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

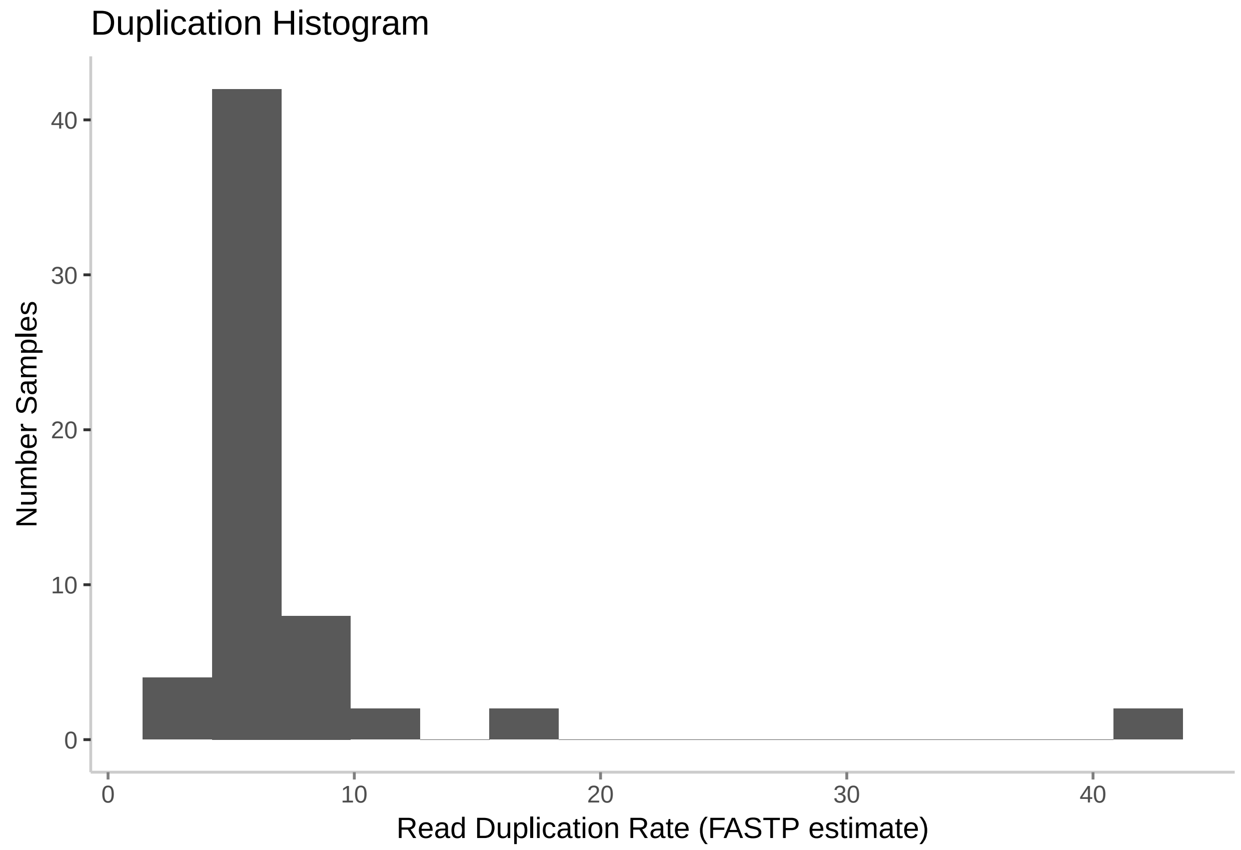
type	minimum	average	maximum
prefiltered	42.3 M	64.7 M	75.7 M
postfiltered	39.4 M	60.6 M	71.1 M
percent retention	92.4	93.7	95.5

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

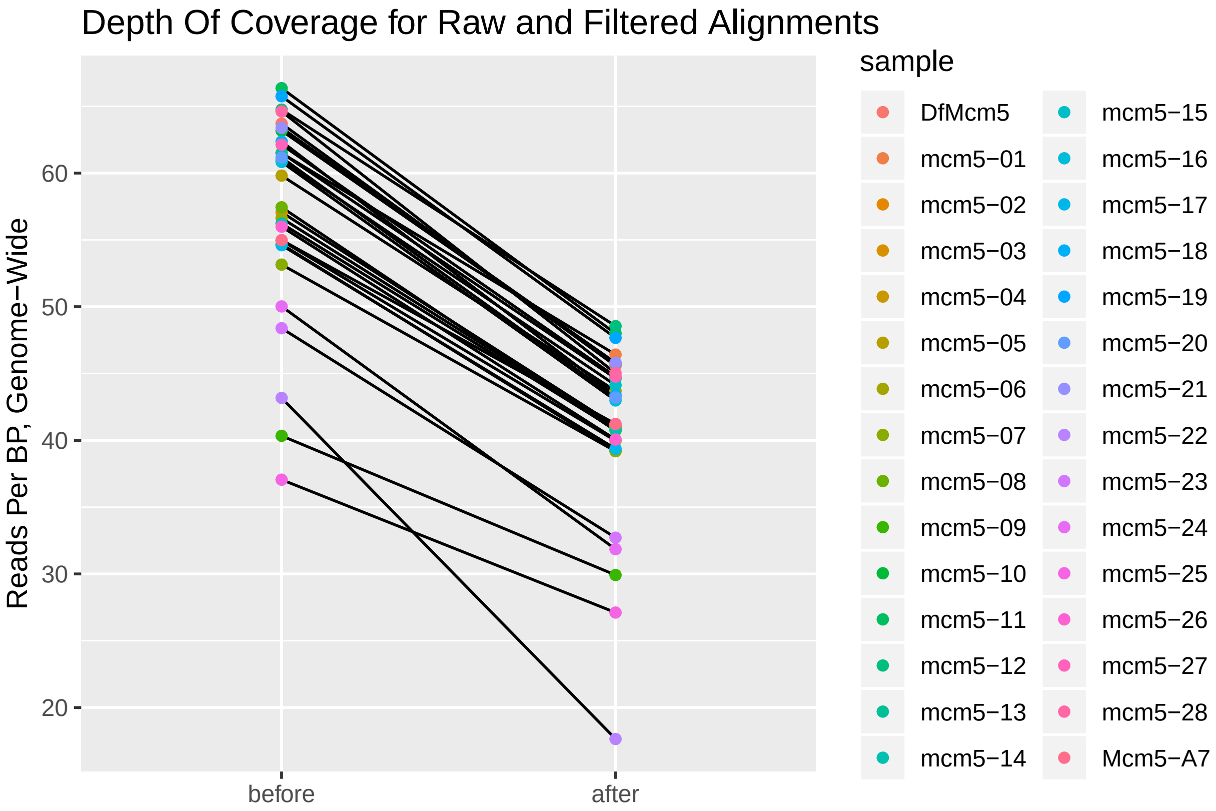
Table 7: Percentage of Reads Retained at Each Step

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.

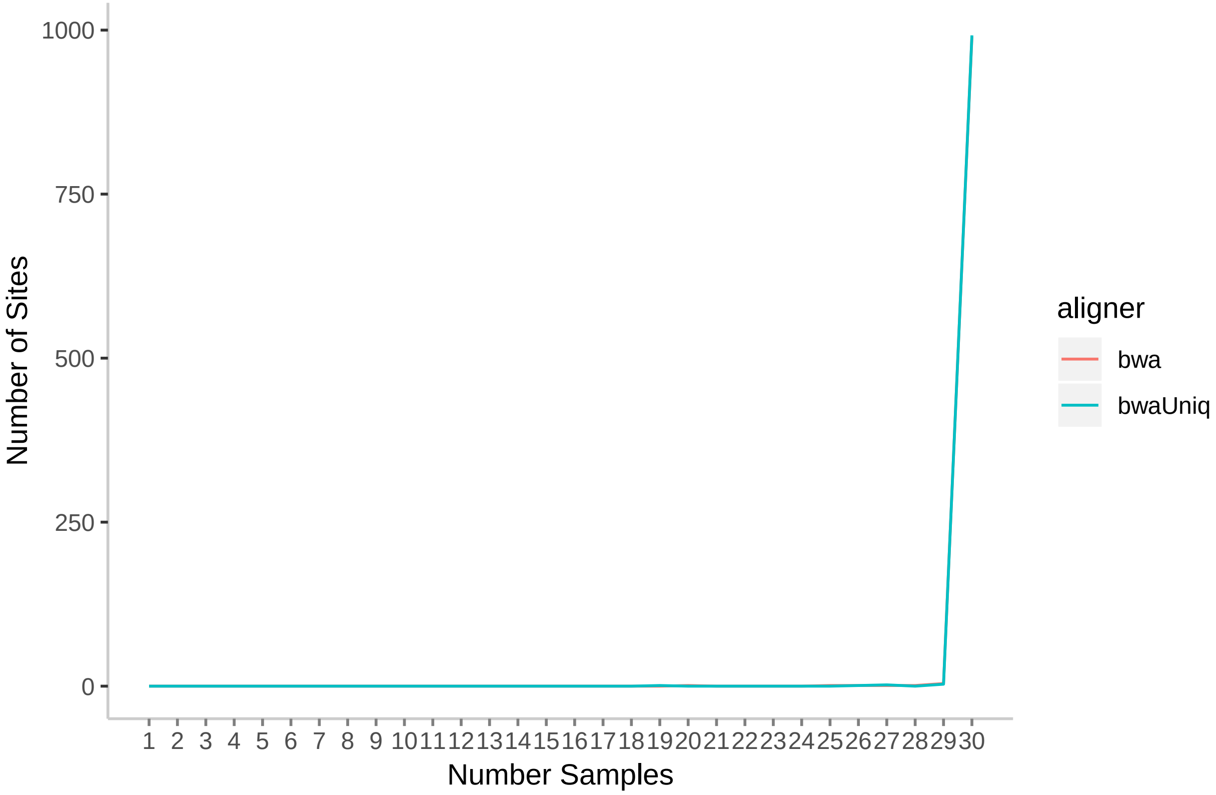
5 March 2019

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)

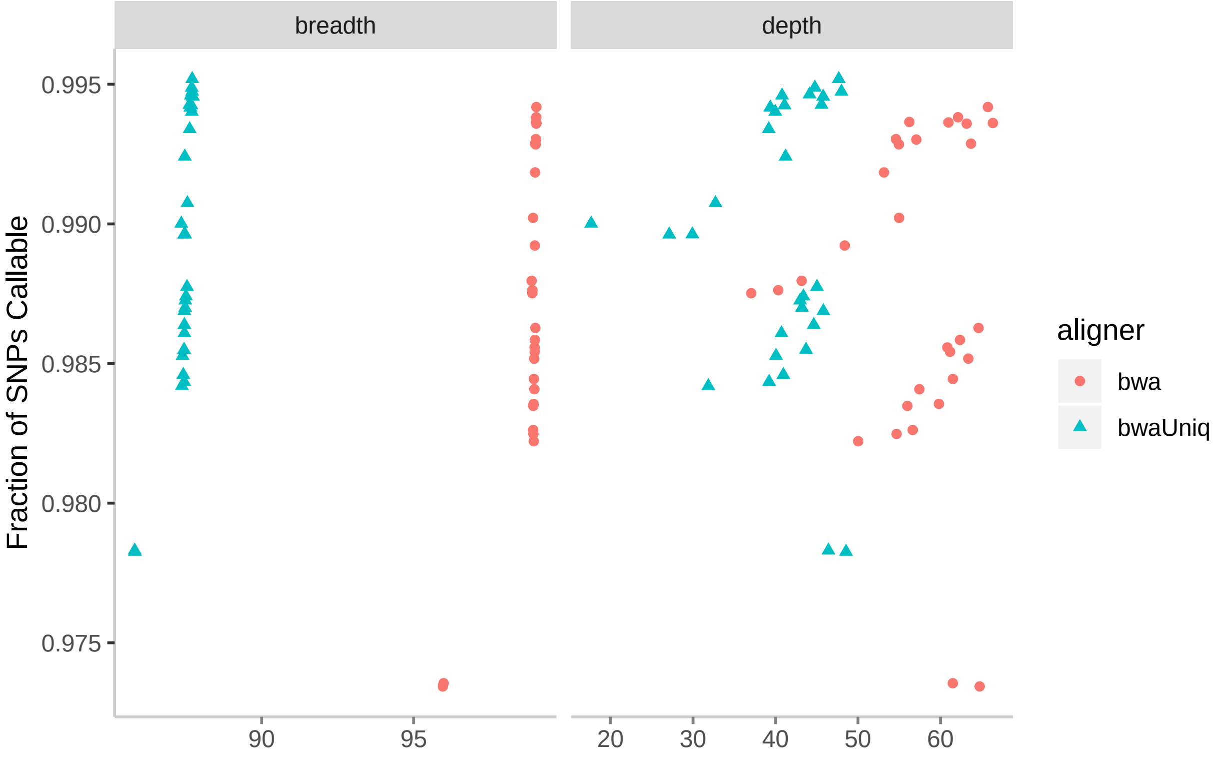
6 March 2019

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

Histogram of SNPs by Number of Samples Called At Site



Jointly Called SNPs Callable per Sample, by Breadth and Depth of Covera



7 March 2019

Might also be good to do a comparison between the two VCFs using vcftools –diff.

This complains: Error: Cannot determine chromosomal ordering of files, both files must contain the same chromosomes to use the diff functions. Found chrUn_DS483679v1 in file 1 and chrUn_DS483680v1 in file 2.

Let's try using the –chr command to limit to main-line chromosomes. . . .

There is also the –diff-site-discordance flag:

"The MATCHING_ALLELES column tells you if the alleles called in file match exactly at that site (i.e the REF and ALT columns are identical in the two files). The N_COMMON_CALLED column tells you the number of individuals at that site that were called in both files (i.e. the individuals in the intersection of the two datasets that don't have missing data ./.). The N_DISCORD column tells you the number of individuals in the intersection that are discordant at that site." -Adam Auton

https://sourceforge.net/p/vcftools/mailman/message/27128665/

also maybe use –diff-indv-discordance then compare individual discordance to e.g. breadth reduction upon BAM filtration

Locii variable in A only:

in B only:

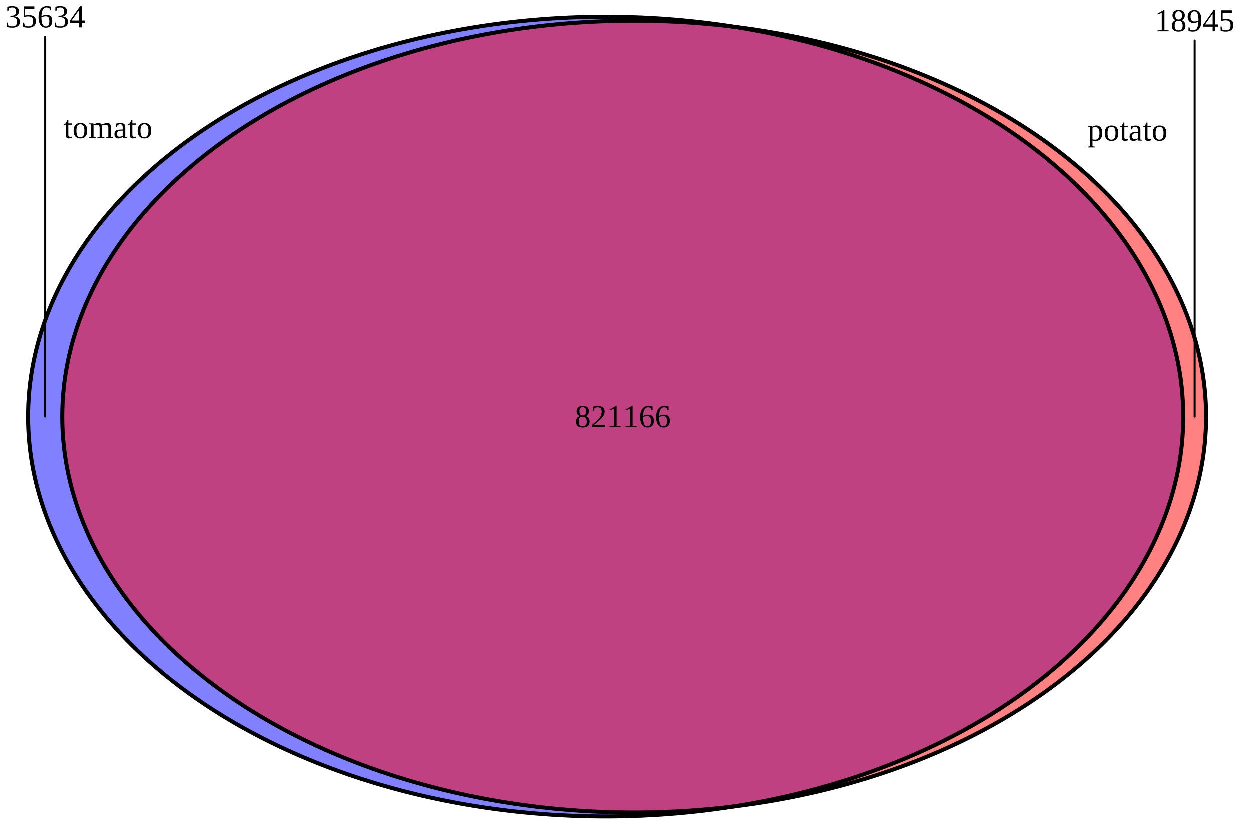
Locii variable in both A and B:

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   X2 = col_character()
## )
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   X2 = col_character()
## )
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   X2 = col_character()
## )
```

https://rstudio-pubs-static.s3.amazonaws.com/13301_6641d73cfac741a59c0a851feb99e98b.html

```
## Loading required package: grid
```

```
## Loading required package: futile.logger
```



```
## (polygon[GRID.polygon.1256], polygon[GRID.polygon.1257], polygon[GRID.polygon.1258], polygon[GRID.pol
```

of the ones in the intersection:

Locii variable in A and B but with at least one discordant individual:

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   X2 = col_character()
## )
```

a total of 7.6637×10^4 sites with at least one discordant individual.