End Joining Signatures - dev

Charlie Soeder

2/15/2019

25 February 2019

Rebuilding, starting with summary stats for the materials/methods section.

Reference genomes

Table 1: Size and Consolidation of Reference Genomes

Reference Genome:	dm6
number_bases	144 M
number_contigs	1.87 k

Sequenced reads

Table 2: Number of Sequenced Samples by Treatment

experimental	sample_count
control	30

Table 3: Sequenced Experimental Samples

name	paired	experimental	source
Mcm5-A7	TRUE	control	dannyMiller
mcm5-28	TRUE	control	dannyMiller
mcm5-27	TRUE	control	dannyMiller
mcm5-26	TRUE	control	dannyMiller
mcm5-25	TRUE	control	dannyMiller
mcm5-24	TRUE	control	dannyMiller
mcm5-23	TRUE	control	dannyMiller
mcm5-22	TRUE	control	dannyMiller
mcm5-21	TRUE	control	dannyMiller
mcm5-20	TRUE	control	dannyMiller
mcm5-19	TRUE	control	dannyMiller
mcm5-18	TRUE	control	dannyMiller
mcm5-17	TRUE	control	dannyMiller
mcm5-16	TRUE	control	dannyMiller
mcm5-15	TRUE	control	dannyMiller
mcm5-14	TRUE	control	dannyMiller
mcm5-13	TRUE	control	dannyMiller
mcm5-12	TRUE	control	dannyMiller
mcm5-11	TRUE	control	dannyMiller
mcm5-10	TRUE	control	dannyMiller
mcm5-09	TRUE	control	dannyMiller
mcm5-08	TRUE	control	dannyMiller

name	paired	experimental	source
mcm5-07	TRUE	control	dannyMiller
mcm5-06	TRUE	control	dannyMiller
mcm5-05	TRUE	control	dannyMiller
mcm5-04	TRUE	control	dannyMiller
mcm5-03	TRUE	control	dannyMiller
mcm5-02	TRUE	control	dannyMiller
mcm5-01	TRUE	control	dannyMiller
DfMcm5	TRUE	control	dannyMiller

Total Starting Reads: $3.88G$ Post-QC Reads: $3.64G$.

Table 4: Read Count and Percent Retention

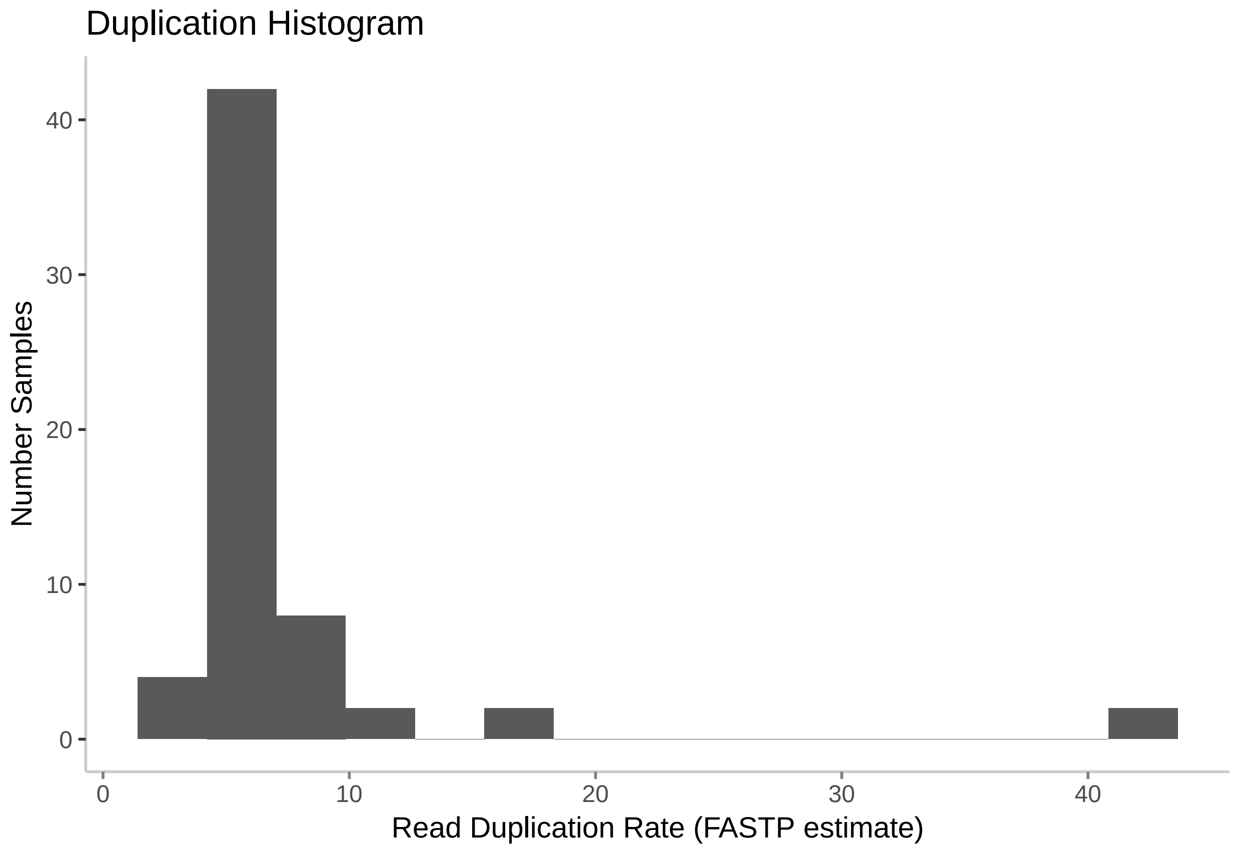
type	minimum	average	maximum
prefiltered	42.3 M	64.7 M	75.7 M
postfiltered	39.4 M	60.6 M	71.1 M
percent retention	92.4	93.7	95.5

This framework is general-purpose enough that it might be a good template............

Dupes:

Table 5: Percentage Duplication

minimum	average	median	maximum
4.1	7.6	5.8	43.5



27 February 2019

Bioinformatics tips on INDEL calling & normalization with DSB background: https://genome.sph.umich.edu/w/images/b/b4/Variant__Calling__and__Filtering__for__INDELs.pdf

5 March 2019

Going to go ahead and recycle BWA-Uniq but may want to change the algorithm later. . . .

Table 6: Read Counts During Alignment & Filtration

measure	minimum	average	median	maximum
filtered__mapped__count	17.5 M	40.5 M	42.1 M	48.5 M
total__mapped__count	38.1 M	58.6 M	60.3 M	68.9 M
total__read__count	39.4 M	60.6 M	62.7 M	71.1 M

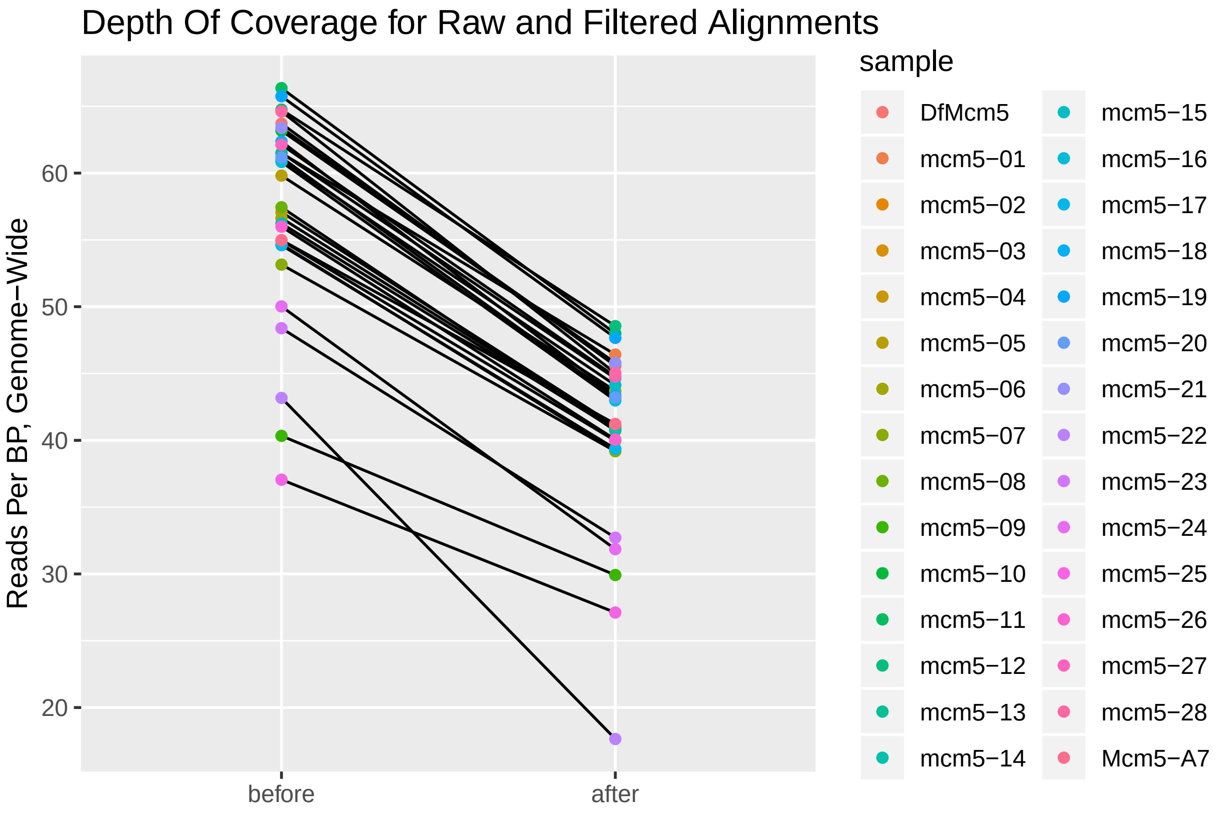
Table 7: Percentage of Reads Retained at Each Step

measure	minimum	average	median	maximum
filter__retention	39.7	68.8	70.1	73.8
mapping__retention	95.8	96.6	96.7	97.3

Depth of coverage:

Table 8: Depth of Coverage Statistics for Raw and Filtered Alignments

step	minimum	average	median	maximum
pre-filtration depth	37.1	57.2	58.6	66.4
post-filtration depth	17.6	40.7	42.1	48.5
depth retention percent	40.9	70.9	72.3	75.5



Breadth of coverage:

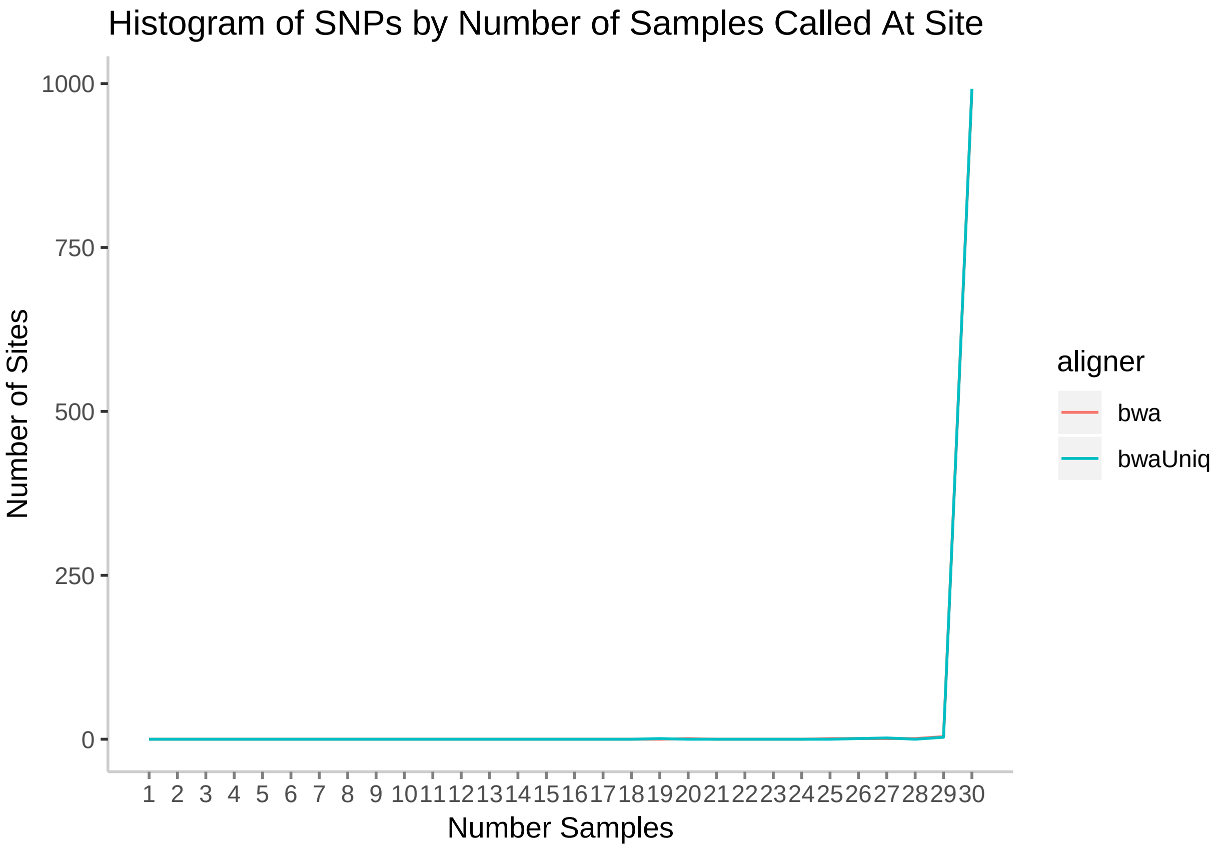
Will run the VCF caller on both BWA and BWA-Uniq; reporting will be reworked since we're interested in indels.

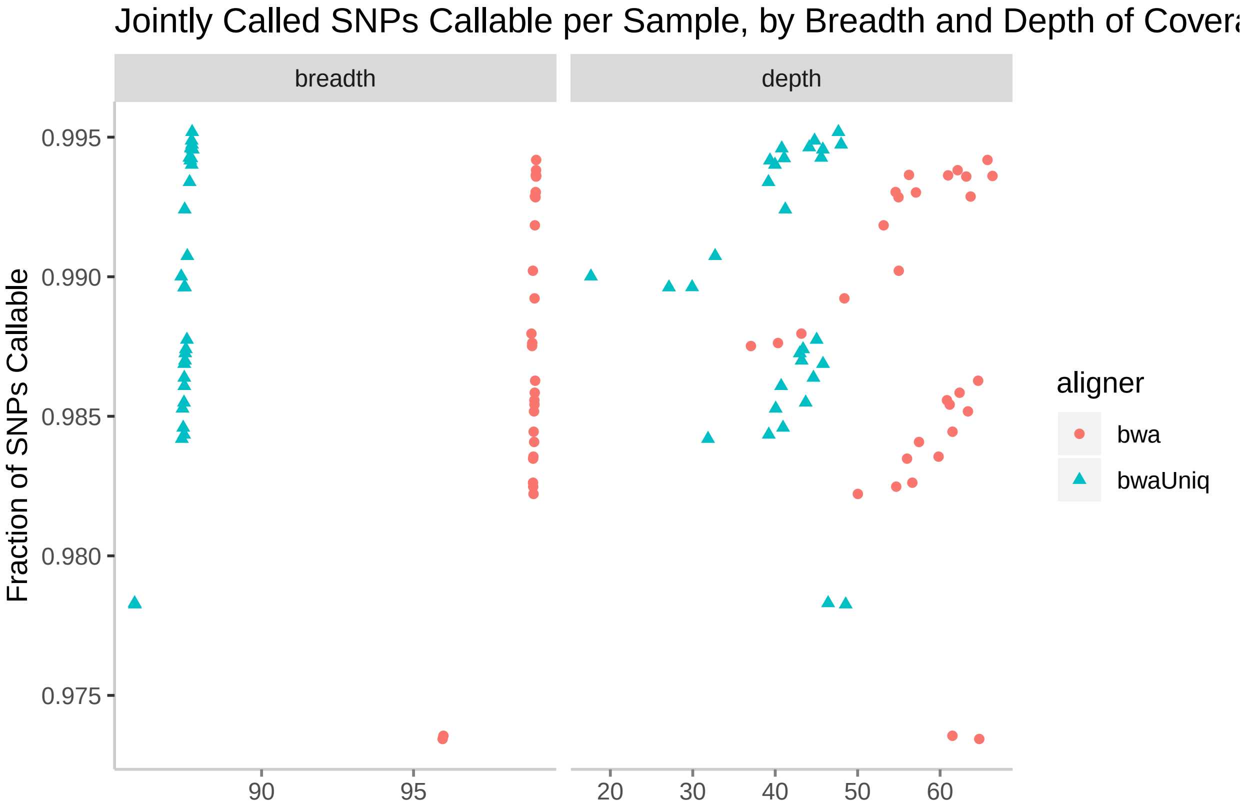
5 March 2019

Doing things a little differently, calling variants from both BWA and BWA-Uniq, then compare the two. (whereas before we used reference genome as a variable)

6 March 2019

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


Histogram of SNPs by Number of Samples Called At Site
Number of Sites
1000
750
500
250
0
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
Number Samples
aligner
bwa
bwaUniq


Jointly Called SNPs Callable per Sample, by Breadth and Depth of Covera
breadth
depth
Fraction of SNPs Callable
0.995
0.990
0.985
0.980
0.975
90
95
20
30
40
50
60
aligner
bwa
bwaUniq