PsiSeq2: The Sequel

Charlie Soeder

2/26/2018

Introduction

PSISeq (Earley and Jones 2011) is a method for uncovering the genetic basis for a phenotype. The basic procedure is to select for a trait while back-crossing to a genomic background lacking that trait, then sequencing and comparing the offspring genome to the parental lines. In particular, this is applied to the drosophila sechellia/simulans group, which can form hybrids and have complex behavioral differences.

Here, the original code and data are repackaged, and tested on several new sequenced lines.

Materials & Methods

data:

- Reference Genomes
  - droSim1, droSec1 (UCSC Genome Browser)
  - dsimr2, dsecr1 (FlyBase)
- DNA-Seq reads
  - (Earley and Jones 2011) reads from selection experiment (SRR303333)
  - New reads frm selection experiment: three biological replicates (10,13,17) with two technical replicates each (A,B)
  - Synthetic reads simulated from the reference genomes using ART
  - Control Sechellia reads from Rich (SucSec) and NCBI (SRR869587, SRR6426002, SRR5860570)
  - Simulans reads from Rich (SucSim)
- Software
  - short-read mappers (BWA, NGM)
  - read simulation (ART)
  - variant calling (Freebayes)
  - misc file manipulation & processing (samtools, bedtools, bedops, vcftools, UCSC liftover)
  - custom python scripts
  - Snakemake workflow
  - R utilities (Markdown, rtracklayer, ggbio, tidyverse)
  - Upstream QA/QC (FASTQC, FASTX-Toolkit)

Include info about depth, read number, etc?

methods:

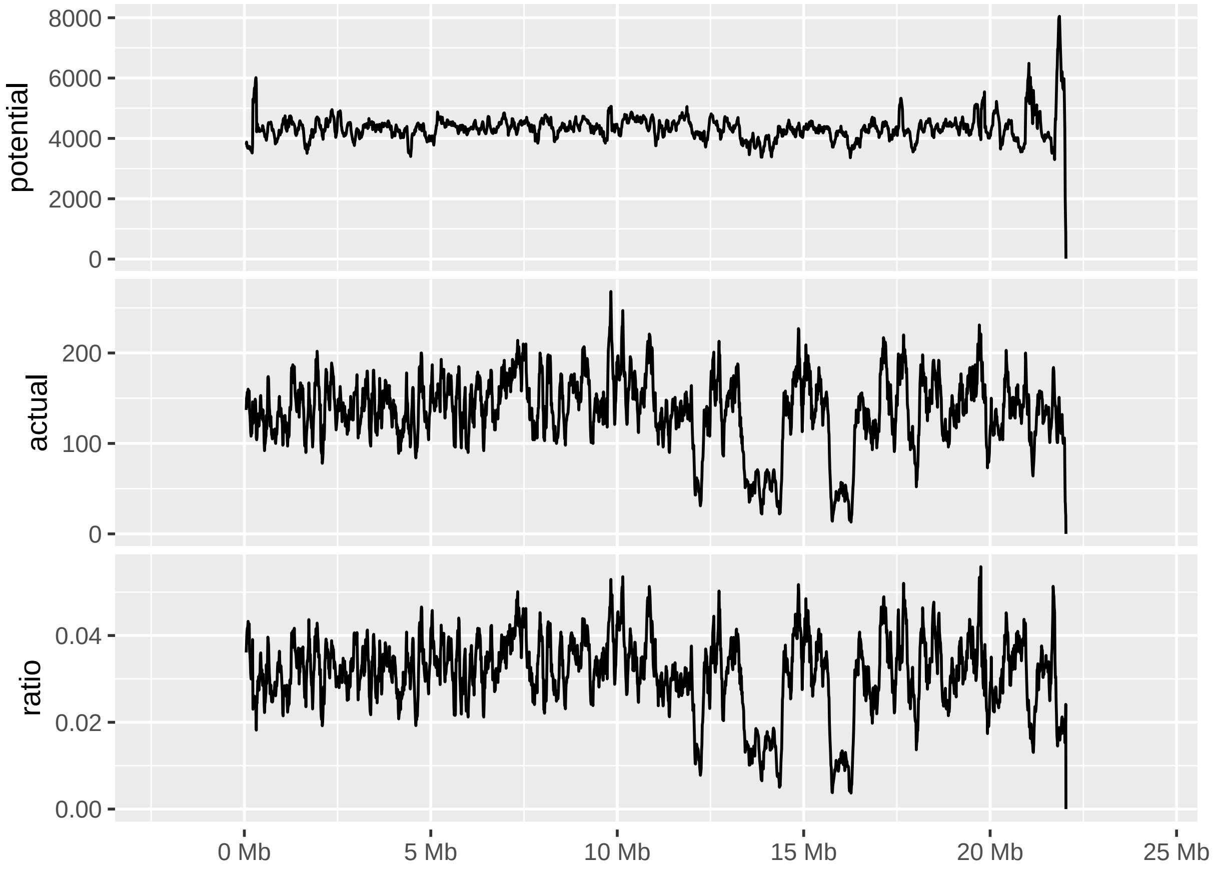
- Basic PSI-Seq algorithm reimplemented and applied to test data
- Various diagnostics run on a artefact on chr2L

Selected lines were produced by mating sechellia with simulans, selecting for avoidance of morinda fruit (ie, for simulans-like behavior), then backcrossing to sechellia.

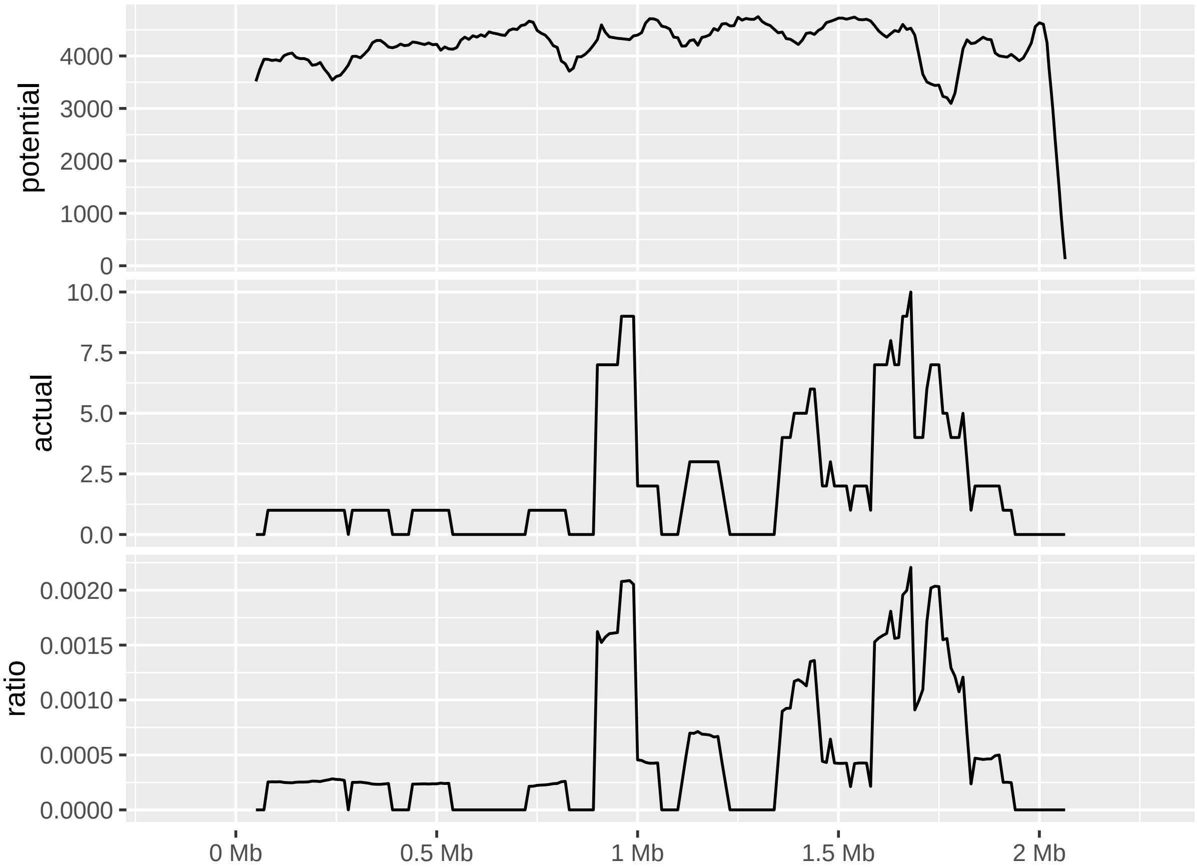
Results

Basic Output

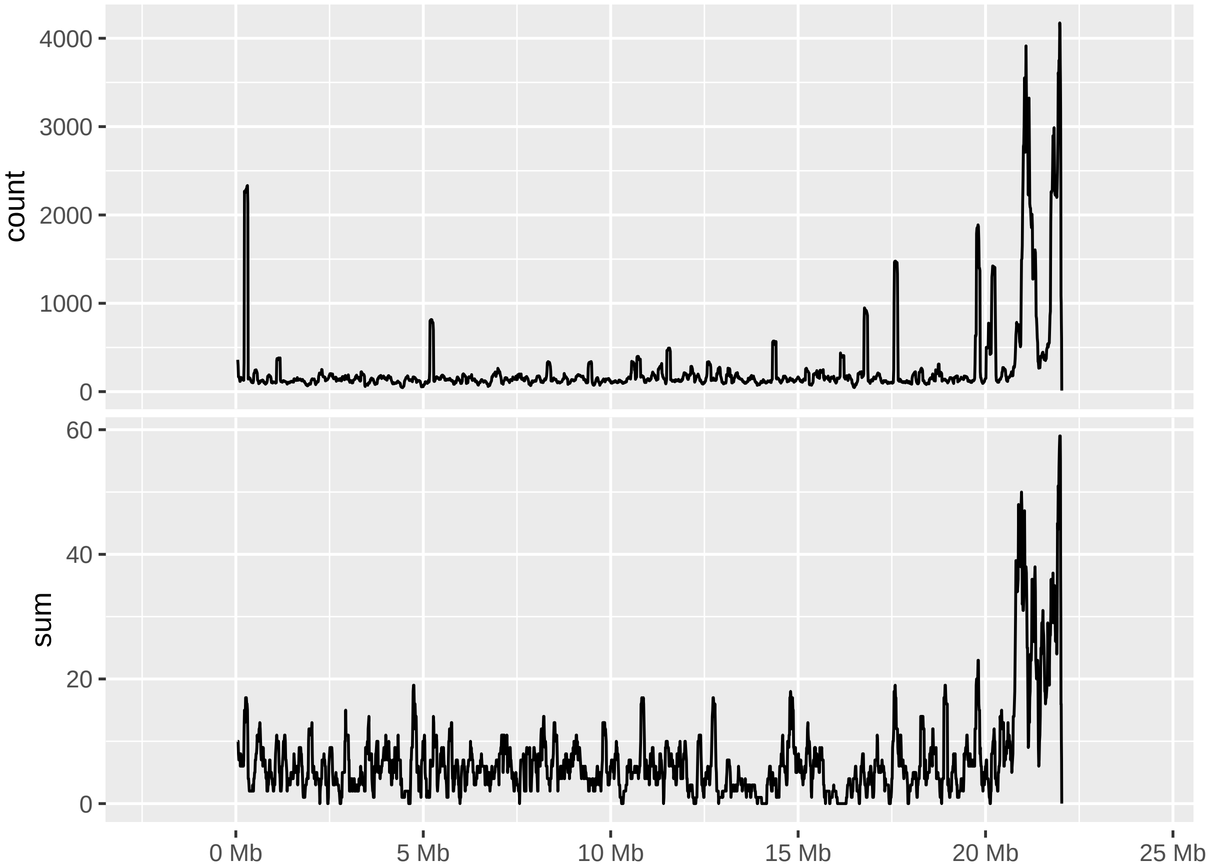
In the first pass of reanalysis, a selected line was mapped to the (UCSC) reference genomes, and the pileups compared. A site was considered to be a “potential” informative SNP if the parental pileup contains a mismatch there. If the offspring shares the same mismatch, that SNP is called “actual”. The genome is split into sliding windows (100kb wide, slid by 10kb), the potential and actual SNPs are tallied per window, and their ratio calculated. This gives a measure of divergence from the reference genome being mapped to.
For example, here is the line mapped to the droSim1 reference genome; synthetic reads were also simulated from the droSec1 reference, since the parent flies themselves were not sequenced. The vertical axis represents SNP count vs droSim1; more shared SNPs between the parental line and the backcrossed line suggests more sechellia character.



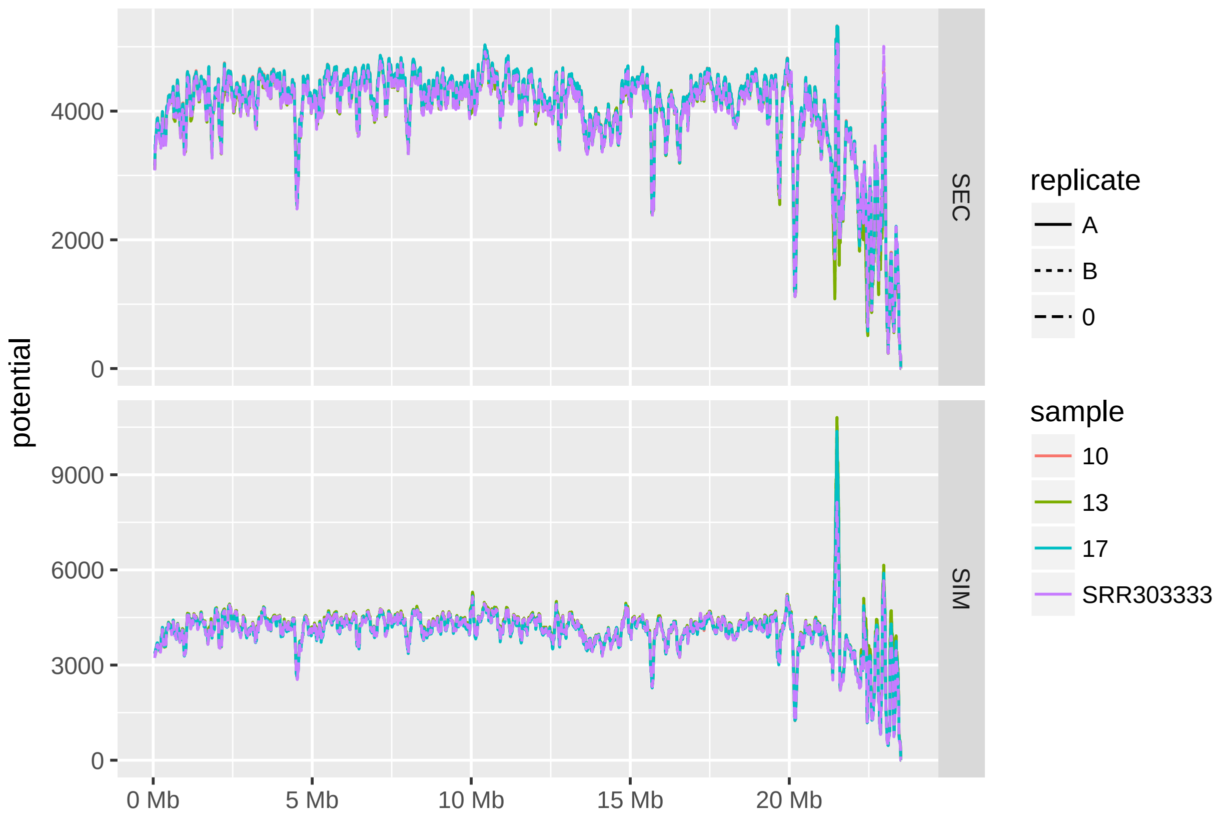
Here is the same analysis, using simulated droSim1 reads against the droSec1 reference:

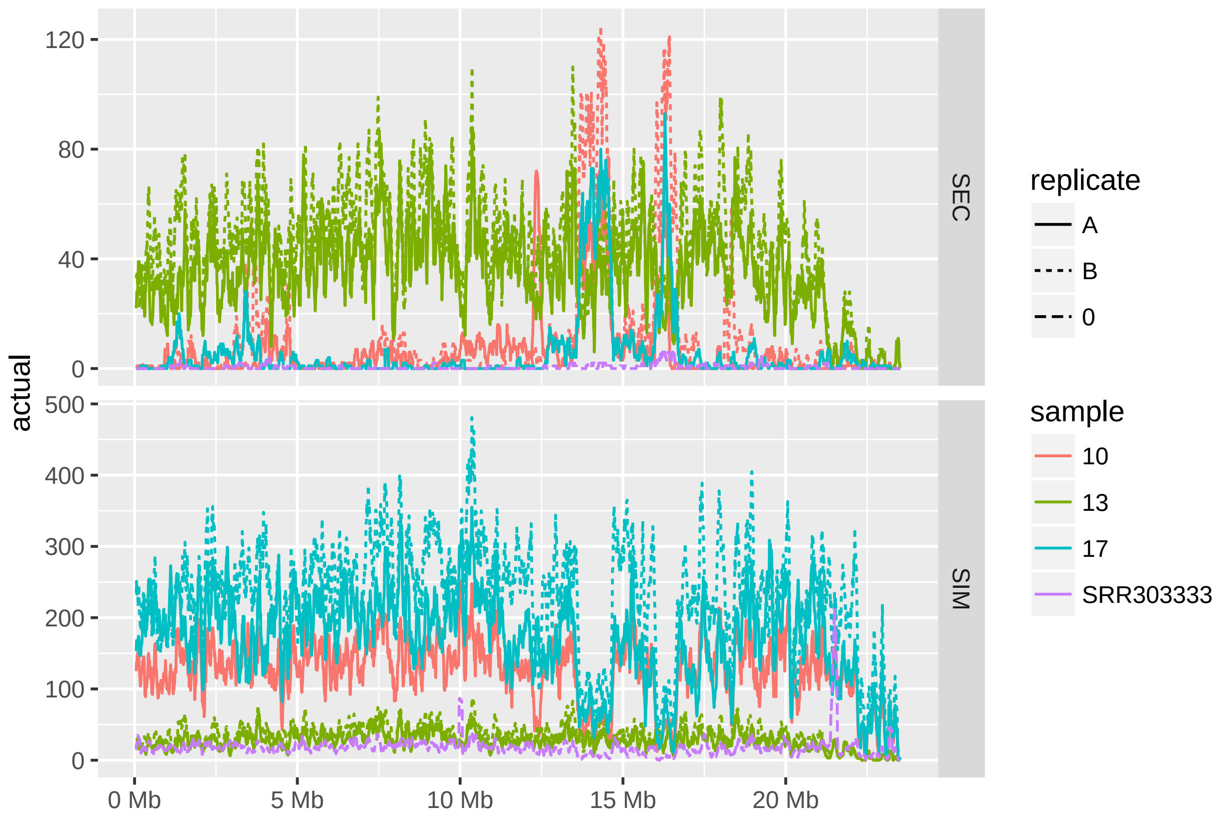


Because the droSec1 genome is relatively fragmented and because the comparisons need to be done on a common coordinate system, the the variants are remapped to the melanogaster dm6 reference using UCSC LiftOver. This is a lossy procedure but the variant sites which don't LiftOver are pretty consistently distributed (with some enhancement near the centromere), and mostly are sites which aren't called as variable in the offspring:

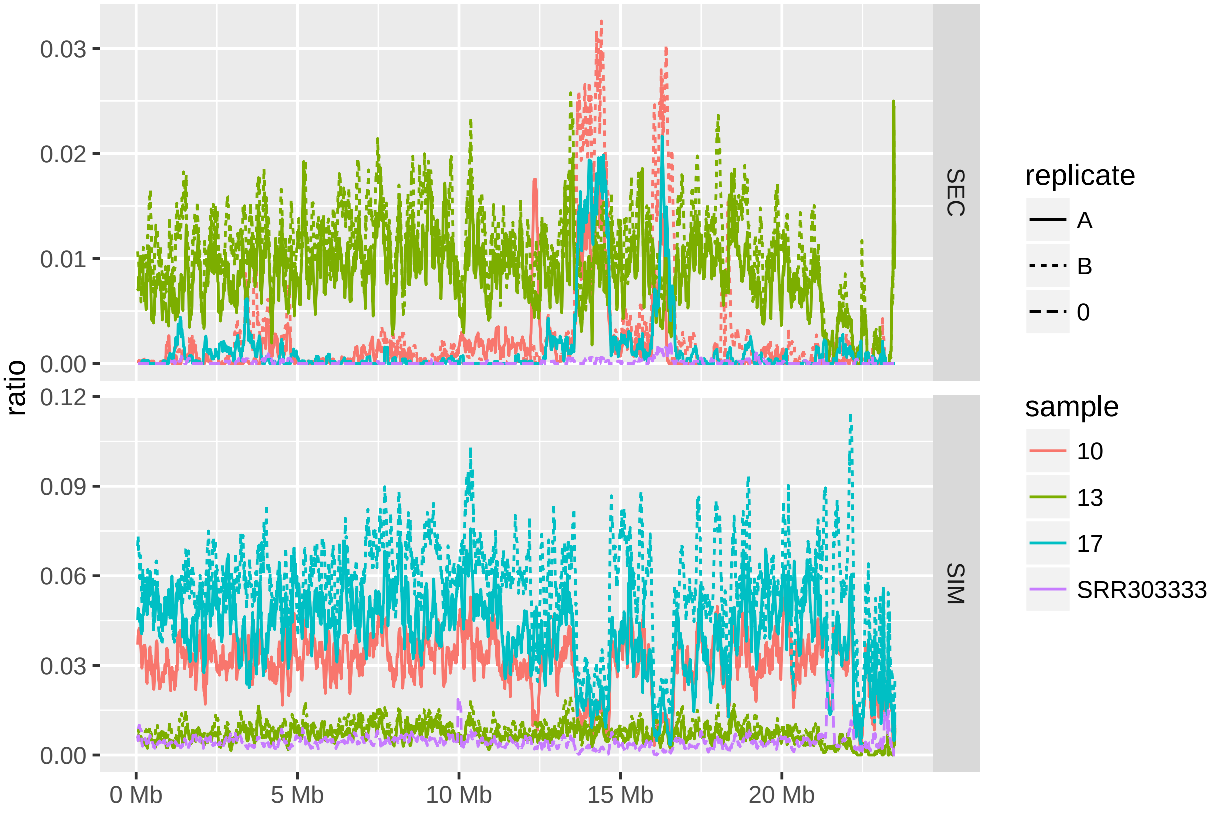


All selected lines were analyzed in this manner:


potential
4000
2000
0
9000
6000
3000
0
0 Mb
5 Mb
10 Mb
15 Mb
20 Mb
SEC
SIM
replicate
A
B
0
sample
10
13
17
SRR303333



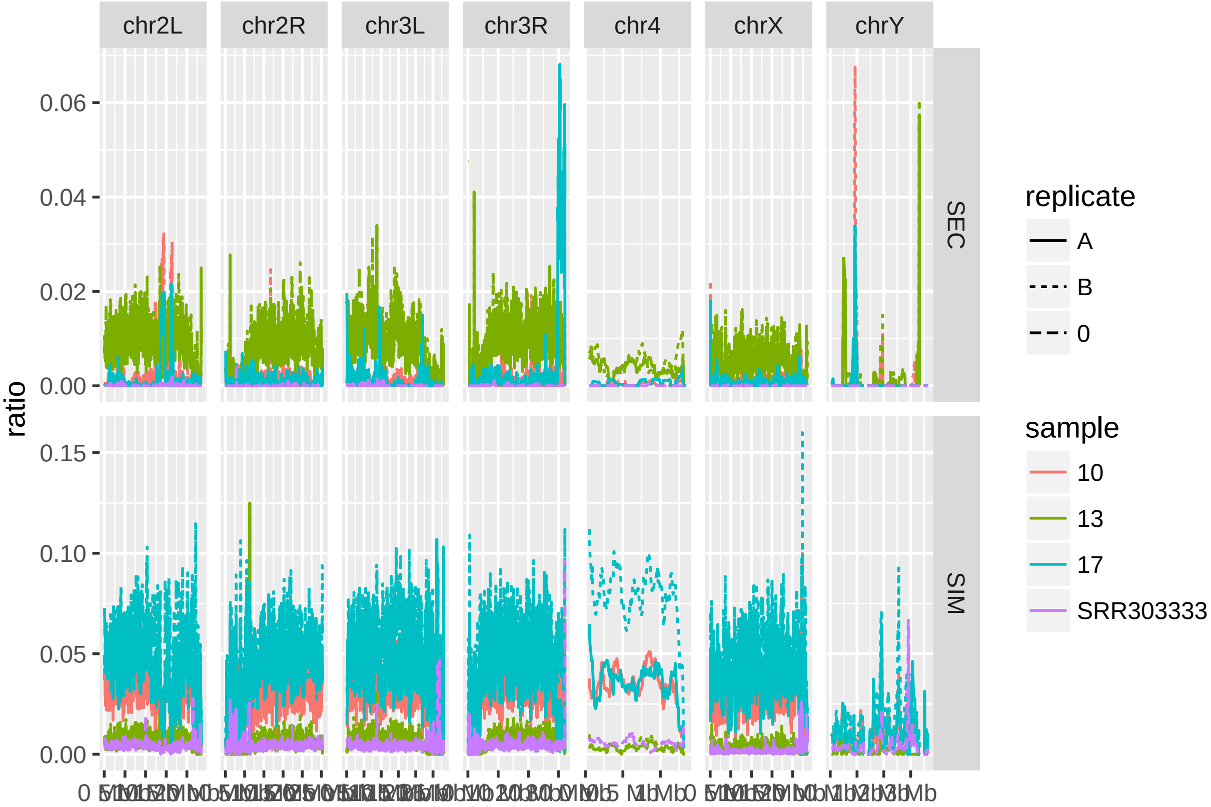
```
## Warning: Removed 18 rows containing missing values (geom_path).
```



In several of the lines there is a clear deflection on chr2L, ~15Mb, which corresponds to an interval identified in (Earley and Jones 2011). However, several do not, including the sample from (Earley and Jones 2011). This may be due to a different parameterization; the earlier code only required a nonzero coverage when comparing the pileups, whereas the reimplementation had a minimum coverage requirement. This was set to a depth of 5 reads for this analysis. The average depth of coverage for the samples behaving as expected was higher than those which were not. (Eg, 10A ~ 4.0, 17A ~ 4.2 vs SRR303333 ~ 3.2, 13A ~ 3.9)

The deflection is in the direction of reduced divergence compared to simulans and increased divergence compared to sechellia, suggesting an enhancement of simulans character and depletion of sechellia character.

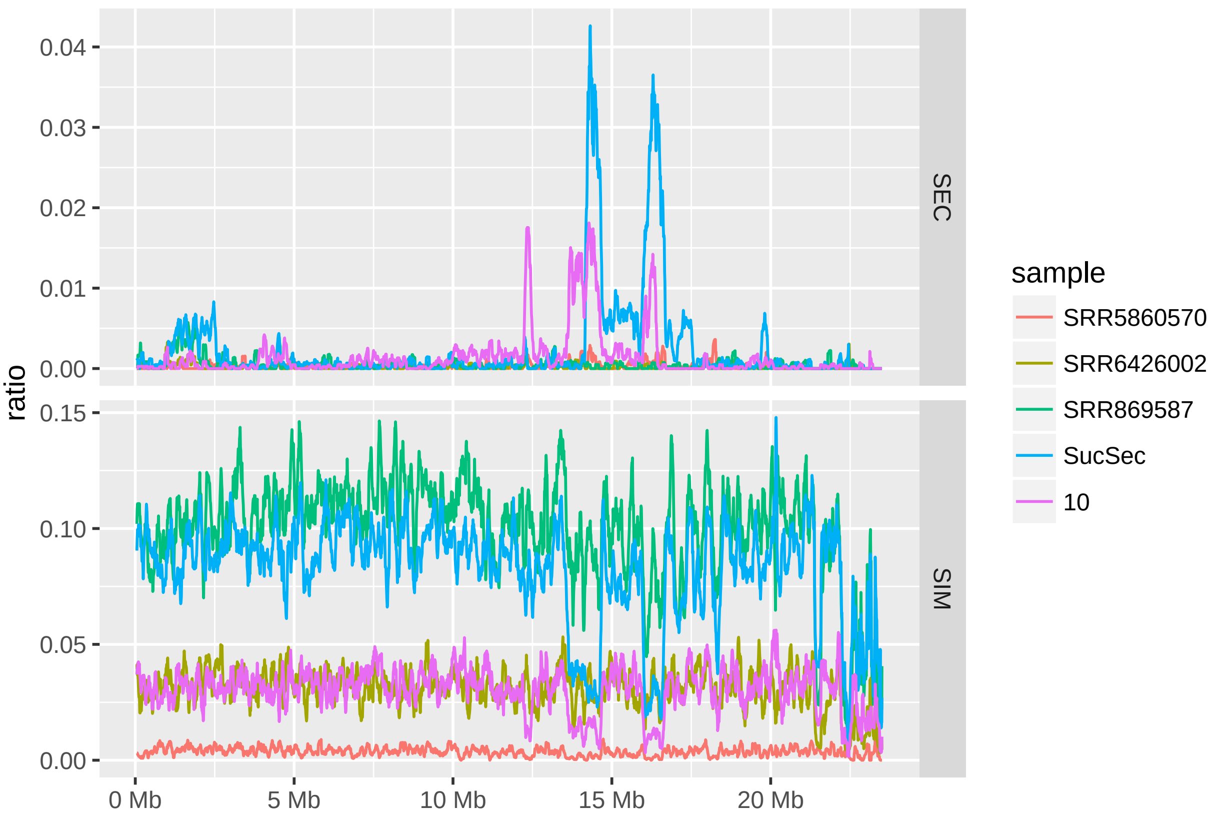
```
## Warning: Removed 11 rows containing missing values (geom_path).
```



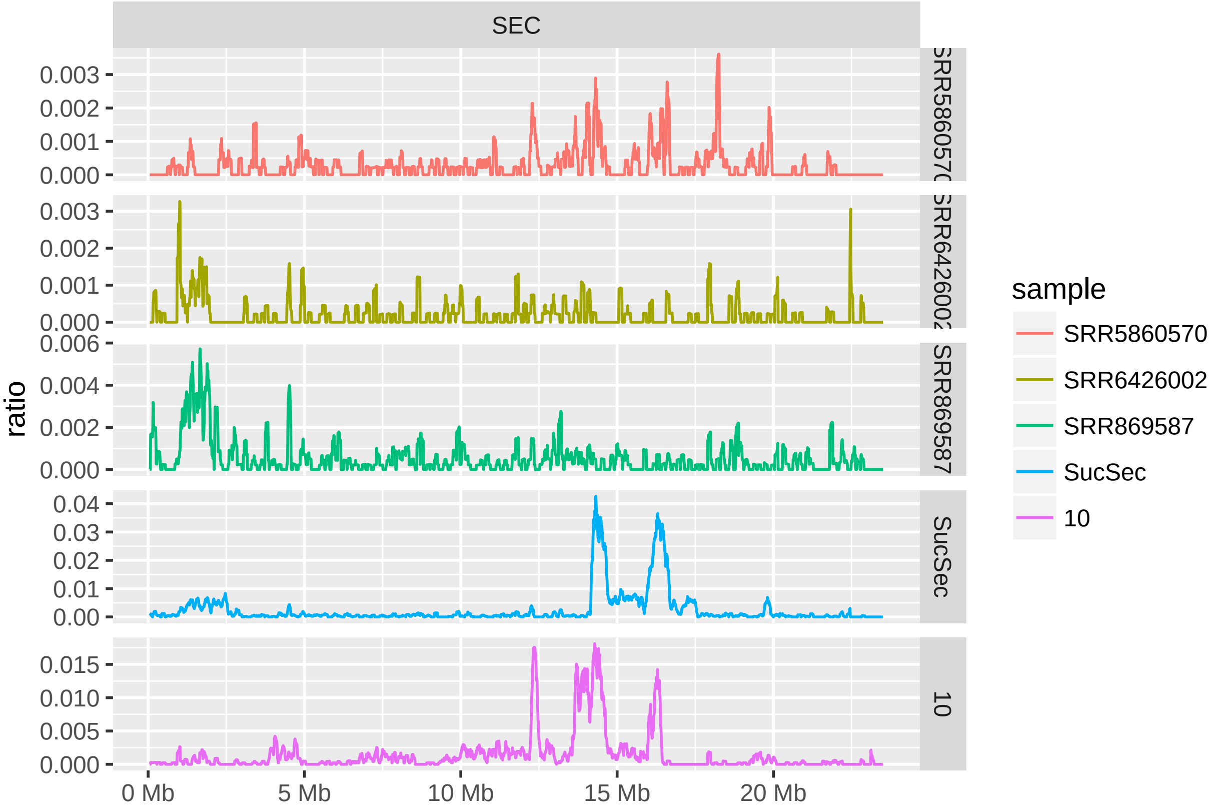
Comparison to control

No control lines were prepared by backcrossing without selection :(Instead, Sechellia reads from a different project (SucSec) or from NCBI (SRR869587, SRR6426002, SRR5860570) were used processed in the same manner as the selection lines.

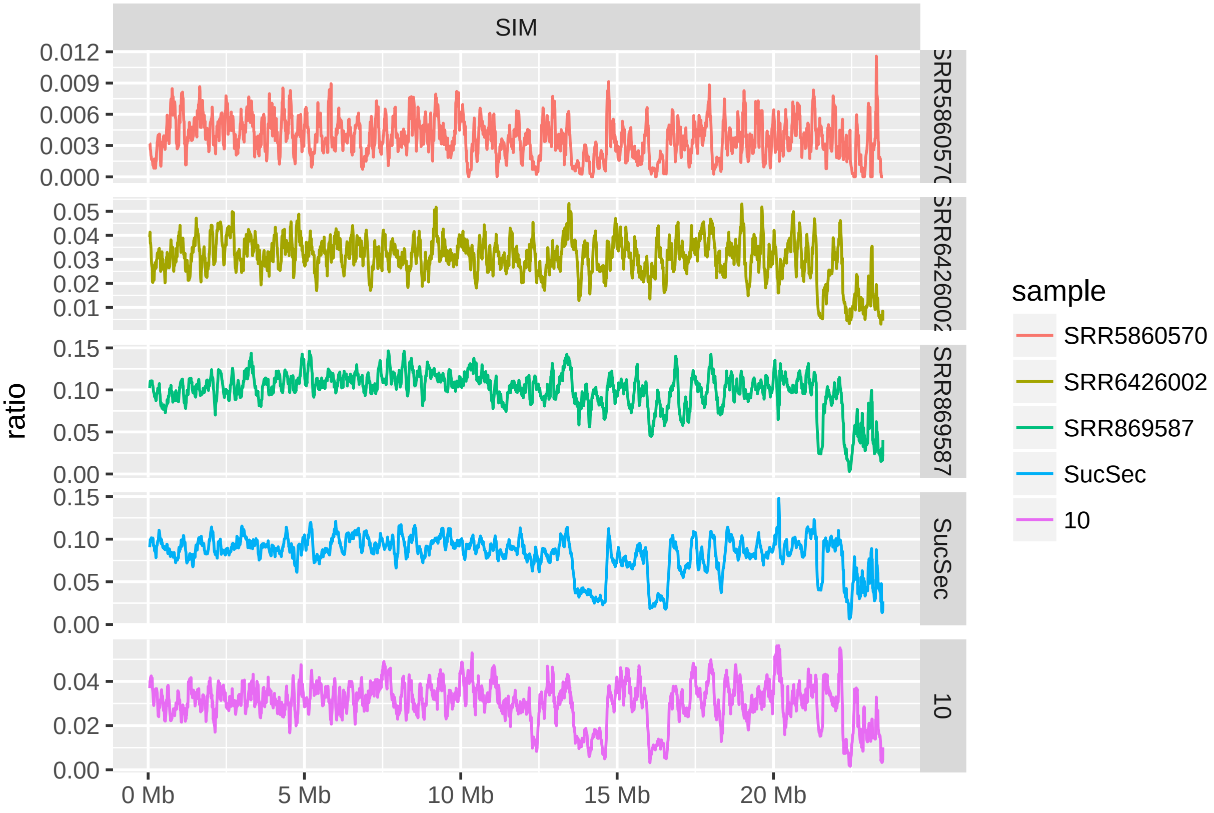
```
## Warning: Removed 10 rows containing missing values (geom_path).
```



```
## Warning: Removed 10 rows containing missing values (geom_path).
```



```
## Warning: Removed 10 rows containing missing values (geom_path).
```

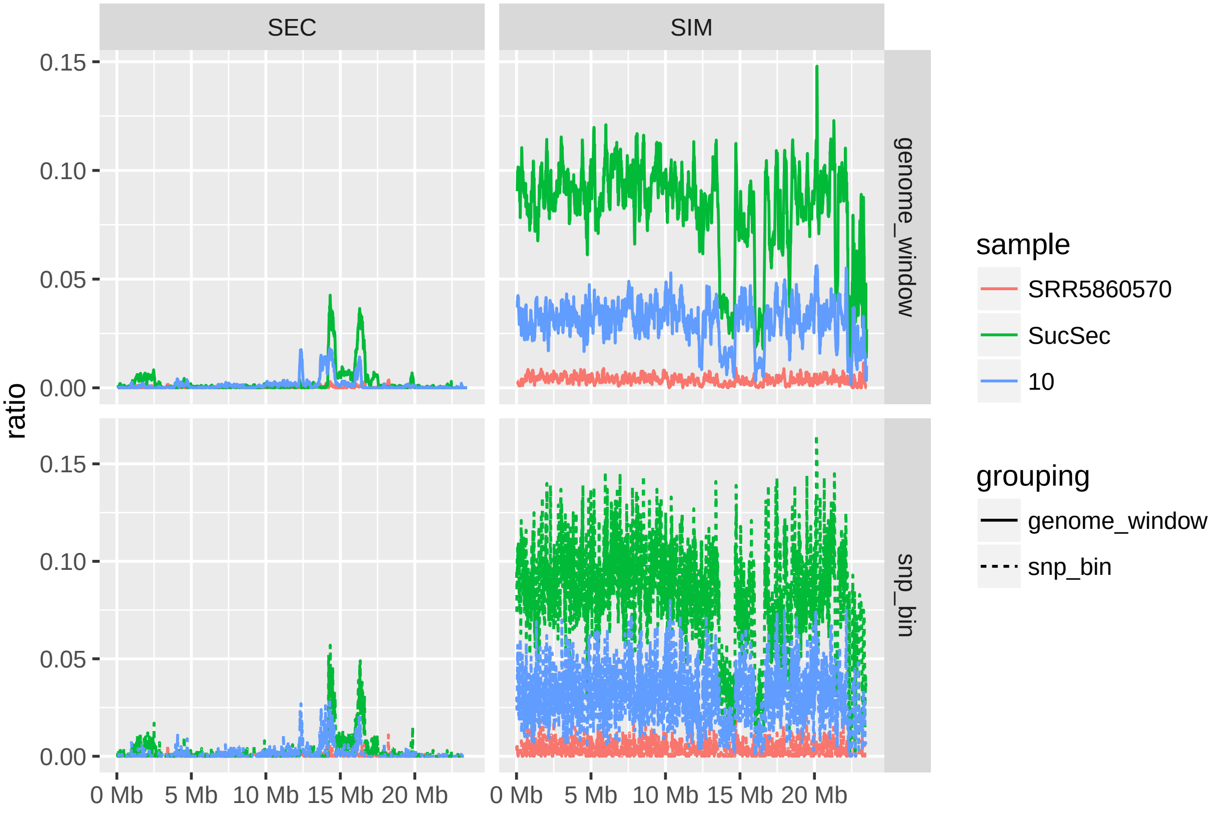


Unfortunately, many of these show the same enhancement of simulans character as the selected lines to some degree, suggesting that this is an artefact. It is especially strong in the SucSec line, which is closely related to the F0 sechellia lines used in the PsiSeq procedure.

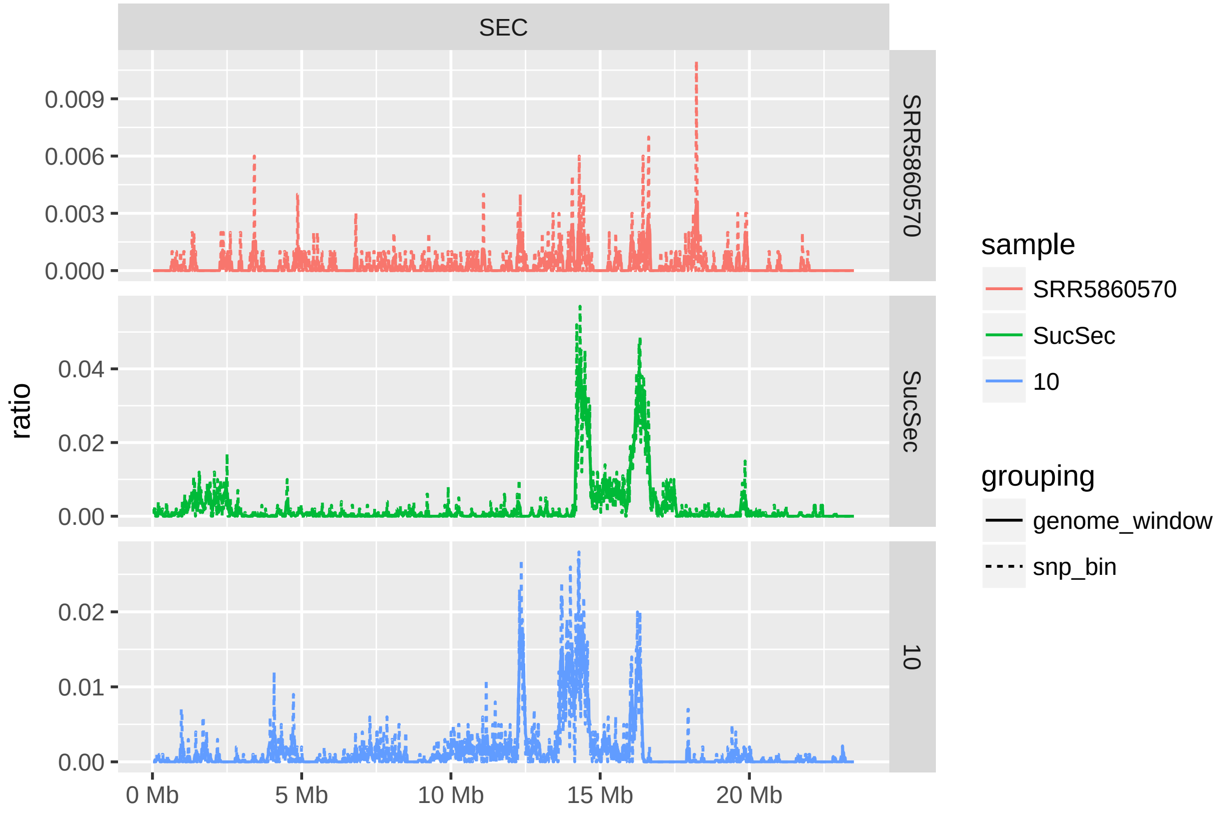
windows vs bins

Variants were counted by dividing the genome into fixed-width windows and counting, which is trivial to implement in bedtools. However, the (Earley and Jones 2011) algorithm used a sliding bins of fixed SNP count. These counting methods were compared and give qualitatively similar results:

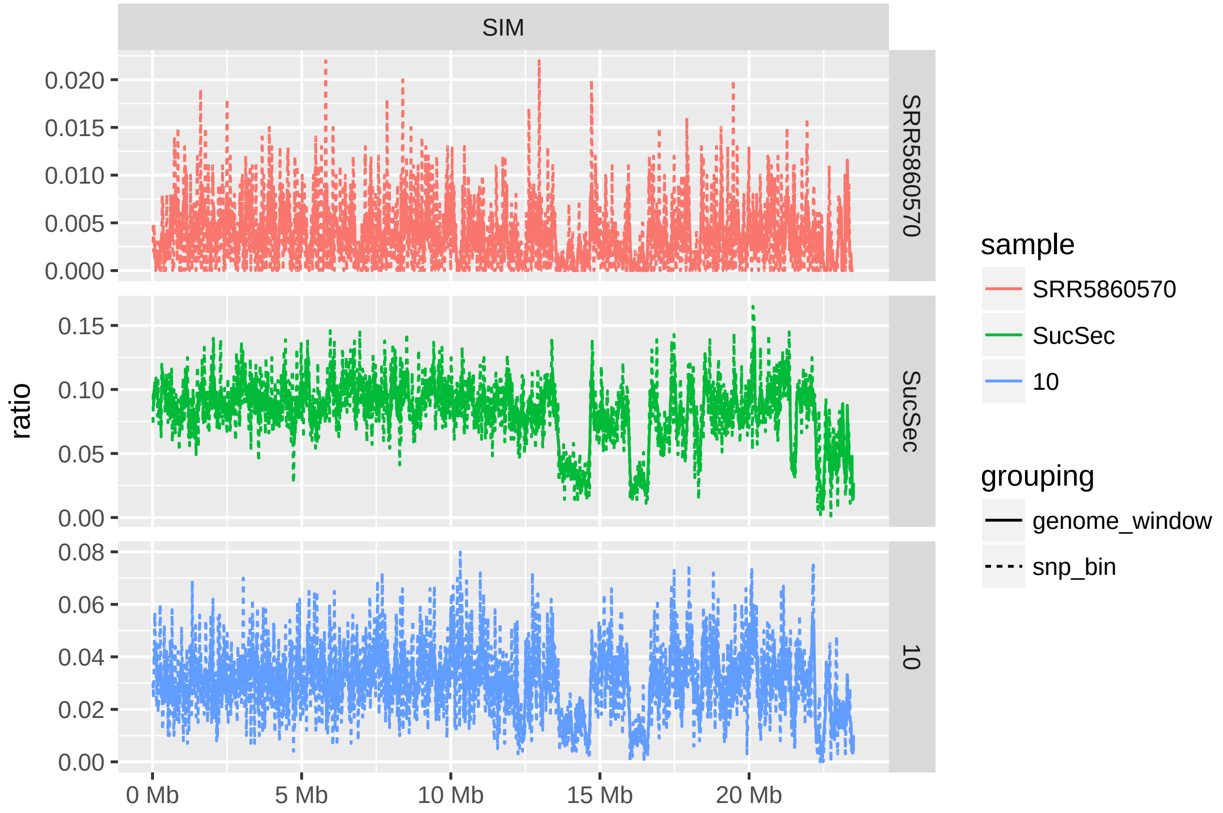
```
## Warning: Removed 6 rows containing missing values (geom_path).
```



```
## Warning: Removed 6 rows containing missing values (geom_path).
```



```
## Warning: Removed 6 rows containing missing values (geom_path).
```

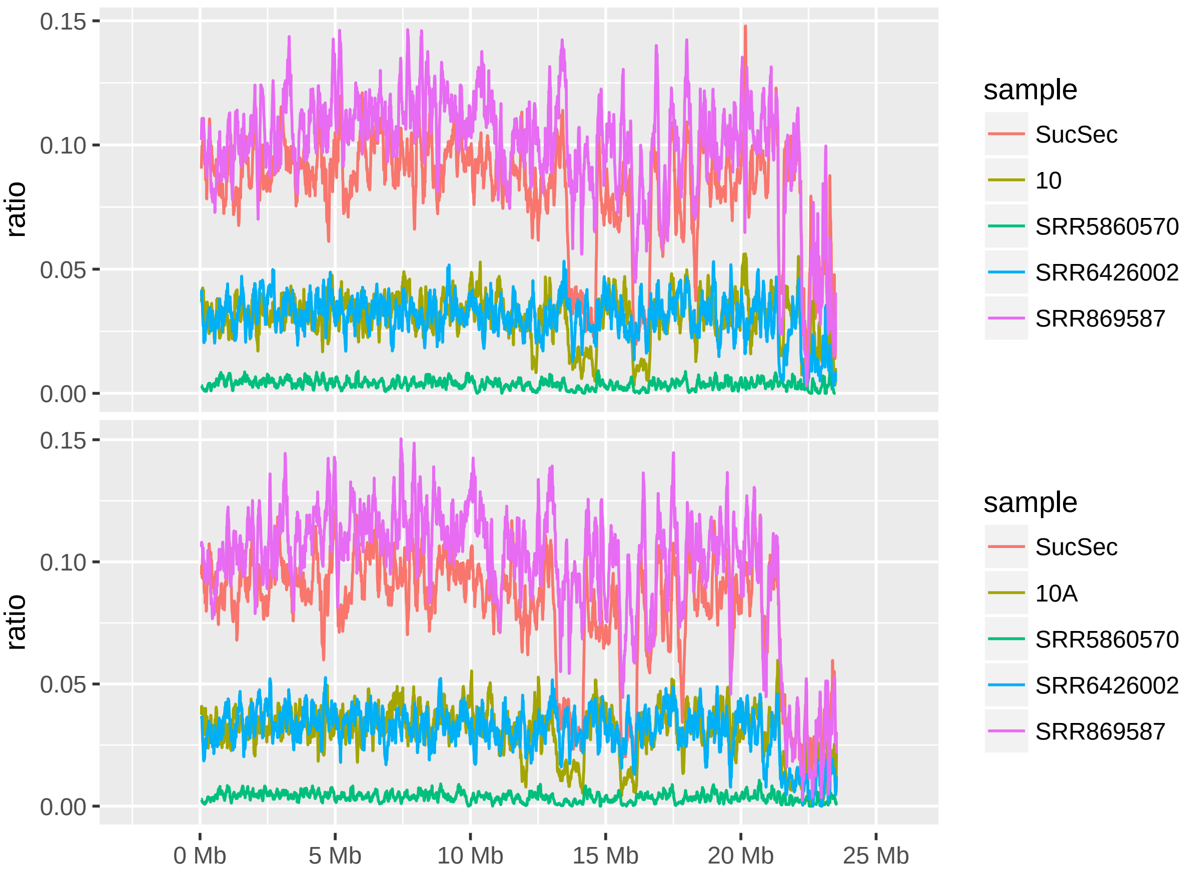


Different genome assemblies

Because of questions about the quality of the droSec1 assembly [cite?] (and droSim1?), the dsecr1 was fragmented and mapped to dsim2. Comparing the chr2L of this with the mapping to droSim1 and liftover to dm6 gives qualitatively simlar results:

```
## Warning in .Seqinfo.mergexy(x, y): The 2 combined objects have no sequence levels in common. (Use
##   suppressWarnings() to suppress this warning.)
```

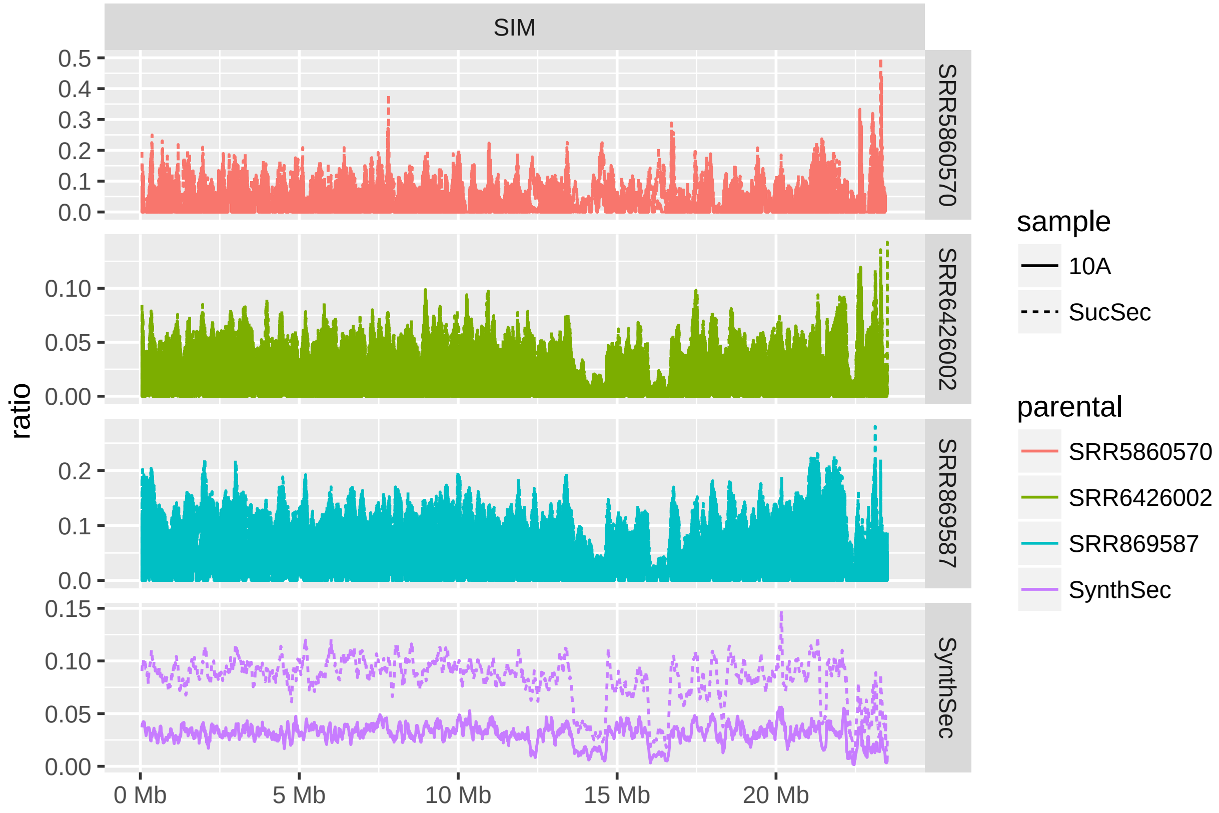
```
## Warning: Removed 10 rows containing missing values (geom_path).
```



Different F0 Lines

The F0 sechellia and simulans flies were not sequenced; synthetic reads made by fragmenting a reference genome (this was done with a custom Perl script in (Earley and Jones 2011) and using the ART read simulator here [cite]). Rich's SucSec and several NCBI lines were used as empirical samples of sechellia genomes. The signal is weaker in these cases (possibly due to depth-of-coverage effects), but these samples recreate the artefact:

```
## Warning: Removed 34 rows containing missing values (geom_path).
```



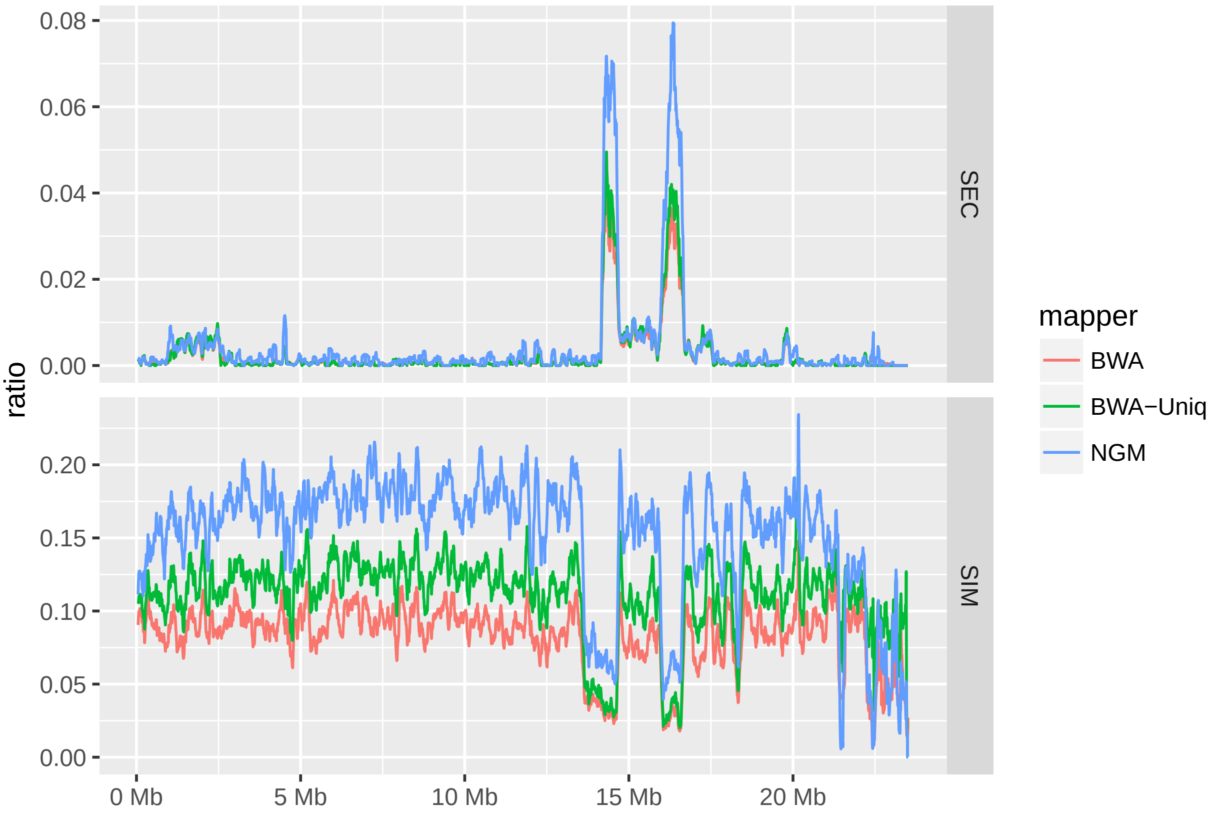
Eliminate coverage and other problems

- no major coverage effects near the artefact
- mapping specificity (check source from sythetic reads)

Different mapping strategies

Like (Earley and Jones 2011), these alignments have been unfiltered BWA output. This might introduce artefacts through multimapping reads or low-quality mappings. A subset of the BWA alignments was gathered by filtering on mapping quality (MAPQ, well-mapped pairs) and mapping uniqueness (no alternative or secondary mappings). To test robustness to aligner algorithm, NGM was also used. The results are qualitatively similar:

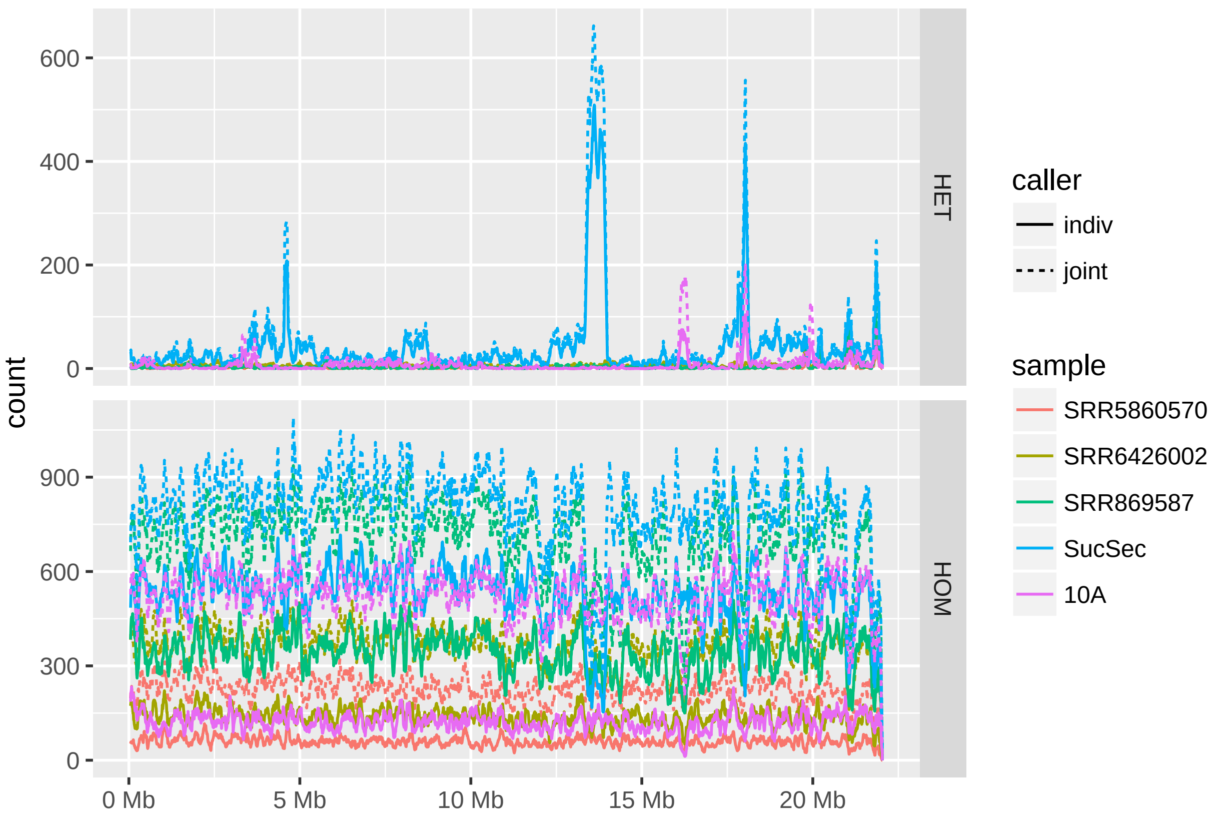
```
## Warning: Removed 6 rows containing missing values (geom_path).
```

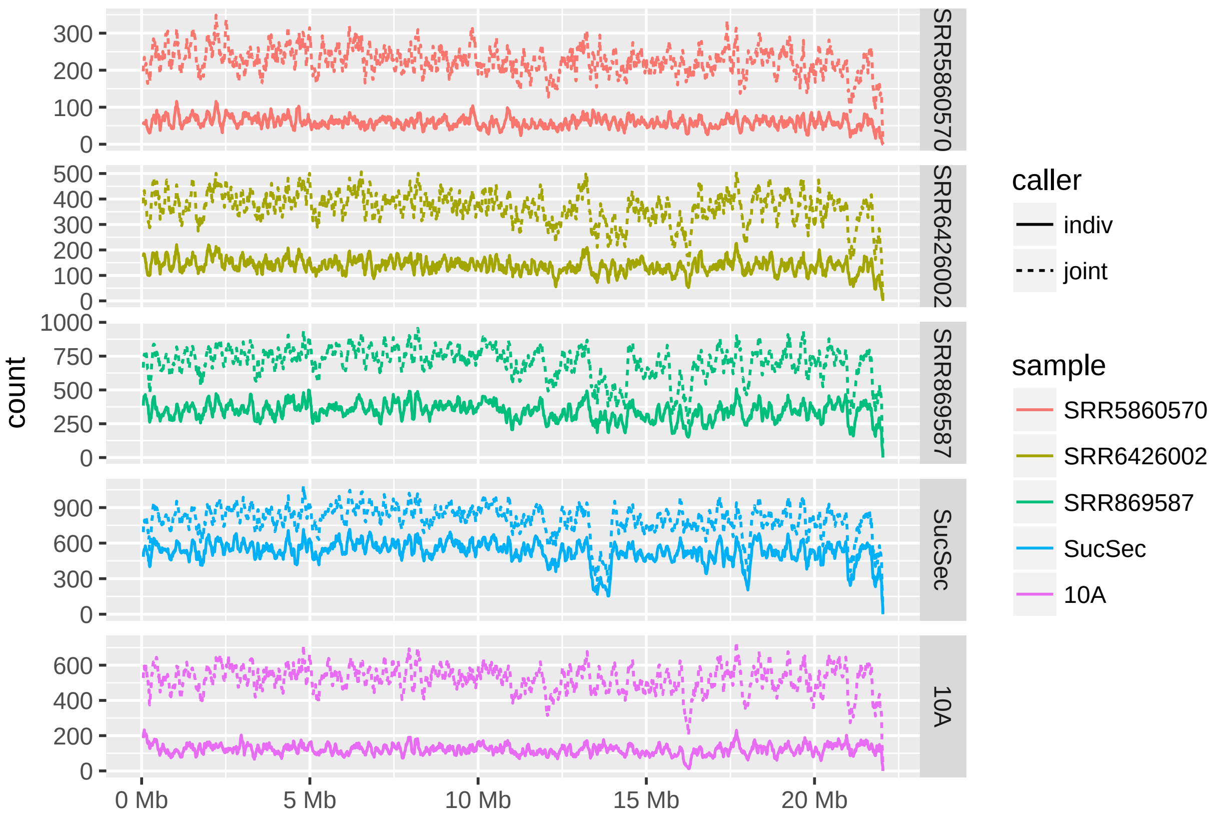


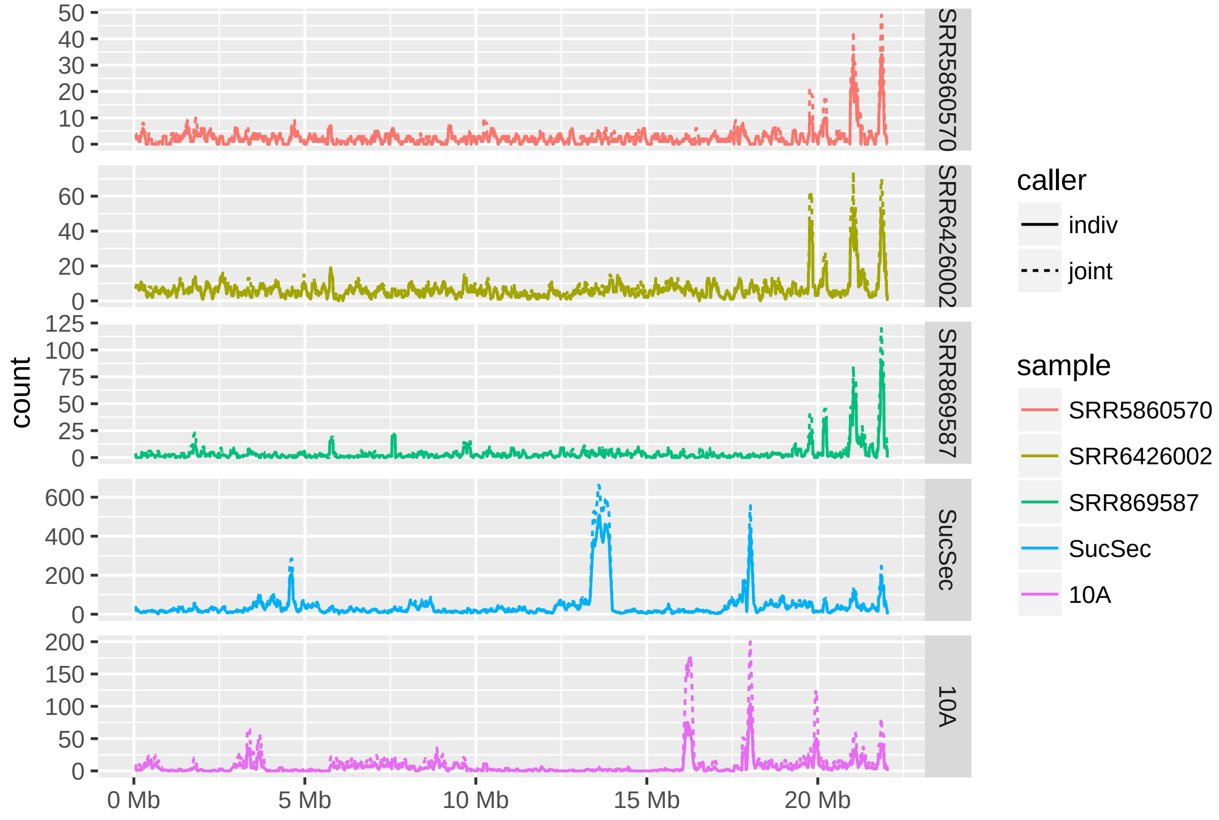
freebayes heterozygosity

The original SNP-calling algorithm was ‘plurality wins’; a site is counted as variable or not, based on whether one base was more common than all others, including the reference. The reimplementation currently requires 90% of the reads to support a non-reference call. Neither approach handles heterozygosity at all. Furthermore, neither the genome fragmenting Perl script in (Earley and Jones 2011) nor ART produce heterozygous reads, and simulating real-world heterozygosity is not trivial, though it has been done (Stephens et al. 2016)(Yuan, Zhang, and Yang 2017) (try these out?).

Freebayes was used to call SNPs from the alignments, both individually and jointly; these were grouped by zygosity and counted by genomic window. They show an enhancement of heterozygosity in the problem region.


count
600
400
200
0
900
600
300
0
HET
HOM
0 Mb
5 Mb
10 Mb
15 Mb
20 Mb
caller
indiv
joint
sample
SRR5860570
SRR6426002
SRR869587
SucSec
10A


count
300
200
100
0
500
400
300
200
100
0
1000
750
500
250
0
900
600
300
0
600
400
200
0
0 Mb
5 Mb
10 Mb
15 Mb
20 Mb
SRR5860570
SRR6426002
SRR869587
SucSec
10A
caller
indiv
joint
sample
SRR5860570
SRR6426002
SRR869587
SucSec
10A



?? what if a variable site is heterozygous for two non-ref alleles??

To Do

- Better exploration of depth-of-coverage effects?
- Use called variants
  - naive, informed
- Simulation
  - mention application to crossovers, eg nicoles flies?
- Rich-Free version?

```
##
## To cite package 'tidyverse' in publications use:
##
##   Hadley Wickham (2017). tidyverse: Easily Install and Load the
##   'Tidyverse'. R package version 1.2.1.
##   https://CRAN.R-project.org/package=tidyverse
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {tidyverse: Easily Install and Load the 'Tidyverse'},
##     author = {Hadley Wickham},
##     year = {2017},
##     note = {R package version 1.2.1},
##     url = {https://CRAN.R-project.org/package=tidyverse},
##   }

##
##   M. Lawrence, R. Gentleman, V. Carey: "rtracklayer: an {R}
##   package for interfacing with genome browsers". Bioinformatics
##   25:1841-1842.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {rtracklayer: an R package for interfacing with
##                   genome browsers},
##     author = {Michael Lawrence and Robert Gentleman and Vincent Carey},
##     year = {2009},
##     journal = {Bioinformatics},
##     volume = {25},
##     pages = {1841-1842},
##     doi = {10.1093/bioinformatics/btp328},
##     url = {http://bioinformatics.oxfordjournals.org/content/25/14/1841.abstract},
##   }

##
## To cite package 'ggbio' in publications use:
##
##   Tengfei Yin, Dianne Cook and Michael Lawrence (2012): ggbio: an
##   R package for extending the grammar of graphics for genomic data
##   Genome Biology 13:R77
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {ggbio: an R package for extending the grammar of graphics for genomic data},
##     author = {Tengfei Yin and Dianne Cook and Michael Lawrence},
##     journal = {Genome Biology},
##     volume = {13},
##     number = {8},
##     pages = {R77},
##     year = {2012},
##     publisher = {BioMed Central Ltd},
##   }
```

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. doi:10.1534/genetics.111.129445.

Stephens, Zachary D., Matthew E. Hudson, Liudmila S. Mainzer, Morgan Taschuk, Matthew R. Weber, and Ravishankar K. Iyer. 2016. "Simulating next-generation sequencing datasets from empirical mutation and sequencing models." *PLoS ONE* 11 (11): 1–18. doi:10.1371/journal.pone.0167047.

Yuan, Xiguo, Junying Zhang, and Liying Yang. 2017. "IntSIM: An Integrated Simulator of Next-Generation Sequencing Data." *IEEE Transactions on Biomedical Engineering* 64 (2): 441–51. doi:10.1109/TBME.2016.2560939.