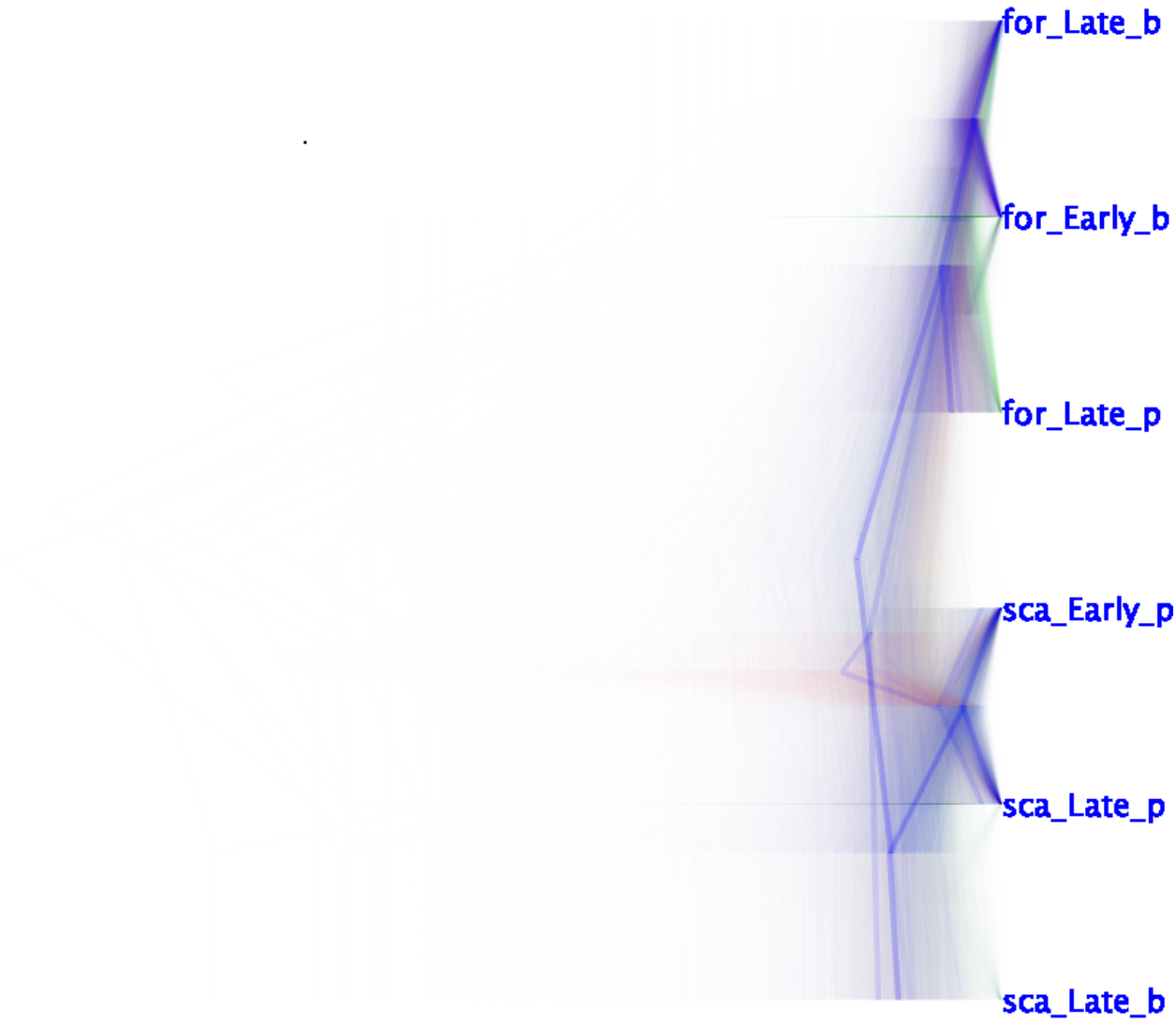


Figure 1: 19,040 Autosomal trees from DensiTree

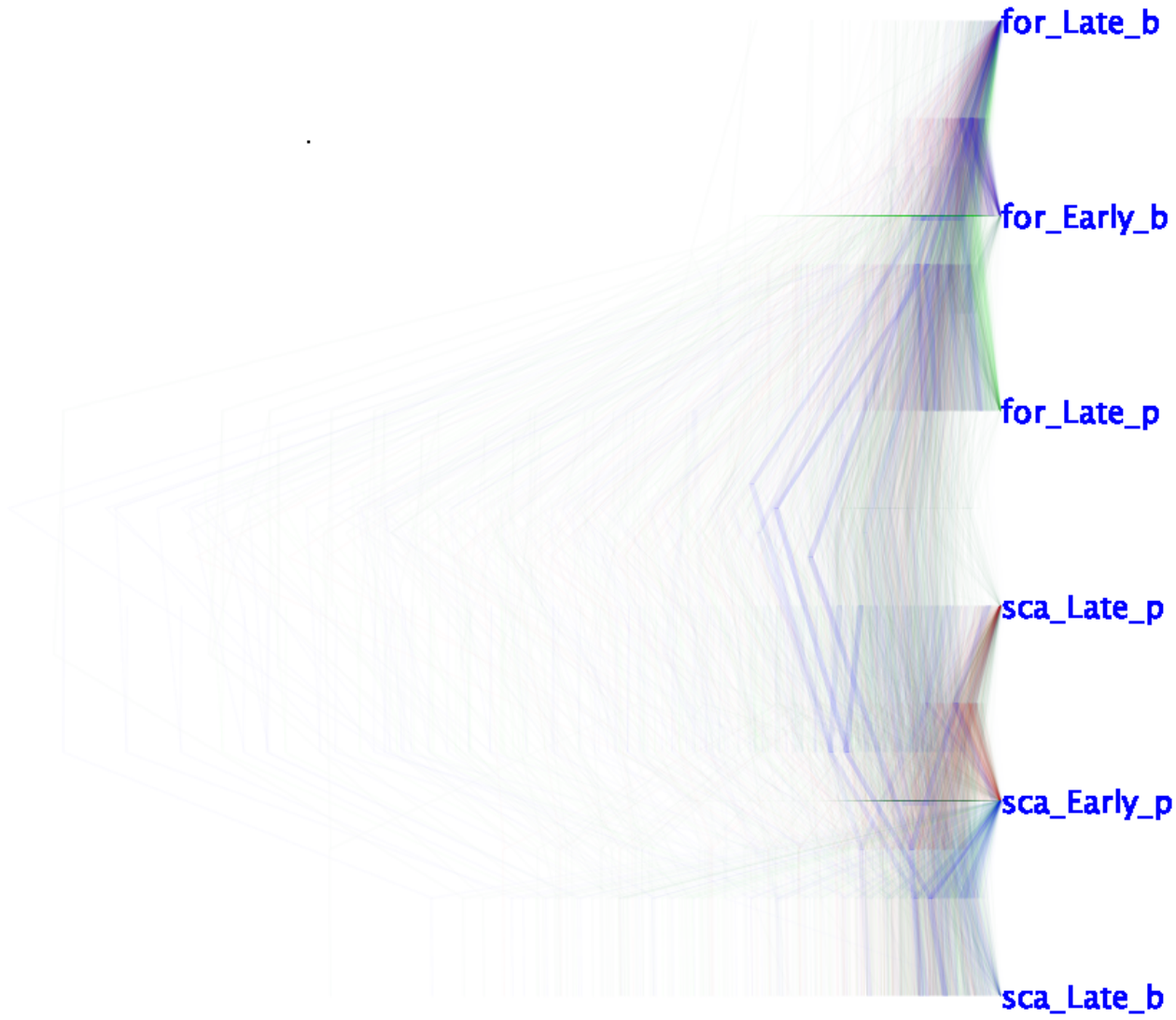
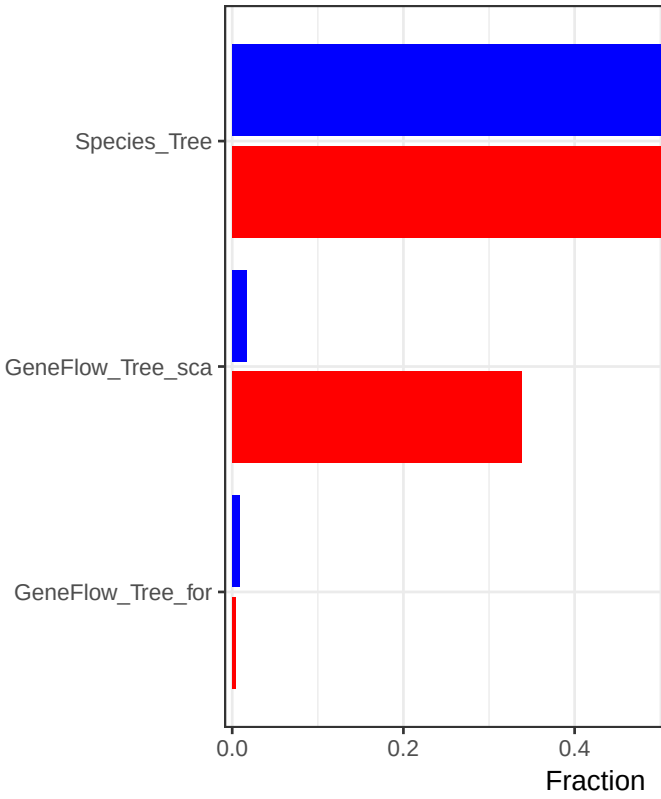


Figure 2: 1,536 Z trees from DensiTree



- Count the number of each rooted toplogy across the genome

```
## # A tibble: 6 x 4
## # Groups:   Var2 [2]
##   Var1              Var2   Freq percent
##   <fct>             <fct> <int>   <dbl>
## 1 GeneFlow_Tree_for Auto     76 0.00399
## 2 GeneFlow_Tree_sca Auto   6449 0.339
## 3 Species_Tree      Auto  11712 0.615
## 4 GeneFlow_Tree_for Z        14 0.00911
## 5 GeneFlow_Tree_sca Z        26 0.0169
## 6 Species_Tree      Z      1304 0.849
```

- Representives of each type of trees
- Recombination rate in the regions of introgression
- exam the relation between tree topology and recombination rate

```
##
##  Welch Two Sample t-test
##
## data:  autosome.species$RR_zebra and autosome.flow$RR_zebra
## t = 6.524, df = 12588, p-value = 7.11e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1517359 0.2820770
## sample estimates:
## mean of x mean of y
## 1.0460707 0.8291643
```

```
##
```

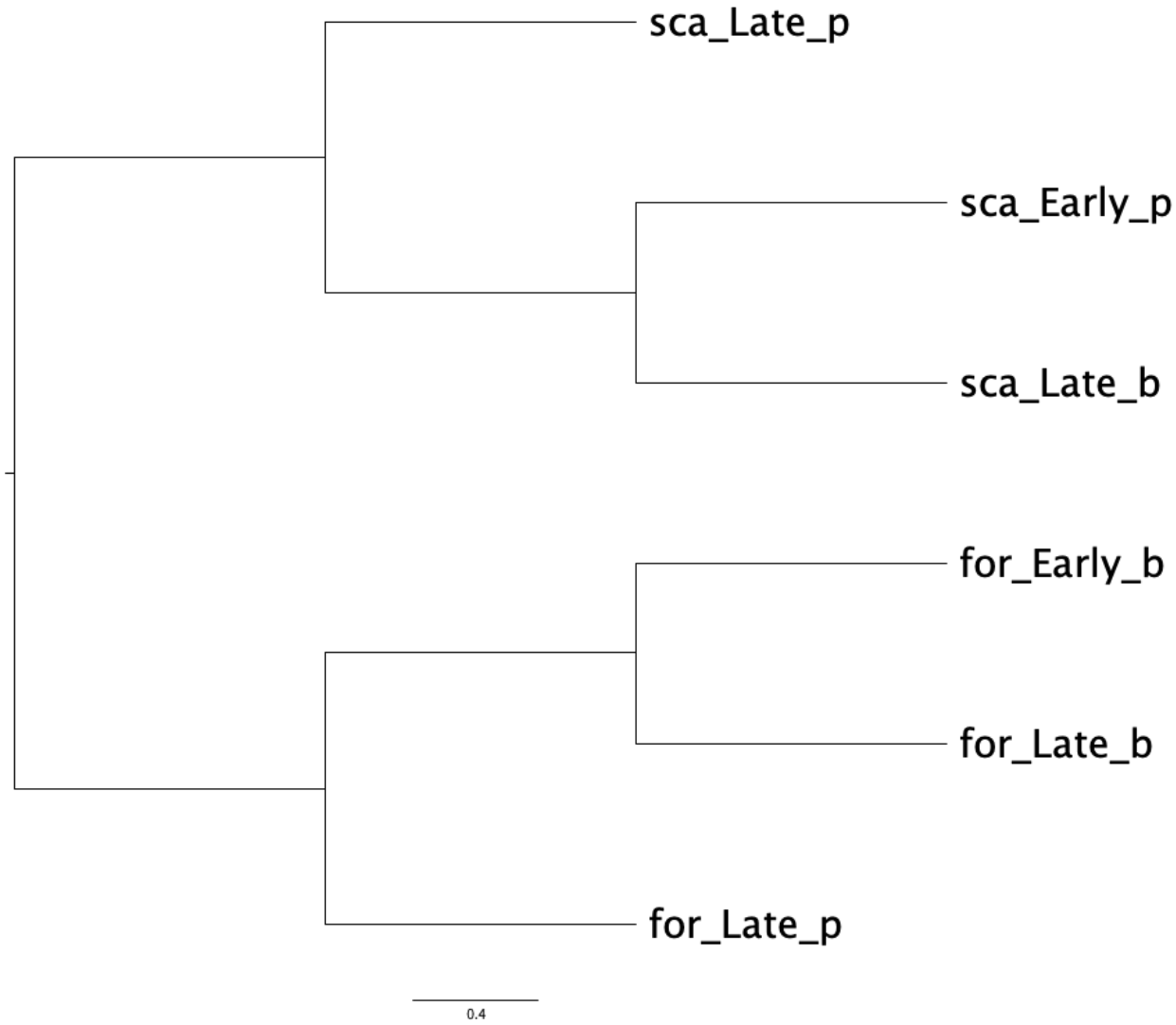


Figure 3: Species Tree

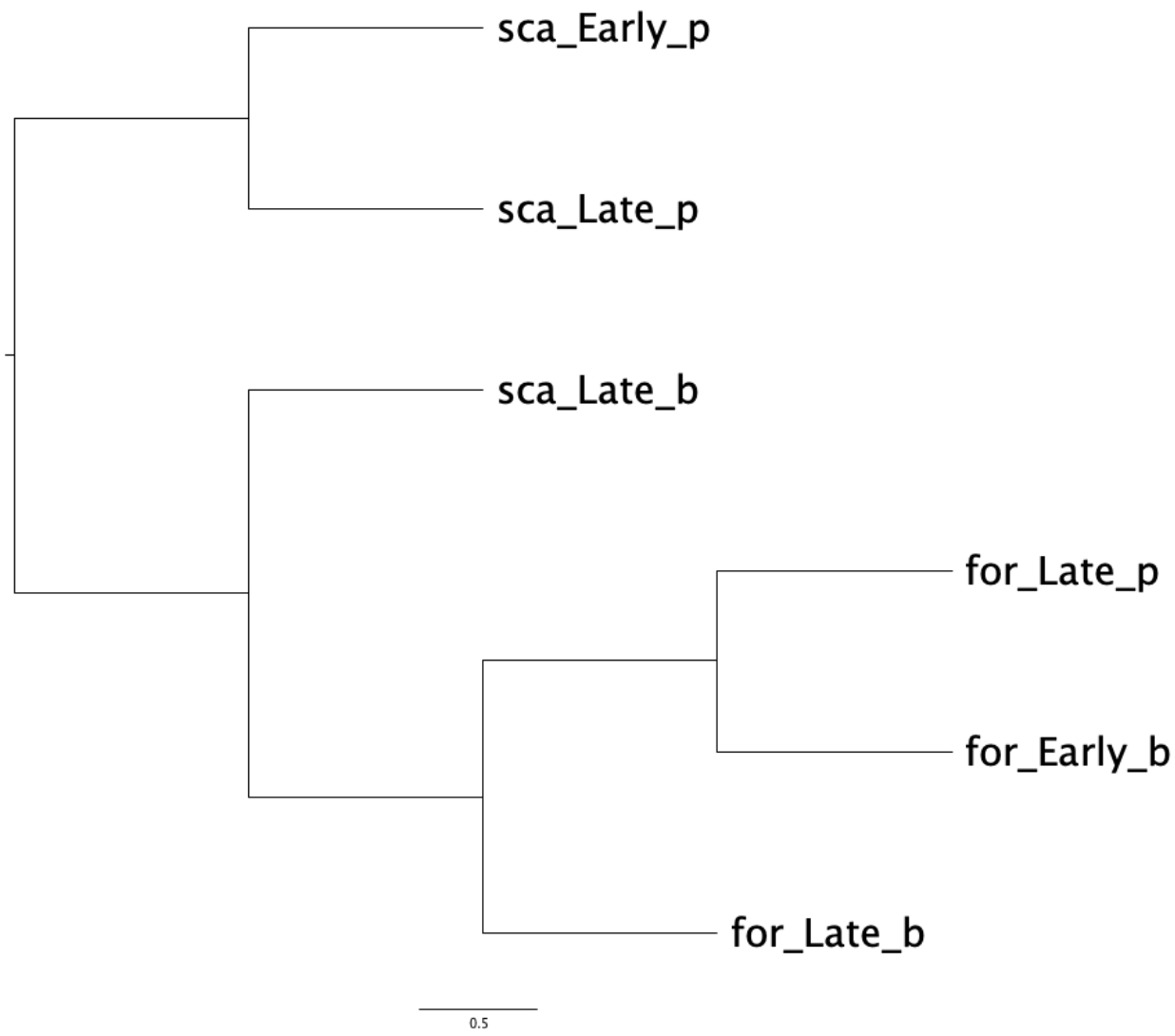


Figure 4: fortis introgressed scandens Late blunt

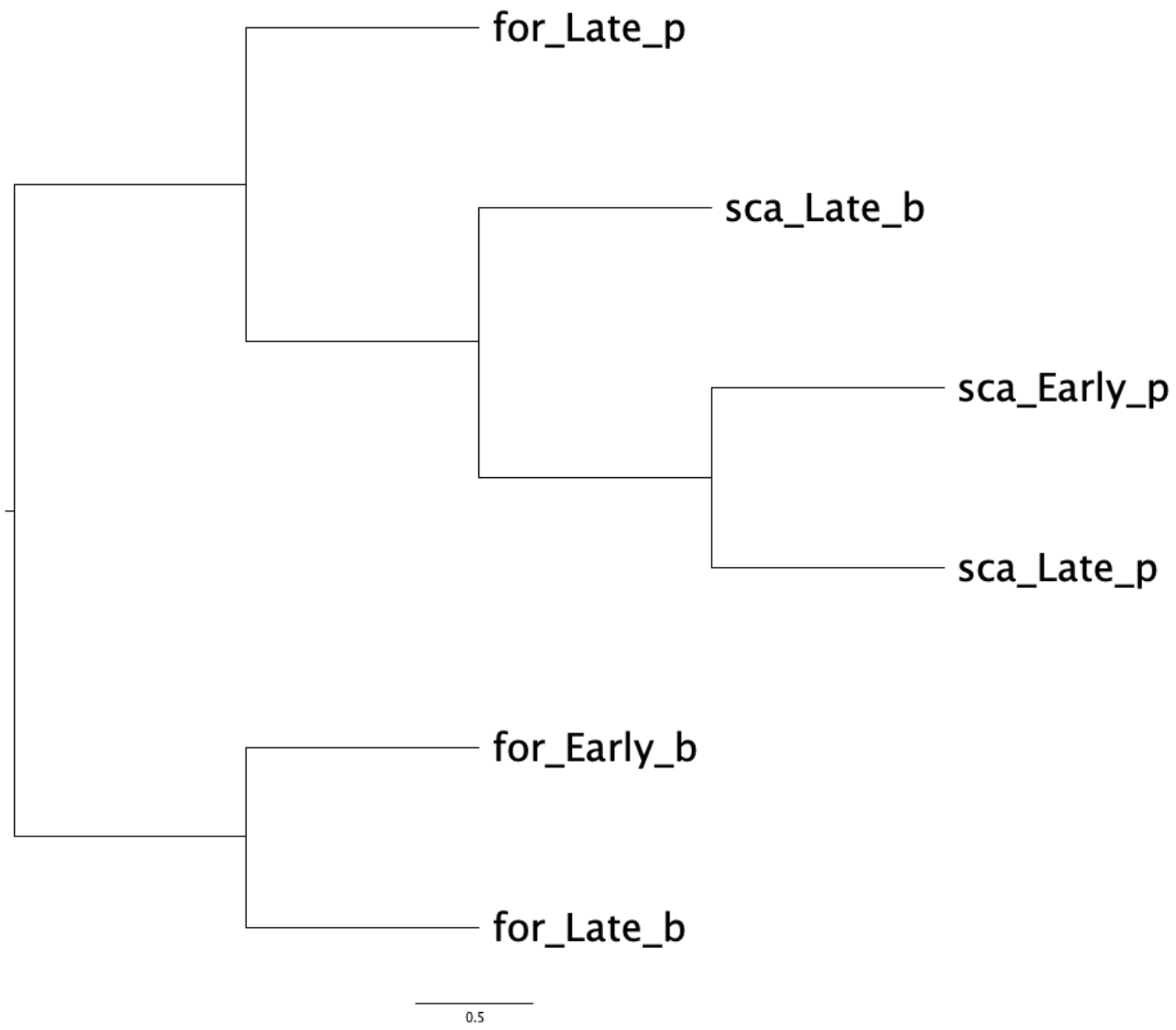


Figure 5: scandens introgressed fortis Late pointed

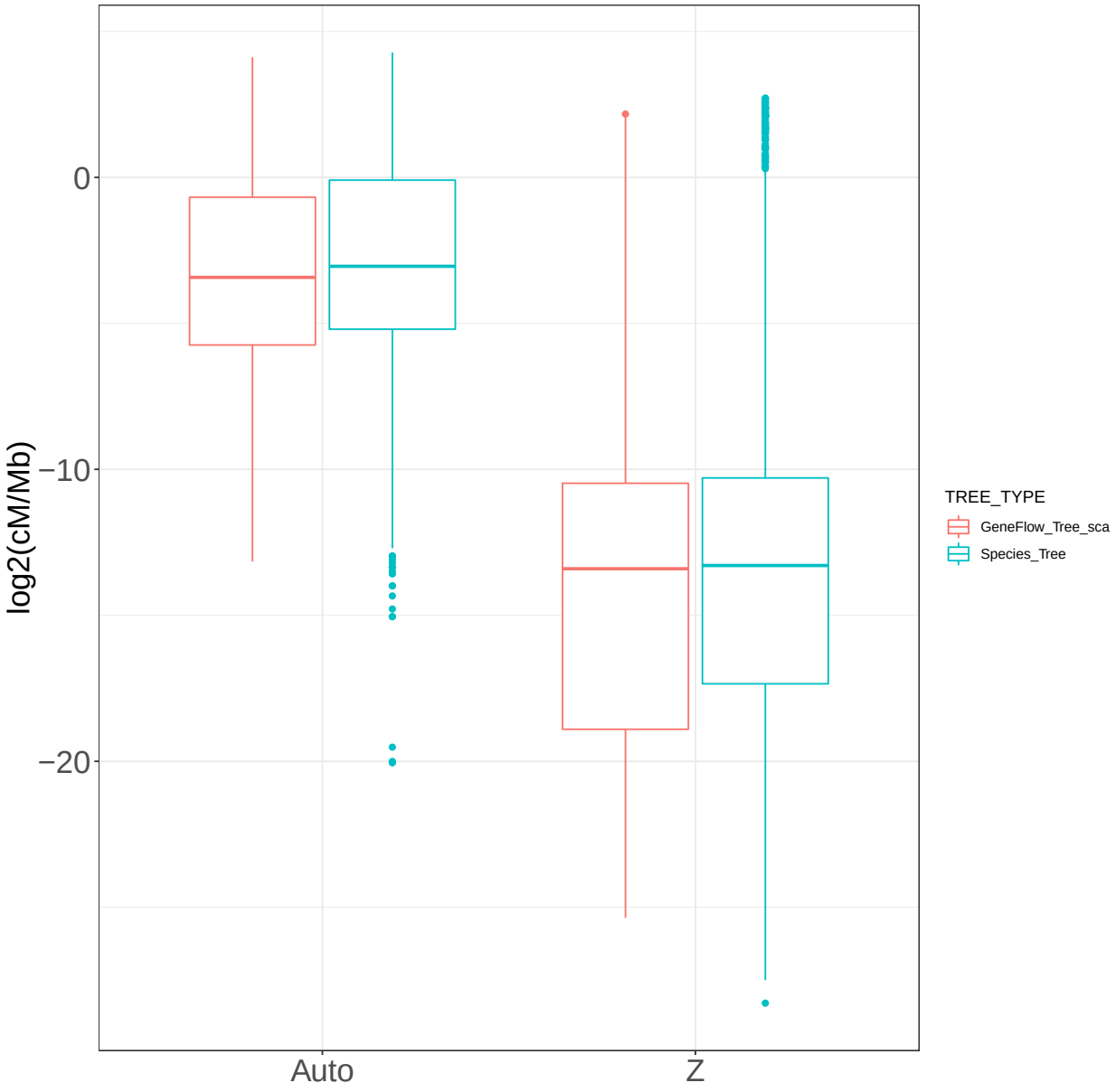
```
##  Wilcoxon rank sum test with continuity correction
##
## data:  autosome.species$RR_zebra and autosome.flow$RR_zebra
## W = 25542000, p-value = 1.719e-13
## alternative hypothesis: true location shift is not equal to 0
```

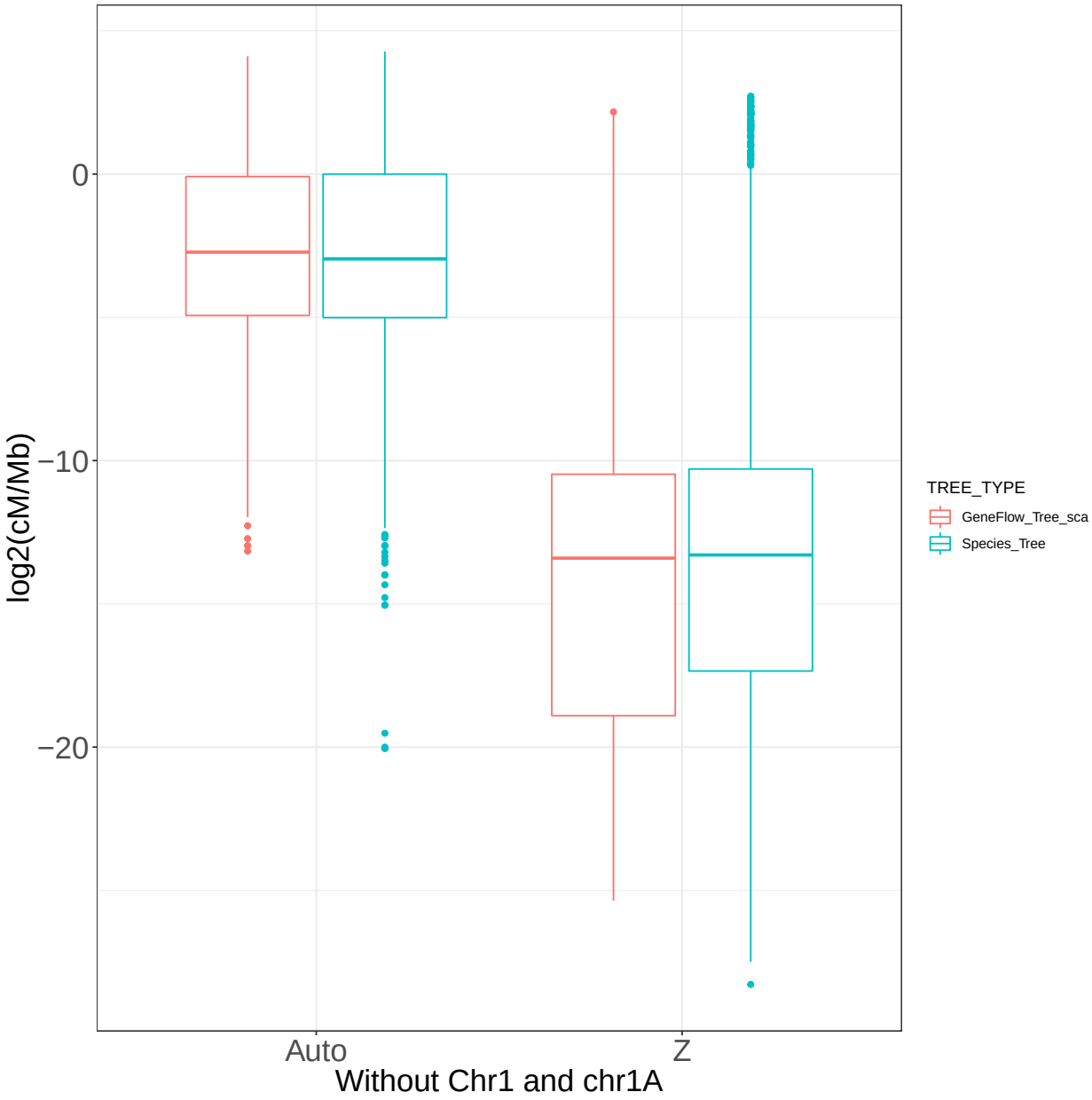
```
##
##  Welch Two Sample t-test
##
## data:  autosome.species[autosome.species$CHR_zebra != "chr1" & autosome.species$CHR_zebra !=  and au
## t = 1.8716, df = 8168.3, p-value = 0.06129
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.003679512  0.159113629
## sample estimates:
## mean of x mean of y
##  1.091190  1.013473
```

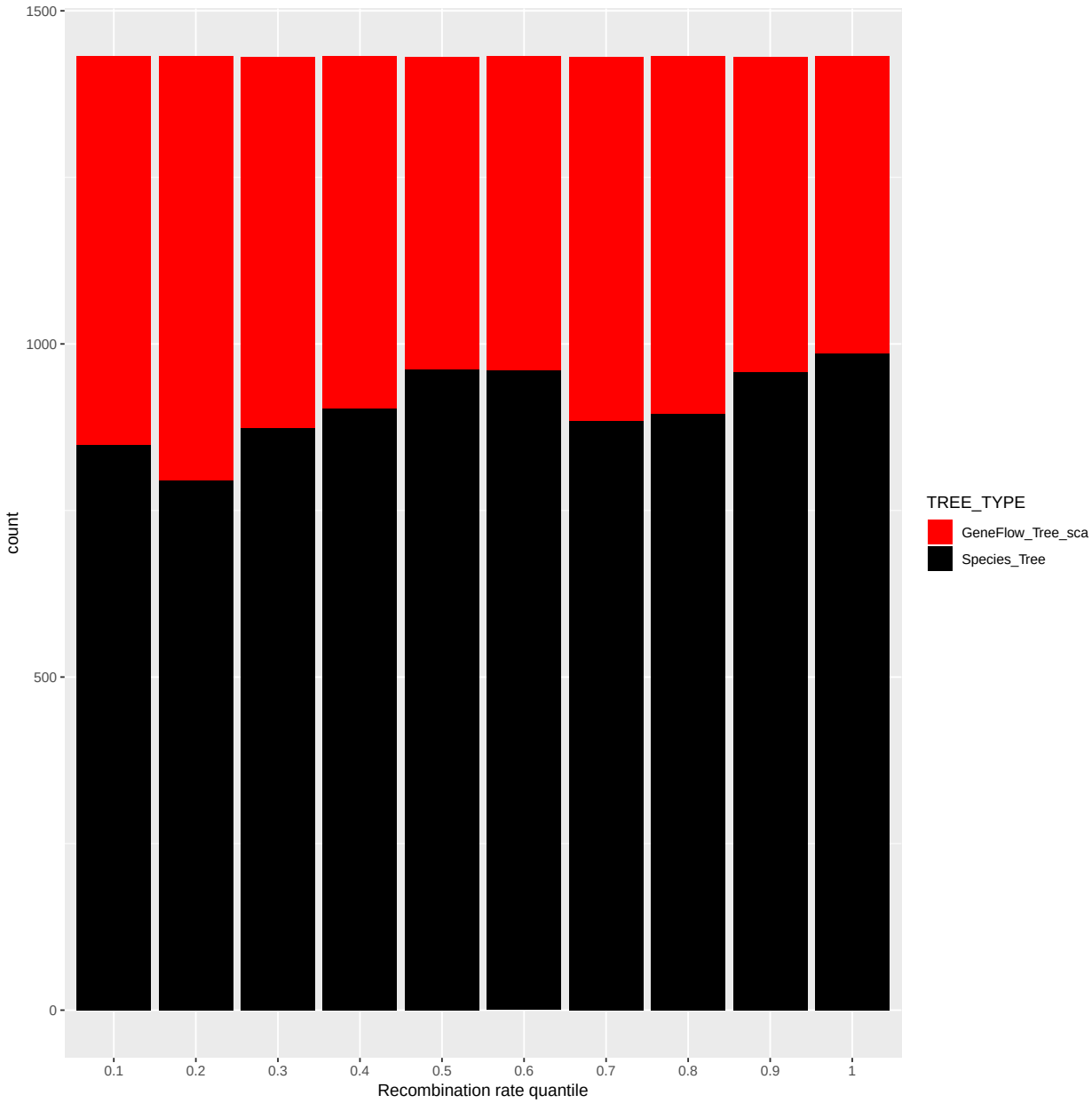
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  autosome.species[autosome.species$CHR_zebra != "chr1" & autosome.species$CHR_zebra !=  and au
## W = 13277000, p-value = 0.2644
## alternative hypothesis: true location shift is not equal to 0
```

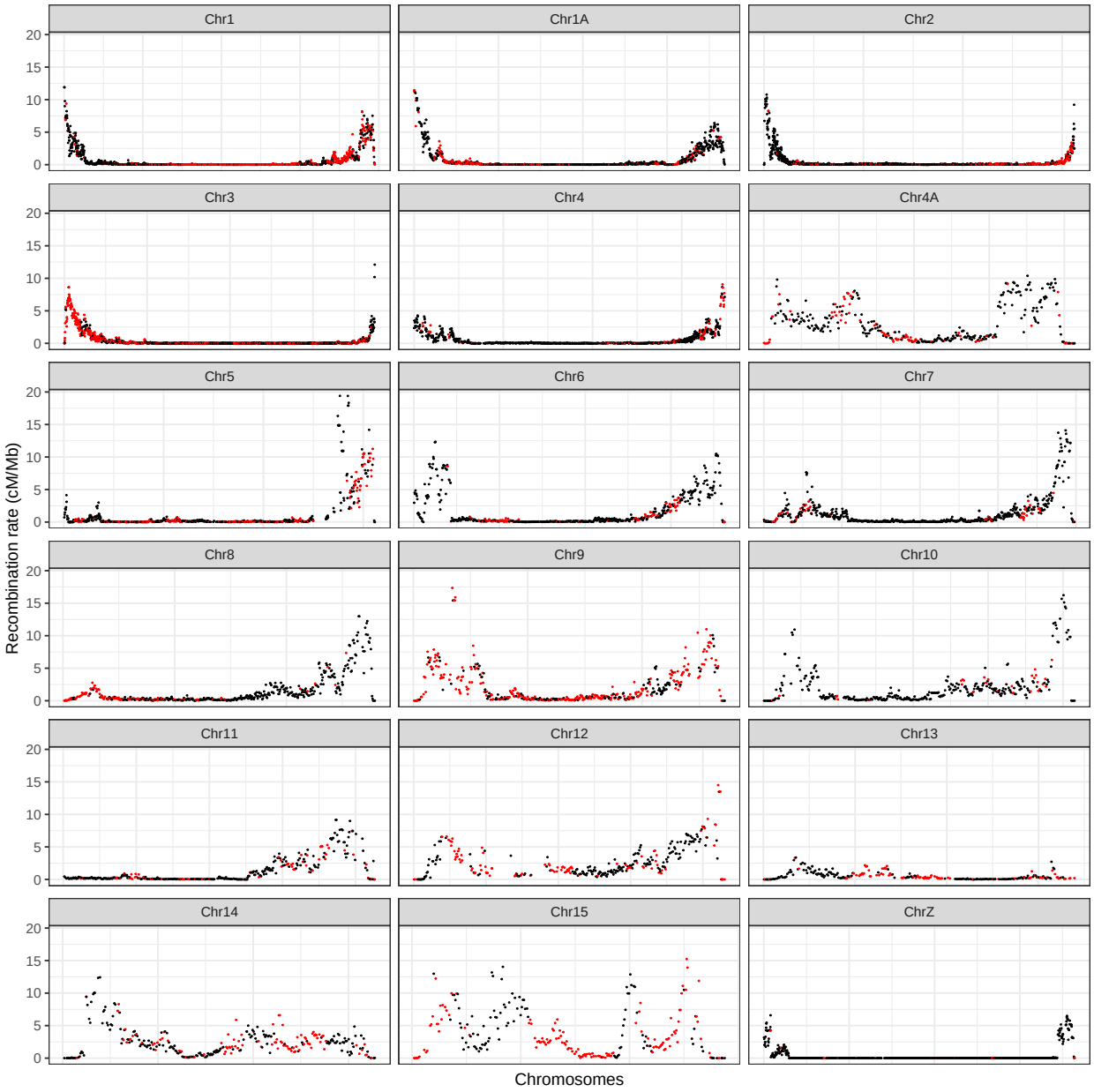
```
##
##  Welch Two Sample t-test
##
## data:  Z.species$RR_zebra and Z.flow$RR_zebra
## t = -0.6865, df = 19.404, p-value = 0.5005
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8544109  0.4319092
## sample estimates:
## mean of x mean of y
## 0.3077909 0.5190418
```

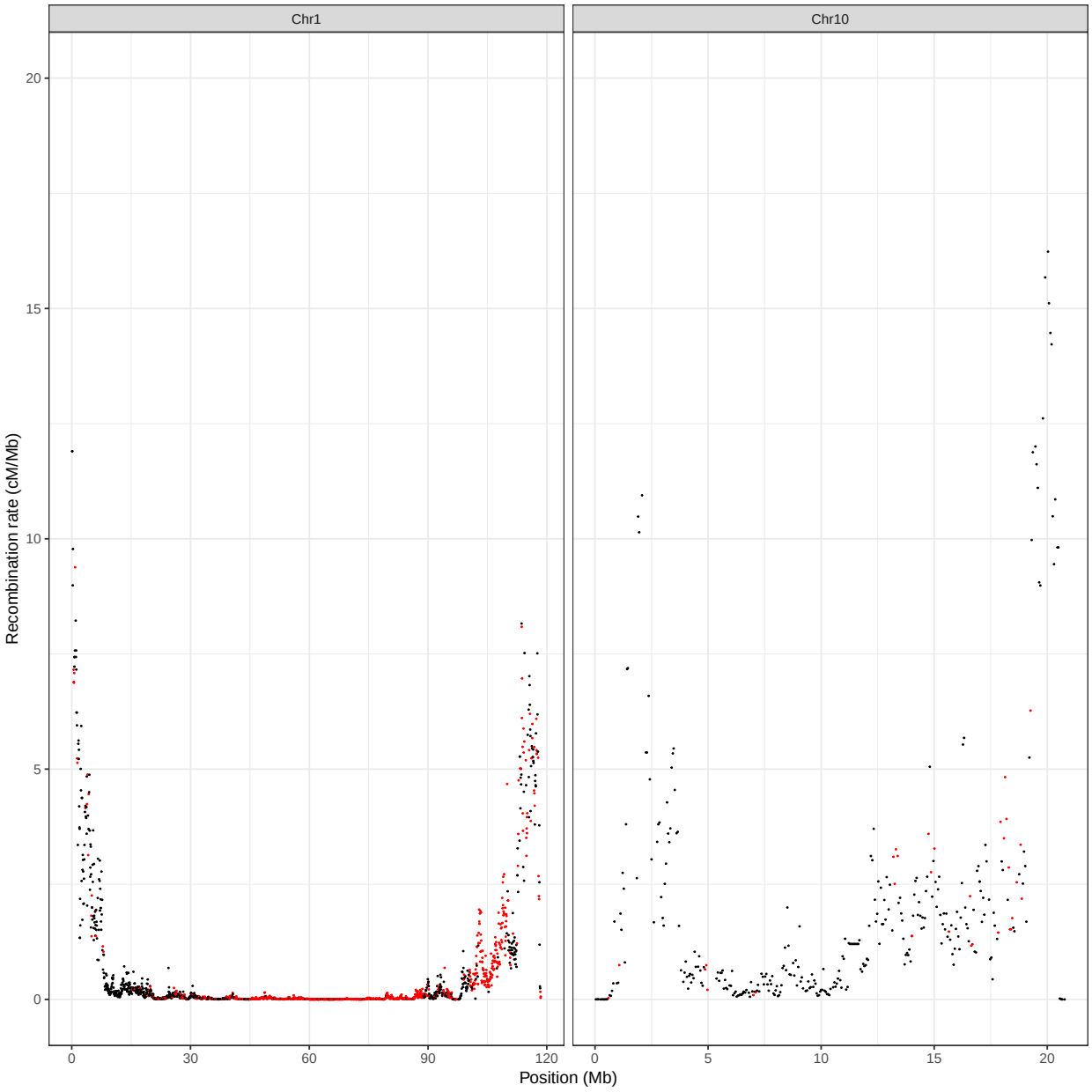
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Z.species$RR_zebra and Z.flow$RR_zebra
## W = 11497, p-value = 0.839
## alternative hypothesis: true location shift is not equal to 0
```


log2(cM/Mb)
0
−10
−20
Auto
Z
TREE_TYPE
GeneFlow_Tree_sca
Species_Tree


0
−10
−20
log2(cM/Mb)
Auto
Z
Without Chr1 and chr1A
TREE_TYPE
GeneFlow_Tree_sca
Species_Tree


1500
1000
500
0
count
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Recombination rate quantile
TREE_TYPE
GeneFlow_Tree_sca
Species_Tree


Chr1
Chr1A
Chr2
Chr3
Chr4
Chr4A
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
ChrZ
Recombination rate (cM/Mb)
Chromosomes



introgression in fortis_Late_pointed

- pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/geneDensity
- These are 66 windows displaying for_Late_pointed introgression, and most of them sit on chr4
- perform GO analysis for these genes
- using the GO database I constructed before: /proj/uppstore2017190/b2012111/private/UserDirectory/fan/phylogenomics/
- TopGO
- Erik suggested to use human GO terms which have better annotation support.

- Download human peptide sequences from ENSEMBL http://www.ensembl.org/biomart/martview/2c334c9e6688c3b744951c664674e003

Correlation between gene flow and gene density

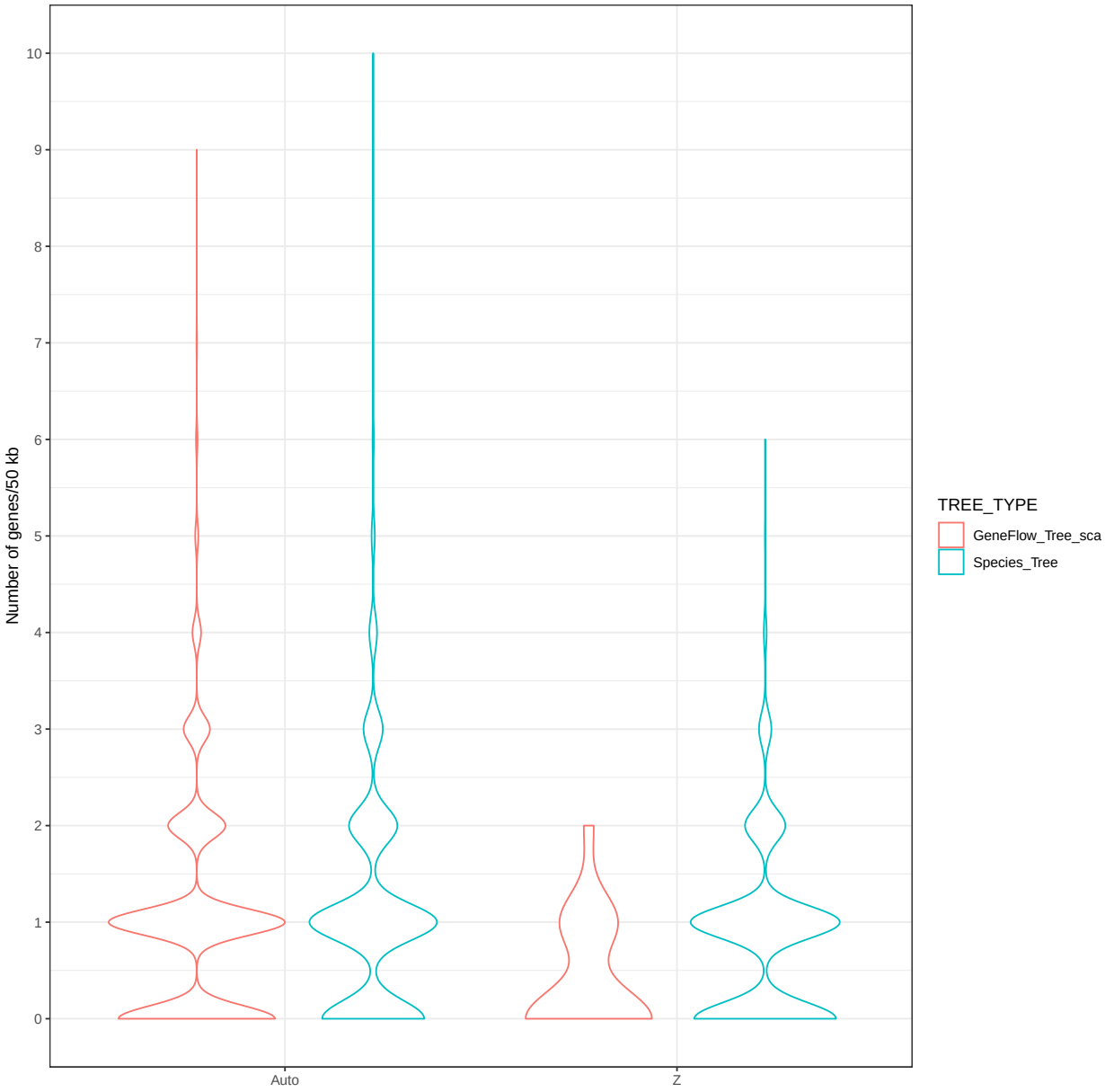
- pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/geneDensity
- Plot the correlation

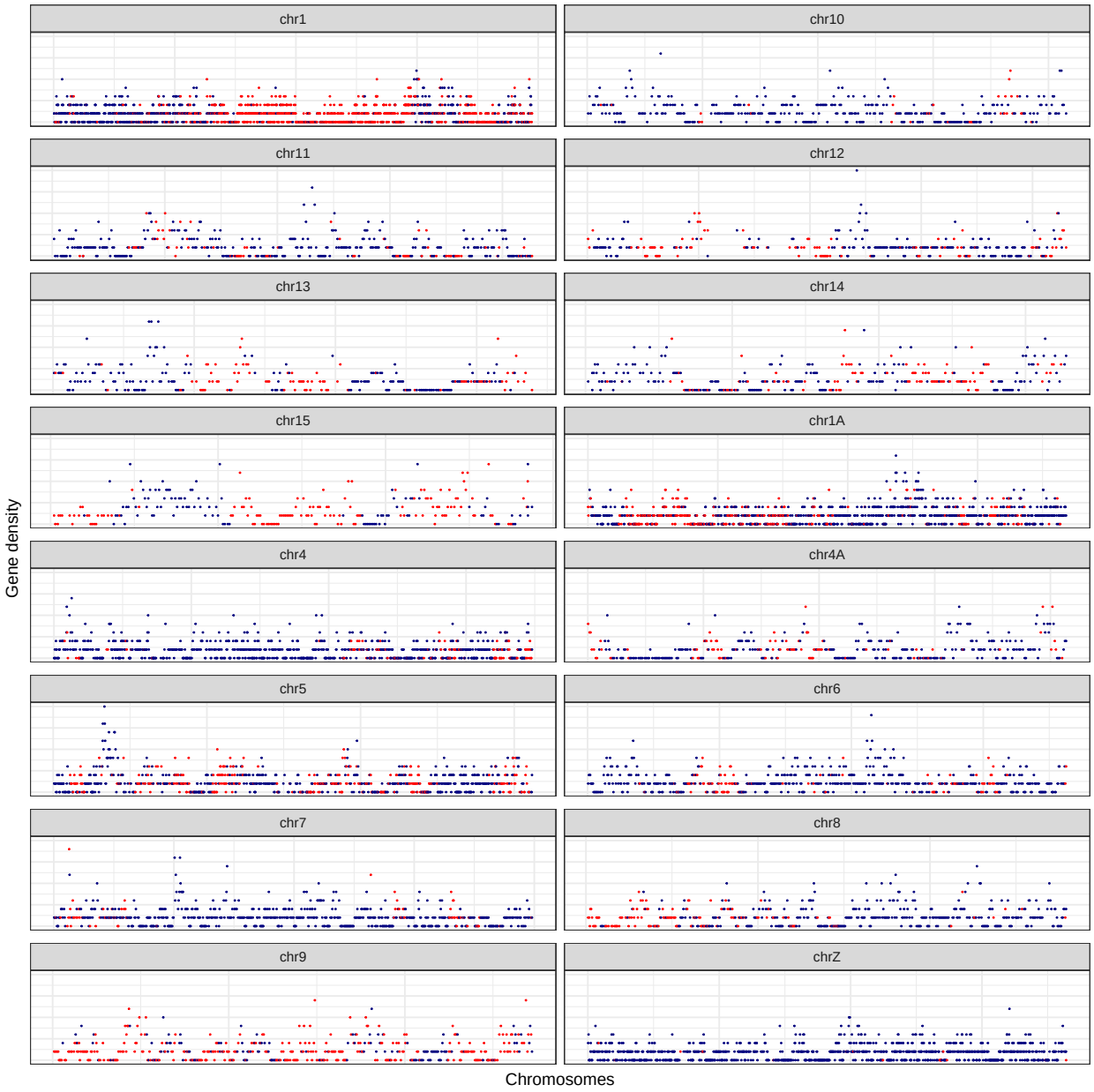
```
##
##  Welch Two Sample t-test
##
## data:  autosome.species$Gene_Density and autosome.flow$Gene_Density
## t = 4.2617, df = 7164.3, p-value = 2.055e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05545001 0.14991184
## sample estimates:
## mean of x mean of y
##  1.112063  1.009383
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  autosome.species$Gene_Density and autosome.flow$Gene_Density
## W = 10907000, p-value = 4.022e-05
## alternative hypothesis: true location shift is not equal to 0
```

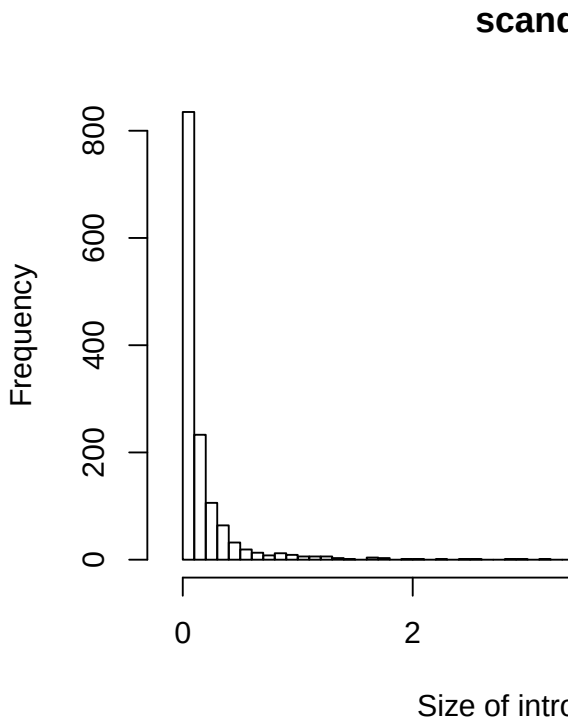
```
##
##  Welch Two Sample t-test
##
## data:  Z.species$Gene_Density and Z.flow$Gene_Density
## t = 3.0467, df = 20.461, p-value = 0.006256
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1313281 0.6990290
## sample estimates:
## mean of x mean of y
## 0.8151786 0.4000000
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Z.species$Gene_Density and Z.flow$Gene_Density
## W = 14242, p-value = 0.02385
## alternative hypothesis: true location shift is not equal to 0
```

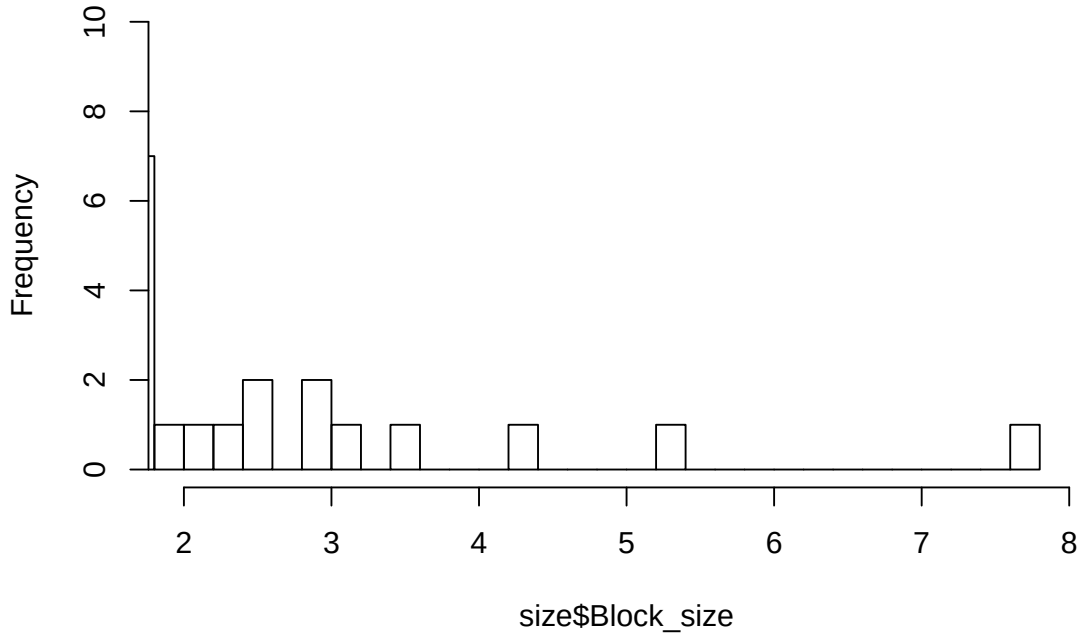

Number of genes/50 kb
10
9
8
7
6
5
4
3
2
1
0
Auto
Z
TREE_TYPE
GeneFlow_Tree_sca
Species_Tree

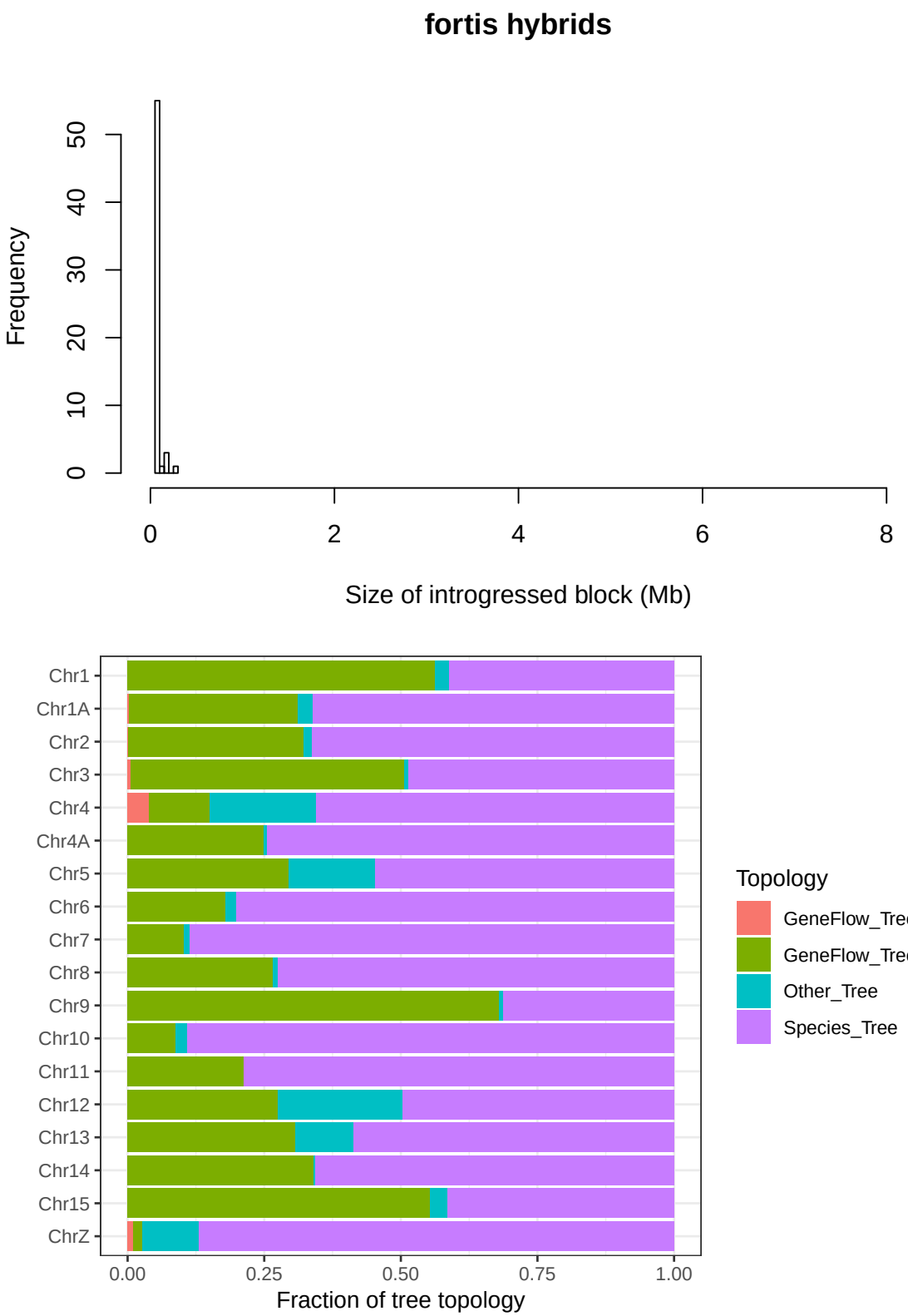


* summarize the fraction of tree toplologies on each chromosomes and calculate the size of introgressed blocks *
pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/TreeScan/Rooted/fraction_blocks
* use darwin's finch's coordinates because lifeover doesn't necessarily map the windows in 50kb intervals



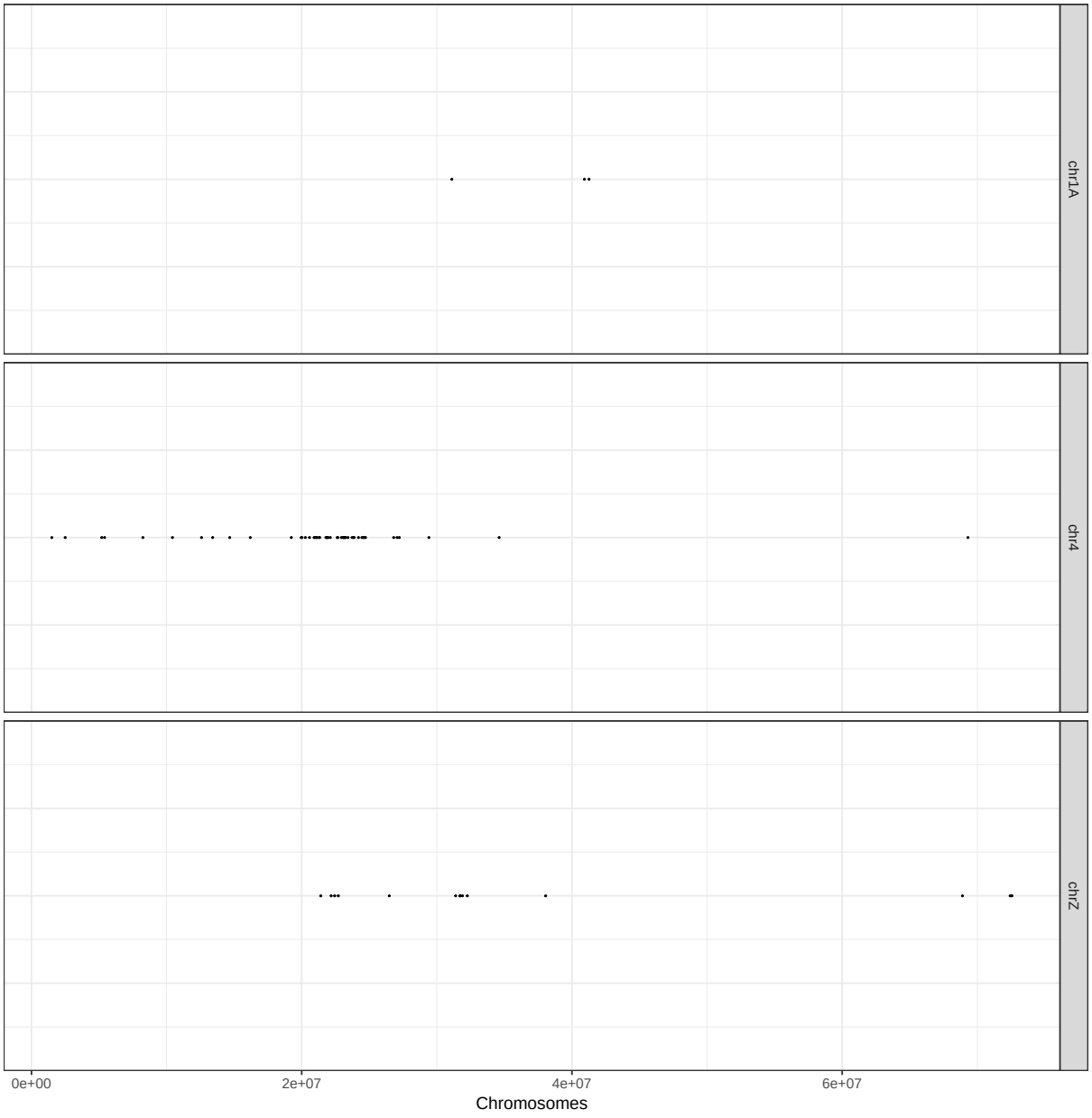
- plot fraction of each tree and size of introgressed blocks per chromosome





- Introgressed regions in fortis hybrids

```
## [1] 66
```



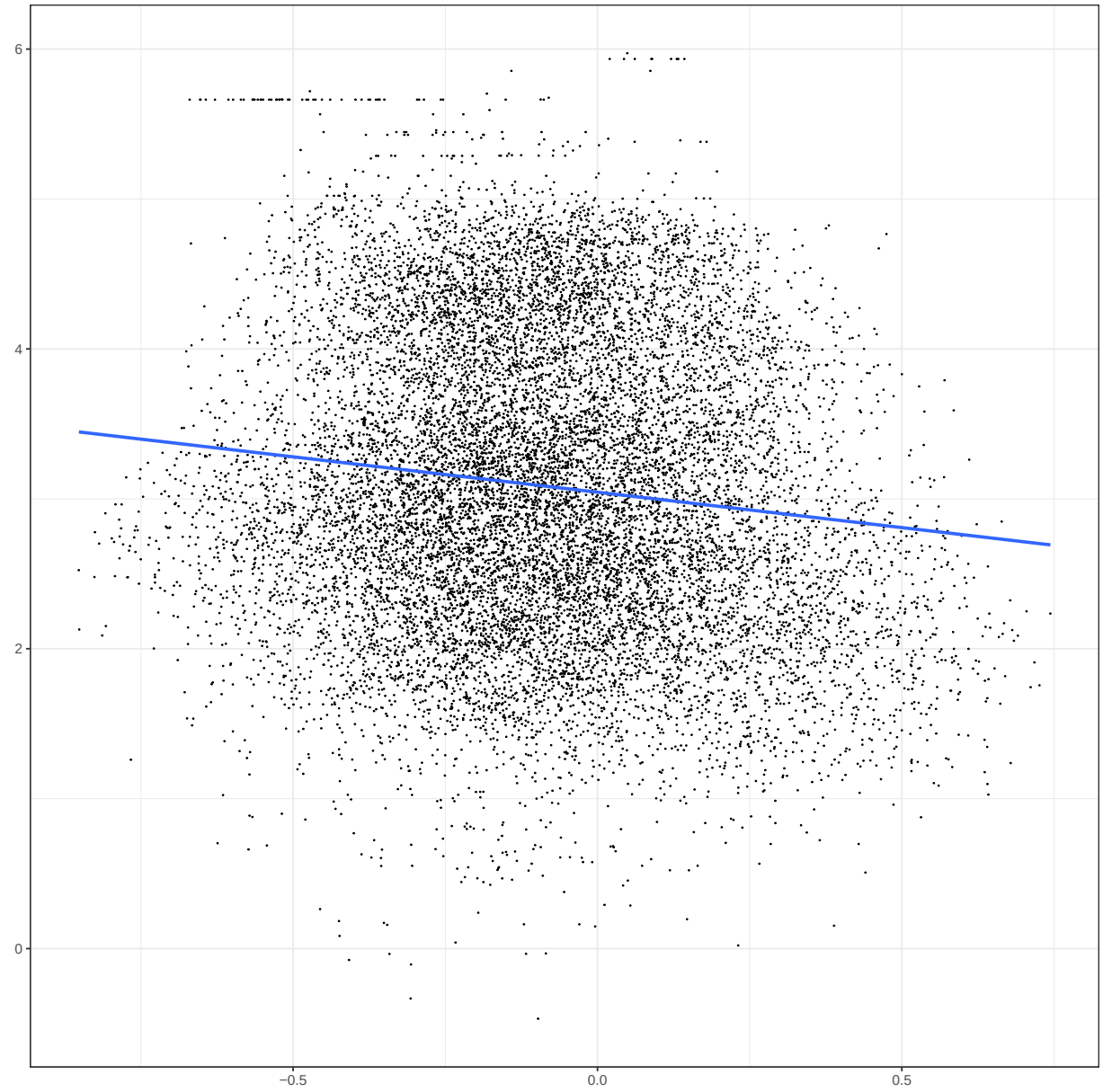
FST ratio

- Matt suggested to calculate the FST between fortis_Early/Late_blunt and fortis_Late_pointed and FST between fortis_Late_pointed and scandens_Early/Late_pointed, and then caluclate the rotio of the FSTs.

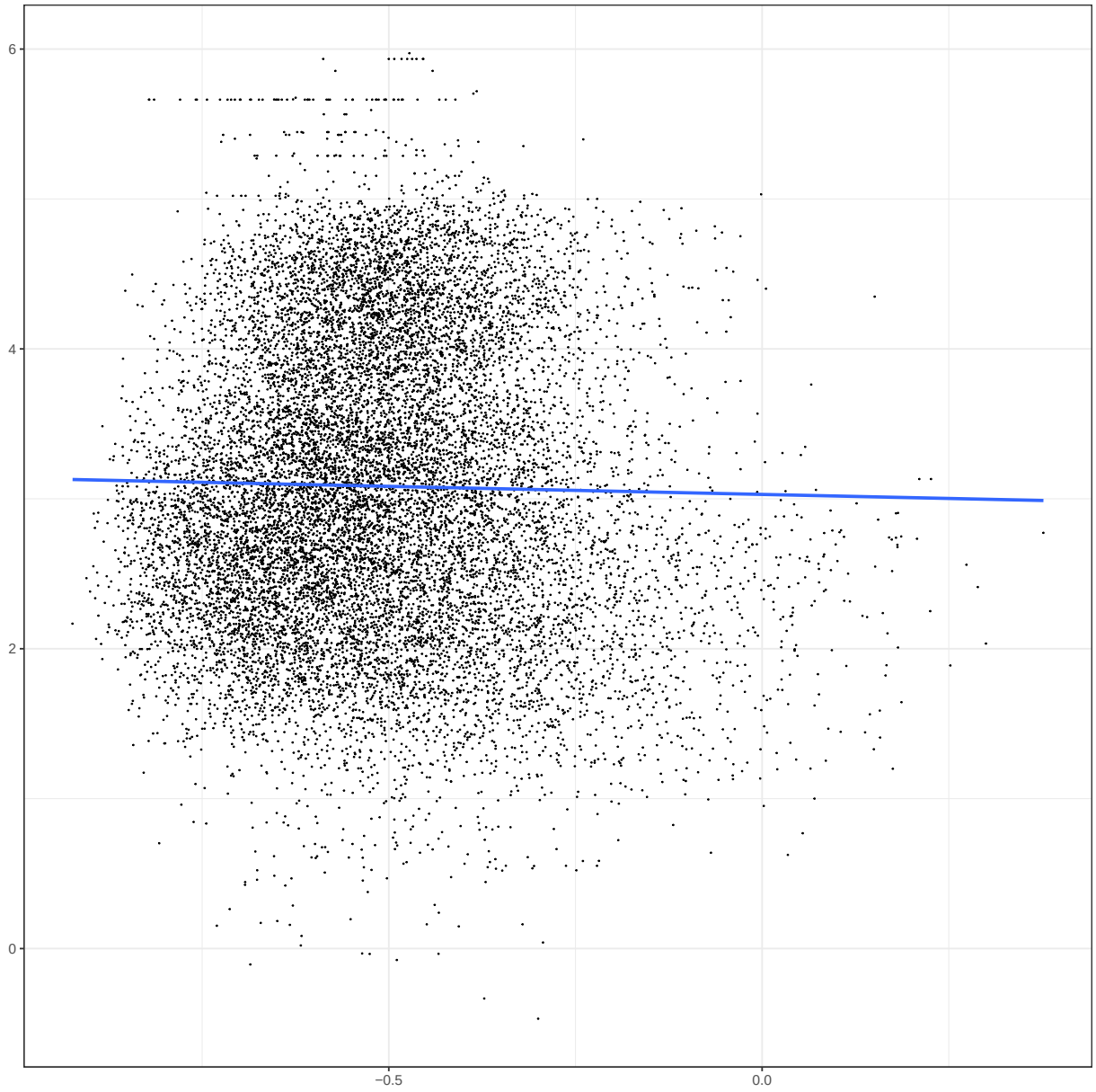
- Measure pooled FST with Popoolation2
- "We furthermore recommend that the pool size (number of individuals in the pool) should be larger than the coverage. This minimizes re-sampling the same allele from a single individual several times."

- pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/FST_ratio
- Plot for all pairwise ZFST
- Plot for all pairwise ZFST on zebrafinch genome
- Ratio of FST. If there is no introgression between fortis_Late_p and sca_Late_p, the ratio of FST(fortis_Early_b.fortis_Late_p):FST(fortis_Late_p:sca_Early_p) should be uniform along the genome. Otherwise, the ratio should be elevated because fortis_Late_p and sca_Late_p are very similar to each other.

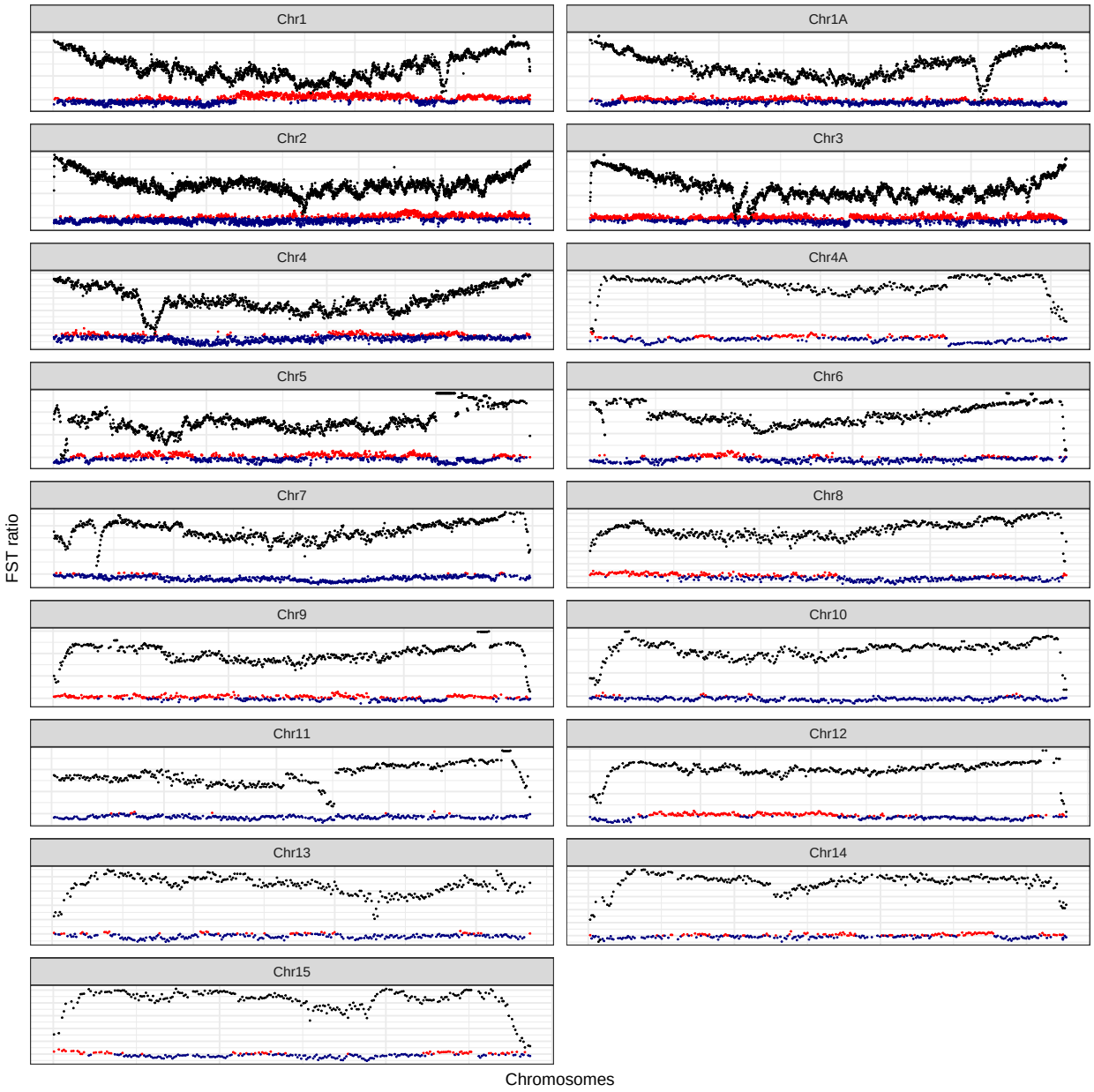
```
## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --

## v tibble  2.0.1     v purrr   0.3.0
## v tidyr   0.8.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Call:
## lm(formula = log10(RR) ~ scandensHybrid, data = FST.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2738 -0.7327 -0.0904  0.7855  2.9576
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.043542   0.008543  356.26   <2e-16 ***
## scandensHybrid -0.469412   0.031782  -14.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.988 on 14887 degrees of freedom
## Multiple R-squared:  0.01444,    Adjusted R-squared:  0.01438
## F-statistic: 218.1 on 1 and 14887 DF,  p-value: < 2.2e-16
```

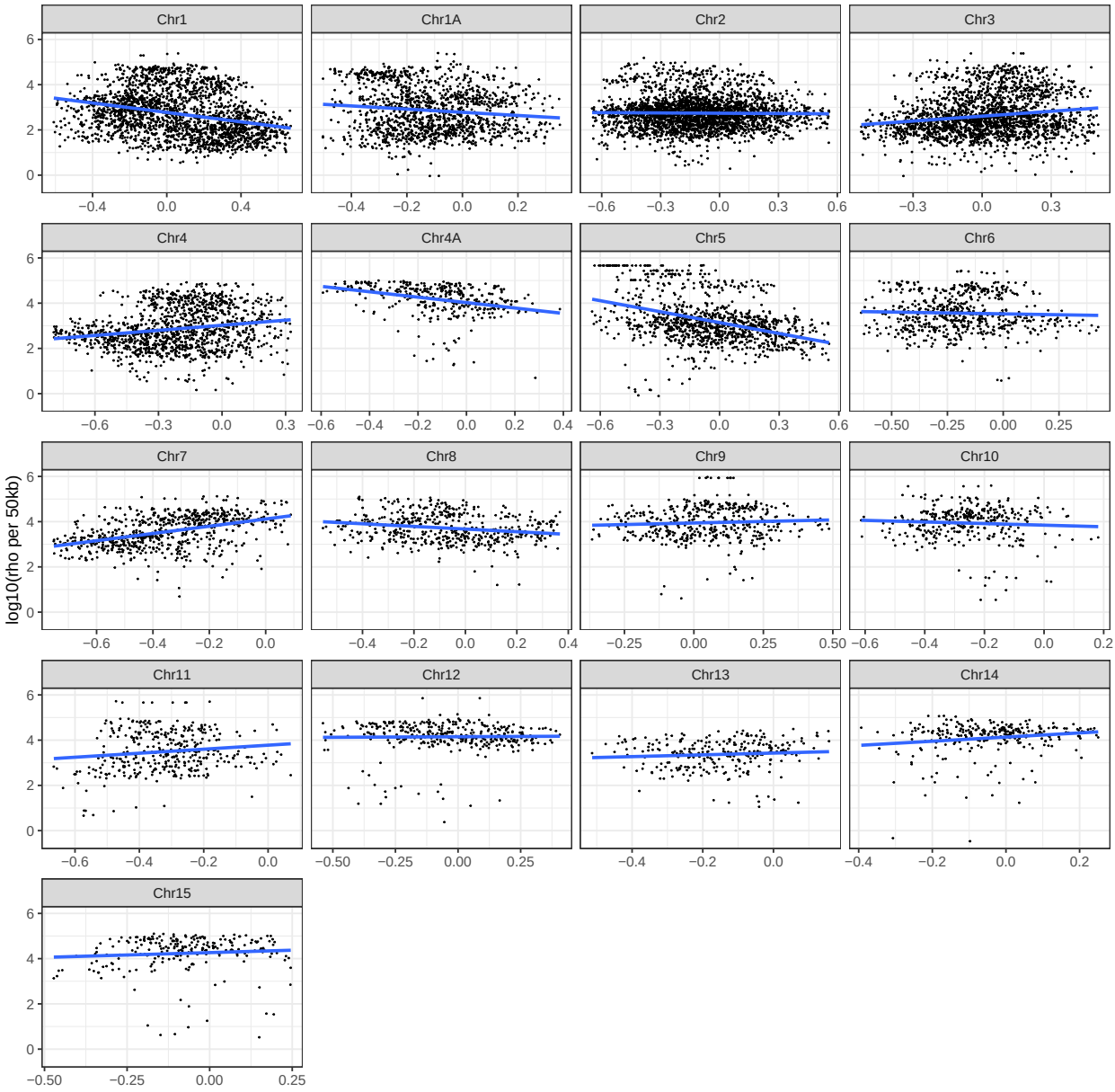



```
##
##  Pearson's product-moment correlation
##
## data:  FST.df$scandensHybrid and FST.df$RR
## t = -9.847, df = 14885, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09638826 -0.06446834
## sample estimates:
##         cor
## -0.08044893
```




```
##
##  Pearson's product-moment correlation
##
## data:  FST.df$fortisHybrid and FST.df$RR
## t = -0.04183, df = 14885, p-value = 0.9666
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01640666  0.01572113
## sample estimates:
##           cor
## -0.0003428534
```


Chr1
Chr1A
Chr2
Chr3
Chr4
Chr4A
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
FST ratio
Chromosomes



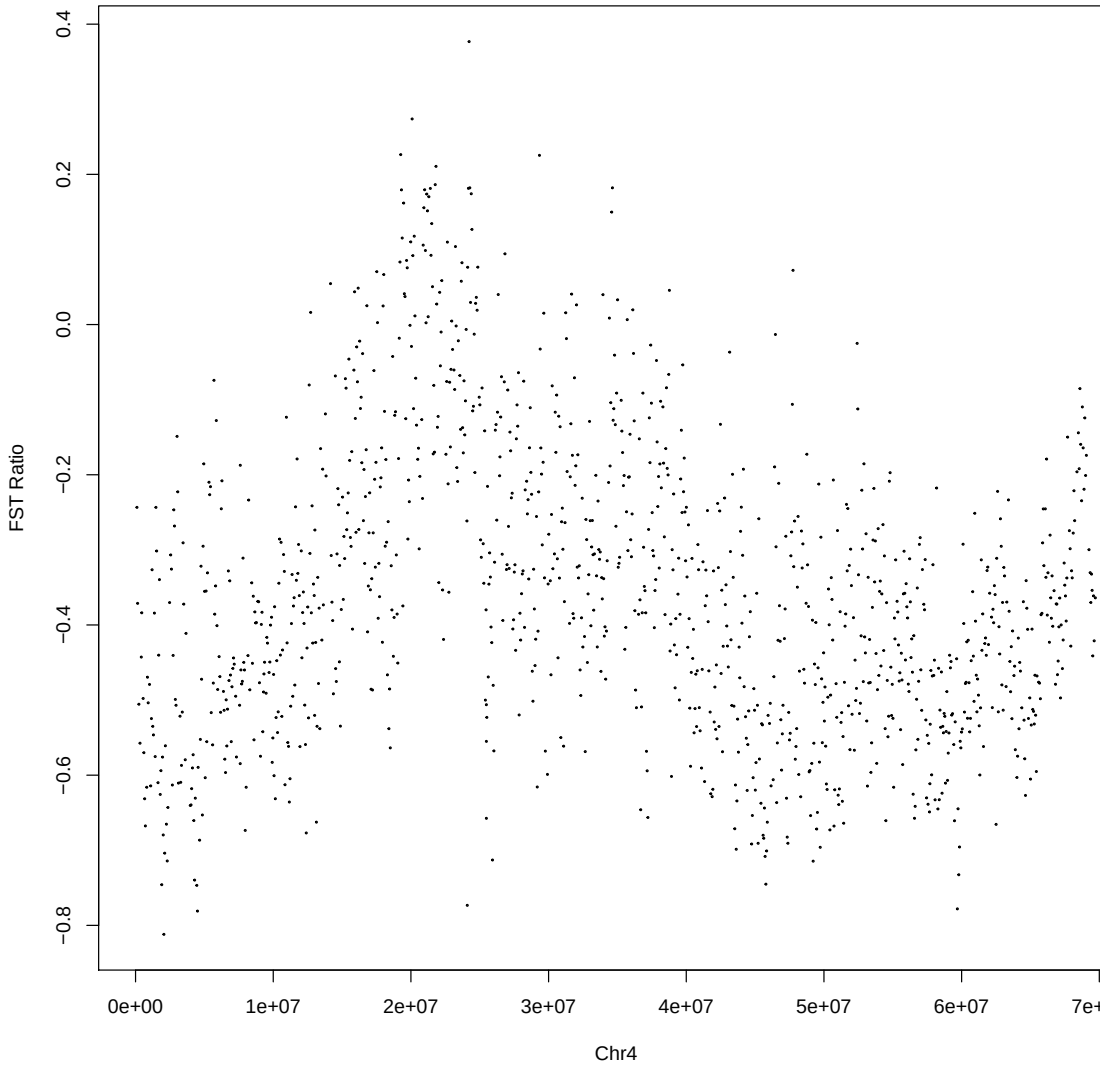

```
## [1] "Chr1"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = -13.919, df = 2014, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3355413 -0.2558763
## sample estimates:
##        cor
## -0.2962239
##
## [1] "Chr1A"
```

```
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = -4.3351, df = 1265, p-value = 1.572e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17489821 -0.06636075
## sample estimates:
##        cor
## -0.1209911
##
## [1] "Chr2"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = -0.79657, df = 2739, p-value = 0.4258
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.05262813  0.02223326
## sample estimates:
##         cor
## -0.01521876
##
## [1] "Chr3"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = 7.6057, df = 2052, p-value = 4.286e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1232141 0.2073484
## sample estimates:
##       cor
## 0.1655825
##
## [1] "Chr4"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = 6.9328, df = 1205, p-value = 6.711e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1409811 0.2495162
## sample estimates:
##       cor
## 0.1958484
##
## [1] "Chr4A"
##
##  Pearson's product-moment correlation
```

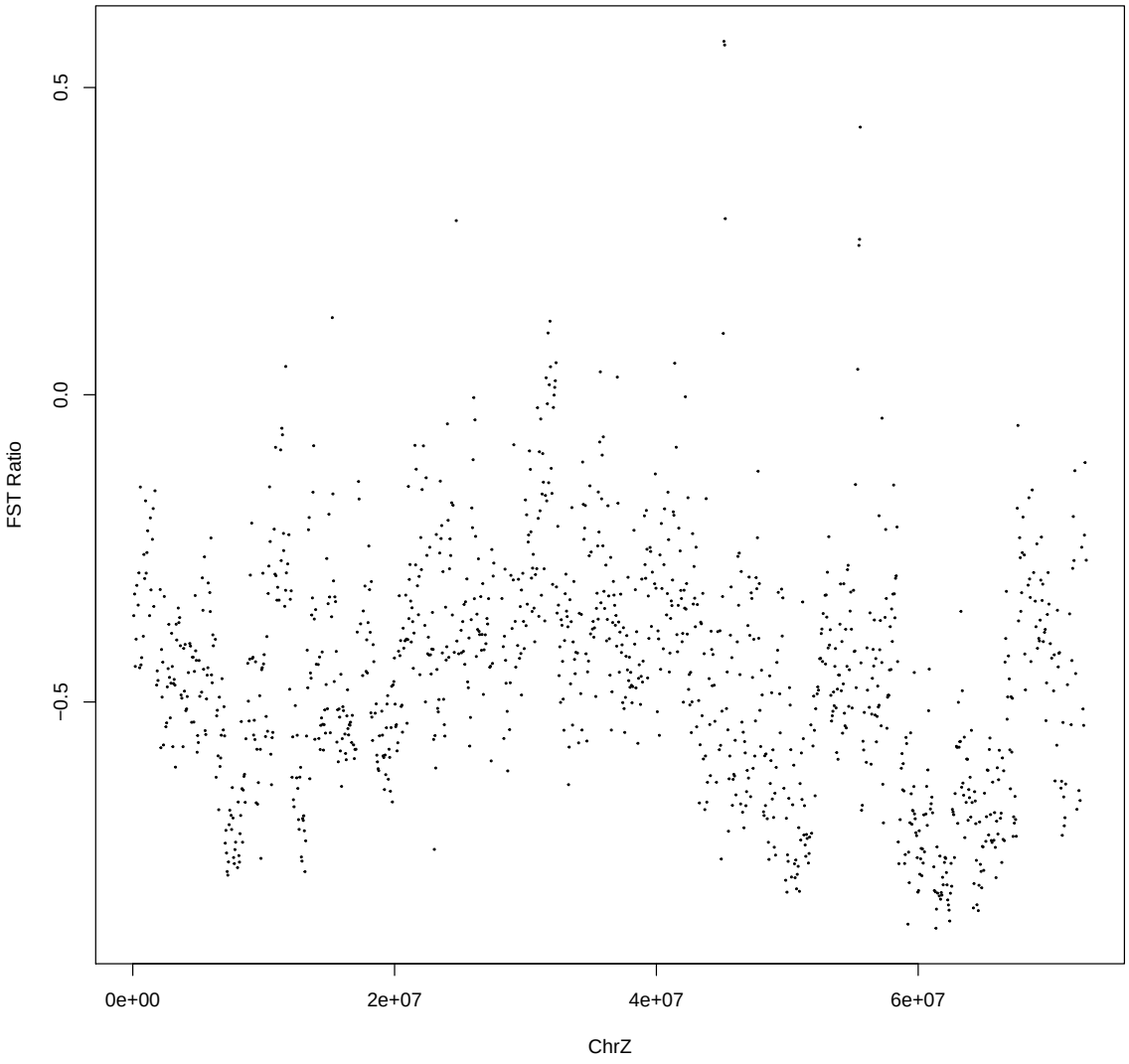
```
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = -6.4042, df = 340, p-value = 5.028e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4195440 -0.2300452
## sample estimates:
##       cor
## -0.3280913
##
## [1] "Chr5"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = -13.866, df = 1056, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4422630 -0.3402329
## sample estimates:
##       cor
## -0.3924546
##
## [1] "Chr6"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = -0.9833, df = 592, p-value = 0.3259
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1204372  0.0401980
## sample estimates:
##       cor
## -0.0403805
##
## [1] "Chr7"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = 12.515, df = 672, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3714475 0.4940684
## sample estimates:
##      cor
## 0.434771
##
## [1] "Chr8"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
```

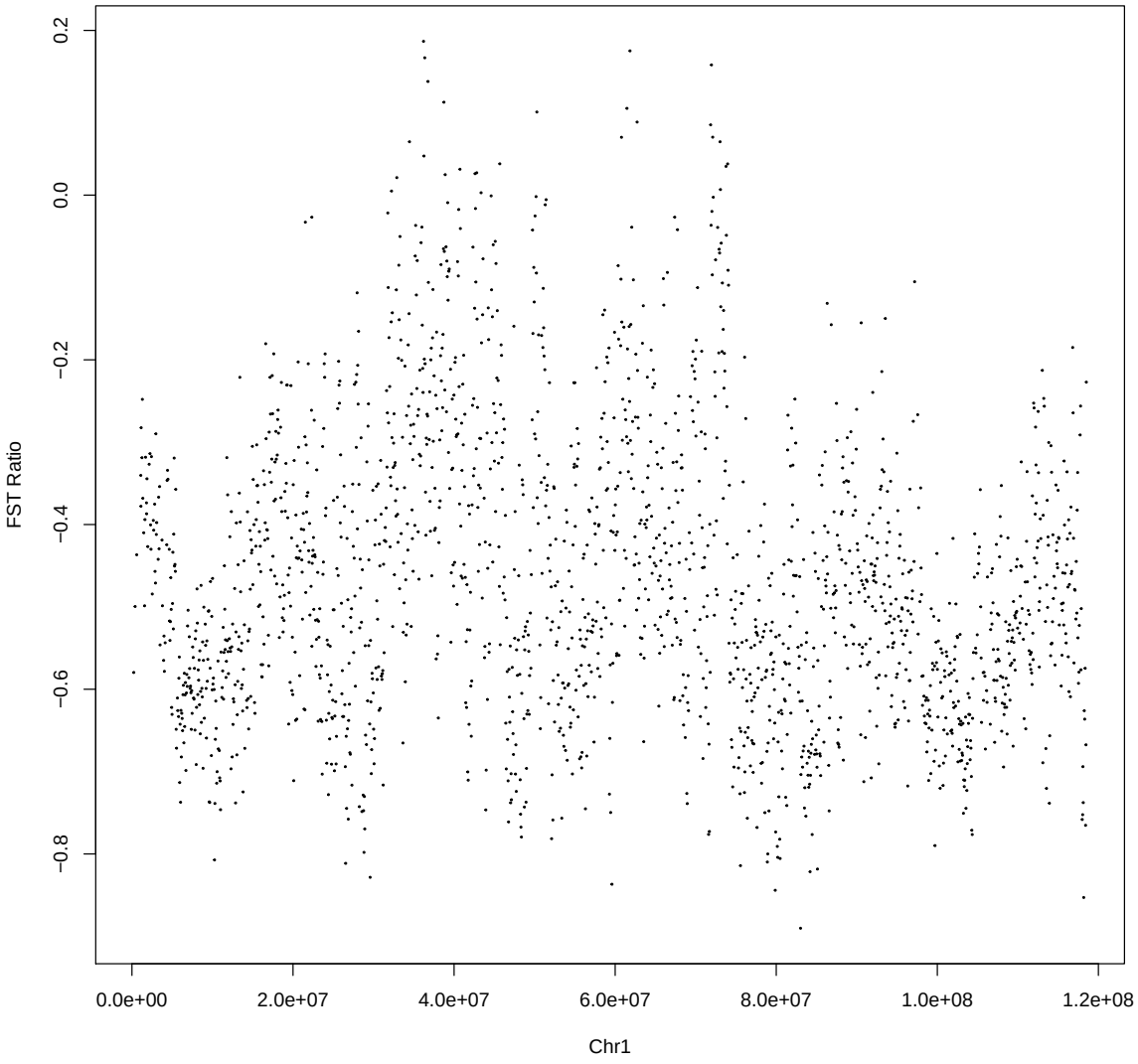
```
## t = -4.4, df = 478, p-value = 1.336e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2818212 -0.1097345
## sample estimates:
##        cor
## -0.1972972
##
## [1] "Chr9"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = 1.2781, df = 429, p-value = 0.2019
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03305605  0.15514568
## sample estimates:
##        cor
## 0.06159227
##
## [1] "Chr10"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = -1.2224, df = 354, p-value = 0.2224
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1676460  0.0393728
## sample estimates:
##         cor
## -0.06483412
##
## [1] "Chr11"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = 2.5768, df = 355, p-value = 0.01037
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03215983 0.23598000
## sample estimates:
##       cor
## 0.1355032
##
## [1] "Chr12"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = 0.28782, df = 356, p-value = 0.7736
## alternative hypothesis: true correlation is not equal to 0
```

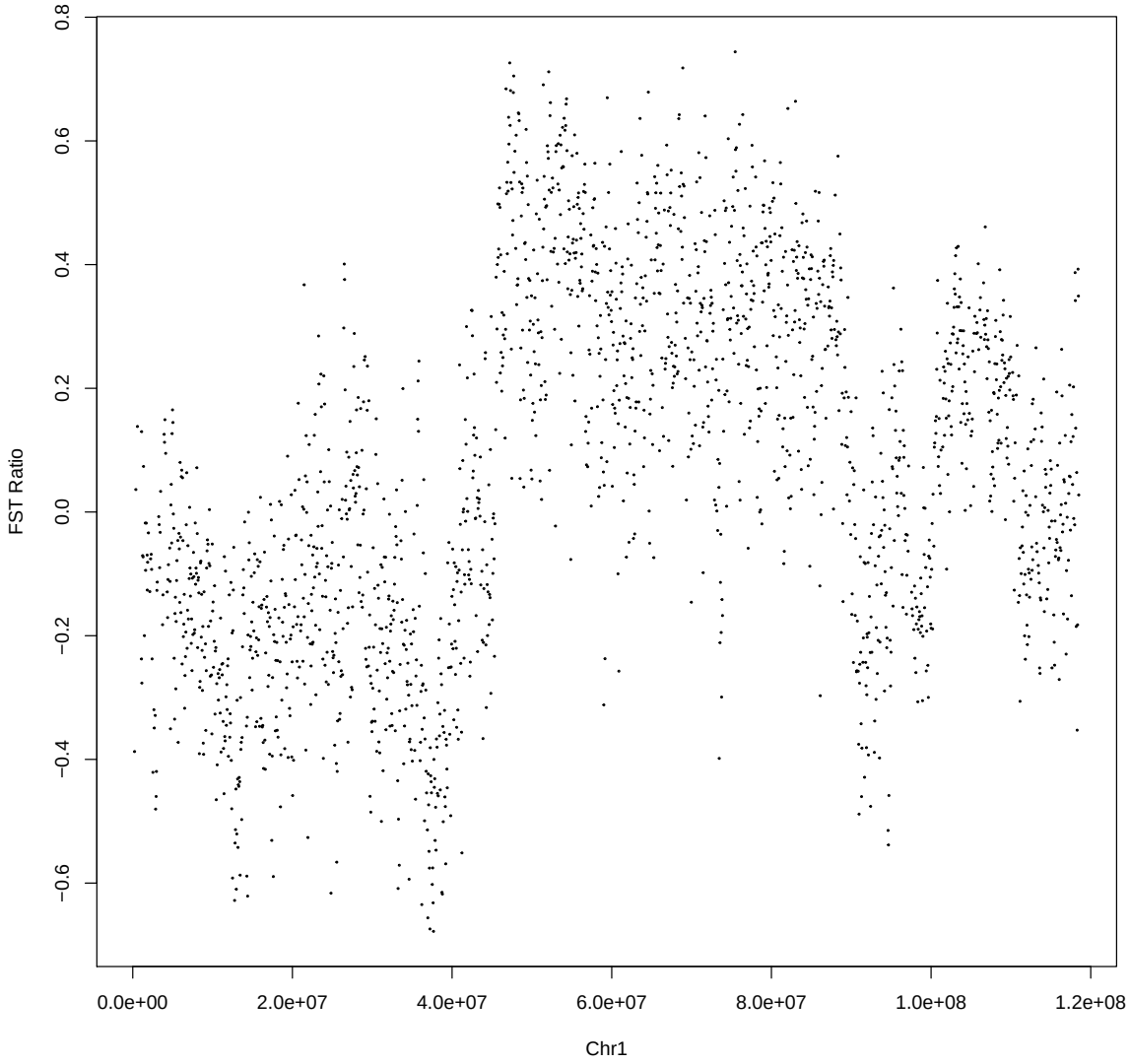
```
## 95 percent confidence interval:
##  -0.08853761  0.11871566
## sample estimates:
##        cor
## 0.01525285
##
## [1] "Chr13"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = 1.4287, df = 273, p-value = 0.1542
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03247031  0.20236644
## sample estimates:
##       cor
## 0.0861445
##
## [1] "Chr14"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = 2.7124, df = 275, p-value = 0.0071
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04440997 0.27406225
## sample estimates:
##       cor
## 0.1614206
##
## [1] "Chr15"
##
##  Pearson's product-moment correlation
##
## data:  tmp$scandensHybrid and log10(tmp$RR)
## t = 1.2282, df = 232, p-value = 0.2206
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04836959  0.20649262
## sample estimates:
##        cor
## 0.08037505
```

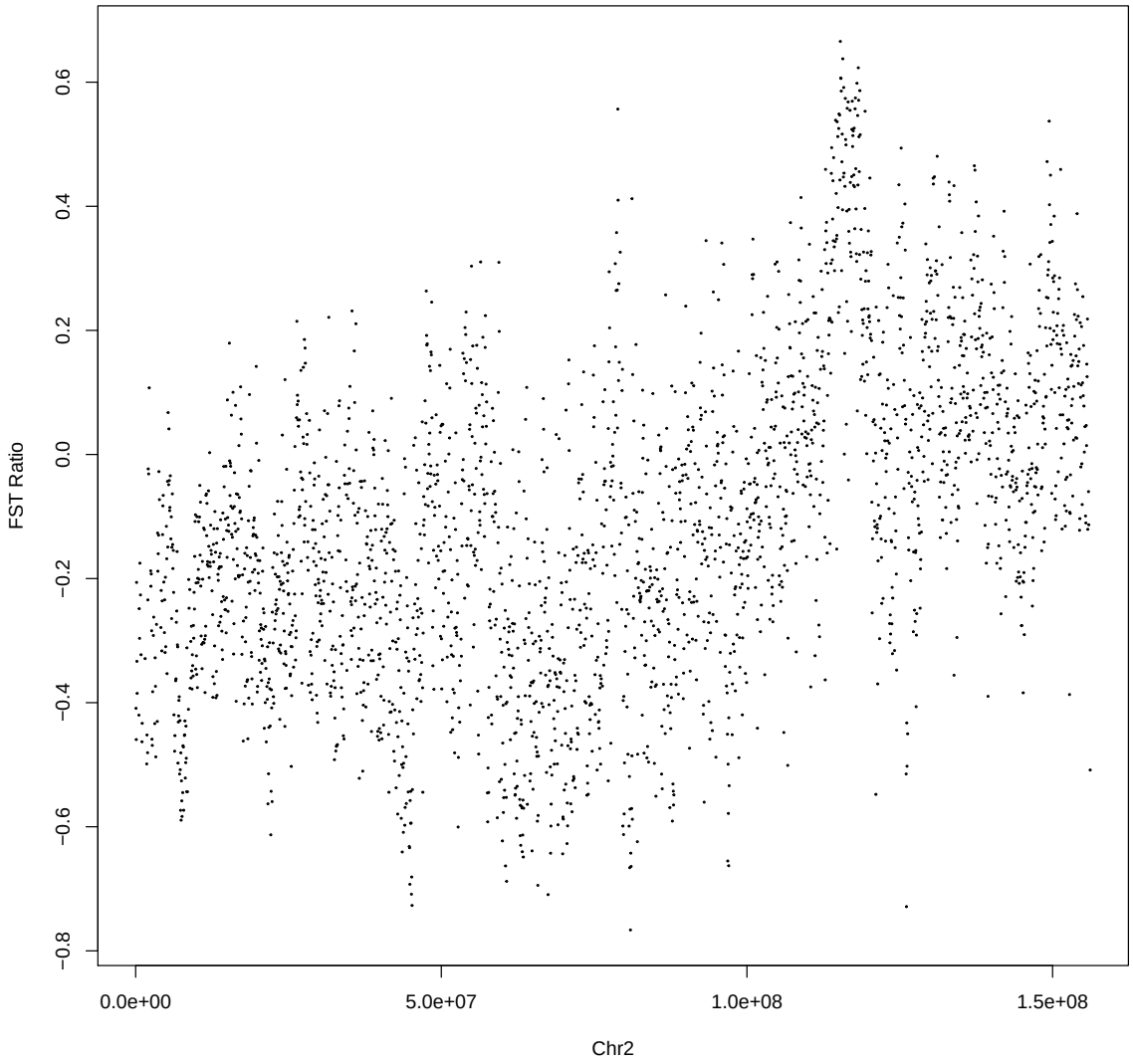


- outliers in the FST ratio


FST Ratio
0.5
0.0
−0.5
0e+00
2e+07
4e+07
6e+07
ChrZ


FST Ratio
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
Chr1


FST Ratio
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
Chr1



- Correlate regions under gene flow with recombination rate in zebra finch
- pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/zebra_RR
- Data is from: http://science.sciencemag.org/content/350/6263/928

Sex estimate

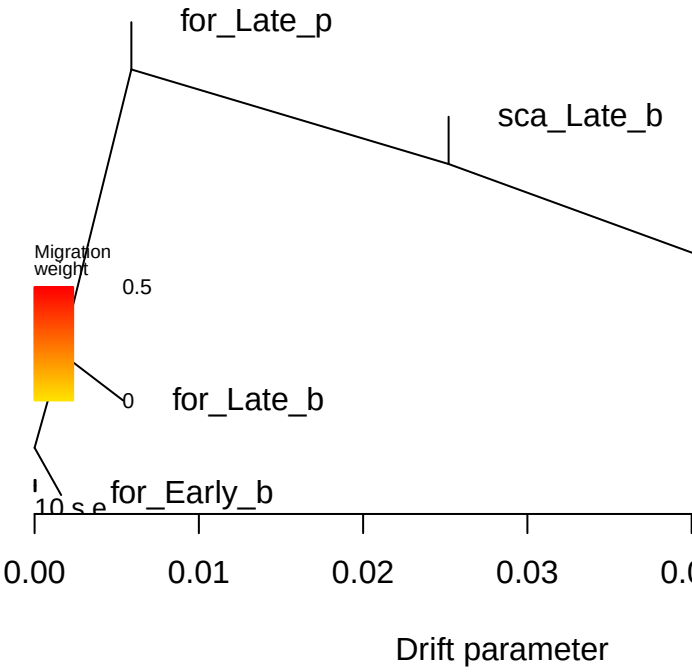
- To determine if there is sex bias in the pools, I checked the depth ratio of Z to autosomes
- pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/Sex_ratio

TreeMix analysis

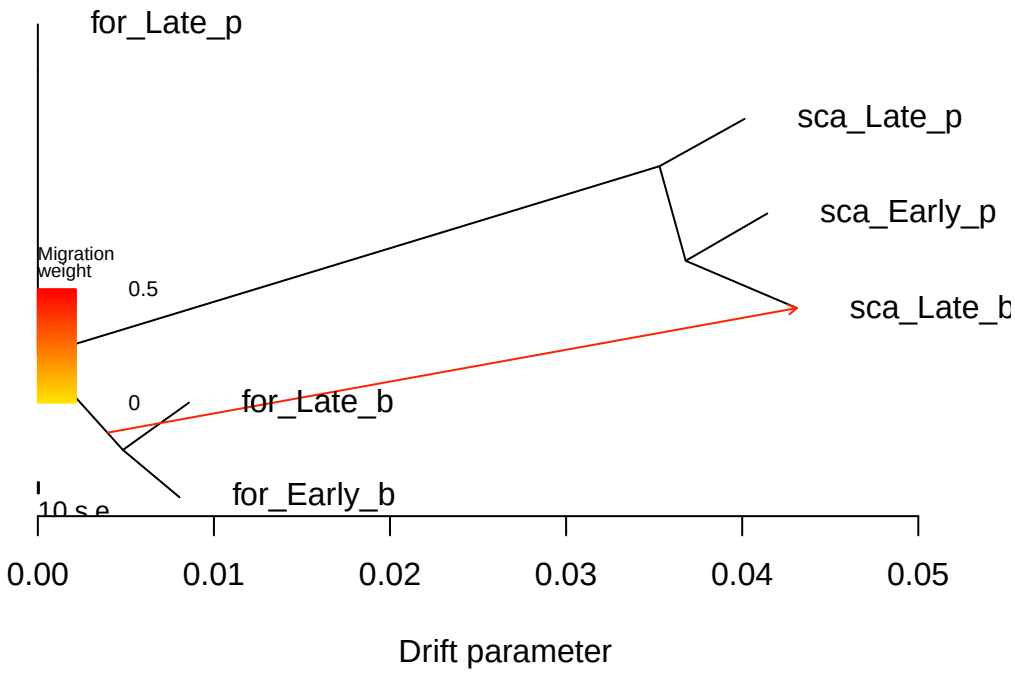
Use TreeMix to infer the pattern of how fortis and scandens populations hybridized
No missing genotype is allowed otherwise TrrMix will throw out errors

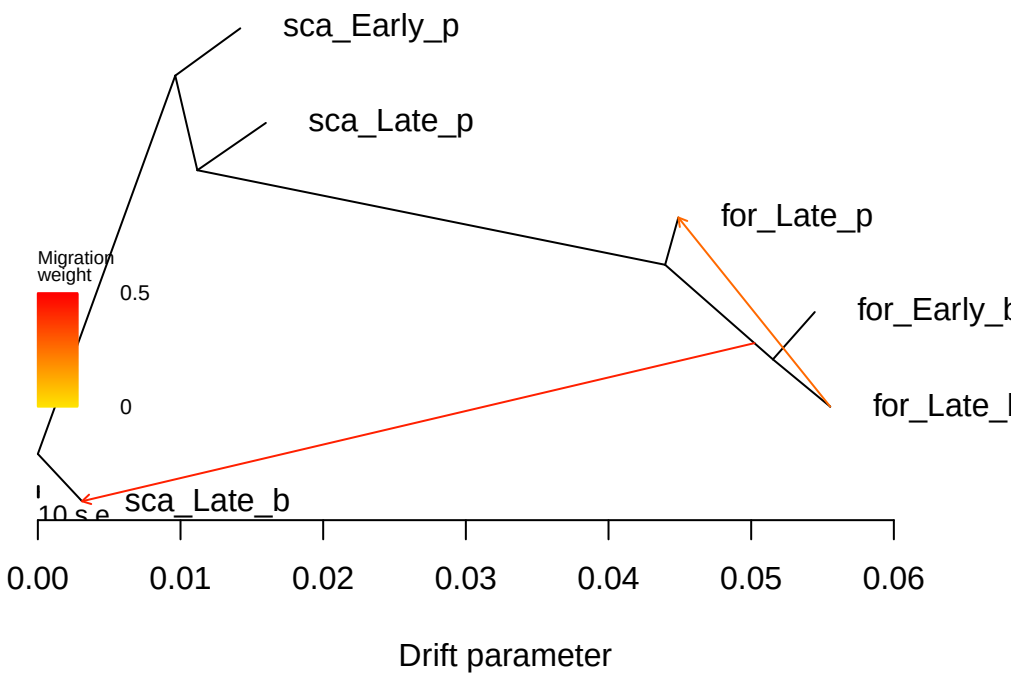
pwd: /proj/uppstore2017190/b2012111_nobackup/private/fan/fortis_scandens_pools/TreeMix

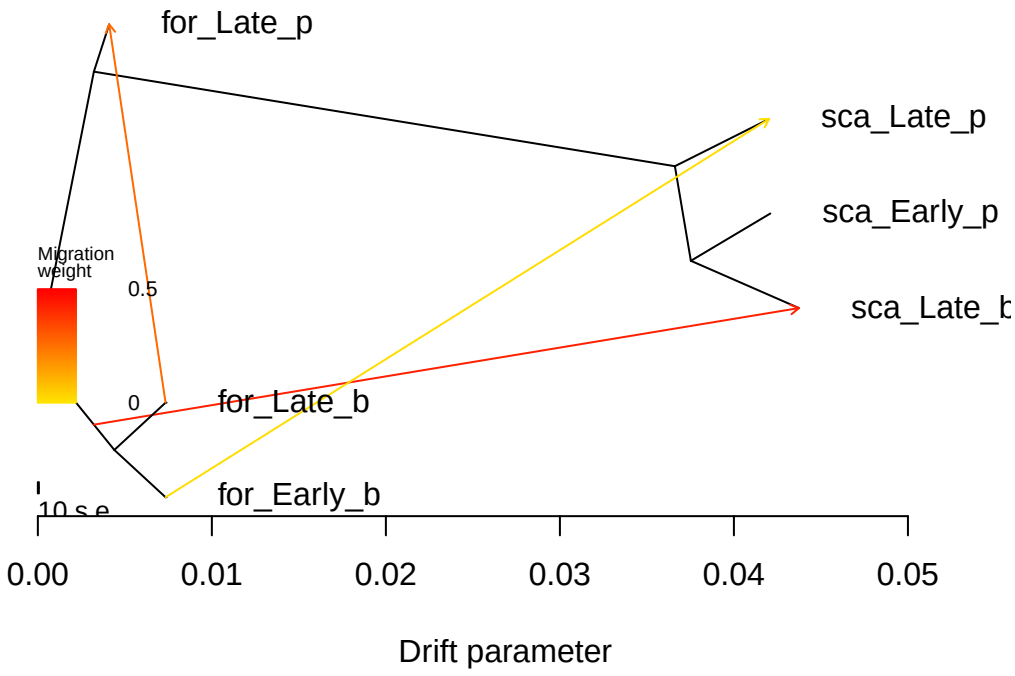
Run time 00:04:51

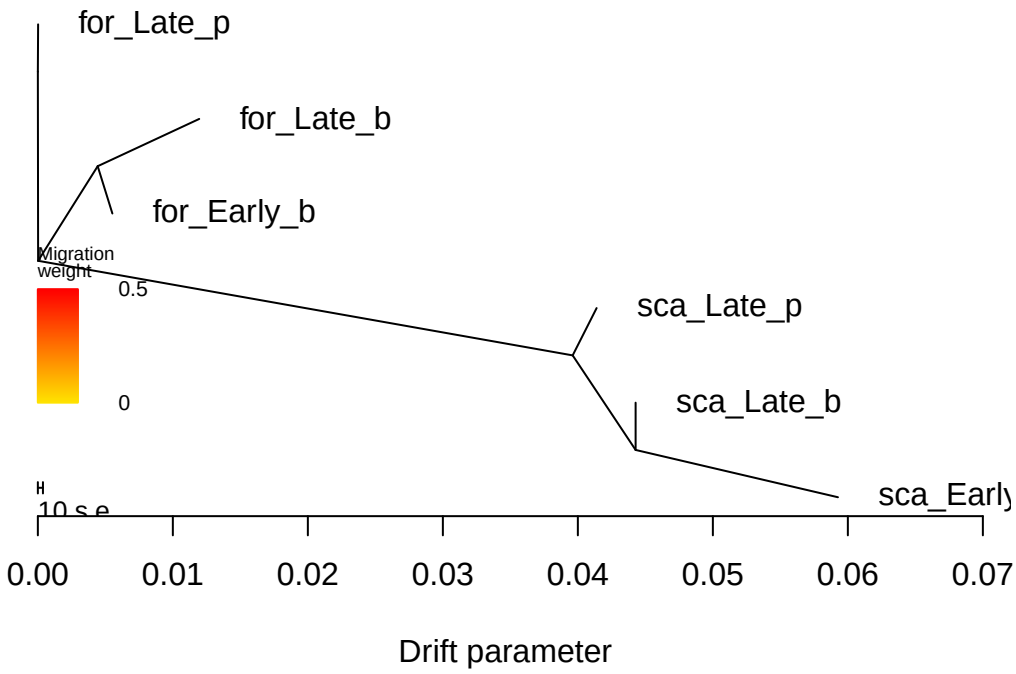


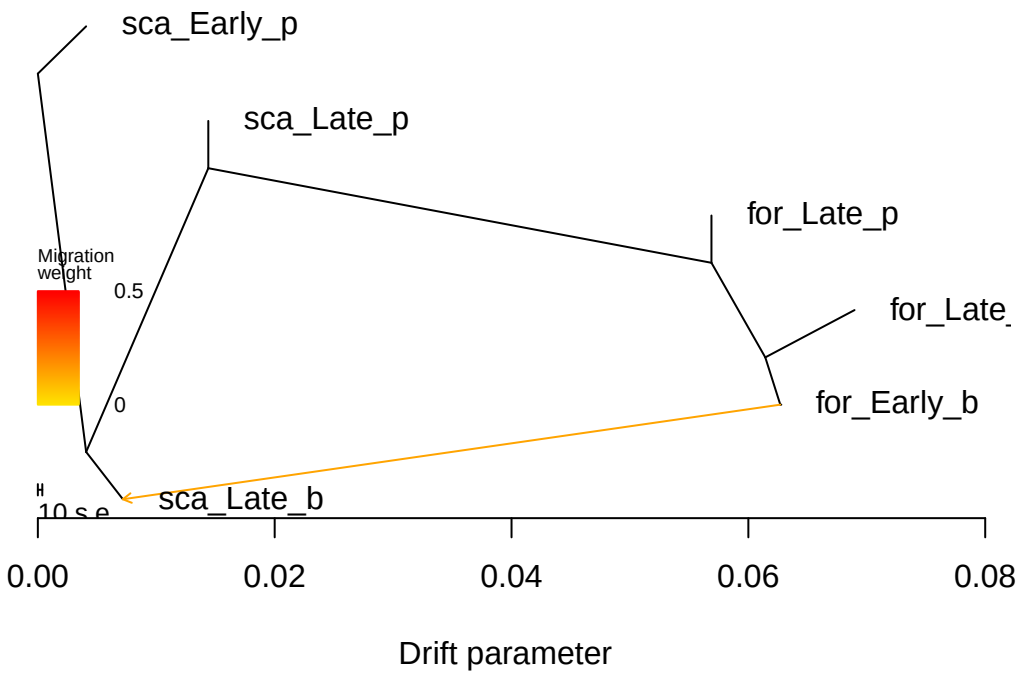
Plot the trees in different -m parameters (migration)

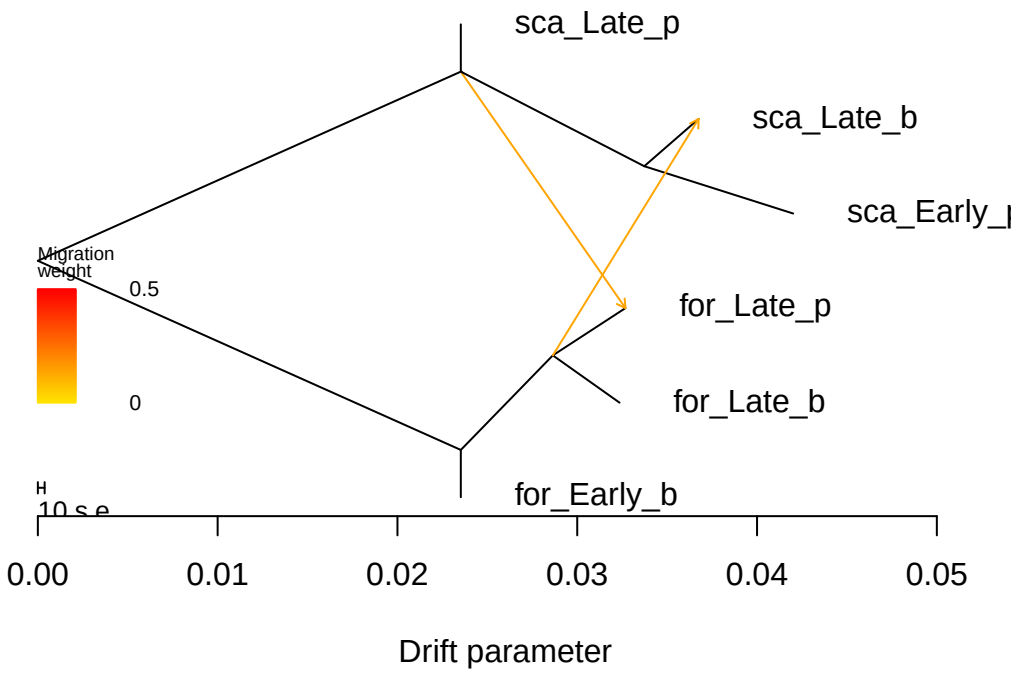

for_Late_p
sca_Late_p
sca_Early_p
Migration
weight
0.5
sca_Late_b
0
for_Late_b
for_Early_b
10 s.e.
0.00
0.01
0.02
0.03
0.04
0.05
Drift parameter

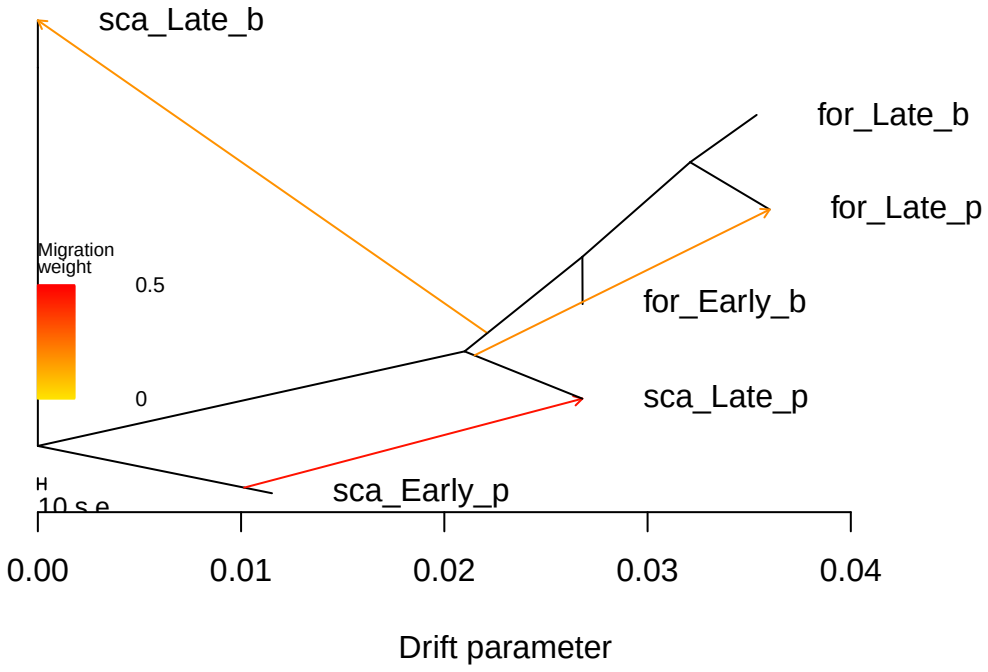

sca_Early_p
sca_Late_p
for_Late_p
for_Early_b
for_Late_b
sca_Late_b
Migration weight
0.5
0
10 s e
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Drift parameter


for_Late_p
sca_Late_p
sca_Early_p
sca_Late_b
Migration
weight
0.5
0
for_Late_b
for_Early_b
10 s.e.
0.00
0.01
0.02
0.03
0.04
0.05
Drift parameter


for_Late_p
for_Late_b
for_Early_b
Migration
weight
0.5
0
sca_Late_p
sca_Late_b
sca_Early
10 s e
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Drift parameter


sca_Early_p
sca_Late_p
for_Late_p
for_Late_
for_Early_b
sca_Late_b
Migration
weight
0.5
0
10 s.e.
0.00
0.02
0.04
0.06
0.08
Drift parameter


sca_Late_p
sca_Late_b
sca_Early_
Migration
weight
0.5
0
for_Late_p
for_Late_b
for_Early_b
0.00
0.01
0.02
0.03
0.04
0.05
Drift parameter


sca_Late_b
for_Late_b
for_Late_p
Migration
weight
0.5
0
for_Early_b
sca_Late_p
sca_Early_p
0.00
0.01
0.02
0.03
0.04
Drift parameter