Simulans-Mauritiana Backcross Analysis for Amanda Moehring/Tom Hsiang

Charlie Soeder

2024-01-29

Contents

1 Introduction

We essentially did introgression of loci linked to interspecies hybrid sterility between sim and mau for ten generations in both directions of backcross, then sequenced. We then looked for which regions of introgressed genome were associated with 10th generation sterile male offspring, compared to fertile male offspring. The data analysis was done ok, but missing some key elements

that we really should add, and I also think there's a lot more precision that could be retrieved out of that data set. (Amanda Moehring, email to Corbin 17 July 2023)

The project aims to map genes associated with interspecies hybrid male sterility between D. simulans and D. mauritiana. We selected using a sterile sperm morphology (we call "needle-eye". NE), selected across 10 generations of backcross by using sisters of sterile males in the backcross crosses. In the 10th generation, approx 50% of males were sterile and 50% fertile, which led us to think it could be a single locus. So we pooled the males of each phenotype into single samples for sequencing. The samples are as follows:

```
mauGFP - wildtype D. mauritania with a GFP-tagged protamine (makes sperm fluoresce green).
simGFP - same as above, but D. simulans
BCM10NE - 10th generation backcross mauritiana males with needle-eye (sterile) sperm
BCM10WT - 10th generation backcross mauritiana males with wildtype sperm
BCS10NE - 10th generation backcross simulans males with needle-eye (sterile) sperm
BCS10WT - 10th generation backcross simulans males with wildtype sperm
```

-Amanda Moehring, email to me 15 Aug 2023

Each pooled sample consisted of 30 individuals.

-Amanda Moehring, email to me 10 Oct 2023

2 Materials, Methods, Data, Software

2.1 Reference Genomes

The droSim1 reference genome was downloaded in FASTA format from UCSC Genome Browser; the UC Irvine mauritiana assembly (GCF_004382145.1) and the Princeton simulans assembly (GCF_016746395.2) were downloaded from NCBI. The droSim1 reference was the best-consolidated. Assemblies for the specific strains being investigated were provided by Tom Hsiang.

Table 1. Size and Consolidation of Reference Genomes

	source	# bases	# contigs
mauritiana backcross			
BCM10NE	moehring lab	$126M$	$47K$
BCM10WT	moehring lab	$125M$	$44K$
simulans backcross			
BCS10NE	moehring lab	$134M$	$173K$
BCS10WT	moehring lab	$144M$	$358K$
drosophila simulans			
droSim1	UCSC Genome Browser	$142M$	18
prinDsim3	NCBI	$132M$	95
simGFP	moehring lab	$137M$	$171K$
drosophila mauritiana			
mauGFP	moehring lab	$144M$	$142K$
ncbiMau	NCBI	$152M$	353

The main chromosomes correspond to the following contigs in the NCBI reference:

```
2L      NC_052520.2
2R      NC_052521.2
3L      NC_052522.2
3R      NC_052523.2
4       NC_052524.2
X       NC_052525.2
```

https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_016746395.2/

2.2 Reference Annotations

NCBI

Table 2. Reference Annotations and their Sizes

	size (bp)				
annot	average	total	total count	genome	source
NCBIsim103	$5.9K$	$93M$	$15.8K$	prinDsim3	UCSC Genome Browser

2.3 Sequenced Reads

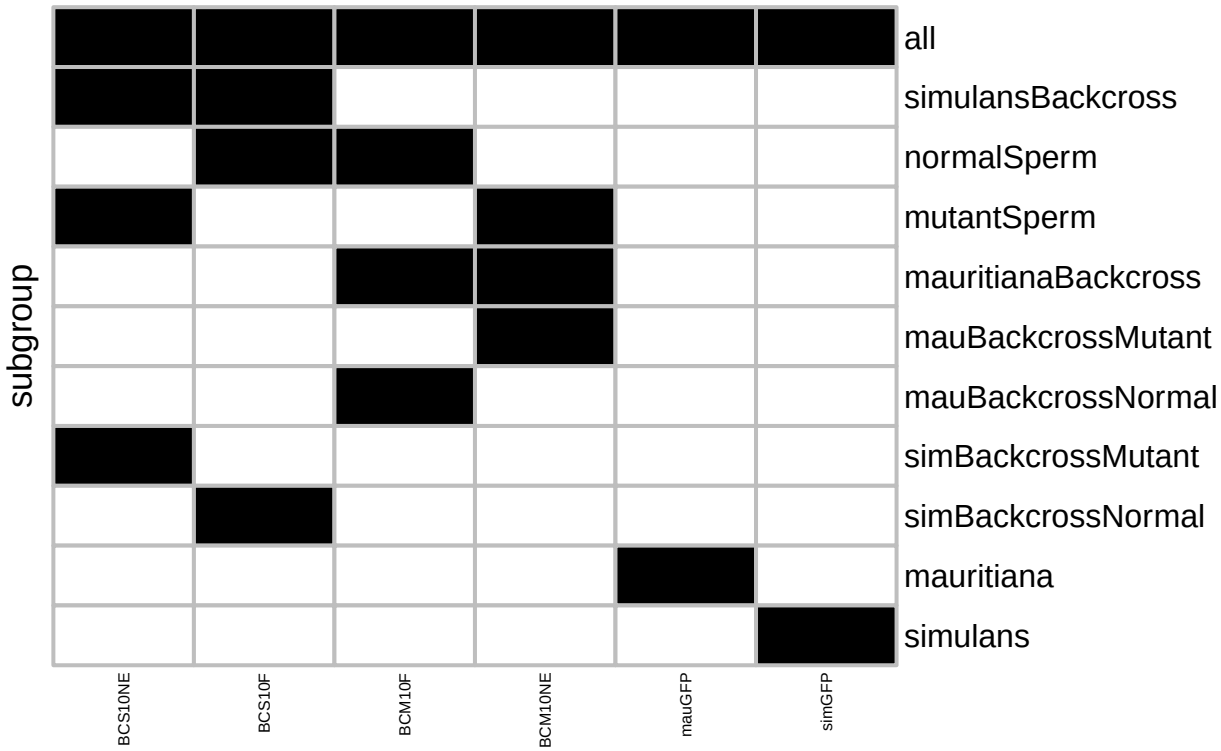
A backcross and introgression experiment was performed, in which simulans-mauritiana crosses were backcrossed to each of their parent species, with and without selection for a mutant phenotype causing male sterility. The offspring of 10-generations of backcrossing were sequenced, as well as the parent stock.

Table 3. Control and Selection Experiments

name	genealogy	genotype
BCM10F	mauritiana backcross	wildtype
BCM10NE	mauritiana backcross	needle eye
BCS10F	simulans backcross	wildtype
BCS10NE	simulans backcross	needle eye
mauGFP	mauritiana	GFP
simGFP	simulans	GFP

PsiSeq 1 and 2 use head-to-head comparisons of individual samples; PopPsiSeq compares subgroups (possibly consisting of a single individual) of samples.

Figure 1. Subgroup Definitions for PopPsiSeq



2.3.1 Pre-processing

These reads were preprocessed with FASTP (Chen et al. 2018) for quality control and analytics.

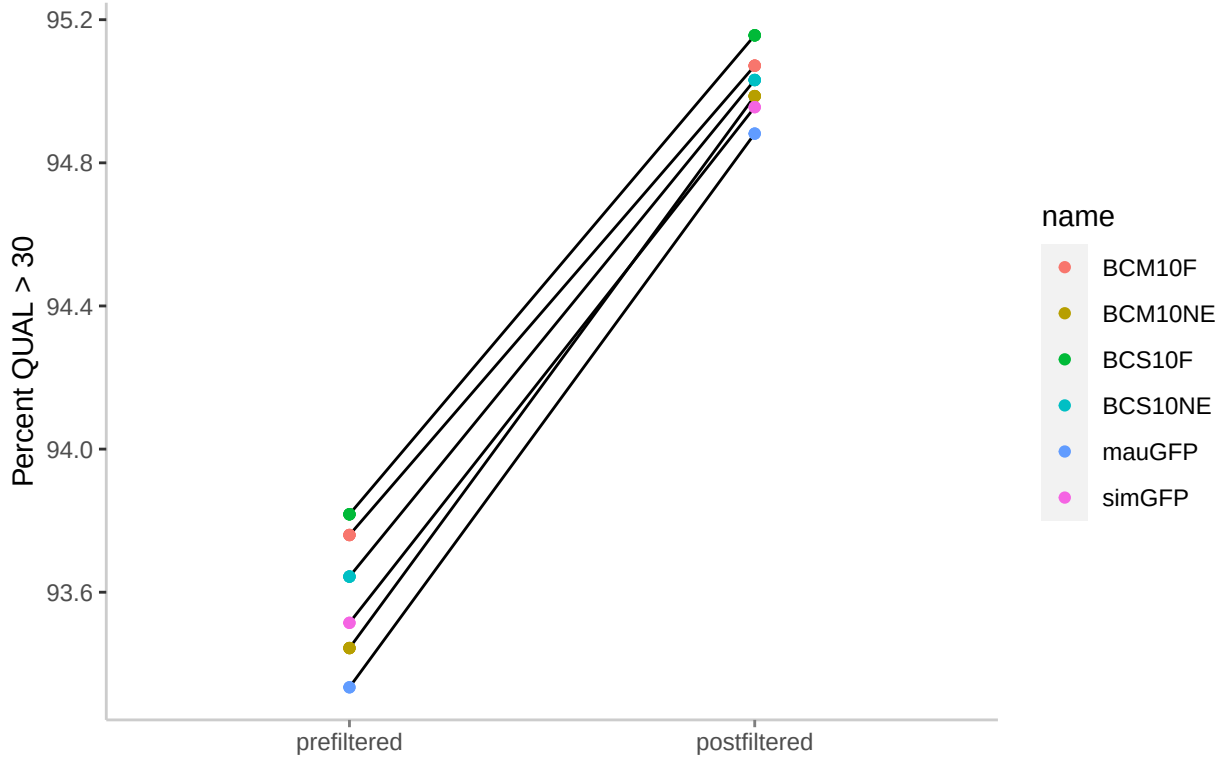
Starting FASTQ files contained a total of $1.22G$ reads; after QC, this dropped to $1.18G$.

Table 4. Read Retention Rate during Preprocessing

	minimum	average	maximum
prefiltered	$31M$	$61M$	$152M$
postfiltered	$30M$	$59M$	$148M$
percent retention	97	97	97

Filtration also increased the read quality, as seen in the increase in the fraction of reads with an average quality score > 30:

Figure 2. Percent of Reads with a mean QUAL > 30

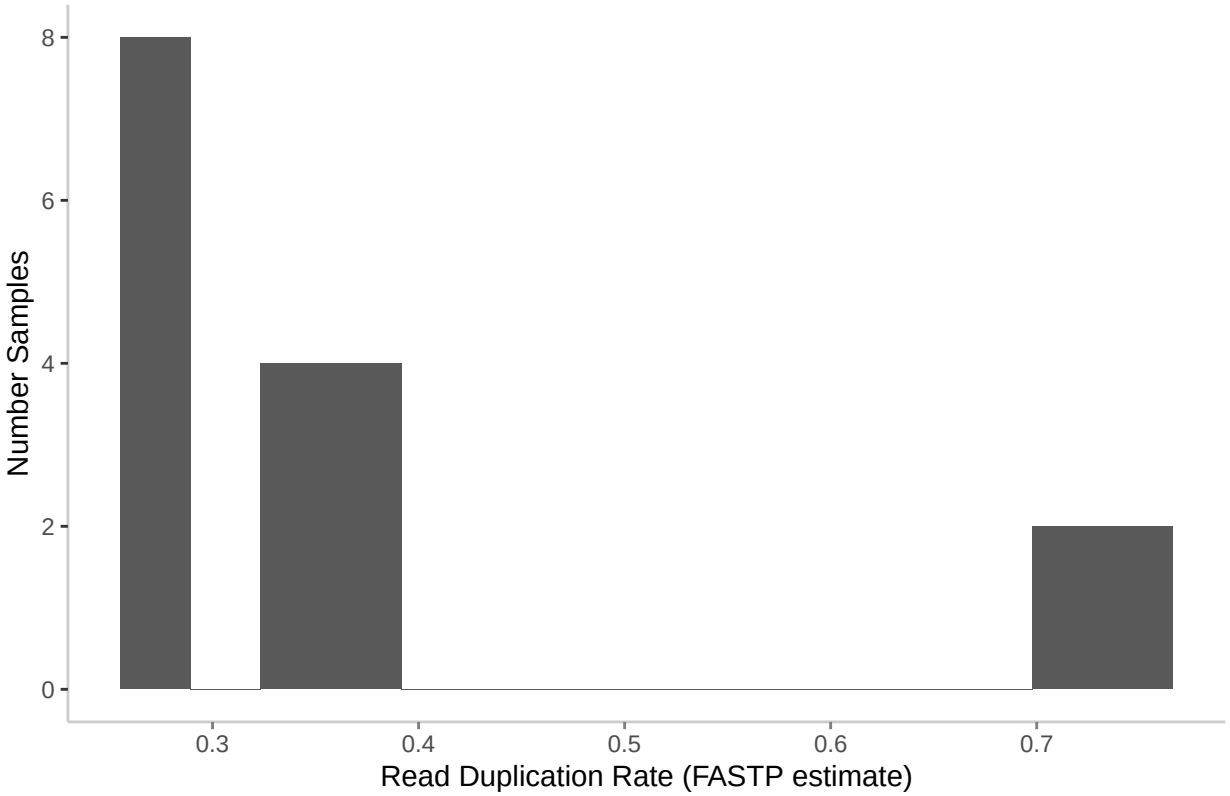


Duplicate reads were also detected; these will be filtered during alignment:

Table 5. Percentage Duplication
FASTP estimate

minimum	average	median	maximum
0.3	0.4	0.3	0.7

Figure 3. Duplication Histogram

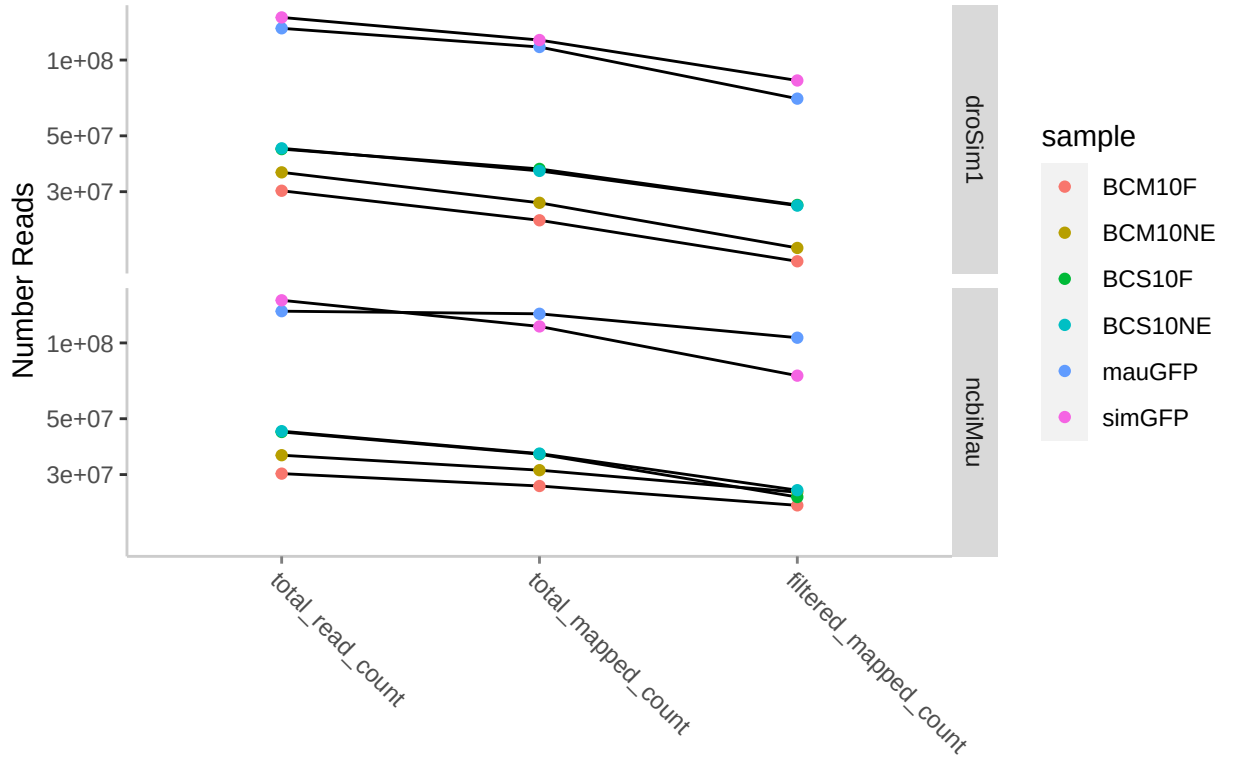


2.4 Mapped Reads

Reads were first mapped to a reference genome using the BWA SAMPE/SE algorithm. Then, the alignment file was filtered for uniqueness (ie, a read must be aligned optimally with no alternative or runner-up hits, `XT:A:U.*X0:i:1.*X1:i:0`), mapping/sequencing quality (`-q 20 -F 0x0100 -F 0x0200 -F 0x0300 -F 0x04`), and deduplication. This filtered alignment is called "bwaUniq".

2.4.1 Read & Alignment Quality

Figure 4. Read Counts by Processing Step: Unmapped, Mapped, Filtered




```
## `summarise()` has grouped output by 'measure'. You can override using the
## `.groups` argument.
```

Table 6. Read Counts During Alignment & Filtration

reference	minimum	average	median	maximum
filtered_mapped_count				
droSim1	15.9M	40.0M	26.5M	82.9M
ncbiMau	22.6M	46.3M	25.8M	104.9M
total_mapped_count				
droSim1	23.1M	59.4M	36.6M	120.1M
ncbiMau	27.0M	62.9M	36.3M	130.5M
total_read_count				
droSim1	30.3M	72.7M	44.4M	147.7M
ncbiMau	30.3M	72.7M	44.4M	147.7M

The fraction of reads retained at each point:

```
## `summarise()` has grouped output by 'measure'. You can override using the
## `.groups` argument.
```

Table 7. Percentage of Reads Retained at Each Step

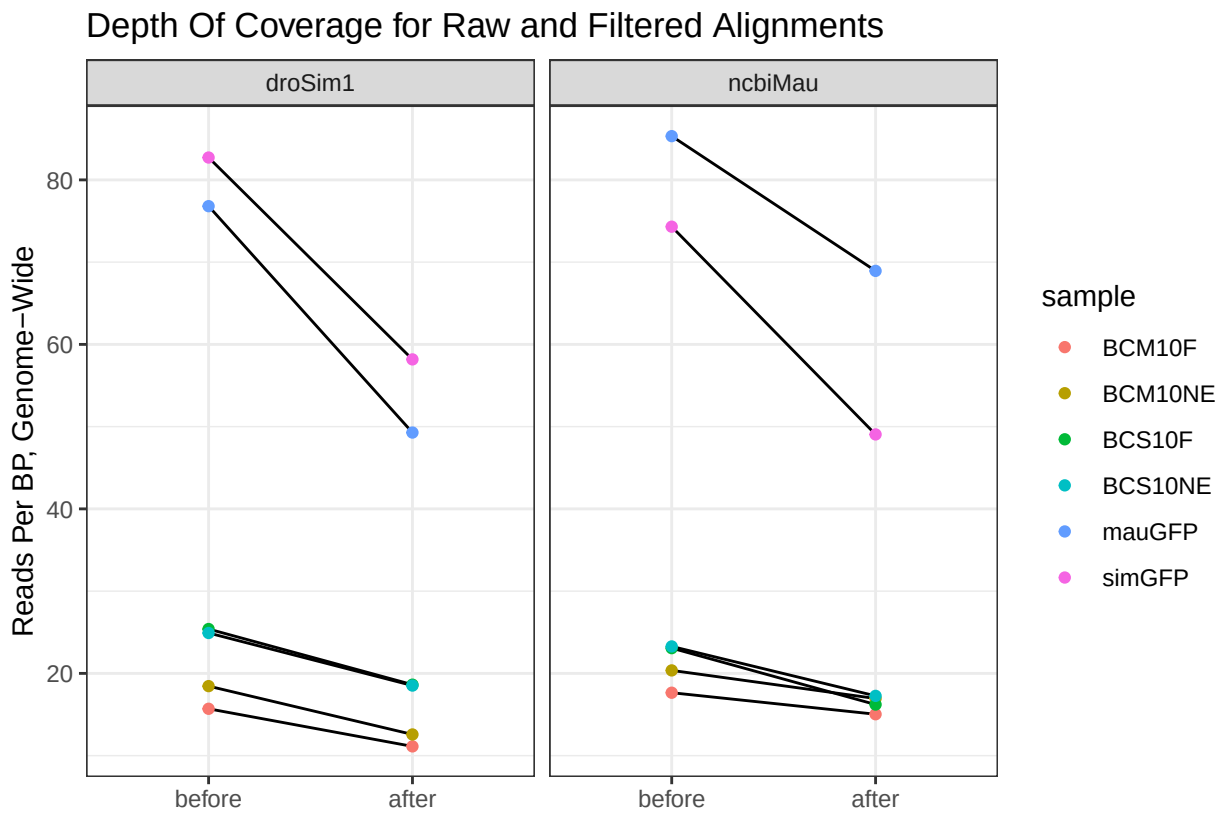
reference	minimum	average	median	maximum
filter_retention				
droSim1	62.4%	68.5%	68.9%	72.8%
ncbiMau	63.7%	74.8%	76.0%	83.8%
mapping_retention				
droSim1	75.7%	80.4%	81.3%	84.3%
ncbiMau	78.7%	86.0%	84.4%	97.6%

2.4.2 Depth & Breadth of Coverage

Depth of coverage, ie, the genome-wide average number of mapped reads per base pair:

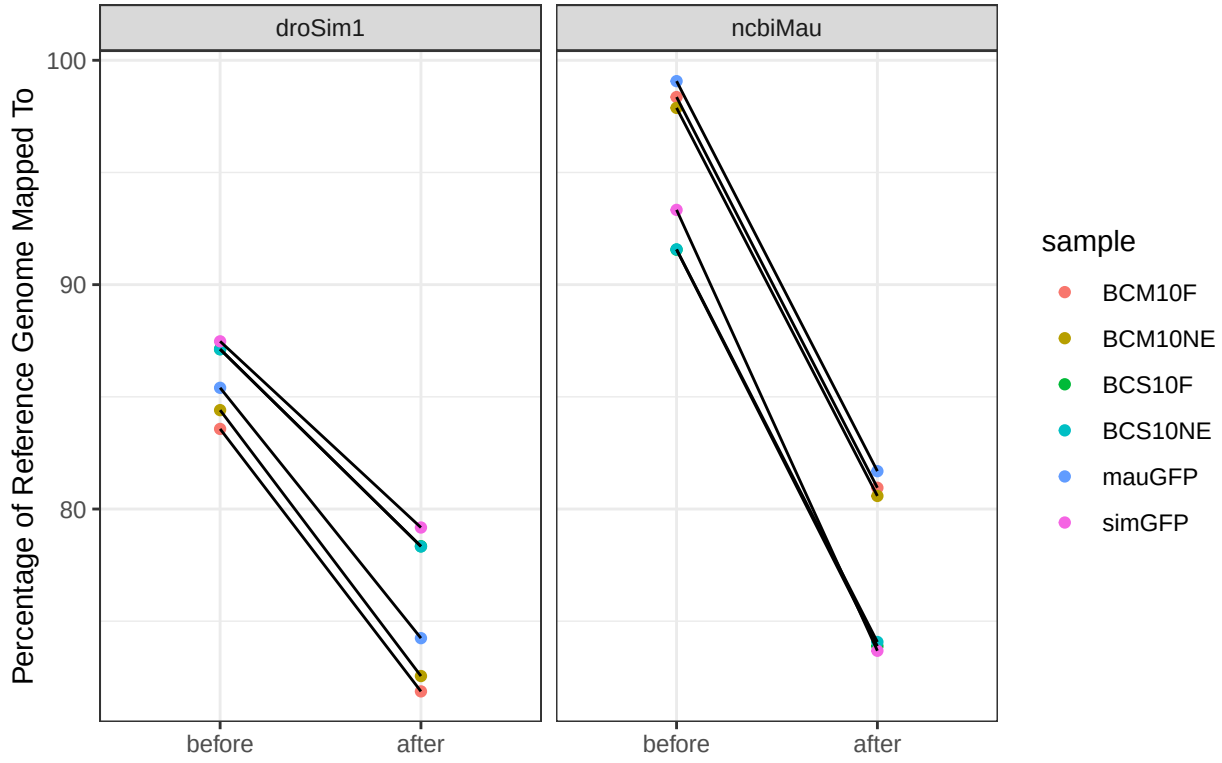
Table 8. Depth of Coverage Statistics for Raw and Filtered Alignments
bwa, bwaUniq

reference	minimum	average	median	maximum
pre-filtration depth				
droSim1	15.7	40.7	25.2	82.7
ncbiMau	17.7	40.7	23.2	85.3
post-filtration depth				
droSim1	11.1	28.1	18.6	58.2
ncbiMau	15.0	30.6	17.1	68.9
depth retention percent				
droSim1	64.2%	70.2%	70.6%	74.3%
ncbiMau	66.0%	76.6%	77.5%	85.2%



Breadth of coverage, ie, the percentage of the genome covered by at least one read:

Figure 6. Depth Of Coverage for Raw and Filtered Alignments



2.5 Called Variants

BWAUniq mappings were used to jointly call variants in VCF format via Freebayes (Garrison and Marth 2012) using standard filters.

expand on smrtFreeBayes

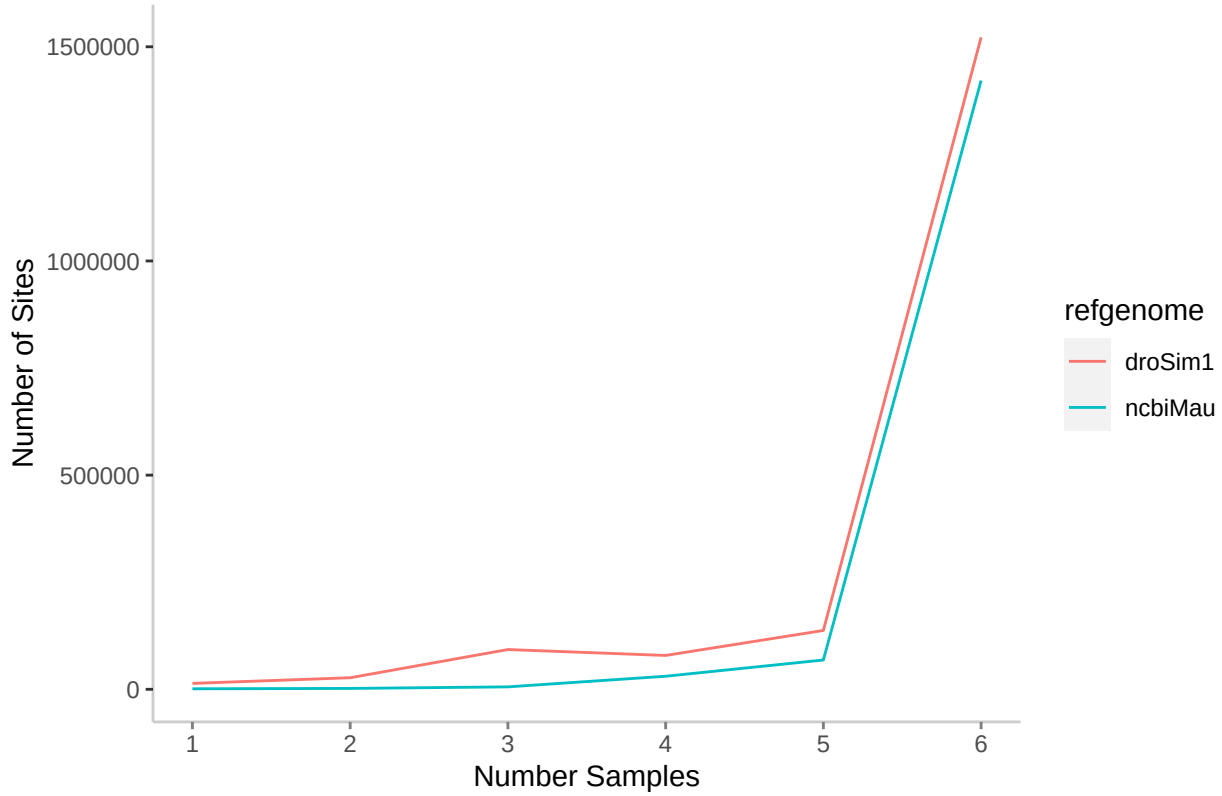
Table 10. SNP count and per-KB SNP rate across all samples

variant caller	refGenome	Genome Size (bp)	# SNPs	SNP rate (per kb)
smrtFreeBayes	droSim1	$142.4M$	$1.6M$	11.1
stdFreeBayes	droSim1	$142.4M$	$1.9M$	13.1
stdFreeBayes	ncbiMau	$152.3M$	$1.5M$	10.0

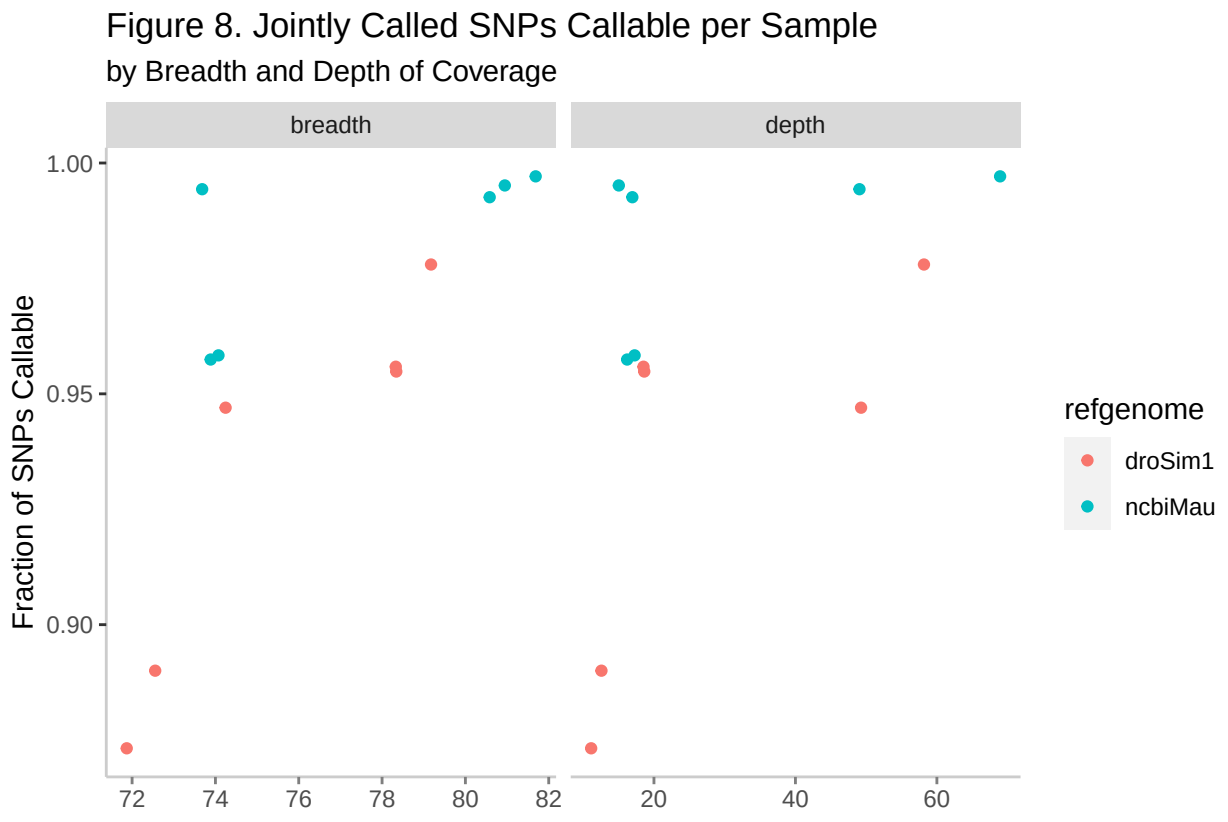
To build this VCF, 6 samples called jointly. However, not all sites were called in all samples (eg, due to coverage differences). The sites had the following group-wide call rate:

```
## Warning: Removed 4 rows containing missing values (‘geom_path()‘).
```

Figure 7. Histogram of SNPs by Number of Samples Called At Site



The fraction of jointly called SNPs which are individually callable:



2.6 PsiSeq Algorithm Details

2.6.1 PsiSeq Classic & PsiSeq2

The first two versions of PsiSeq were similar: reads are aligned to a reference genome and the alignments are compared directly using the pileups. Sites were identified which differed in the parent alignments; ancestry was inferred at each such site in the offspring, according to which parents' allele was present, and assigned a 1 for one parent and a zero for the other. These were then averaged by window: a window with a high average score will be enriched in ancestry from parent 1.

The PsiSeq1 workflow published in Earley and Jones (2011), PsiSeq, uses a perl script to directly compare read alignments (as mpileup files) between "control" and selected treatments. The scripts included have minor modifications over the original, such as handling edge cases without error.

The original algorithm from Earley and Jones (2011) used mpileups as input; in this way it compares the alignment of the backcrossed hybrid and of one parent species, against the other parent species' reference genome.

Each line of an mpileup is individually examined. For sites meeting basic criteria (eg, the reference base is defined), the aligned bases at the corresponding site are tallied and the base with the most supporting reads is considered the genotype at that site:

The hybrid alignment is first examined, and a dictionary built containing these site calls. Next, the parent alignment is examined. The genotype at each site is called; if the hybrid has a genotype recorded at this site in the dictionary, they are compared and the site is scored 1 if they are the same and zero otherwise. Notably, if the hybrid has no such site recorded, the site is scored zero.

PsiSeq2 was written as an update the the earlier software. This included reorganizing it into a Snakemake (Rahmann and Ko (2017),) workflow and shifting the emphasis away from simulated reads. It also aimed to give more nuanced and finely controlled comparison of the alignments. However, its development was ultimately overtaken by PopPsiSeq.

2.6.2 PopPsiSeq and Allele Frequency Shift

The alignment comparison scripts of PsiSeq 1 and 2 are essentially variant callers which identify sites which mismatch the reference genome to a large degree. This was reasonable for its time but has been rendered obsolete by the development of more sophisticated variant calling algorithms; PopPsiSeq was built around FreeBayes (Garrison and Marth (2012)). This allows the comparison of groups (eg, replicates of a treatment) whereas earlier comparisons were on an individual basis.

Earlier comparisons were also based on the presence or absence of a fixed variant. By working on a population level, PopPsiSeq is able to use difference in allele frequency between groups (of which fixation is an extreme case). This will hopefully increase statistical power and allow examination of eg polygenic traits.

Once the SNPs were called, the VCF file was split into subsets. The simGFP and mauGFP variants which are treated as ancestral populations. Other subsets represent experimental treatments, such as all simulans backcrosses, all needle-eye mutants, or the one needle-eye simulans backcross.

For each SNP still meeting minimum requirments (biallelic, at most one missing sample)<- *clean this up*, the subgroup-wide allele frequency was calculated. Using the simGFP and mauGFP as ancestral, the distance to the simulans frequency and the mauritiana frequency calculated for each SNP, for each subset. The per-window average shift was then calculated.

Here is a hypothetical example: suppose that at a given site in the genome, 75% of alleles in the mauritiana population are T and 25% are A. Suppose in the simulans population, it's 25% T and 75% A. Now, the allele frequency is tallied in three different subgroups:

In the first subgroup, 100% of alleles are T. This subgroup would have a mau-ward shift of +0.25 and a sim-ward shift of -0.75.

In the second subgroup, 50% of alleles are T. This subgroup would have a mau-ward shift of -0.25 and a sim-ward shift of -0.25.

In the third subgroup, 0% of alleles are T and all are A. This subgroup would have a mau-ward shift of -0.75 and a sim-ward shift of +0.25.

2.7 Gene Lists

A curated list of genes of interest was provided by Amanda Moehring (5 Dec 2023):

```
Lgr4 (CG34411)
Bap60 (CG4303)
DNAlig4 (CG12176)
Fer3HCH (CG4349)
Nna1 (CG44533)
CG2692
CG10996
lncRNA:CR45622 (CR45622)
gce (CG42739)
mh (CG9203)
CG32820
CG32819
Dhc16F (CG7092)
```

CG15373, CG17450
tilB (CG14620)
p-cup (CG12993)
pcm (CG3291)
r-cup (CG10998)
Nup153 (CG4453)
Ulp1 (CG12359)
wupA (CG7178)
Ste (FBgn0003523)
Ada3 (CG7098)
CG15446
CG4318

Of these, most were easily converted from melanogaster to simulans, using www.orthodb.org, and their gene identifier in the NCBI annotation retrieved from https://www.ncbi.nlm.nih.gov/gene/:

```
Lgr4 (CG34411)    GD15902   LOC6725833
Bap60 (CG4303)    GD15903   LOC6725831
DNAlig4 (CG12176)       GD27483   LOC27207332
Fer3HCH (CG4349)        GD15908   LOC6725819
Nna1 (CG44533)      GD17141   LOC6725884
CG2692      GD24898   LOC6736171
CG10996     GD15872   LOC6725893
gce (CG42739)       GD17204   LOC6726007
mh (CG9203      27209153    LOC27209153
Dhc16F (CG7092)     GD24683   LOC6739854
CG15373     GD17382   LOC6726322
tilB (CG14620)      GD10326   LOC6733227
p-cup (CG12993)     GD17357   LOC6726287
pcm (CG3291)        GD27168   LOC27207018
r-cup (CG10998)     LOC6727351    LOC6727351
Nup153 (CG4453)     GD24841   LOC6740169
wupA (CG7178)       GD24482   LOC6740188
Ada3 (CG7098)       GD27178   LOC27207028
CG15446     GD24437   LOC6740264
CG4318      GD15906   LOC6725826
```

A handful of genes were refractory: `lncRNA:CR45622` did not appear to have expression or a known transcript in simulans. Stellate, `Ste`, appears to be a family rather than a single gene (use the family as a gene list of its own?) `Ulp1` seems to be associated with two different genes in simulans; `LOC6726423` looks to be the right one. `CG32820`,`CG32819`,`CG17450` are tandem duplicates which are all listed as orthologs of the simulans `LOC6726613`; only one is included in the final list. Finally, gooseberry-neuro, `CG2692`, was found to be on chr2R in melanogaster; so was its ortholog in simulans. The list is meant to be genes on chrX with expression in male reproductive tissue, so this gene was excluded.

```
LOC6725833  GD15902 Lgr4(CG34411)
LOC6725831  GD15903 Bap60(CG4303)
LOC27207332 GD27483 DNAlig4(CG12176)
LOC6725819  GD15908 Fer3HCH(CG4349)
LOC6725884  GD17141 Nna1(CG44533)
LOC6725893  GD15872 CG10996
LOC6726007  GD17204 gce(CG42739)
LOC27209153 27209153    mh(CG9203
LOC6726613  GD15491 CG32820
LOC6739854  GD24683 Dhc16F(CG7092)
LOC6726322  GD17382 CG15373
```

```
LOC6733227  GD10326 tilB(CG14620)
LOC6726287  GD17357 p-cup(CG12993)
LOC27207018 GD27168 pcm(CG3291)
LOC6727351  LOC6727351  r-cup(CG10998)
LOC6740169  GD24841 Nup153(CG4453)
LOC6726423  120285358/GD15595   Ulp1(CG12359)
LOC6740188  GD24482 wupA(CG7178)
LOC27207028 GD27178 Ada3(CG7098)
LOC6740264  GD24437 CG15446
LOC6725826  GD15906 CG4318
```

With this gene list in hand, the corresponding genomic locii were extracted by pulling from the gene annotation GTF these genes and in particular the lines with the "gene" tag in the feature field. These were converted to BED format and extended by 10kb in each direction using the bedtools (Quinlan and Hall 2010) slop utility. These locii were used as intervals for averaging the PopPsiSeq frequency the way windows are for whole-genome scans. To sample the background genomic distribution, a gene list's locii were shuffled across their containing contigs without overlap to generate pseudolocii intervals; ie,

```
bedtools shuffle -chrom -noOverlapping -i {input.gene_bed} -g {fai} > {output.shuff_bed}
```

Ten such shuffles were generated per list. The genetic background was generated similarly, by shuffling the gene annotation and picking the first genes with the same chromosome distribution.

3 Results

3.1 ~~PsiSeq Classic~~

~~Above is the PsiSeq1 analysis, in which all samples have been compared to simGFP. The vertical axis is the fraction of the sites where the simGFP sample differs from the reference genome, and the sample under analysis also shares this variant allele. As expected, simGFP has a high similarity with itself across the genome, whereas mauritiana and mauritiana backcrosses have low similarity.~~

~~Here's a closer look, restricted to the 3L chromosome and just the simulans backcrosses. Two features stand out: a region ~8-10Mb where the wild-type backcross has decreased similarity to simGFP, and a region ~15-20 Mb where the needle-eye backcross has decreased similarity relative to the wild-type backcross.~~

~~Finally, here are the comparisons with both simGFP and mauGFP. It appears that only the second feature near the end of the chromosome arm is the decrease in simulans similarity accompanied by an increase in mauritiana similarity.~~

3.2 ~~PsiSeq2~~

~~Above is the PsiSeq2 analysis, in which all samples have been compared to simGFP. The vertical axis is the fraction of the sites where the simGFP sample differs from the reference genome, and the sample under analysis also shares this variant allele. As expected, simGFP has a high similarity with itself across the genome, whereas mauritiana the backcrosses' similarity is closer to that of mauGFP. The simulans backcrosses have two distinct behaviors: a "background" in which their similarity is close to that of simGFP, and a "disturbance" on chr3R and intervals on chr2L and chr2R; these generally correspond to similar regions in the PsiSeq1 analysis. As in PsiSeq1, the wild-type backcross is more similar to droSim1 than the needle-eye backcross. What is notable about this analysis is that the backcrosses are actually less similar to droSim1 than mauGFP in the disturbance regions.~~

~~Another feature worth mentioning is the deflection on chr2L, ~13-15Mb. This also appears in PsiSeq1 though it is much less pronounced. This looks very similar to one of the peaks which was originally identified in Earley~~

~~and Jones (2011) as a region of interest re: sechellia and simulans speciation; however on closer inspection its actual origins were murky and although it could well be of interest might not mean what we originally thought.~~

~~Here's a closer look, restricted to the 3L chromosome and just the simulans backcrosses. As in the PsiSeq1 analysis, there is a region ~15-20 Mb where the needle-eye backcross has decreased similarity relative to the wildtype backcross; however, both backcrosses experience a drop in similarity below that of the mauGFP control.~~

~~There is no dramatic deflection in the ~8-10Mb like in the PsiSeq1; however, there is what looks like a local anomaly in the mauGFP genome, where there is an extended interval of heightened similarity to droSim1. This may be causing an artefact in PsiSeq1.~~

~~The low similarity of the backcrosses to both parental strains are explained by the fact that PsiSeq2 requires high agreement between reads (ie, it requires the sites to be fixed in both parents and the offspring) whereas the classic algorithm would sample stochastically. However, if we look at the fraction of sites which are heterozygous in each sample, there is still considerable heterozygosity in the hybrids and to a lesser extent the simulans F0~~

3.3 Allele Frequency Shift (PopPsiSeq)

Allele frequencies were calculated for all samples and the difference between backcross AF and sim/mauGFP AF was calculated, then summed and averaged by 100kB window.

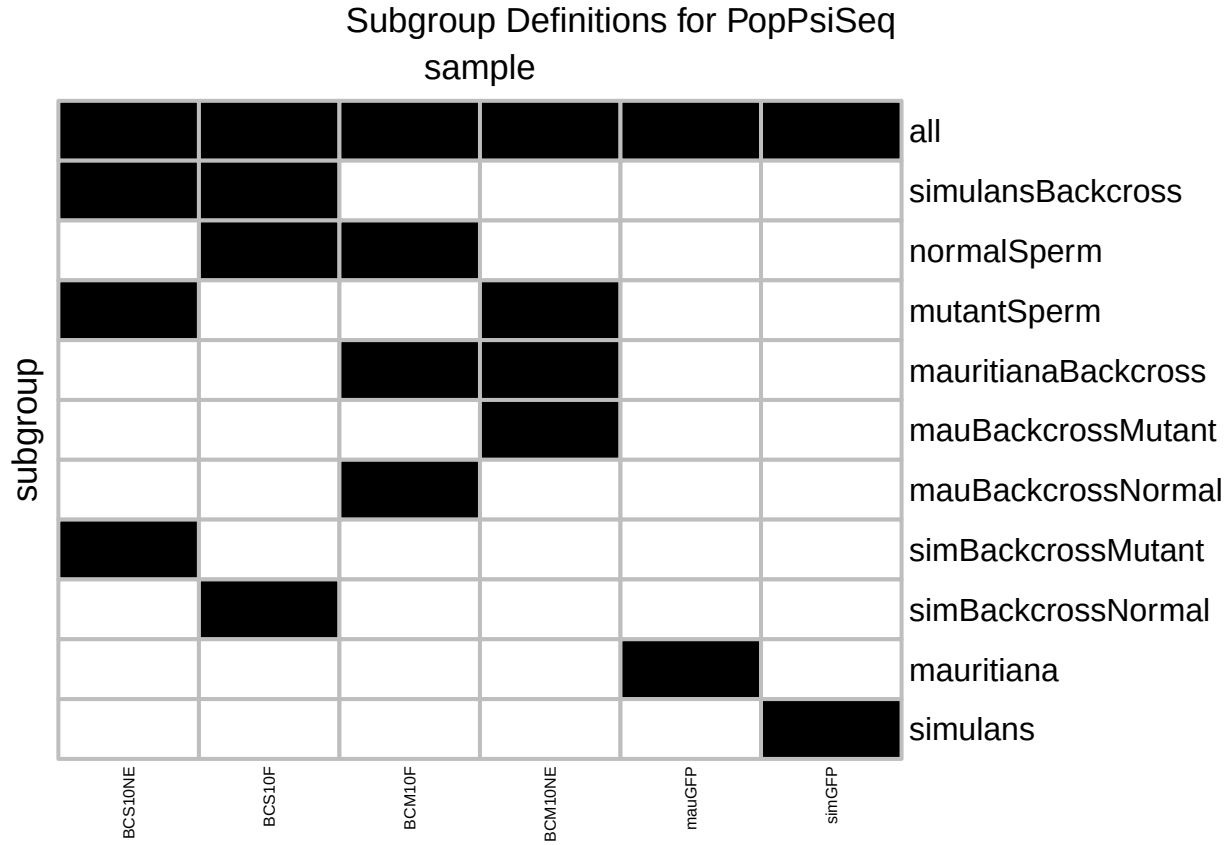
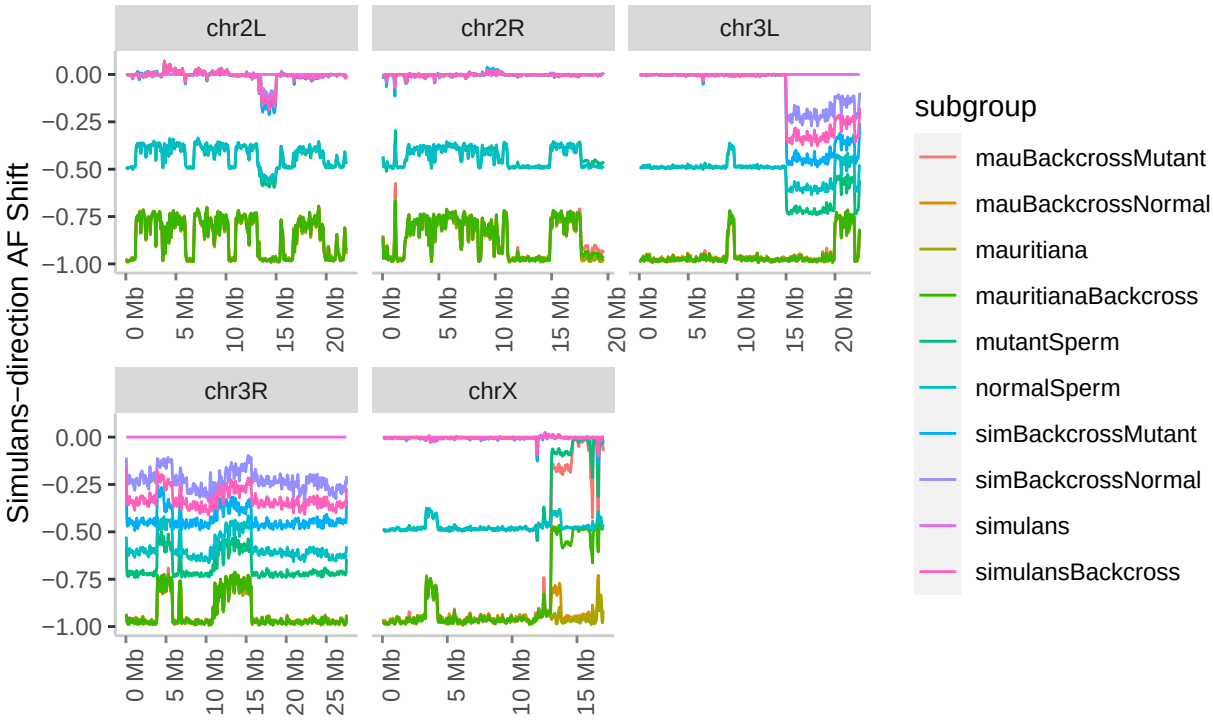


Figure 16. Shift in Allele Frequency

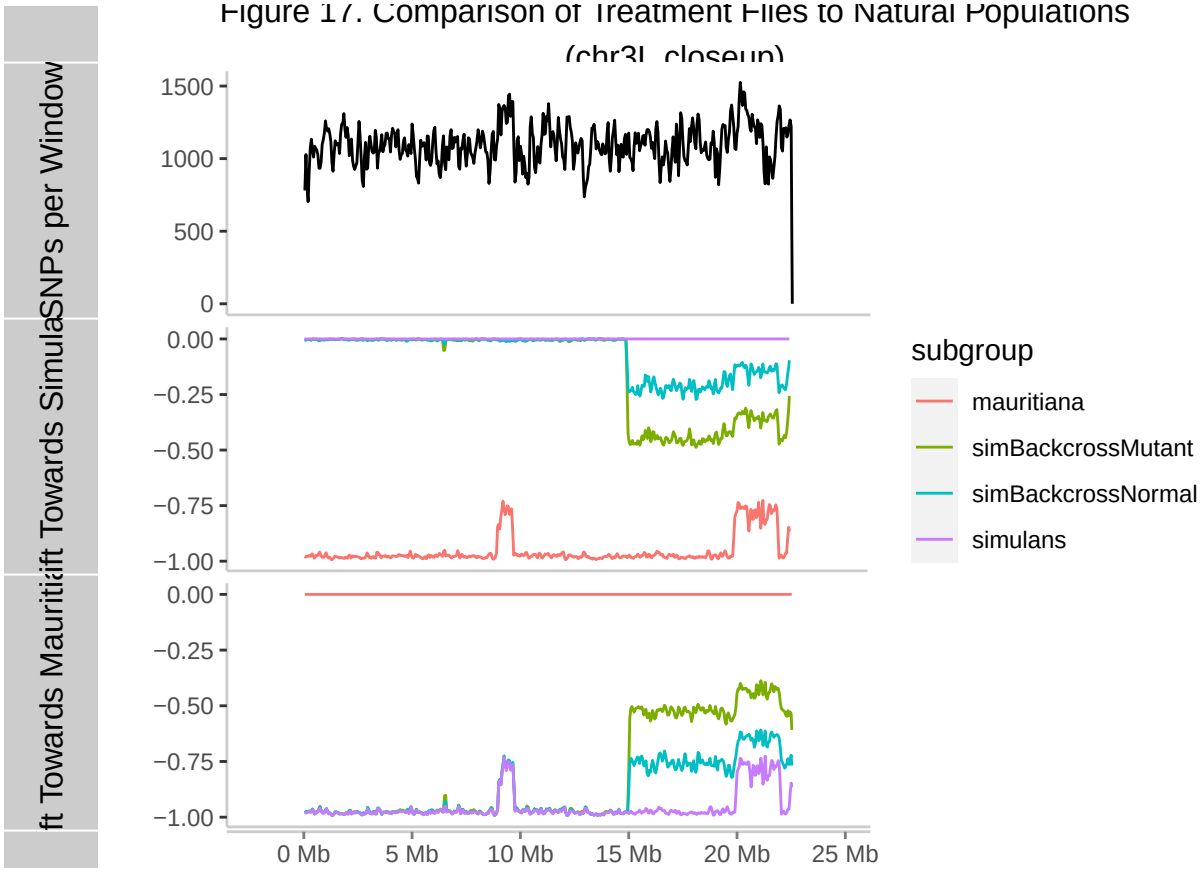
Towards Simulans Allele




```
## pdf
##   2
```

Above is the output of the PopPsiSeq analysis for all subgroups. There's a lot going on, so let's take a closer look at the simulans backcrosses on chr3L

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
```




```
## pdf
##   2
```

Here's the chr3L arm with the simulans backcrosses, simulans, and mauritiana. The top panel shows the number of informative SNPs per window. The middle panel shows the average shift towards the simulans allele frequency. The simGFP control in purple is at a baseline of zero, because it's already at the simulans allele frequency! The change in AF for the mauGFP control (red) is negative, as expected. For the most part it is near a value of -1, indicating that most of the SNPs in these region have one allele fixed in simulans and the other allele fixed in maurtiana. There are segments where the mauGFP is elevated above the -1 boundary; this indicates regions where there are many SNPs which have different allele frequencies between simulans and mauritiana, but aren't fixed in both. It is possible that a sample could be lower or higher than one of the parent strains, if alleles are not fixed in one or both, and alelle frequency becomes more extreme than one or the other (eg, if simulans alleles are specifically selected for they might rise to a higher frequency than in the unselected population) However, this is not observed here.

The bottom panel is interpreted similarly, except the change in allele frequency is towards that of MauGFP.

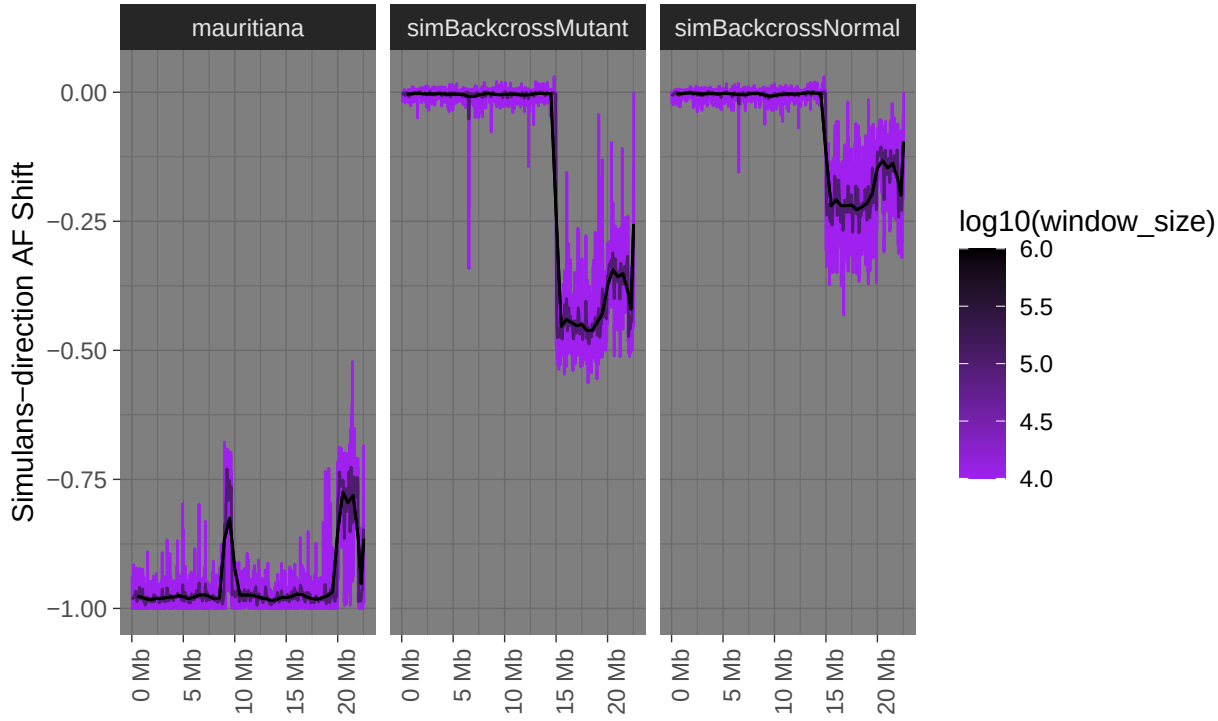
The simulans backcrosses both show values bounded by the simGFP and mauGFP. For most of the chromosome, both backcrosses are mauritiana-like: there would be close to zero change in allele frequency to make the alleles match mauGFP, and the alles would have to be essentially reversed to match simGFP. The same general region ~15-20Mb stands out as in the other methods. Here, both backcrosses are somewhat more simulans-like. The needle-eye mutant is the more simulans-like of the two, being almost heterozygous on average between simGFP and mauGFP; the wildtype is has about half as much simulans character.

There is again an artefact ~8-10Mb, apparently a reduction of allele fixation in one or both of simGFP and mauGFP. The backcrosses behave identically in this blip.

3.3.1 Window Parameters

Figure 18a. Impact of Smoothing Parameters: Genome Window Size

shift averaged over rolling windows 10kB,100kB,500kB, and 1MB wide

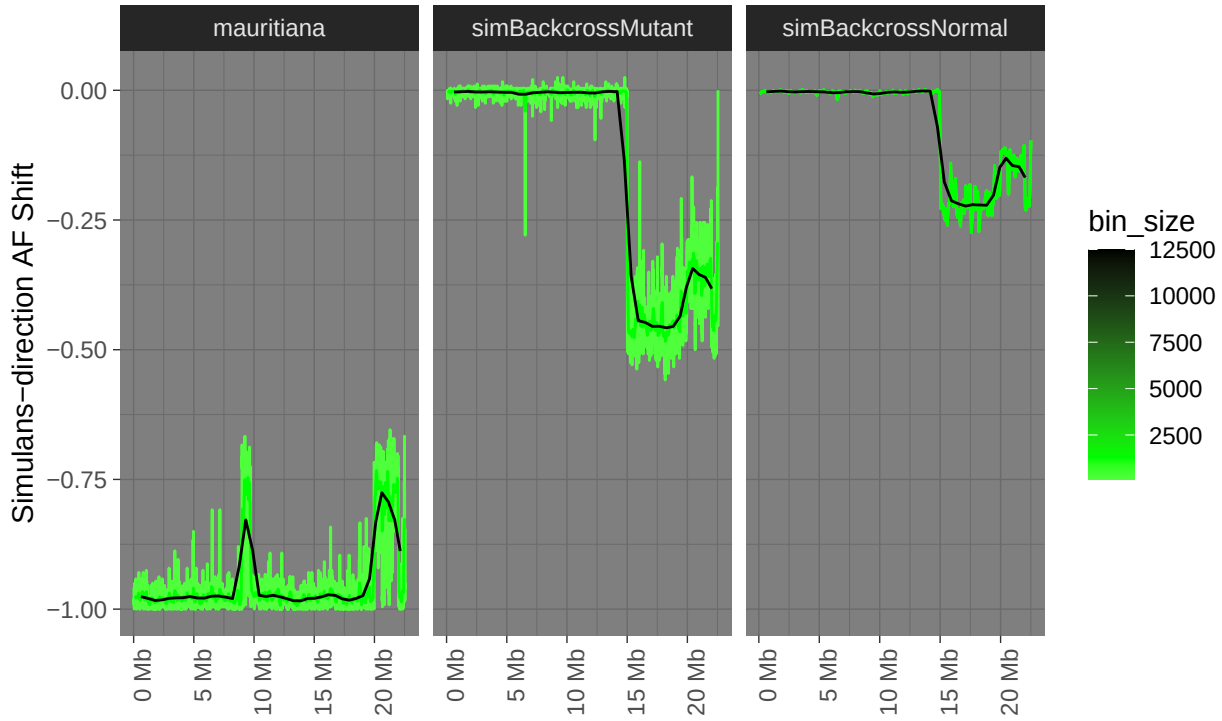



```
## pdf
##   2
```

at ~12.5 Snp per kb, the standard window is about 1250 SNPs wide...

Figure 18b. Impact of Smoothing Parameters: SNP Bin Size

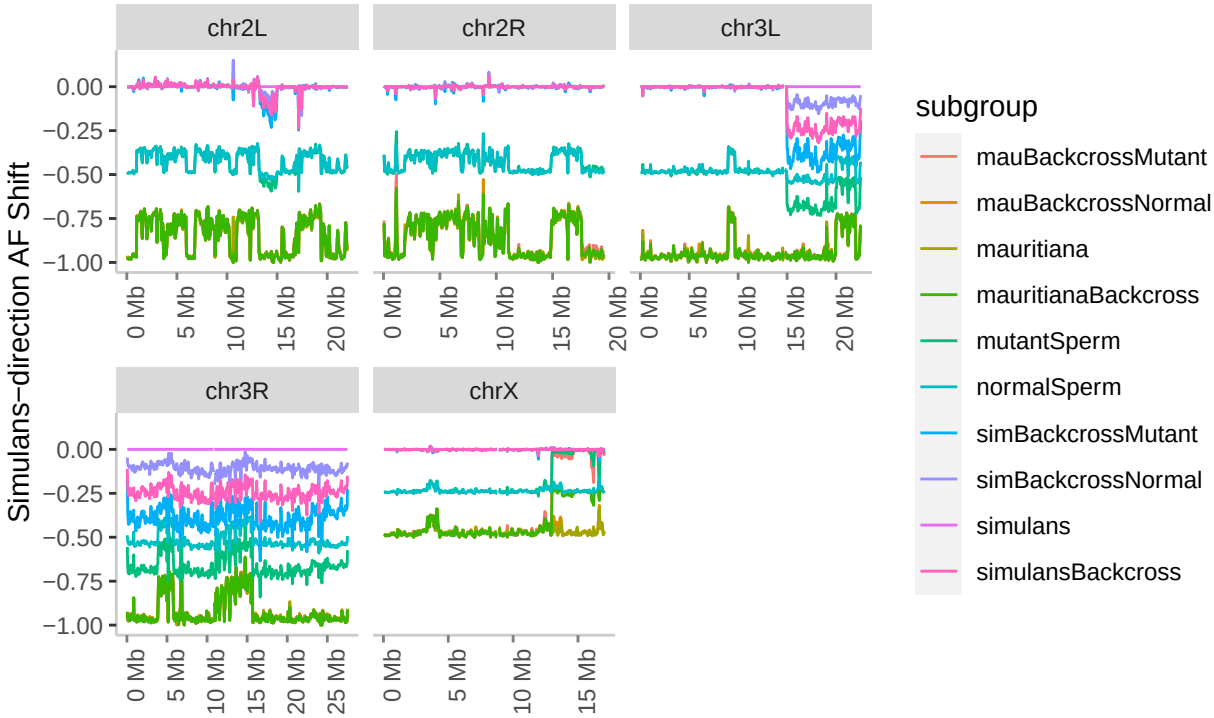
shift averaged over rolling bins 120,1250,12500 SNPs wide




```
## pdf
##   2
```

3.3.2 smrtFreebayes

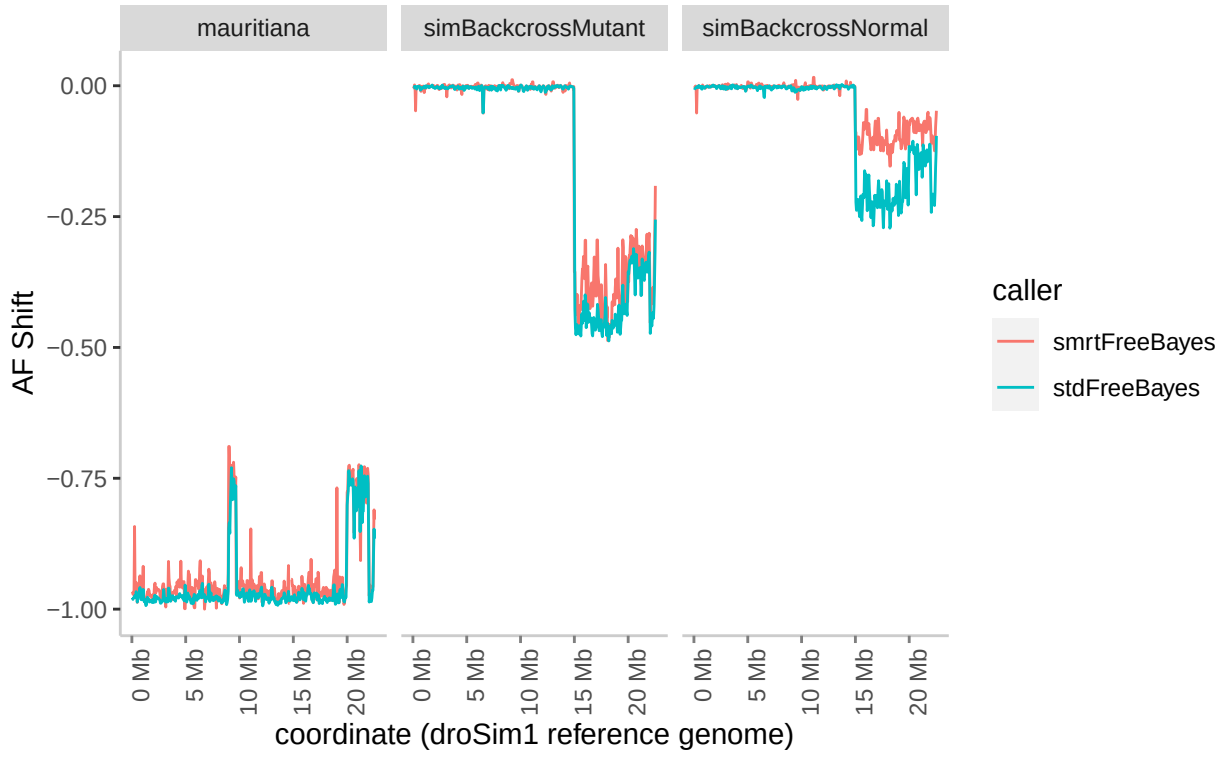
Figure 19. Shift in Allele Frequency (smrtFreeBayes)
Towards Simulans Allele




```
## pdf
##   2
```

Figure 19. PopPsiSeq: Impact of Variant Calling Strategy

Allele Frequency Shift towards SimGFP




```
## pdf
##   2
```

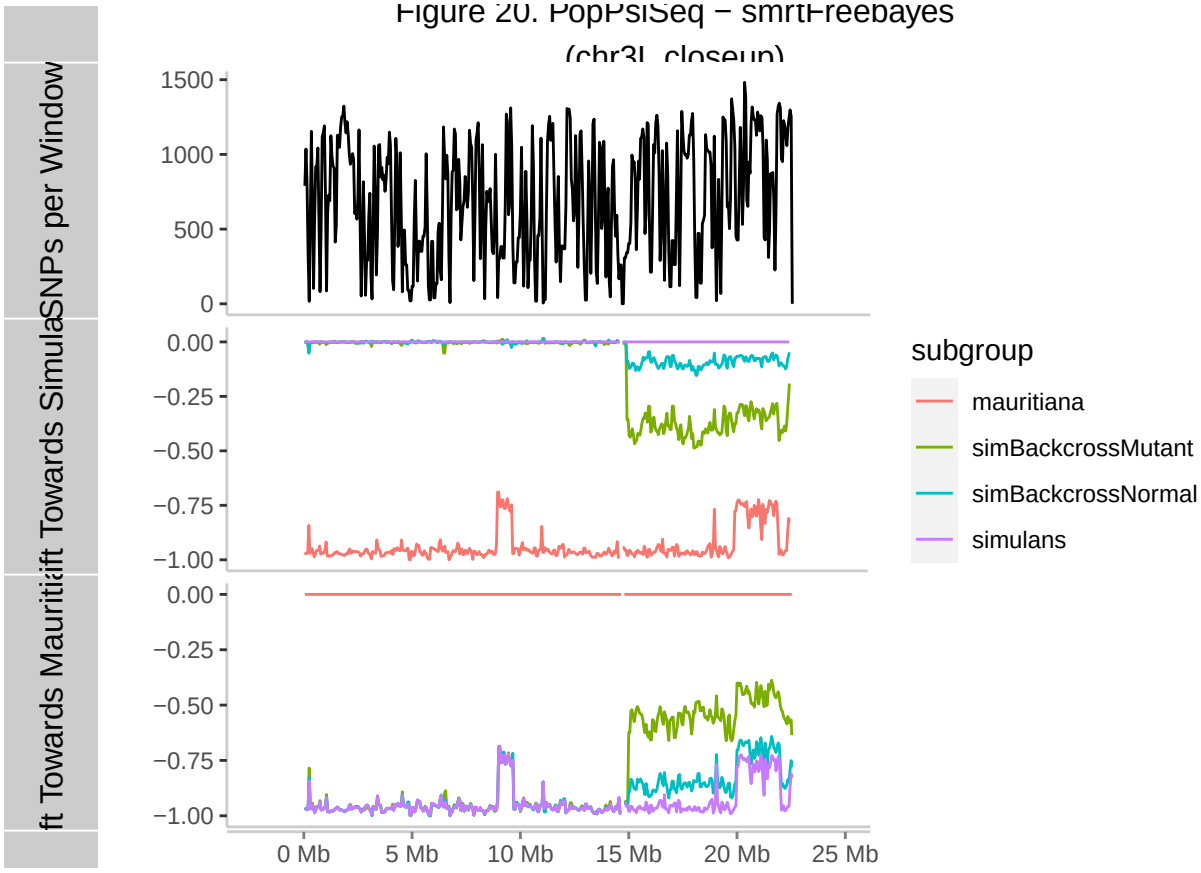
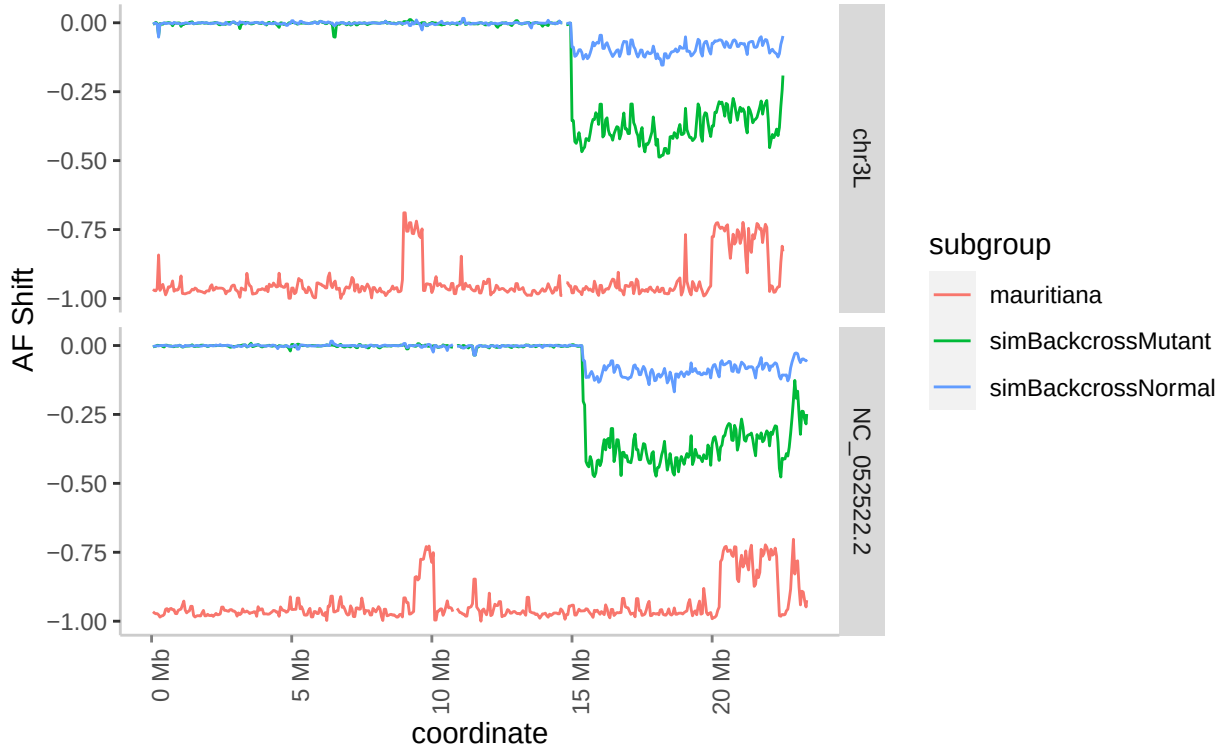


Figure 20: PopPsiSeq − smrtFreebayes
(chr3L closeup)
SNPs per Window
ft Towards Simula
ft Towards Mauriti
1500
1000
500
0
0.00
−0.25
−0.50
−0.75
−1.00
0 Mb
5 Mb
10 Mb
15 Mb
20 Mb
25 Mb
subgroup
mauritiana
simBackcrossMutant
simBackcrossNormal
simulans

```
## pdf
##   2
```

3.3.3 Moving to Princeton Reference Genome

Figure 21. PopPsiSeq: Comparison of Simulans Reference Genomes (drc

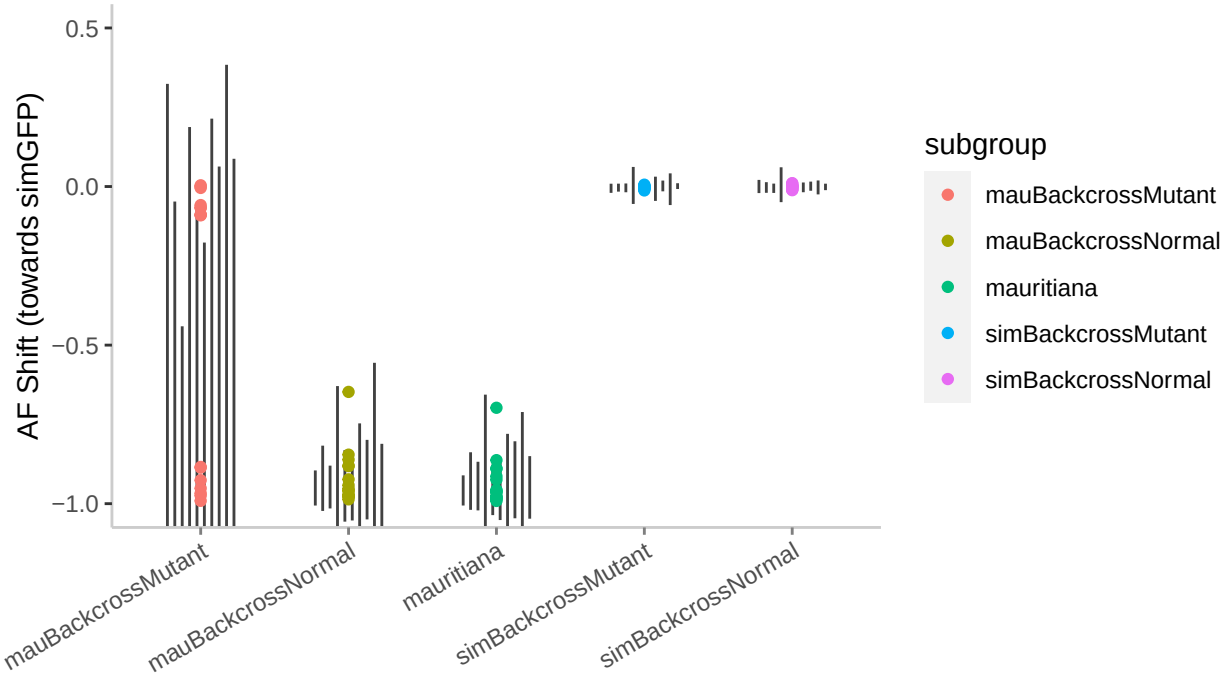



```
## pdf
##   2
```

3.4 Gene list

Figure 22. Allele Frequency Shift for Genes of Interest

curated list of genes on the X with expression in male sex tissue

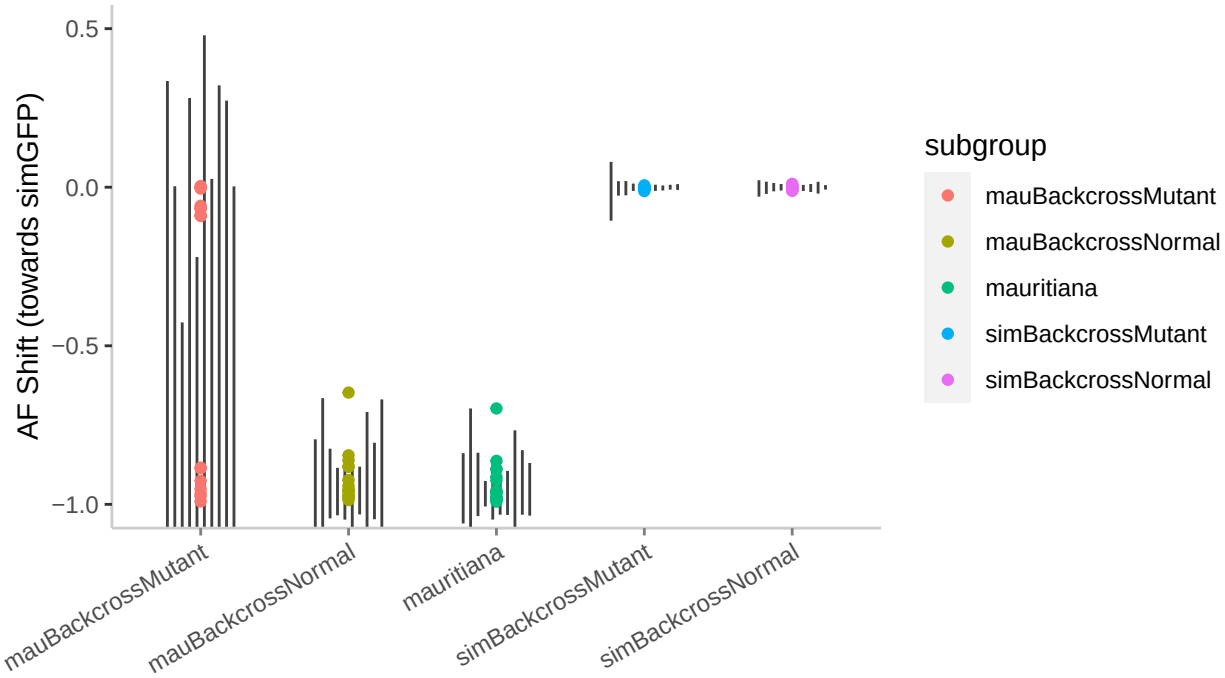


with 10 replicates of shuffled non−overlapping, same−size intervals

```
## pdf
##   2
```

Figure 22. Allele Frequency Shift for Genes of Interest

curated list of genes on the X with expression in male sex tissue



with 10 replicates of resampled genes from annotation

```
## pdf
##   2
```

4 Discussion

5 References

5.1 Software

```
##
## Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R,
## Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E,
## Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi
## K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the
## tidyverse." _Journal of Open Source Software_, *4*(43), 1686. doi:
## 10.21105/joss.01686 (URL: https://doi.org/10.21105/joss.01686).
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Welcome to the {tidyverse}},
##     author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostin
##     year = {2019},
##     journal = {Journal of Open Source Software},
```

```
##     volume = {4},
##     number = {43},
##     pages = {1686},
##     doi = {10.21105/joss.01686},
##   }
##
## To cite the 'knitr' package in publications use:
##
##   Yihui Xie (2023). knitr: A General-Purpose Package for Dynamic Report
##   Generation in R. R package version 1.42.
##
##   Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition.
##   Chapman and Hall/CRC. ISBN 978-1498716963
##
##   Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible
##   Research in R. In Victoria Stodden, Friedrich Leisch and Roger D.
##   Peng, editors, Implementing Reproducible Computational Research.
##   Chapman and Hall/CRC. ISBN 978-1466561595
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.
##
## To cite package 'yaml' in publications use:
##
##   Shawn P Garbett, Jeremy Stephens, Kirill Simonov, Yihui Xie, Zhuoer
##   Dong, Hadley Wickham, Jeffrey Horner, reikoch, Will Beasley, Brendan
##   O'Connor, Gregory R. Warnes, Michael Quinn and Zhian N. Kamvar
##   (2023). yaml: Methods to Convert R Data to YAML and Back. R package
##   version 2.3.7. https://CRAN.R-project.org/package=yaml
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {yaml: Methods to Convert R Data to YAML and Back},
##     author = {Shawn P Garbett and Jeremy Stephens and Kirill Simonov and Yihui Xie and Zhuoer Dong a
##     year = {2023},
##     note = {R package version 2.3.7},
##     url = {https://CRAN.R-project.org/package=yaml},
##   }
##
## ATTENTION: This citation information has been auto-generated from the
## package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

O. Tange (2018): GNU Parallel 2018, Mar 2018, ISBN 9781387509881, DOI https://doi.org/10.5281/zenodo.1146014

Bibliography

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. "Fastp: An ultra-fast all-in-one FASTQ preprocessor." *Bioinformatics* 34 (17): i884–90. https://doi.org/10.1093/bioinformatics/bty560.

Earley, Eric J., and Corbin D. Jones. 2011. "Next-generation mapping of complex traits with phenotype-based selection and introgression." *Genetics* 189 (4): 1203–9. https://doi.org/10.1534/genetics.111.129445.

Garrison, Erik, and Gabor Marth. 2012. "Haplotype-based variant detection from short-read sequencing," July. http://arxiv.org/abs/1207.3907.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A flexible suite of utilities for comparing genomic features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Rahmann, Sven, and Johannes Ko. 2017. "Snakemake a scalable bioinformatics workflow engine" 28 (19): 2520–22. https://doi.org/10.1093/bioinformatics/bts480.