Sea lamprey RAPTURE dataset characterization

Sard et al. 2019

July 2019

Contents

Introduction and Setup

RAD capture (RAPTURE) provides a means of rapidly genotyping hundreds to thousands of standardized restriction site associated (RAD) loci. Here we provide a detailed characterization of a panel of 12,435 RAD loci developed for sea lamprey management and research. The input file for all the following analyses is appendedLoci, which is available in the research compendium (R package) described below.

Install SeaLampreyRapture package

The SeaLampreyRapture package includes all data necessary for performing the following analyses. It can be download from GitHub using the following R commands.

```
options(repos=structure(c(CRAN="http://cran.r-project.org")))
install.packages("devtools")
library(devtools)
install_github("ScribnerLab/SeaLampreyRapture")
```

Load required packages

Interactively run /analysis/installPackages.R to ensure all necessary packages are available

```
library(adegenet)
library(vcfR)
library(hierfstat)
library(tidyverse)
library(gridExtra)
library(ggthemes)
library(SeaLampreyRapture)
library(quantsmooth)
data(SeaLampreyRapture)
```

Sequencing profile

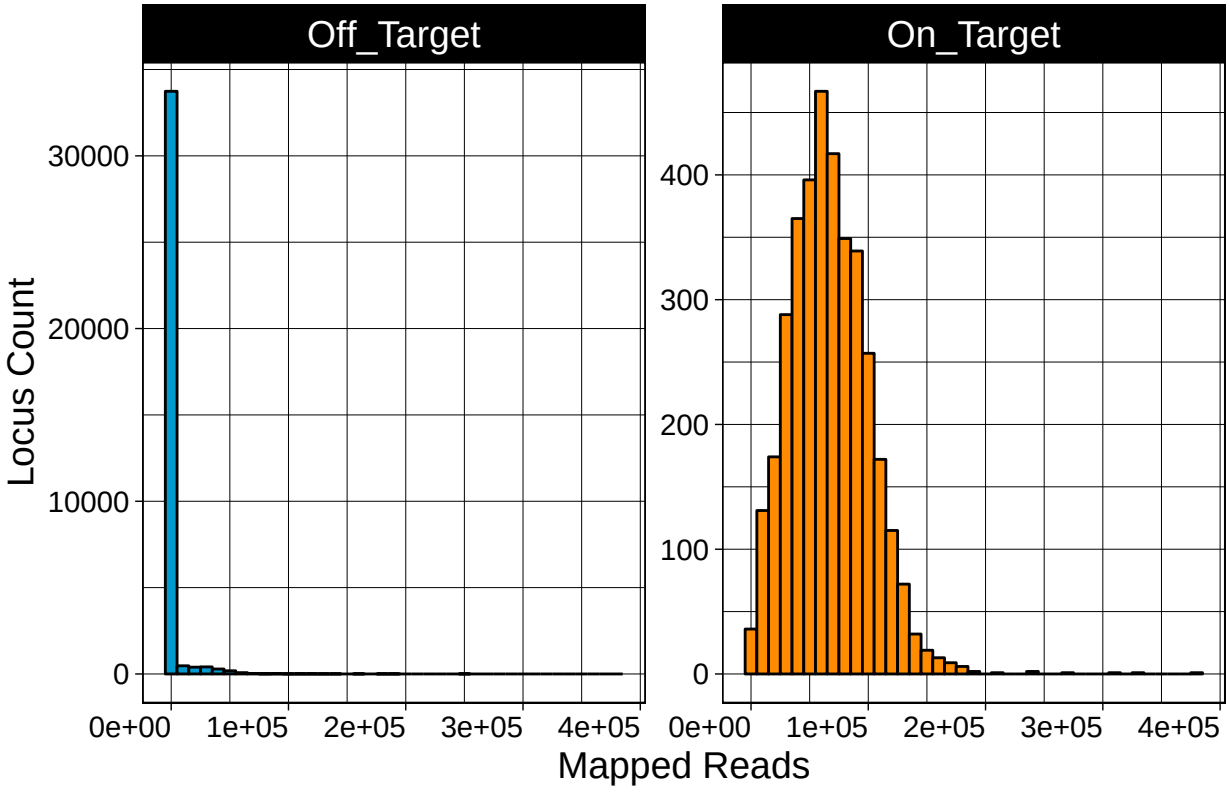
The following section characterizes the sequencing results.

On/off target reads per locus

The number of mapped read pairs for RAD loci targeted with rapture (On_Target) versus RAD loci that were not targeted (Off_Target). Overall, 80.6% of reads aligned to targeted loci even though they only made up less than 10% of all RAD loci.

```
p <- ggplot(onTarget_readCount, aes(x=ReadCount, fill = T)) +
  geom_histogram(binwidth = 10000, color = "black") +
  scale_fill_manual(values=c("deepskyblue3", "darkorange"))+
  facet_wrap(~T, scales = "free_y")

p + ggtitle("") +
  xlab("Mapped Reads") + ylab("Locus Count") +
  theme_linedraw() +
  theme(strip.text.x = element_text(size = 14, angle = 0)) +
  theme(legend.position="none",
        text = element_text(size=14),
        axis.text.x = element_text(angle=0, hjust=1))
```



Allele balance

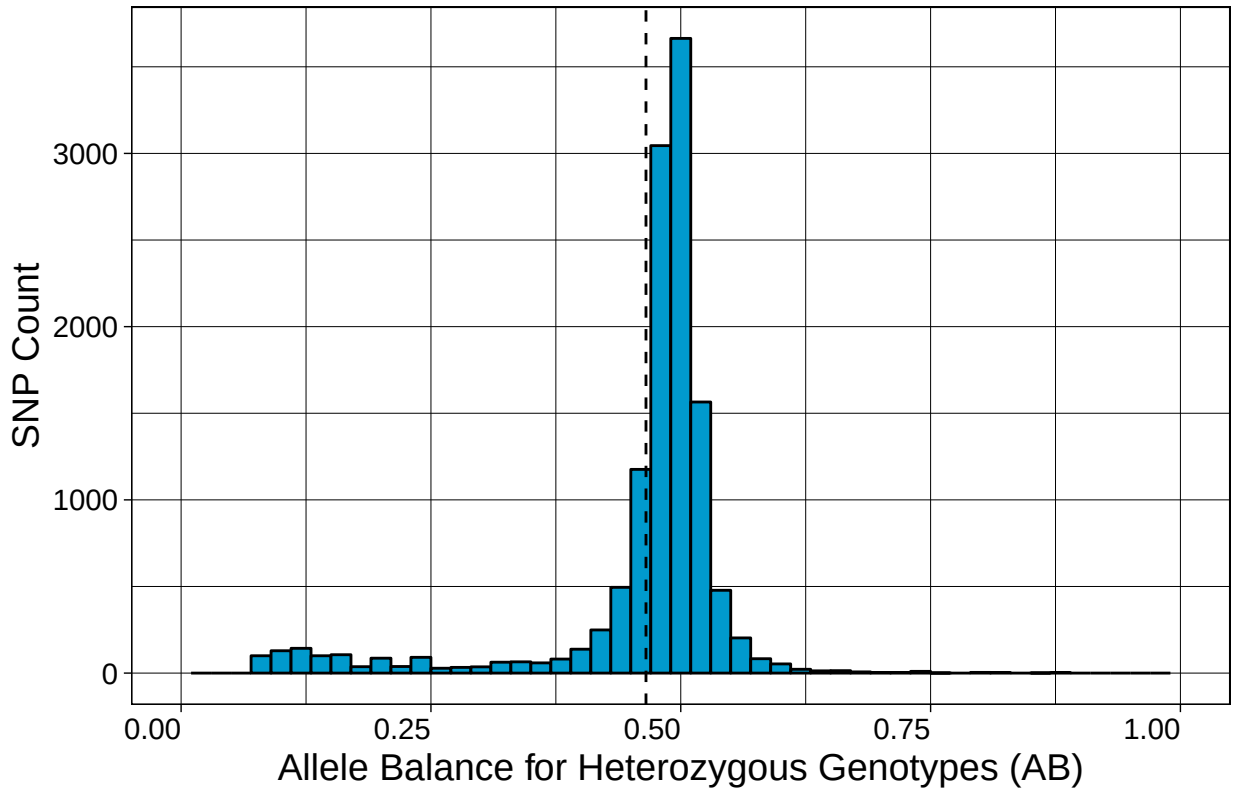
Allele read balance (AB) calculated for all SNPs in manuscript Table 1. Allele balance is the proportion of reads supporting the reference allele for heterozygous genotypes at a given locus. We expect values to be

near 0.5 for diploid loci that are not confounded by sequencing or mapping errors.

```
s1 <- subset(appendixLoci, allele_balance < 0.7)
s2 <- subset(s1, allele_balance > 0.3)

p <- ggplot(appendixLoci, aes(x=allele_balance)) +
  geom_histogram(fill="deepskyblue3", color = "black", binwidth = 0.02) +
  xlim(c(0,1)) +
  geom_vline(xintercept = mean(appendixLoci$allele_balance), linetype = "dashed")

p + ggtitle("") +
  xlab("Allele Balance for Heterozygous Genotypes (AB)") + ylab("SNP Count") +
  theme_linedraw()+
  theme(legend.position="none",
        text = element_text(size=14),
        axis.text.x = element_text(angle=0, hjust=1))
```



Target density

```
## Need input file from Seth, I think.
```

Chromosome plots

Greater than 10 mb

Distribution of target RAD loci across the 44 largest scaffolds (>10 mb) in the sea lamprey genome. Each horizontal black line represents a scaffold and each vertical blue line represents the location of a targeted

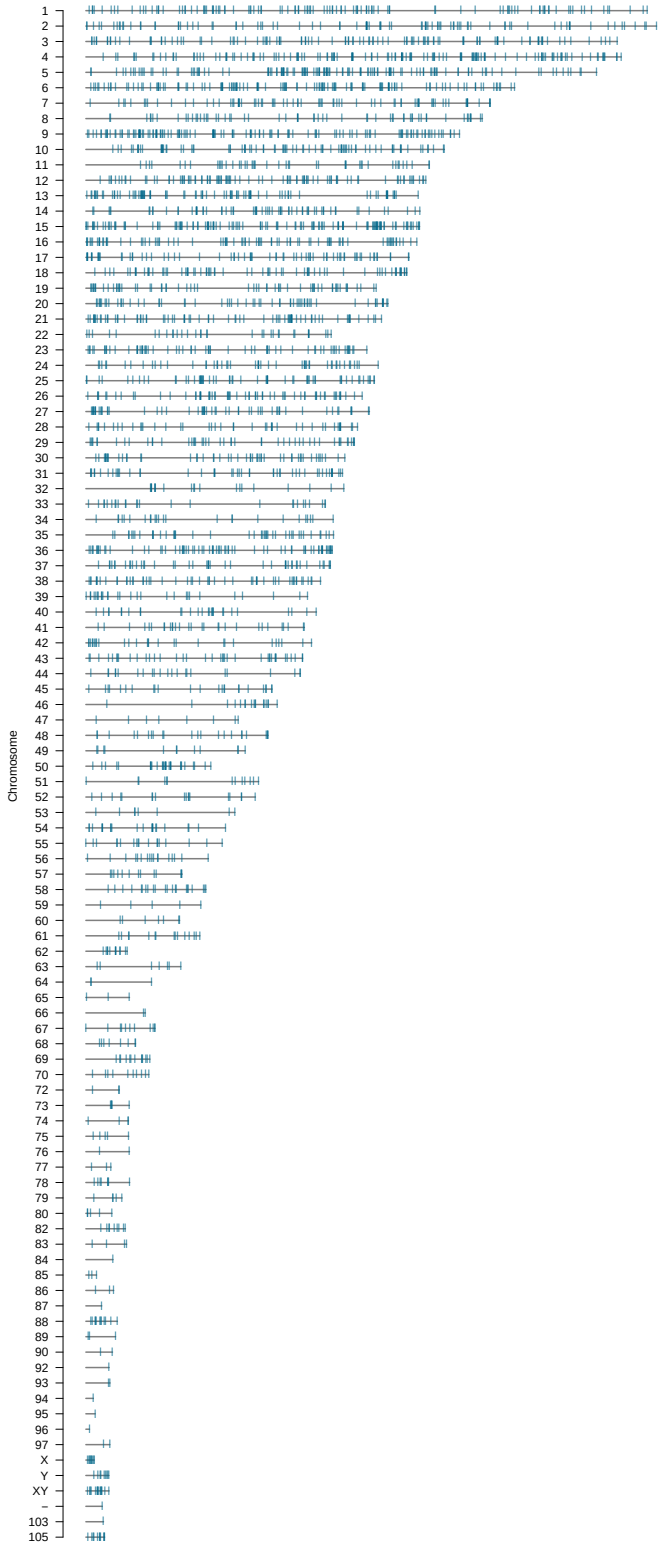
locus. A total of 2844 of 3446 targeted loci map to these scaffolds (82.5%).

```
dat45 <- subset(targets.chrpos, targets.chrpos$CHR < 45)
chrompos <- prepareGenomePlot(dat45, cols = "grey50", paintCytobands = TRUE, bleach = 0, topspace = 1, c
points(chrompos[,2],chrompos[,1]+0.05,pch="|", cex = 0.75, col="deepskyblue4")
```

Greater than 1 mb

Distribution of target RAD loci across 100 scaffolds greater than 1 MB in length. Each horizontal black line represents a scaffold and each vertical blue line represents the location of a targeted locus. A total of 3316 of 3446 targeted loci map to these scaffolds (96.22%).

```
dat106 <- subset(targets.chrpos, targets.chrpos$CHR < 106)
chrompos <- prepareGenomePlot(dat106, cols = "grey50", paintCytobands = TRUE, bleach = 0, topspace = 1,
points(chrompos[,2],chrompos[,1]+0.05,pch="|", cex = 0.75, col="deepskyblue4")
```


Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
72
73
74
75
76
77
78
79
80
82
83
84
85
86
87
88
89
90
92
93
94
95
96
97
X
Y
XY
–
103
105

Genetic variation

The following section characterizes the genetic variation detected using the bait panel

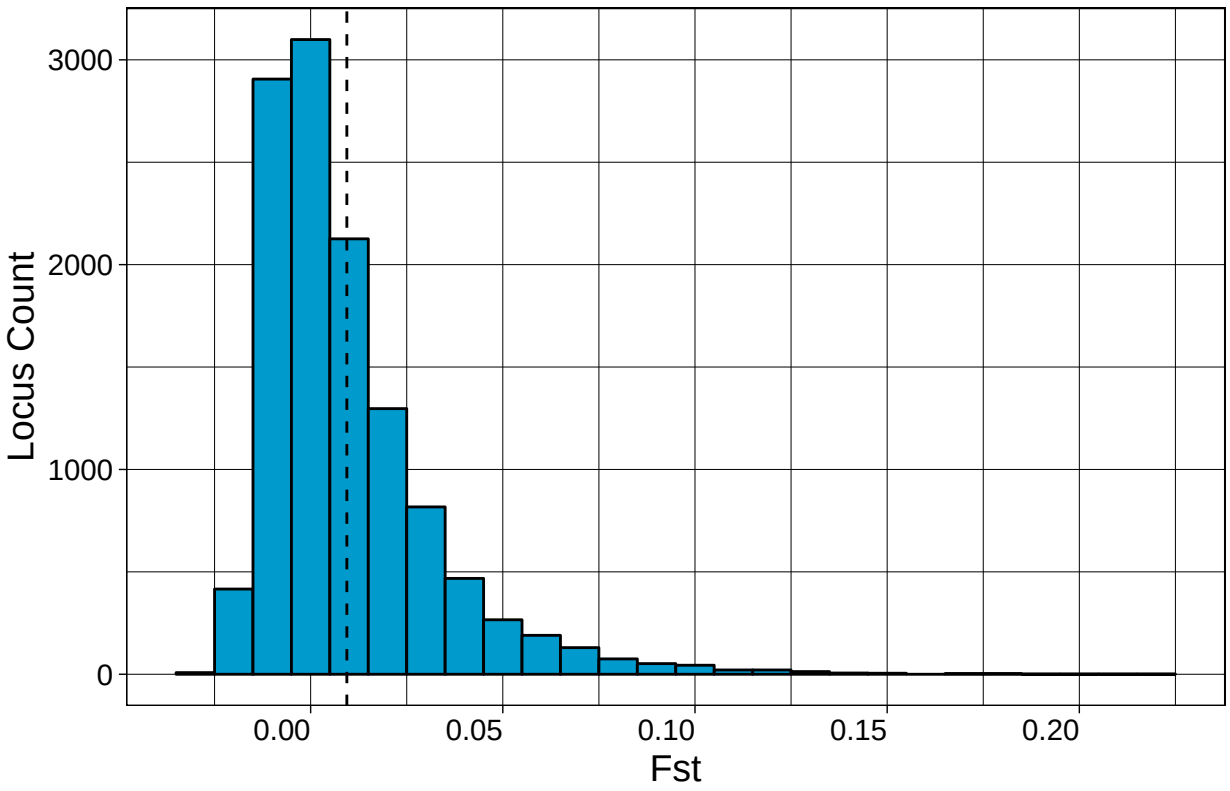
Histogram of F_{ST} values per locus

Distribution of F_{ST} values for 11,970 SNP loci genotyped in sea lamprey at five spawning sites. The dashed vertical line indicates the mean F_{ST} value.

```
dat.fst <- appendixLoci %>% drop_na(fst_nei73_heirfstat)
Fst <- dat.fst$fst_nei73_heirfstat
mf <- mean(Fst)

p <- ggplot(dat.fst, aes(x=fst_nei73_heirfstat)) +
  geom_histogram(fill="deepskyblue3",color = "black", binwidth = 0.01) +
  geom_vline(xintercept = mf, linetype = "dashed")

p + ggtitle("") +
  xlab("Fst") + ylab("Locus Count") +
  theme_linedraw()+
  theme(legend.position="none",
        text = element_text(size=14),
        axis.text.x = element_text(angle=0, hjust=1))
```



F_{IS} values per locus, per population

Distributions of F_{IS} generally centered around zero for 11,970 SNP loci genotyped in sea lamprey at five spawning sites.

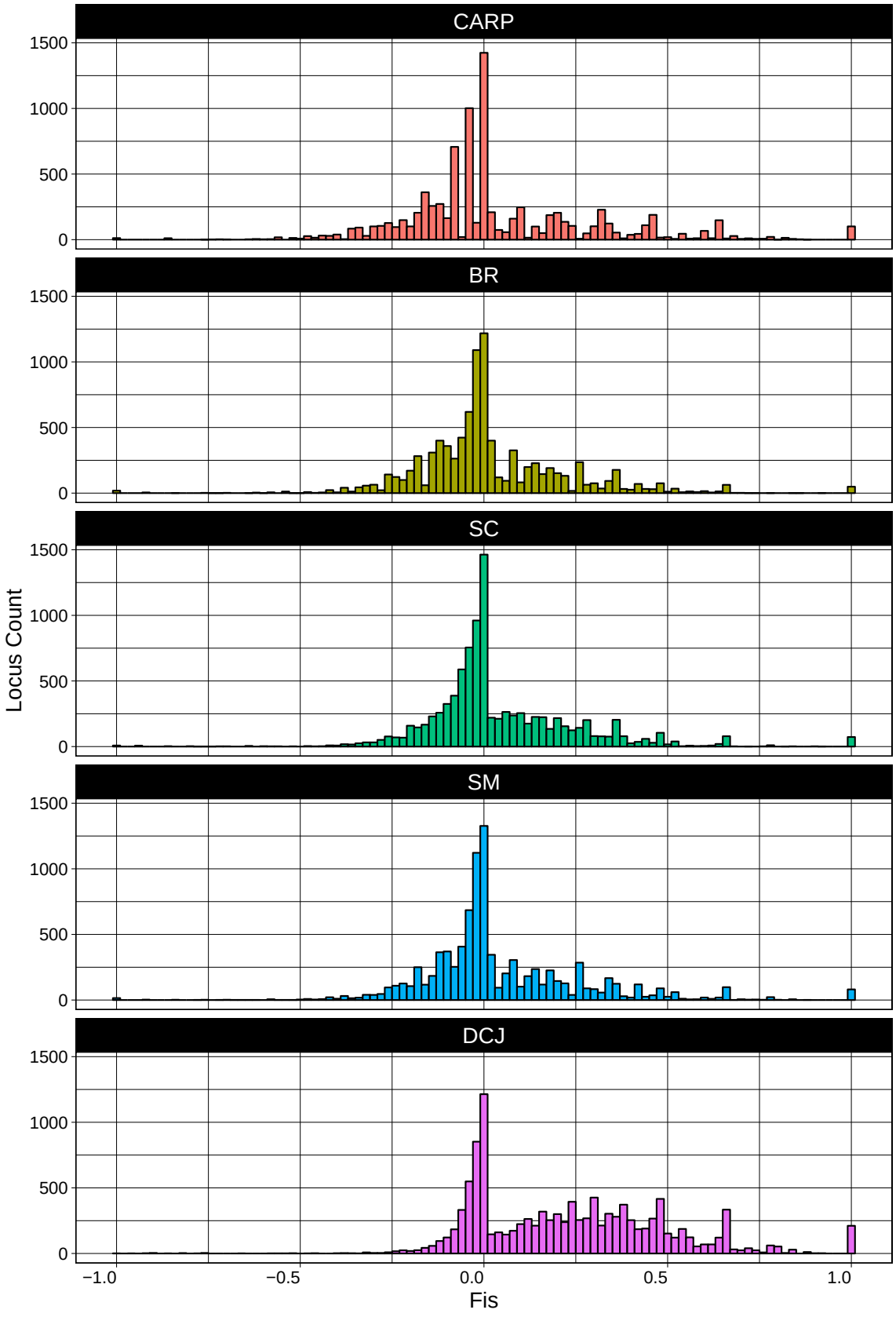
```
CARP <- as.data.frame(cbind(appendixLoci$Fis_CARP, rep(x = "CARP", nrow(appendixLoci))))
BR <- as.data.frame(cbind(appendixLoci$Fis_BR, rep(x = "BR", nrow(appendixLoci))))
SC <- as.data.frame(cbind(appendixLoci$Fis_SC, rep(x = "SC", nrow(appendixLoci))))
SM <- as.data.frame(cbind(appendixLoci$Fis_SM, rep(x = "SM", nrow(appendixLoci))))
DCJ <- as.data.frame(cbind(appendixLoci$Fis_DCJ, rep(x = "DCJ", nrow(appendixLoci))))

names(CARP) <- c("Fis", "Pop")
names(BR) <- c("Fis", "Pop")
names(SC) <- c("Fis", "Pop")
names(SM) <- c("Fis", "Pop")
names(DCJ) <- c("Fis", "Pop")

FisTable <- rbind(CARP, BR, SC, SM, DCJ)
FisTable$Fis <- as.numeric(as.character(FisTable$Fis))

p <- ggplot(FisTable, aes(x=Fis, fill = Pop)) +
  geom_histogram(binwidth = 0.02, color = "black") +
  facet_wrap(~Pop,nrow = 5)

p + ggtitle("") +
  xlab("Fis") + ylab("Locus Count") +
  theme_linedraw() +
  theme(strip.text.x = element_text(size = 14, angle = 0)) +
  theme(legend.position="none",
        text = element_text(size=14),
        axis.text.x = element_text(angle=0, hjust=1))
```


CARP
BR
SC
SM
DCJ
Locus Count
0
500
1000
1500
−1.0
−0.5
0.0
0.5
1.0
Fis

Minor allele frequencies per locus, per population

Distributions of minor allele frequencies for 11,970 SNP loci genotyped in sea lamprey at five spawning sites varied among populations.

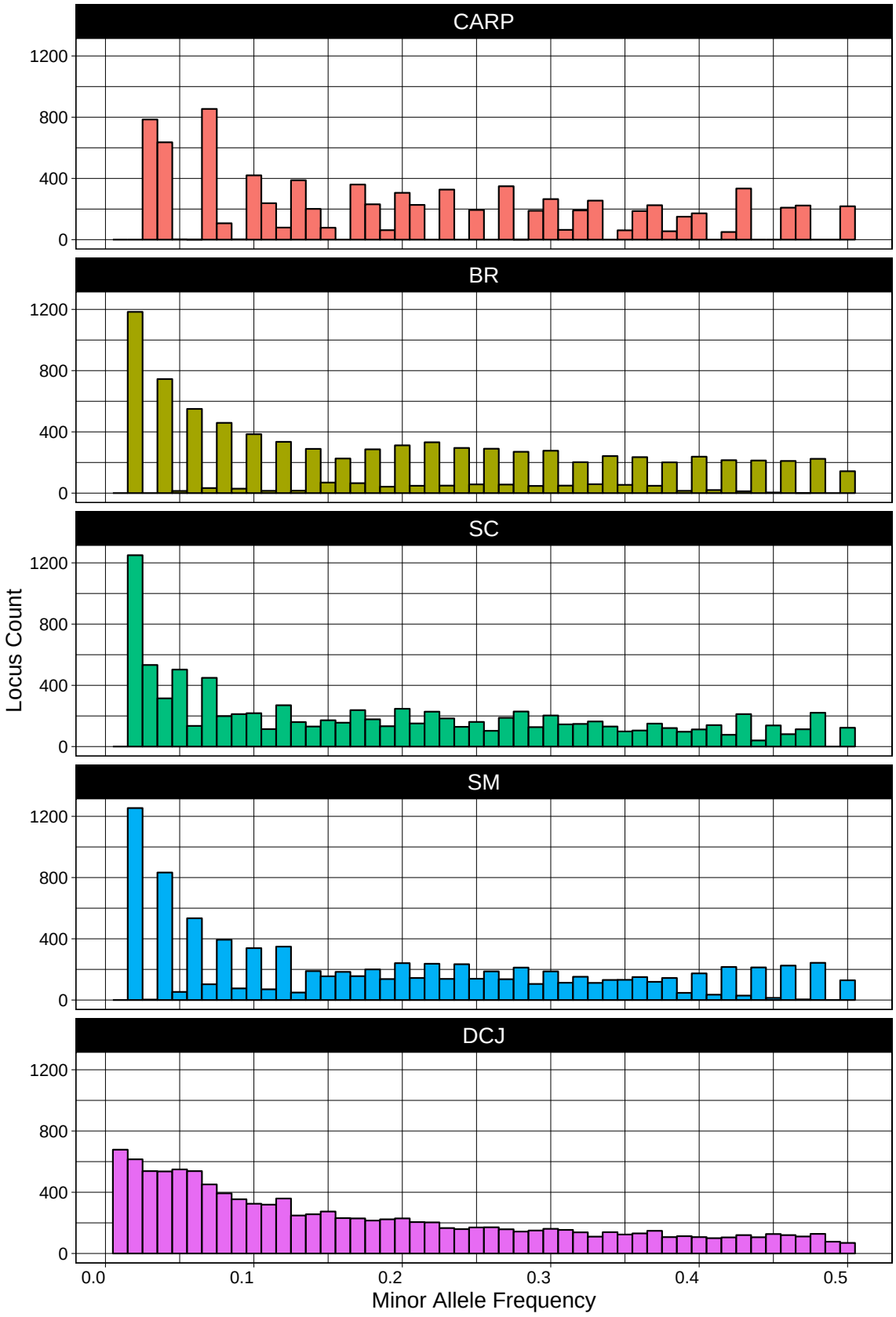
```
CARP <- as.data.frame(cbind(appendixLoci$MAF_CARP, rep(x = "CARP", nrow(appendixLoci))))
BR <- as.data.frame(cbind(appendixLoci$MAF_BR, rep(x = "BR", nrow(appendixLoci))))
SC <- as.data.frame(cbind(appendixLoci$MAF_SC, rep(x = "SC", nrow(appendixLoci))))
SM <- as.data.frame(cbind(appendixLoci$MAF_SM, rep(x = "SM", nrow(appendixLoci))))
DCJ <- as.data.frame(cbind(appendixLoci$MAF_DCJ, rep(x = "DCJ", nrow(appendixLoci))))

names(CARP) <- c("MAF", "Pop")
names(BR) <- c("MAF", "Pop")
names(SC) <- c("MAF", "Pop")
names(SM) <- c("MAF", "Pop")
names(DCJ) <- c("MAF", "Pop")

MAFTable <- rbind(CARP, BR, SC, SM, DCJ)
MAFTable$MAF <- as.numeric(as.character(MAFTable$MAF))
MAFTable <- subset(MAFTable, MAF < 1 & MAF > 0)

p <- ggplot(MAFTable, aes(x=MAF, fill = Pop)) +
  geom_histogram(binwidth = 0.01, color = "black") +
  facet_wrap(~Pop,nrow = 5)

p + ggtitle("") +
  xlab("Minor Allele Frequency") + ylab("Locus Count") +
  theme_linedraw() +
  theme(strip.text.x = element_text(size = 14, angle = 0)) +
  theme(legend.position="none",
        text = element_text(size=14),
        axis.text.x = element_text(angle=0, hjust=1))
```


CARP
BR
SC
SM
DCJ
Locus Count
0
400
800
1200
0.0
0.1
0.2
0.3
0.4
0.5
Minor Allele Frequency