PopPsiSeq Summary

Charlie Soeder

11/20/2018

Contents

1 Introduction

Explain motivation, overview of PsiSeq and PsiSeq2

Population-based approach, rather than ancestral

2 Materials, Methods, Data, Software

2.1 Reference Genomes

The droSim1 and droSec1 reference genomes were downloaded in FASTA format from UCSC Genome Browser. These were in the 140-170Mb range, with the droSec1 relatively unconsolidated:

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

Table 1: Size and Consolidation of Reference Genomes

Reference Genome:	dm6	droSec1	droSim1
number_bases	144 M	167 M	142 M
number_contigs	1.87 k	14.7 k	18

(add a by-chromosome breakdown for droSim and a histogram for droSec?)

2.2 Sequenced Reads

A backcross and introgression experiment was performed, in which simulans females were mated with sechellia males, and the hybrid offspring were selected for avoidance of morinda odorants. The offspring were sequenced after 15 rounds of backcrossing and introgression (Earley and Jones 2011). One sample was sequenced in this experiment; a follow-up experiment generated three more samples with two replicates each. As a control, several wild-type sechellia sequences were downloaded from NCBI:

Table 2: Sequenced Experimental Samples

name	paired	experimental	source
SRR5860570	TRUE	control	NCBI
SRR303333	FALSE	selection	EarlyJones2011
17B	TRUE	selection	EarlyJones2013
17A	TRUE	selection	EarlyJones2013
10B	TRUE	selection	EarlyJones2013
10A	TRUE	selection	EarlyJones2013

For population-wide data, wild D. simulans and D. sechellia flies were captured and sequenced by Daniel Matute:

Table 3: Number of Sequenced Samples per Species

species	sample_count
drosophila sechellia	4
drosophila simulans	4

2.2.1 Pre-processing

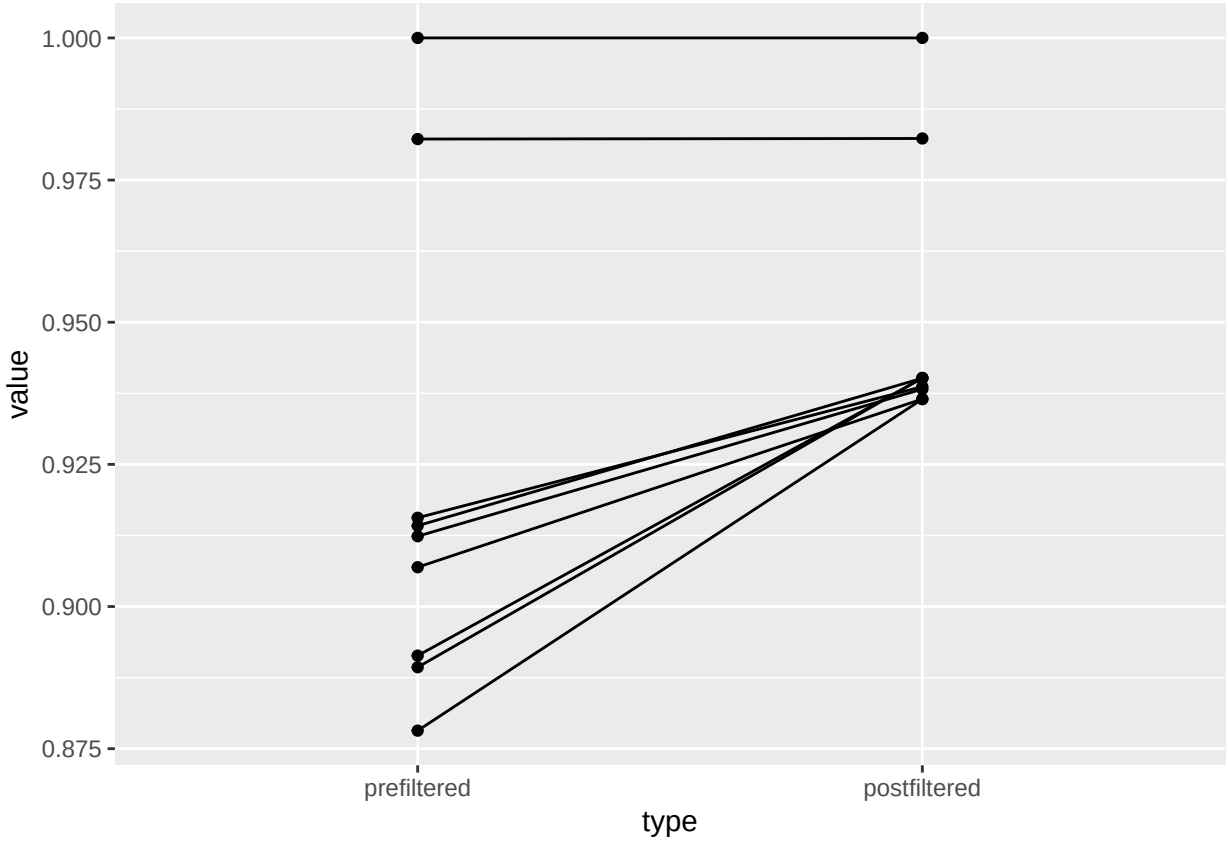
These reads were preprocessed with FASTP (S. Chen et al. 2018) for quality control and analytics.

Starting FASTQ files contained a total of $400M$ reads; after QC, this dropped to $379M$.

Table 4: Read Count and Percent Retention

type	minimum	average	maximum
prefiltered	1.48 M	44.5 M	85.4 M
postfiltered	1.4 M	42.1 M	81.1 M
percent retention	93.3	95.6	100

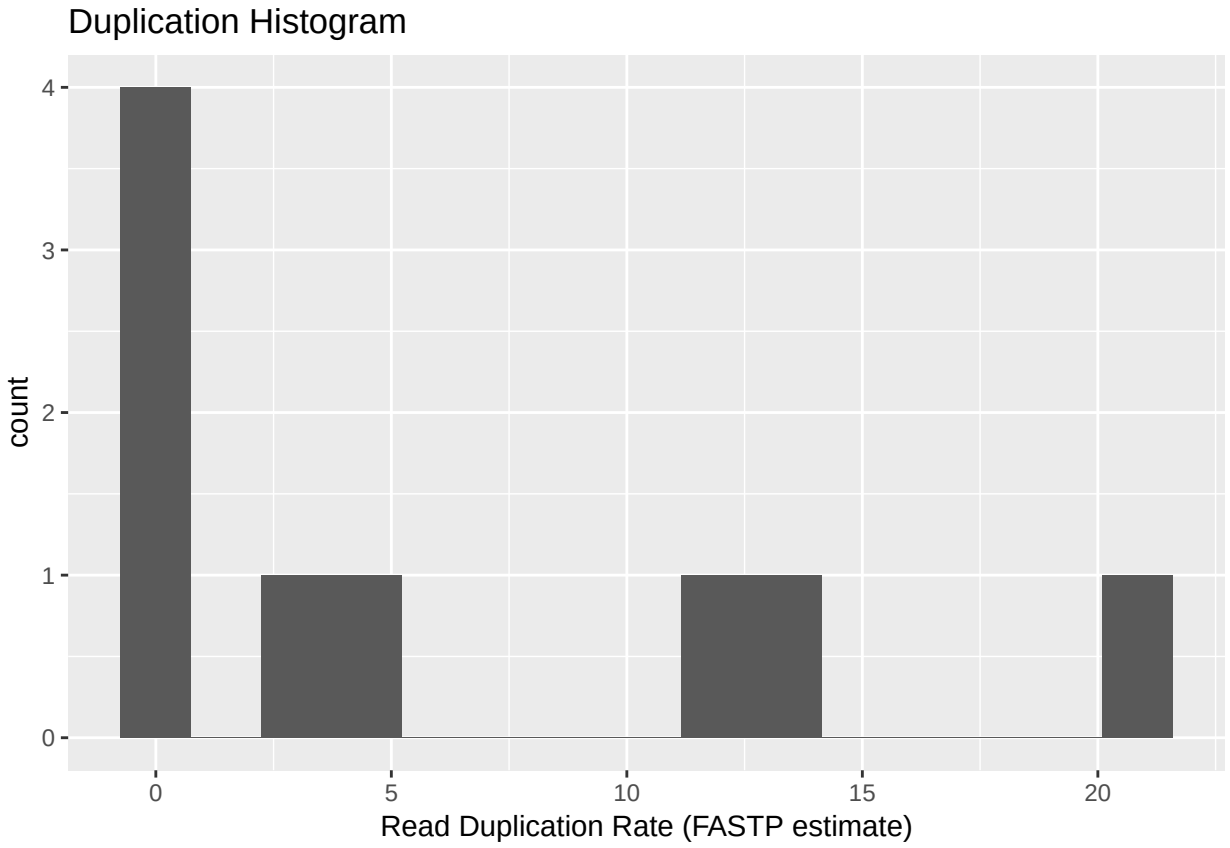
Filtration also increased the read quality, as seen in the increase in the fraction of reads with an average quality score > 30:



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

Table 5: Percentage Duplication

minimum	average	median	maximum
0.1	6	2.6	20.9



2.3 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.

2.3.1 Read & Alignment Quality

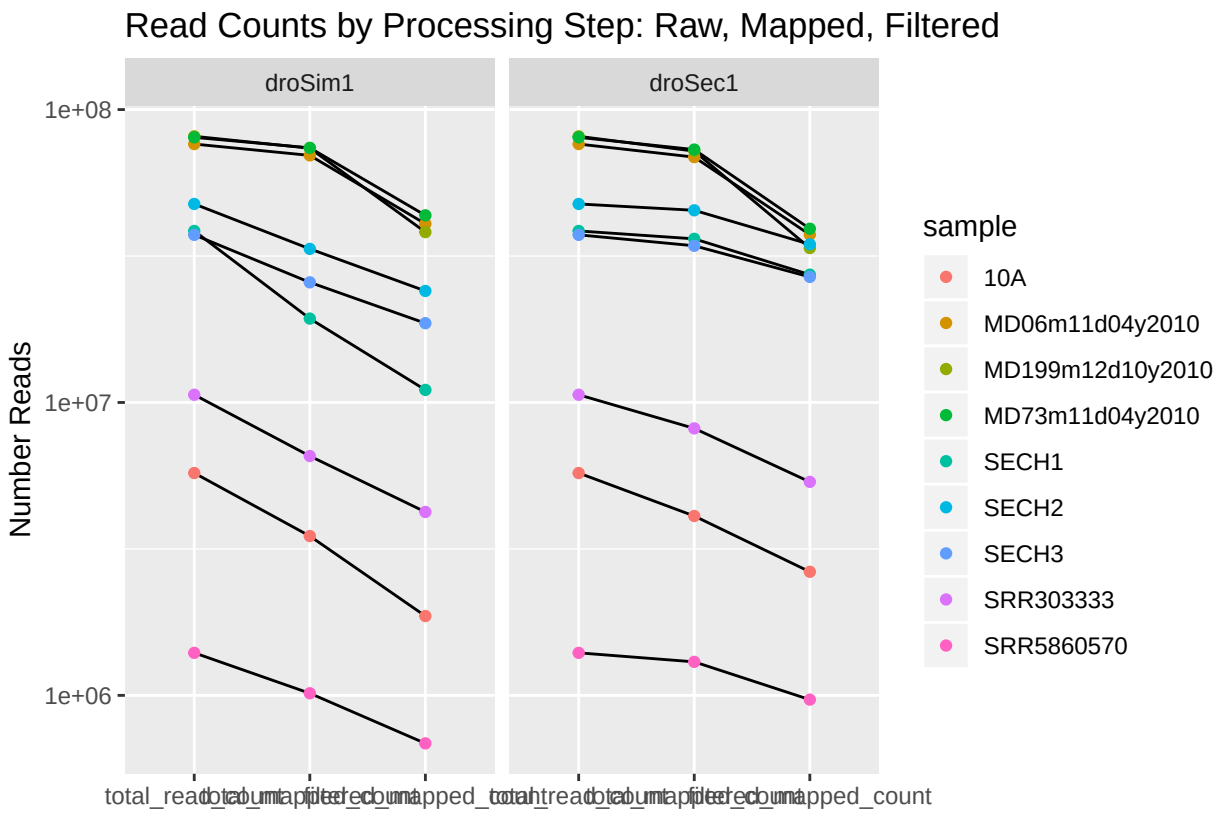


Table 6: Read Counts During Alignment & Filtration

measure	minimum	average	median	maximum
filtered_mapped_count	686 k	21.7 M	25.4 M	43.5 M
total_mapped_count	1.02 M	36.1 M	33.9 M	74 M
total_read_count	1.4 M	42.1 M	38.5 M	81.1 M

The fraction of reads retained at each point:

Table 7: Percentage of Reads Retained at Each Step

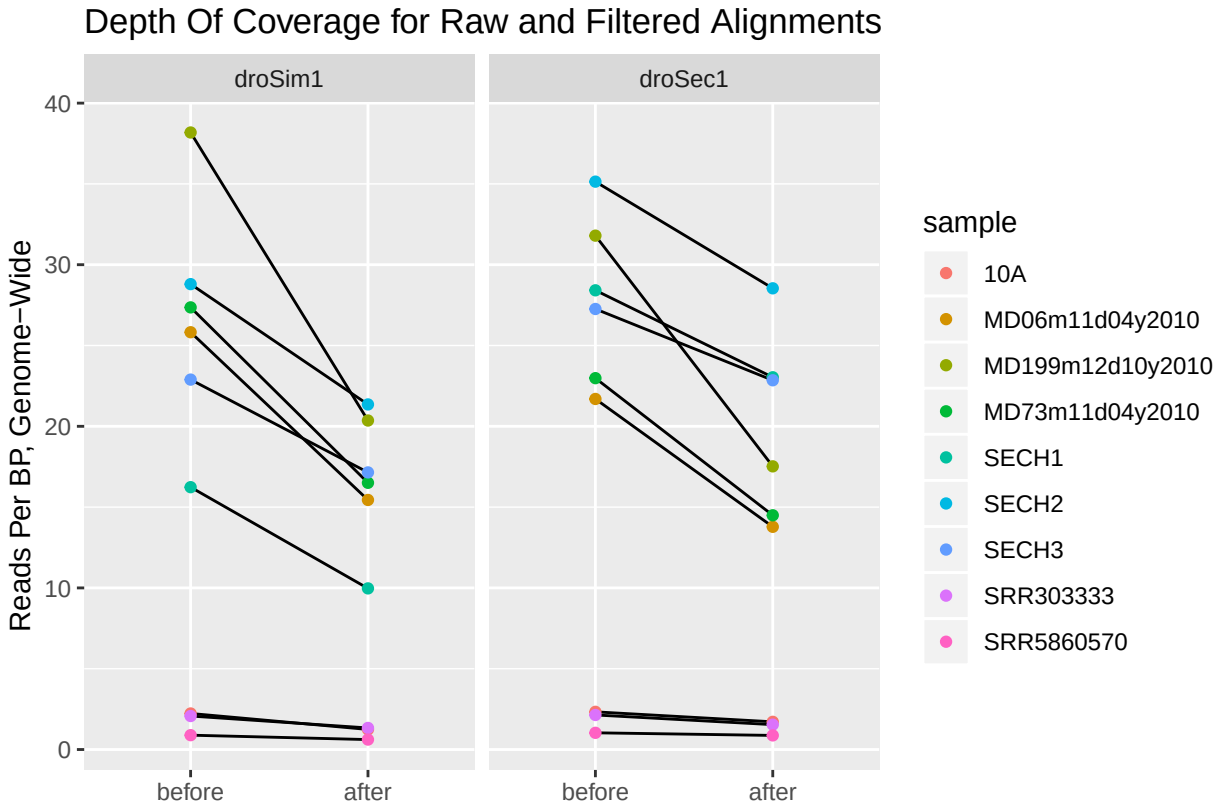
measure	minimum	average	median	maximum
filter_retention	46.7	63.6	64.5	78.3
mapping_retention	50.2	80.6	89.6	95.1

2.3.2 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

step	minimum	average	median	maximum
pre-filtration depth	0.9	18.7	22.9	38.2
post-filtration depth	0.6	12.7	15.0	28.5
depth retention percent	53.3	68.4	66.9	84.4



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

2.4 Called Variants

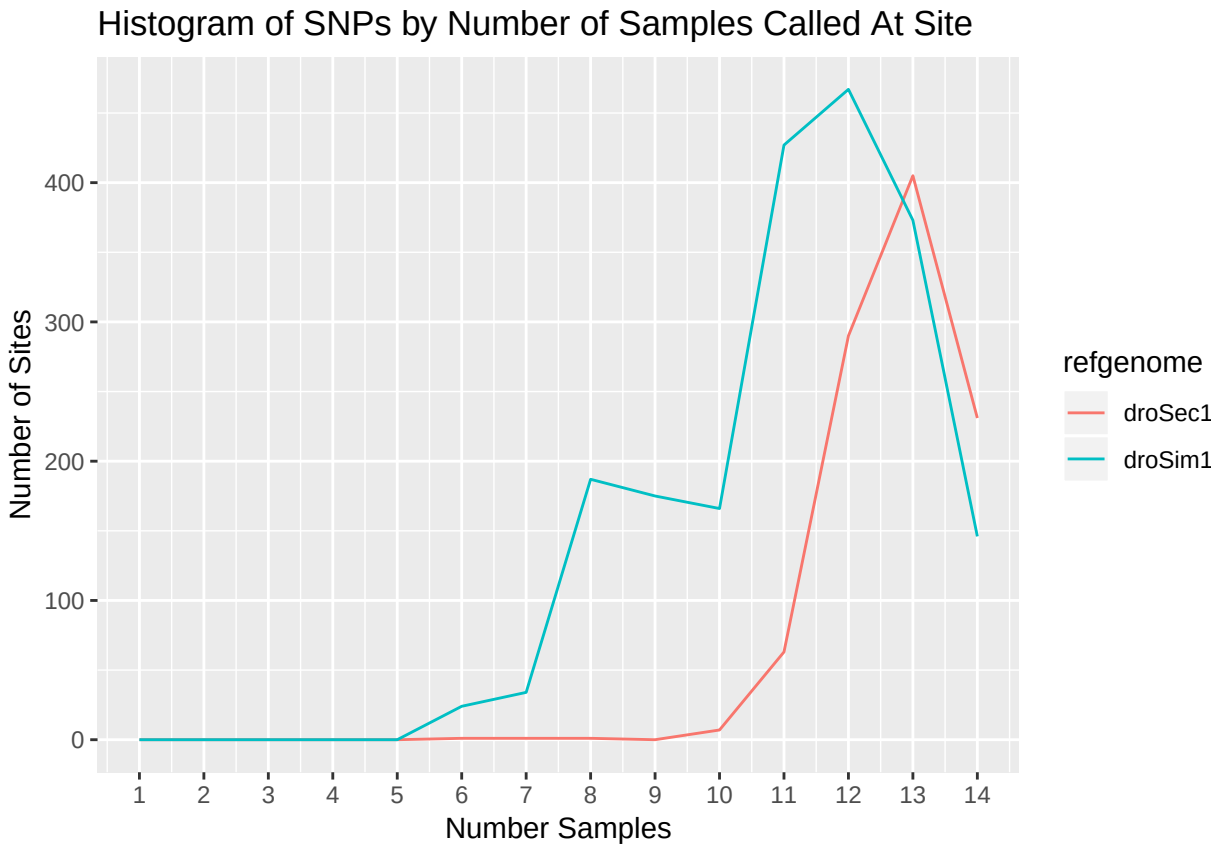
BWAUniq mappings were used to jointly call variants in VCF format via Freebayes (Garrison and Marth 2012) using standard filters.

Table 9: SNP count and per-KB SNP rate across all samples

reference genome	Genome size (bp)	total SNP count	SNPs per kB
droSec1	167 M	2.44 M	14.7
droSim1	142 M	2.7 M	18.9

To build this VCF, 14 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:

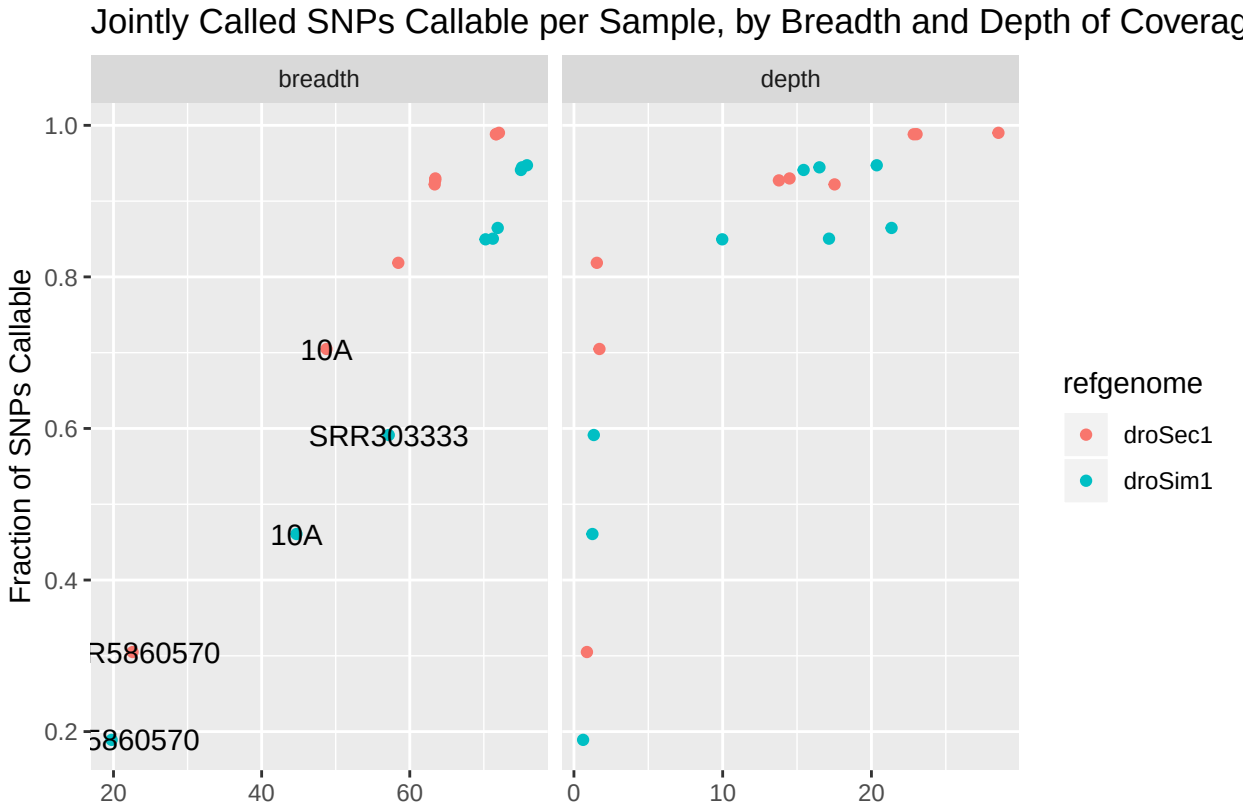
```
## Warning: Removed 4 rows containing missing values (geom_path).
```



The fraction of jointly called SNPs which are individually callable:

```
## Warning: Column `refgenome`/`reference` joining factors with different
## levels, coercing to character vector
```

```
## Warning: Column `refgenome`/`reference` joining character vector and
## factor, coercing into character vector
```



3 Results

4 References

4.1 Software

```
##
## 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},
##   }

##
## To cite the 'knitr' package in publications use:
##
```

```
##   Yihui Xie (2018). knitr: A General-Purpose Package for Dynamic
##   Report Generation in R. R package version 1.20.
##
##   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:
##
##   Jeremy Stephens, Kirill Simonov, Yihui Xie, Zhuoer Dong, Hadley
##   Wickham, Jeffrey Horner, reikoch, Will Beasley, Brendan O'Connor
##   and Gregory R. Warnes (2018). yaml: Methods to Convert R Data to
##   YAML and Back. R package version 2.2.0.
##   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 = {Jeremy Stephens and Kirill Simonov and Yihui Xie and Zhuoer Dong and Hadley Wickham an
##     year = {2018},
##     note = {R package version 2.2.0},
##     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")'.
```

Bibliography

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. "Fastp: An ultra-fast all-in-one FASTQ preprocessor." *Bioinformatics* 34 (17): i884–i890. doi: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. doi: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.