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

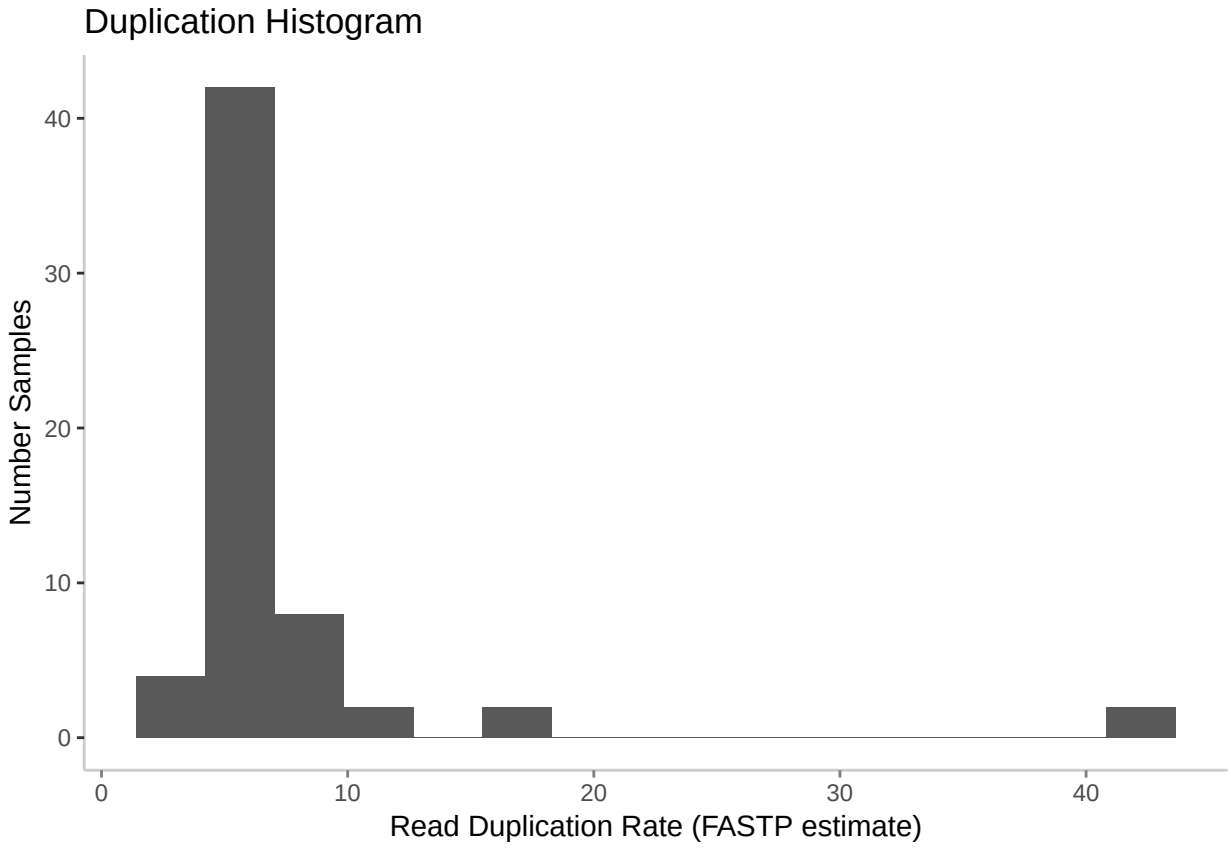
type	minimum	average	maximum
prefiltered	42.3 M	64.7 M	75.7 M
postfiltered	39.4 M	60.6 M	71.1 M
percent retention	92.4	93.7	95.5

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

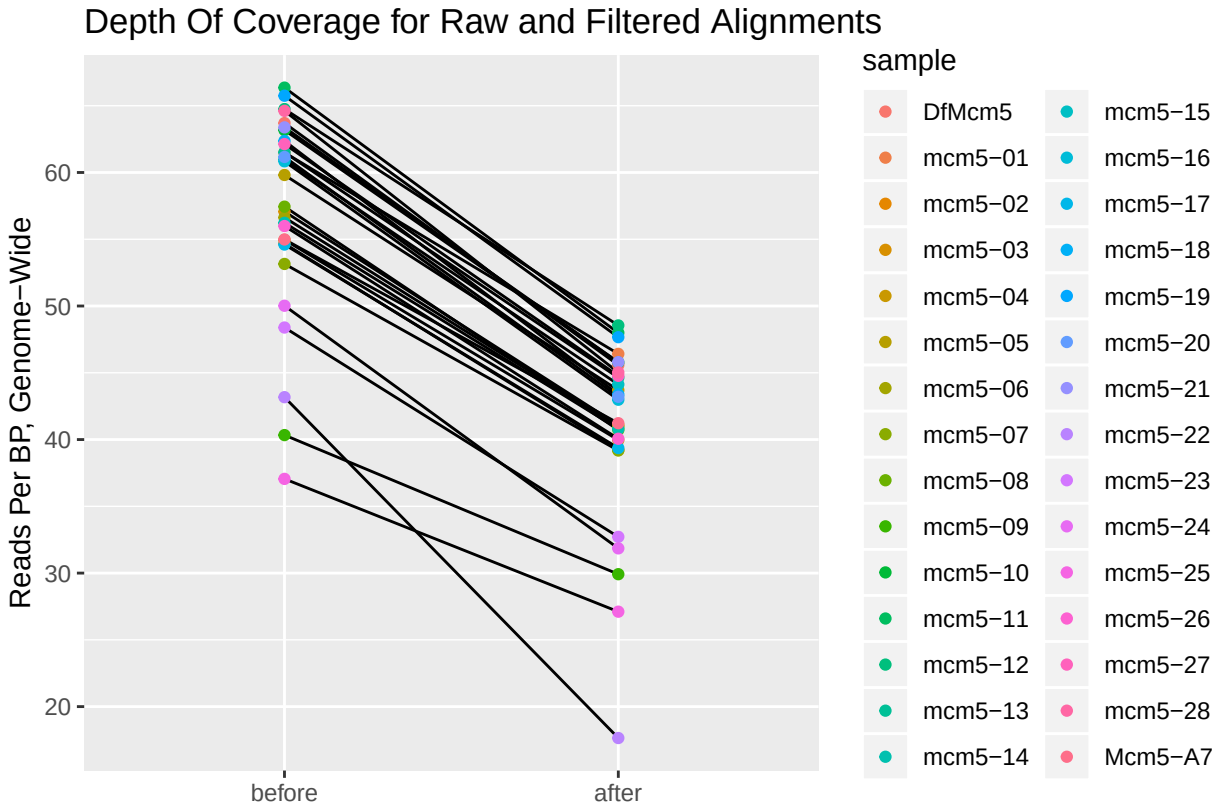
Table 7: Percentage of Reads Retained at Each Step

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:

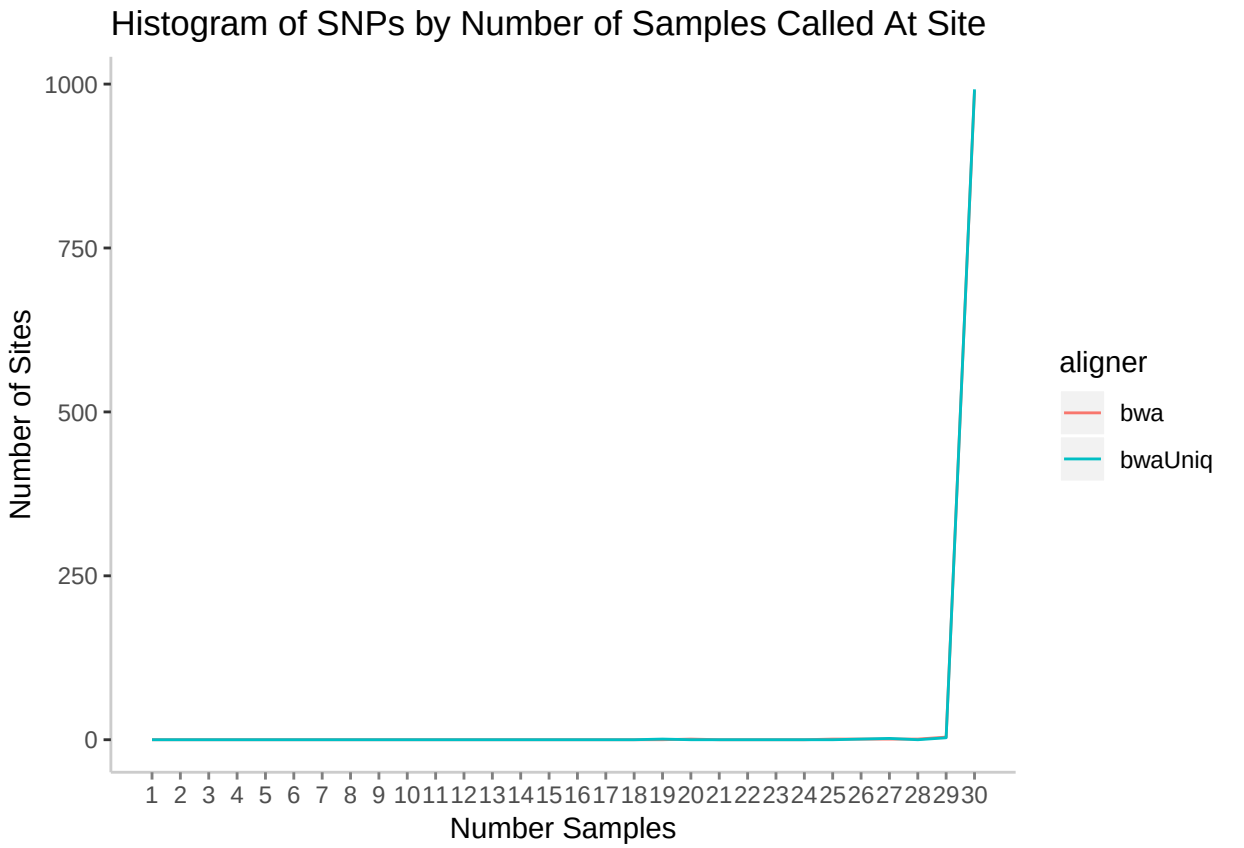
Will run the VCF caller on both BWA and BWA-Uniq; reporting will be reworked since we're interested in indels.

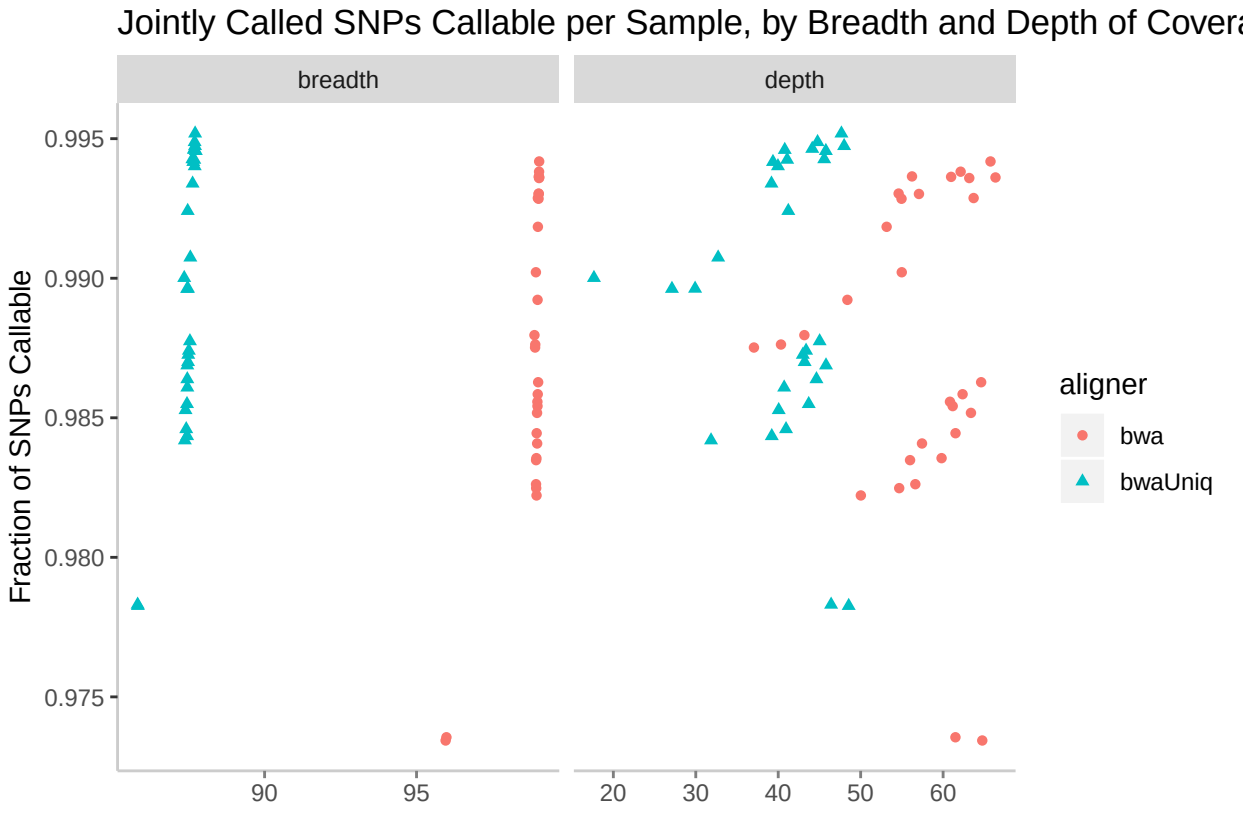
5 March 2019

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

7 March 2019

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

```
vcftools --vcf variants/all_samples.vs_dm6.bwa.vcf --diff variants/all_samples.vs_dm6.bwaUniq.vcf --dif
```

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. . . .

```
vcftools --vcf variants/all_samples.vs_dm6.bwa.vcf --diff variants/all_samples.vs_dm6.bwaUniq.vcf --dif
```

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:

```
cat dev/meta/VCFs/comparisonTest.diff.sites | tail -n +2 | awk '{if($3 == 1)print;}' | cut -f 1 | uniq
```

```
cat dev/meta/VCFs/comparisonTest.diff.sites | tail -n +2 | awk '{if($3 == 1)print;}' | cut -f 1 | uniq
```

in B only:

```
cat dev/meta/VCFs/comparisonTest.diff.sites | tail -n +2 | awk '{if($3 == 2)print;}' | cut -f 1 | uniq
```

```
cat dev/meta/VCFs/comparisonTest.diff.sites | tail -n +2 | awk '{if($3 == 2)print;}' | cut -f 1 | uniq
```

Locii variable in both A and B:

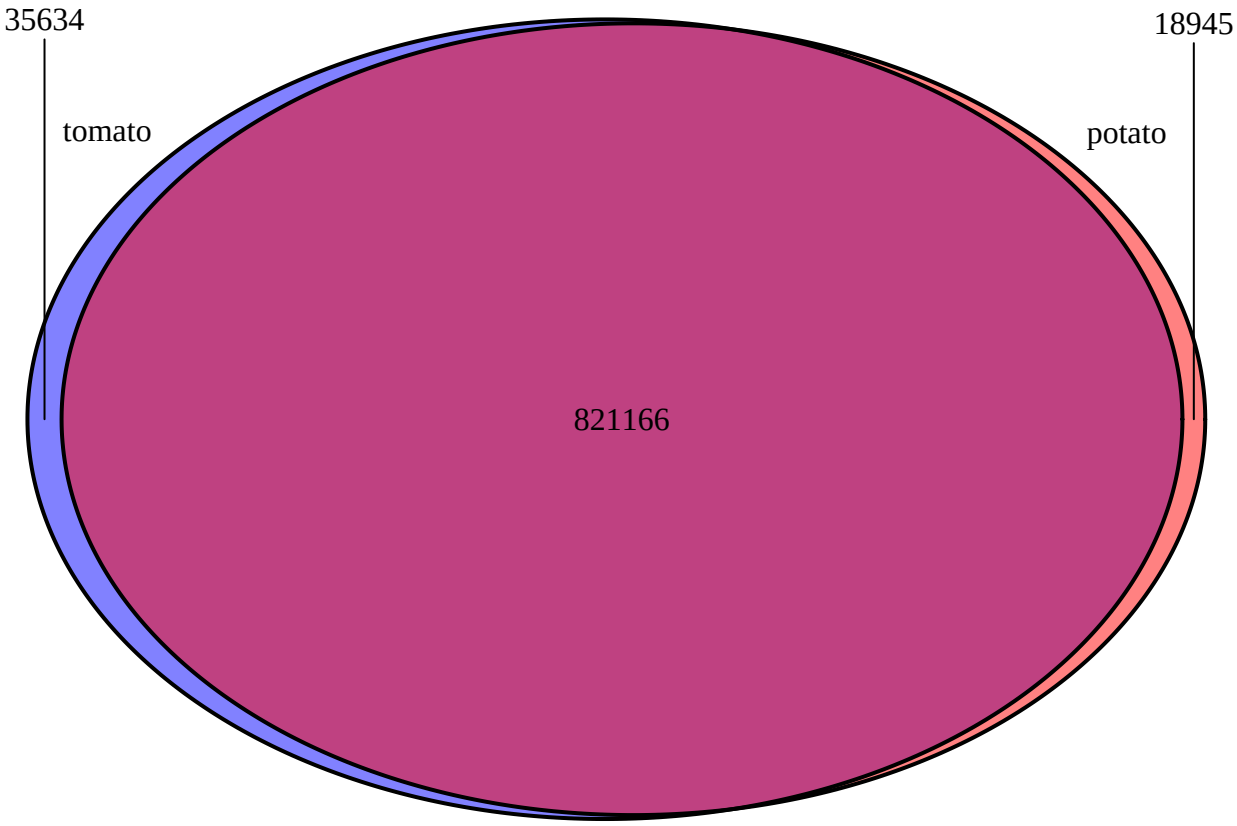
```
cat dev/meta/VCFs/comparisonTest.diff.sites | tail -n +2 | awk '{if($3 == "B")print;}' | cut -f 1 | unic
```

```
cat dev/meta/VCFs/comparisonTest.diff.sites | tail -n +2 | awk '{if($3 == "B")print;}' | cut -f 1 | unic
```

```
## 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:

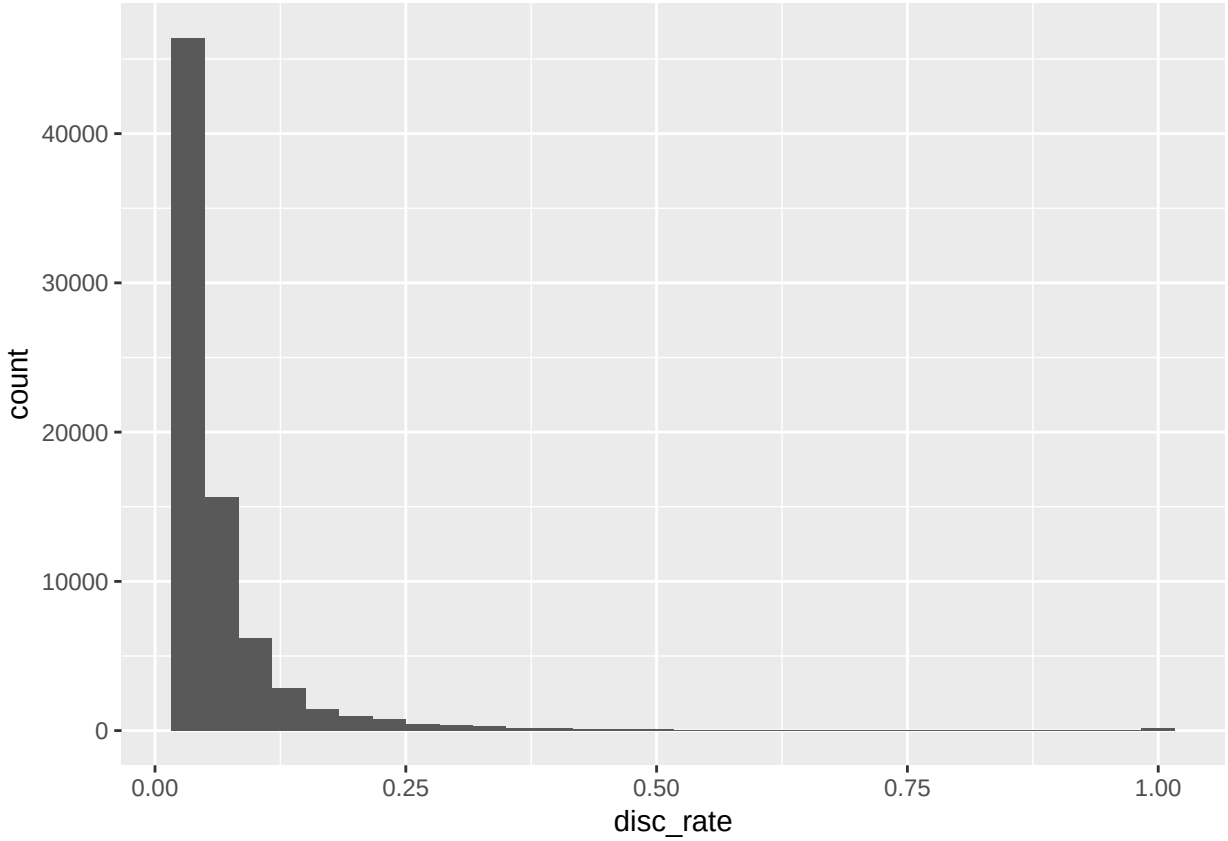
Locii variable in A and B but with at least one discordant individual:

```
## 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.

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   X2 = col_double(),
##   X3 = col_character(),
##   X4 = col_double(),
##   X5 = col_double(),
##   X6 = col_double(),
##   X7 = col_double()
## )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



OK so that's some VCF comparison infrastructure; how to implement? Here we are comparing VCFs from two different alignment/calling strategies but ultimately we'll probably want to compare subgroups of all_samples. Do I use the same rule for both? It's possible that down the line when an alignment/calling strategy is decided upon, I won't want something so general...

8 March 2019

Looking at INDELs within the control subset...

```
vcftools --vcf variants/all_samples.vs_dm6.bwaUniq.vcf --keep-only-indels --freq --out test
vcftools --vcf variants/all_samples.vs_dm6.bwaUniq.vcf --keep-only-indels --counts --out test
```

hmmm, we're presumably looking for individual-specific alleles; there could potentially be more than one such allele at a given site! (these show up in the VCF as eg 0/2)

The INDELs .... will they be het or hom??? Presumably Het, since the break repair only happens on one strand....

Finding all the variants from the above count tallies with only one chromosome in the population carrying; awk + grep for more filtration.

```
cat test.frq.count | grep -P ':1[$,\t]'
```

10 March 2019

Background: Figure 1 of (McVey and Lee 2008) and https://en.wikipedia.org/wiki/Non-homologous_end_joining

It looks like NHEJ and MMEJ are prone to forming heterozygous indels near the DSB site. (am i right that MMEJ at least will form flanking indels?)

Are the DSBs random across the genome? What are the odds that the same site would be struck by DSB and repair error twice in the population (ie, 30 flies)? if it's very low, we need to look for indels which are heterozygous for an indel in one individual but homozygous in everyone else.

One data subtelty here is that the called variants are vs the ref genome so the ancestral genotype is presumably the one with the higher (f~1.0 in approx isogenic pop) allele frequency. Thus a reversion to het dm6 reference (0/1) from a population of hom variants (1/1) would be a candidate site. Also, since a site which is different from the reference may later be altered via DSB repair (ie, in a 1/1 population a 1/2 genotype with low AF on 2 is a candidate.)

figs to make: * histogram of minor allele frequency for indel variants. Many will have a MinAF of 0 (ie, fixed differences); filter these? others will have a small minAF corresponding to e.g. 1 chromosome in the population.

if there is a flanking behavior this should be easy to pick out: do a histogram of intrachromsomal distance among variants and then to a freq_poly plot binned by genotype (ie, hom vs het). Look for heterozygotes which are neighbors.

From (Miller et al. 2016): "Drosophila oocytes experience ~11–17 DSBs per meiosis that are restricted to the euchromatin ... How the position of these DSBs is determined and their fate (whether they become COs or NCOs) is poorly understood." also discusses recombinational hotspots; evidence is against them??

Table S2 for crossover sites identified in (Miller et al. 2016); Table S3 for noncrossover sites.

From Danny Miller 11 Jan 2019:

*First, I called SNPs using GATK. From this output I isolated novel deletions for both chromosome X and 2. I re-wrote this script tonight and re-ran it and found four likely novel deletions that I hadn't seen before:

mcm5-12, chrX:12825436, GAAA deletion mcm5-21, chr2R:12737873, A deletion mcm5-22, chr2R:8597548, T deletion mcm5-24, chr2L:21664793, TATATA deletion*

his main finding:

I really expected to find lots of deletions suggesting repair of DSBs via NHEJ, but I didn't. This finding is consistent with the 95-ish single genomes I sequenced from homozygous c3g females where I also failed to find any deletions.

So, maybe another figure would be deletions vs. the major allele? eg, histogram of indel change in nucleotides.

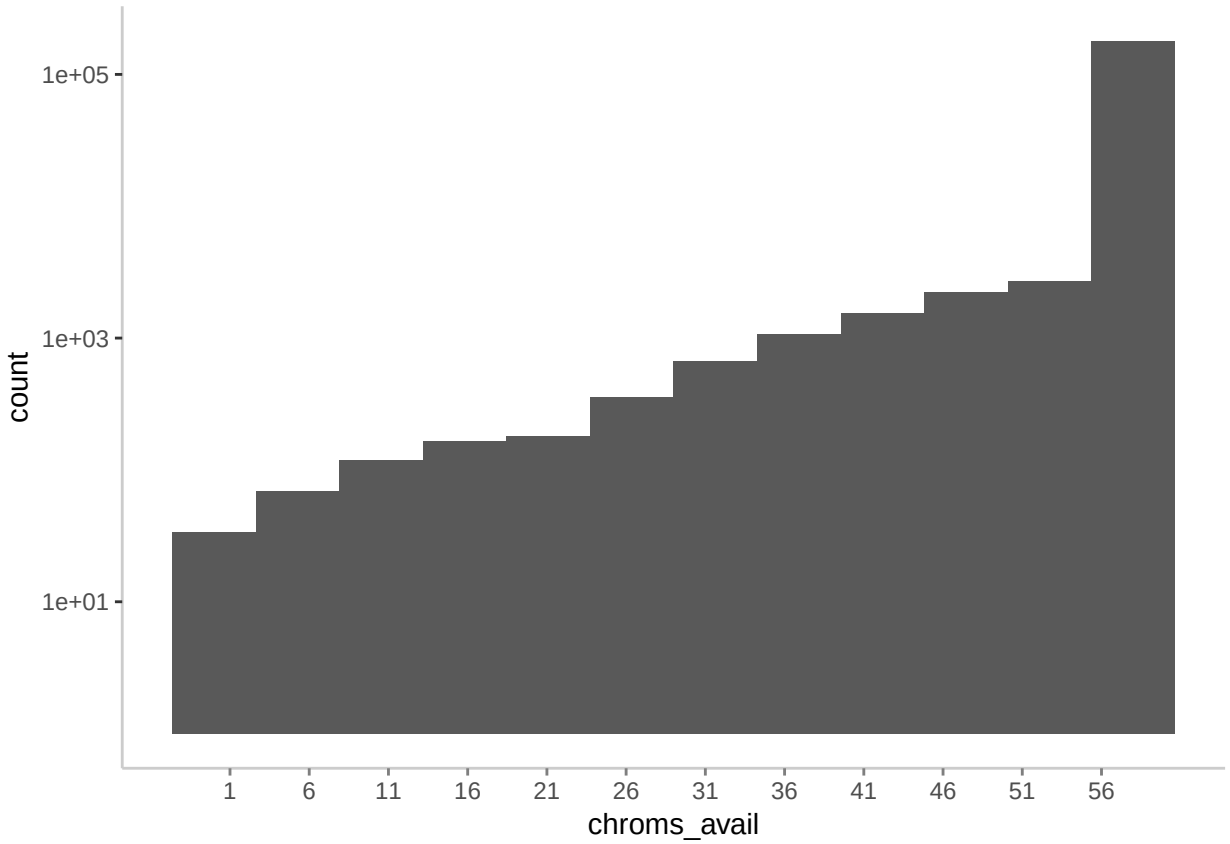
11 March 2019

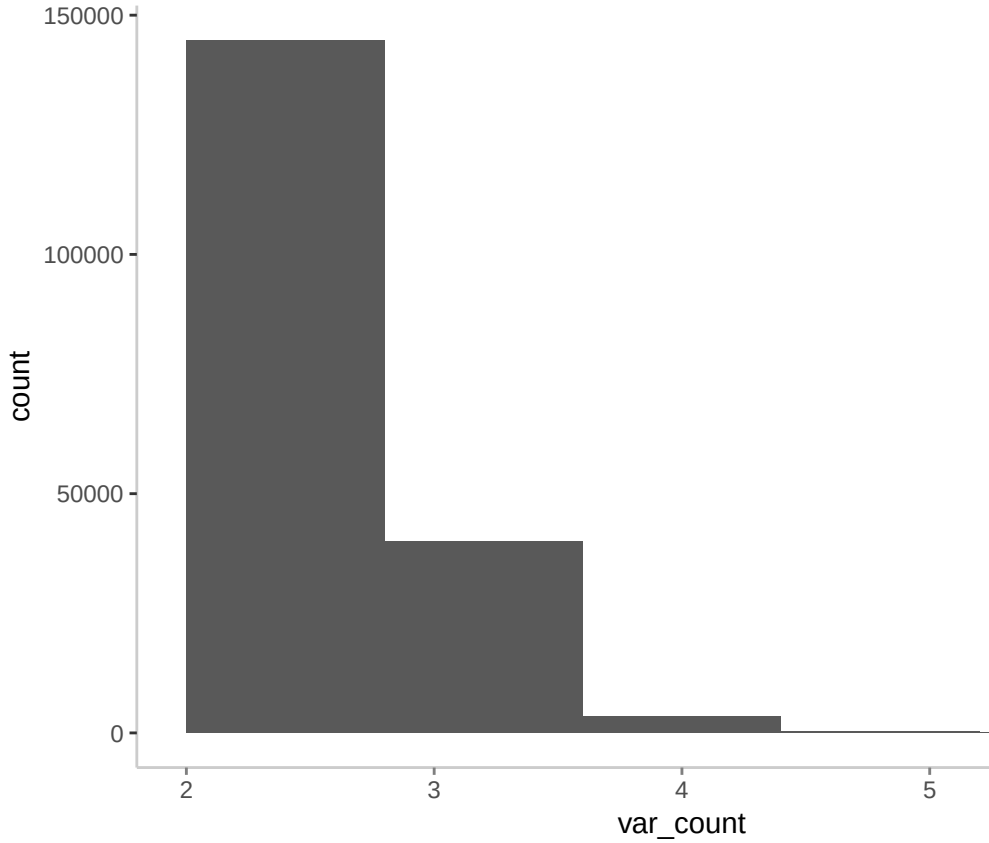
```
vcftools --vcf variants/all_samples.vs_dm6.bwaUniq.vcf --keep-only-indels --freq --out indelFrq.test

 vcftools --vcf variants/all_samples.vs_dm6.bwaUniq.vcf --keep-only-indels --counts --out indelCount.tes
cat indelCount.test.frq.count  | tail -n +2 | tr ":" "\t" | nl -n ln | head -n 1000 > dev/indelCount.tes
```

importing a CSV with a variable number of columns: https://stackoverflow.com/questions/18922493/how-can-you-read-a-csv-file-in-r-with-different-number-of-columns

Filter to main-line autosomes?

Just some top-level stuff: how many chroms avail per site?

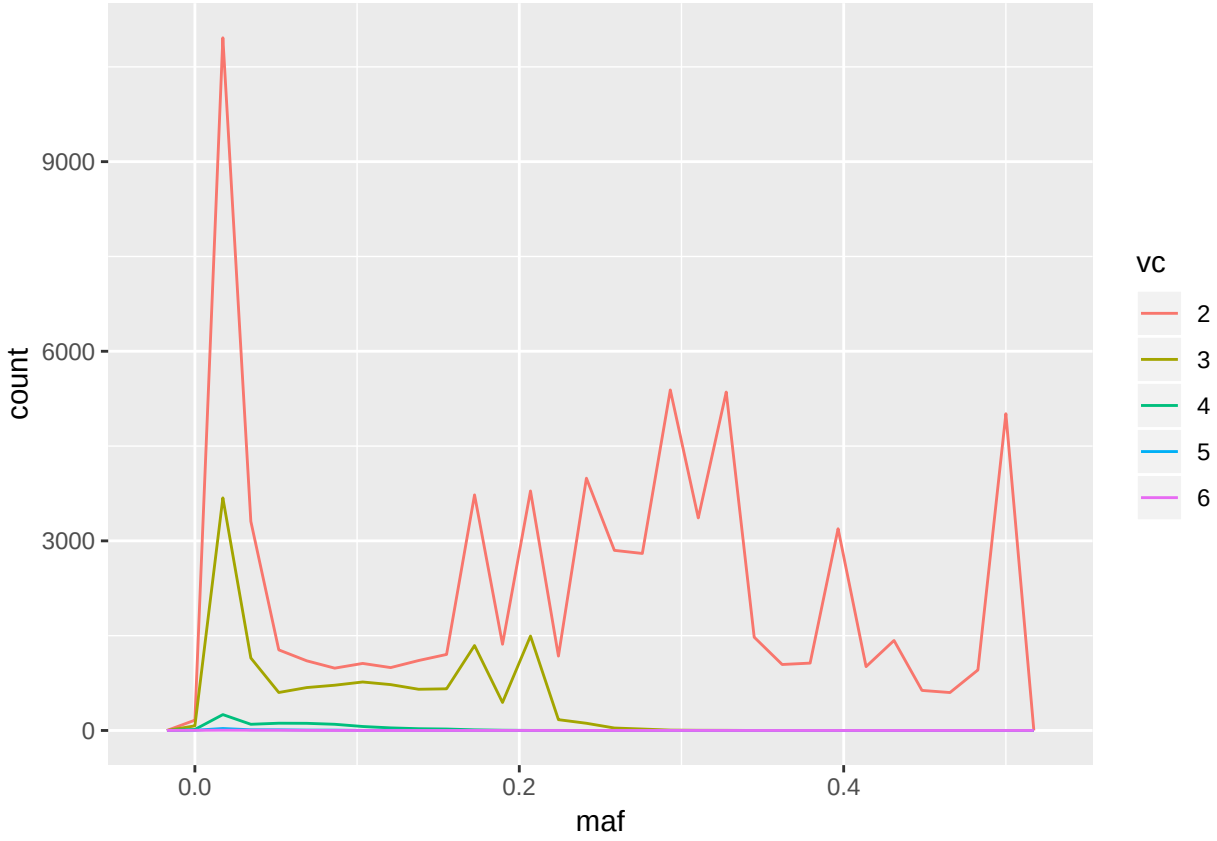




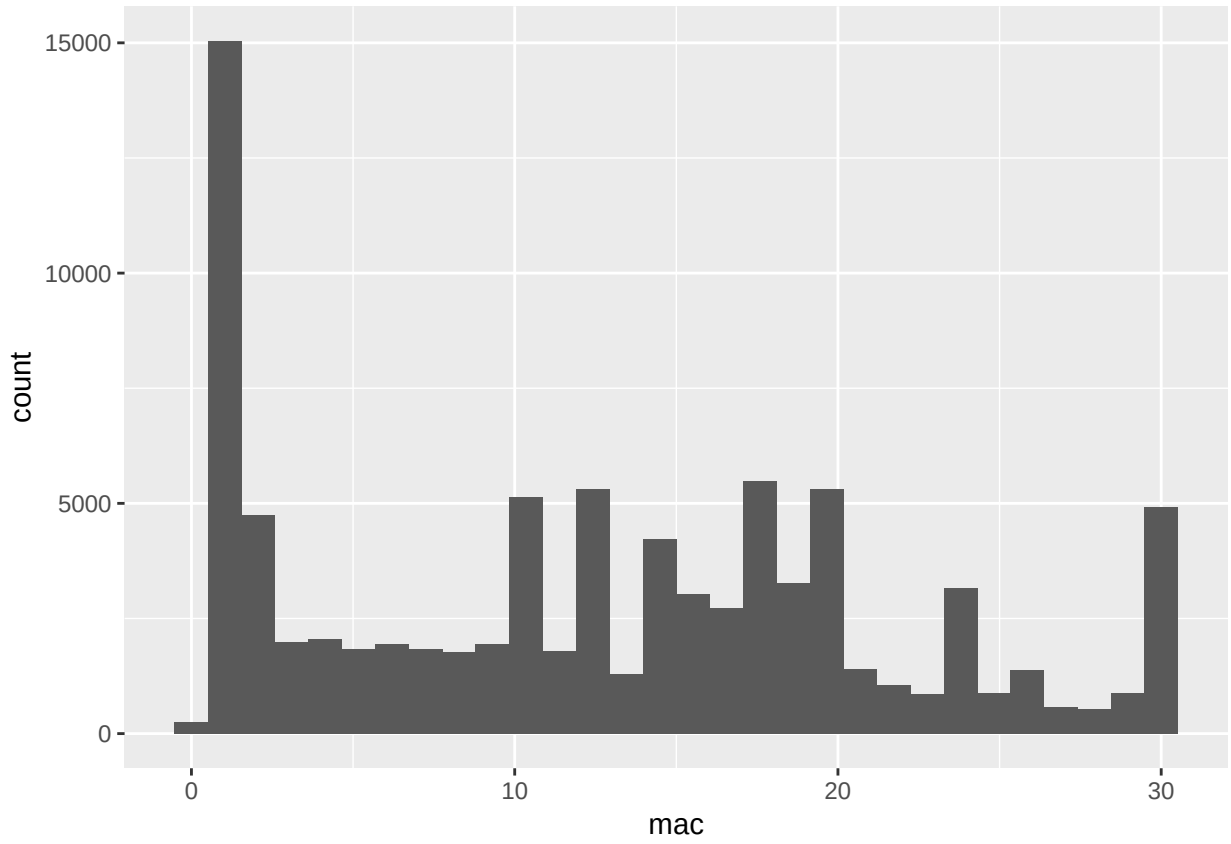
Distribution of variants per variant site?

Minimum allele frequency by site:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Let's start with the simplest case: biallelic sites with all 60 chromosomes available and allele count of 1 (ie, one het indiv.) hmmm, a number of MNPs showing up; let's try to filter those out for a really simple case.

```
grep "#" variants/all_samples.vs_dm6.bwaUniq.vcf > variants/all_samples.vs_dm6.bwaUniq.vcf.noMNP.tmp

vcftools --vcf variants/all_samples.vs_dm6.bwaUniq.vcf --keep-only-indels --recode --recode-INFO-all --s



vcftools --vcf variants/all_samples.vs_dm6.bwaUniq.vcf.noMNP.tmp --keep-only-indels --counts --out indel

cat indelCount.test.frq.count  | tail -n +2 | tr ":" "\t" | nl -n ln | head -n 1000 > dev/indelCount.tes
```

```
## [1] 10332
```

Hmmm, that seems like a lot. Let's get all the mainline chroms and look at them

```
cat indelCount.test.frq.count  | tail -n +2 | tr ":" "\t" | nl -n ln | grep -v Un | grep -v rand  > dev/
```

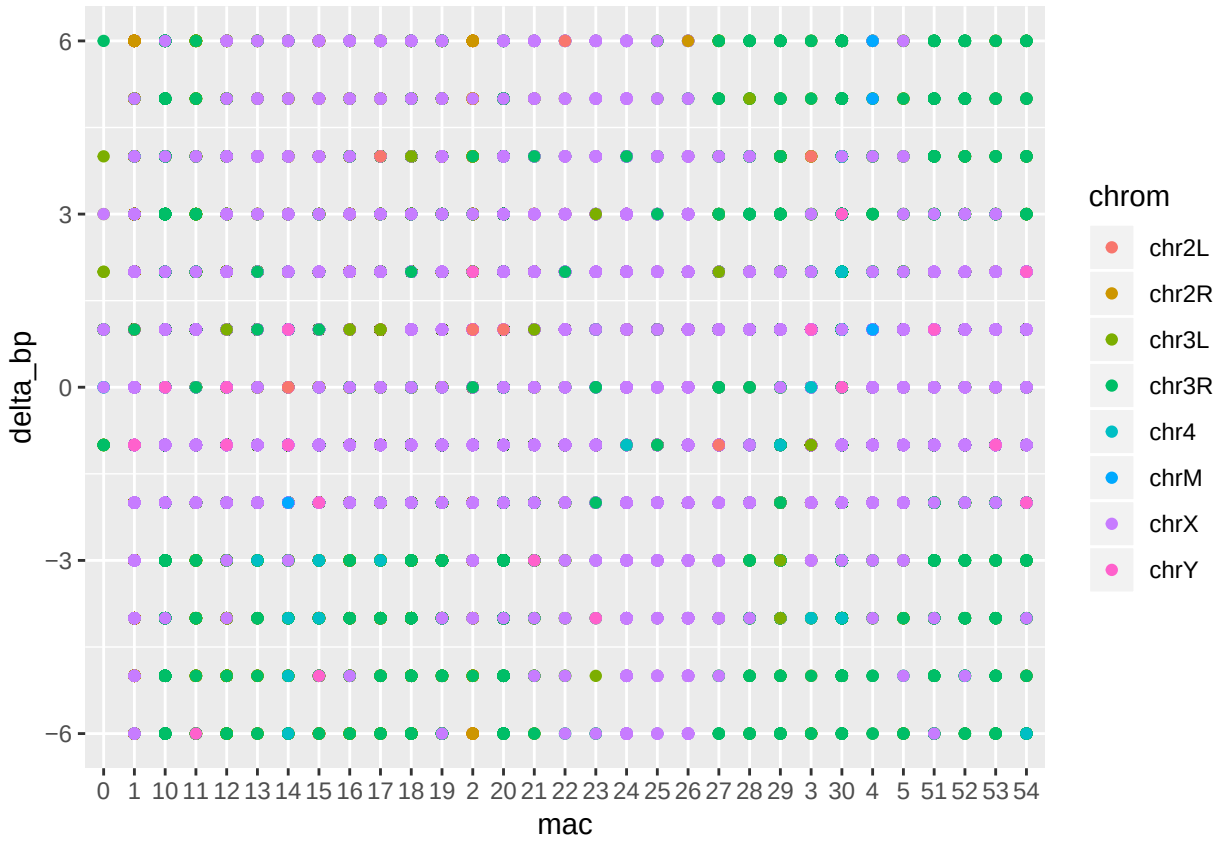
Hmm, yes 10k+ is a lot???

Corbin & Talia suggest a histogram of delta-bp for each indel...

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

spread by multiple columns: https://stackoverflow.com/questions/30592094/r-spreading-multiple-columns-with-tidyr

calculate delta-bp relative to the larger allele count or if it's 50/50, use the reference (allele 1)



ugh who knows, I'll figure it out tomorrow

13 March 2019

Ah, the variants with AC == 0 are examples of fixed differences vs. the dm6 reference.

It looks like the variants with delta_bp == 0 are cases of ??recognized SNPs?? being counted as indels?? what??

Maybe something like

```
cat variants/all_samples.vs_dm6.bwaUniq.vcf | ~/modules/vcflib/bin/vcfallelicprimitives | vcftools --vc
```

hmmm yes, this seems right.

maybe add some depth-filtering to improve genotype calls? avg depth is pretty high so it could be set stringent. The –minDP 10 flag in vcftools will do this (sets the called genotype to ./.; –counts then tallies this GT as uncalled.)

maybe something like

```
cat variants/all_samples.vs_dm6.bwaUniq.vcf | vcftools --vcf -  --minDP 10 --recode --recode-INFO-all -
```

??

Pi Day 2019

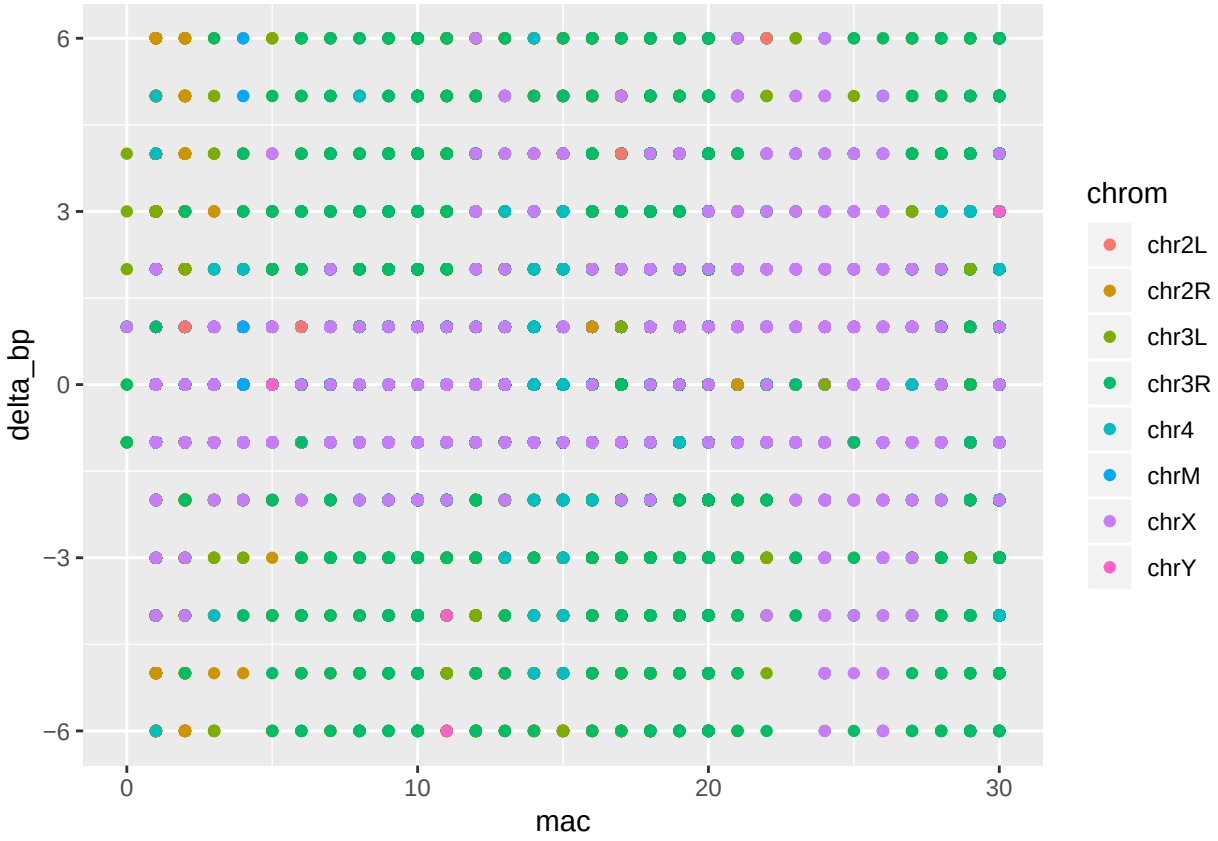
Hmmm, looks like vcfallelicprimatives breaks up MNPs and such, possibly unwanted. Maybe shunt the MNPs down a different path? "complex" repeats it doesn't seem to touch.

modding the tomato.count.frq.count file from yesterday...

```
cat variants/all_samples.vs_dm6.bwaUniq.vcf | vcftools --vcf -  --minDP 10 --recode --recode-INFO-all --
```

```
cat tomato.count.frq.count | tail -n +2 | tr ":" "\t" | nl -n ln | grep -v Un | grep -v rand  > dev/toma
```

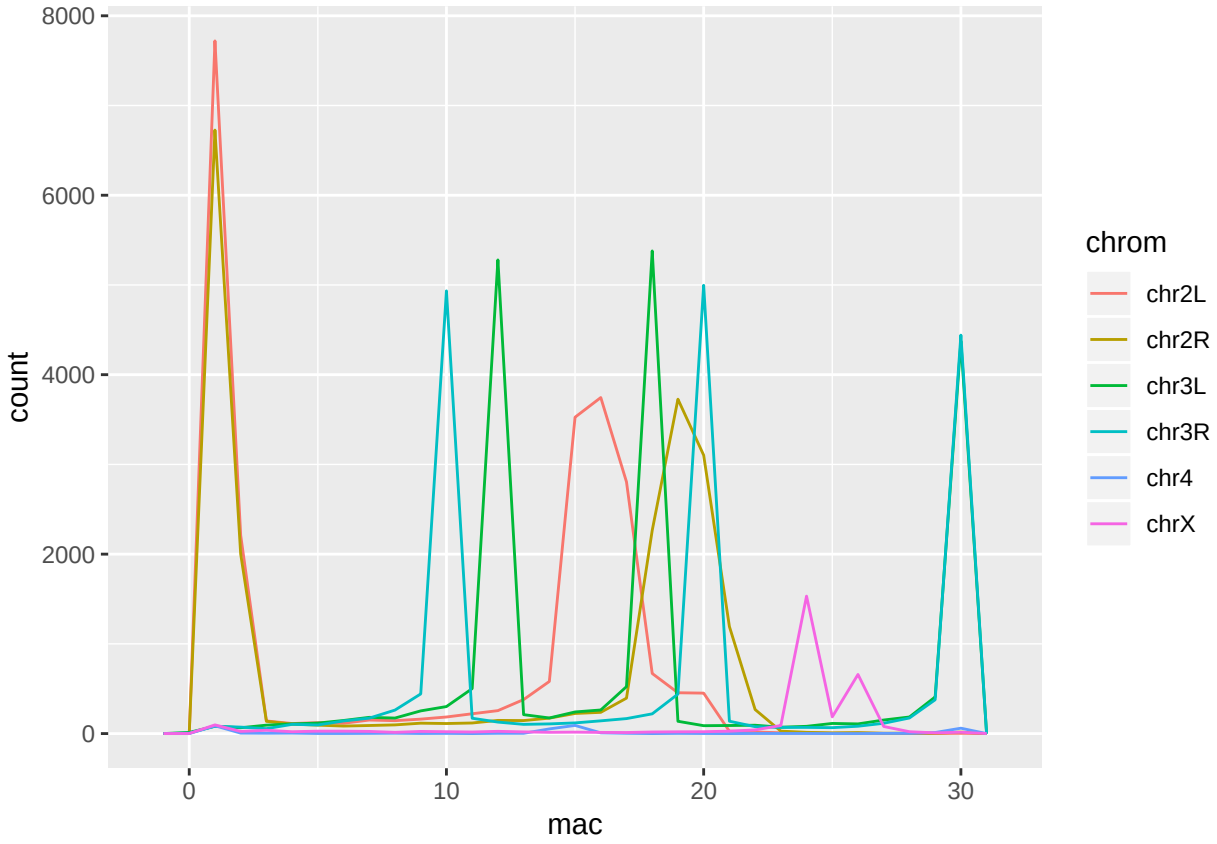
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```



Ok, I see part of the problem. the unite() pipe segment recasts the column as a string apparently?

15 March 2019

Apparently these are the offspring of a single mating between two flies? That makes the sharp peaks of the histogram more reasonable, but not the lopsided chromosomal distribution:



Getting some weird errors on the proposed pipeline so going through stepwise:

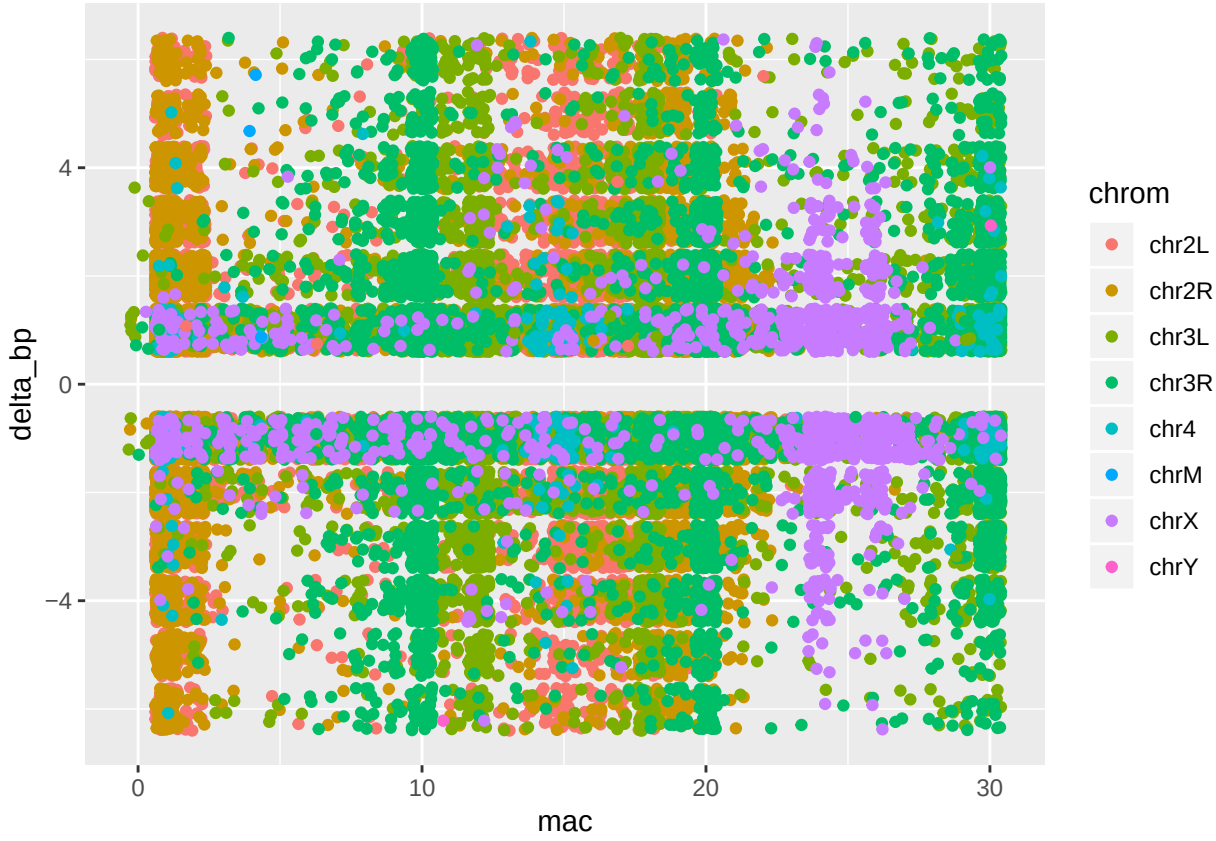
```
cat variants/all_samples.vs_dm6.bwaUniq.vcf | vcftools --vcf -  --minDP 10 --recode --recode-INFO-all --

 cat test.depthFiltered.recode.vcf | grep -v "TYPE=complex" | ~/modules/vcflib/bin/vcfallelicprimitives
```

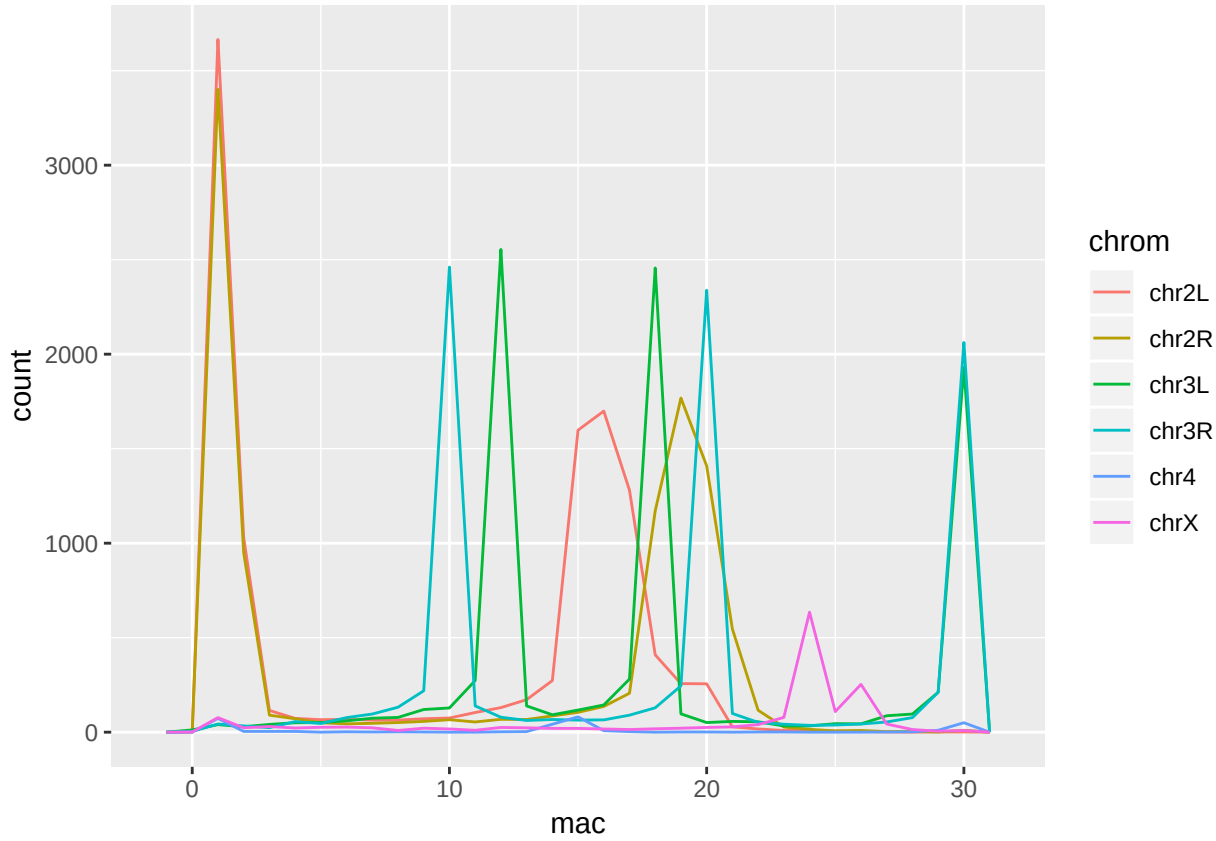
hmm it looks like the sites which are broken down by vcfallelicprimitives are in a different format than the SNPs.

Also, vcfallelicprimitiives reclassifies some variants as complex so might as well exclude them afterwards....

```
vcftools --vcf test.depthFiltered.simplified.primitive.vcf --keep-only-indels --recode --recode-INFO-all

vcftools --vcf test.depthFiltered.simplified.primitive.indelOnly.vcf   --counts --out test.depthFiltered

cat test.depthFiltered.simplified.primitive.indelOnly.count.frq.count | tail -n +2 | tr ":" "\t" | nl -n
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```


delta_bp
4
0
−4
0
10
20
30
mac
chrom
chr2L
chr2R
chr3L
chr3R
chr4
chrM
chrX
chrY


count
3000
2000
1000
0
0
10
20
30
mac
chrom
chr2L
chr2R
chr3L
chr3R
chr4
chrX

ok so this looks like some data sanity problems are fixed (eg no indels of length 0) but there are still some issues (eg the bizarre lopsided distribution of allele frequency across chromosomes. . . .)

Returning to the Danny Miller calls:

> mcm5-12, chrX:12825436, GAAA deletion mcm5-21, chr2R:12737873, A deletion mcm5-22, chr2R:8597548, T deletion mcm5-24, chr2L:21664793, TATATA deletion

Slopped these locations 100bp either way and intersected with the VCF:

```
cat misc/dannysCalls.prebed | cut -f 2 -d " " | tr -d "," | tr ":" "\t"  | awk '{print $1, $2-100, $2+10

edtools intersect -wa -wb -a misc/dannysCalls.bed -b test.depthFiltered.simplified.primitive.indelOnly.v
```

The first, chrX:12825436, appears in the VCF but is called as a homozygote (both BWA an BWA-Uniq alignments agree here!)

The second, an A deletion at chr2R:12737873, appears in the VCF, appears to be heterozygous in two different flies (mcm5-21,mcm5-22). Also, there are two insertions (+A, +AA) at this site as well in the samples. (DfMcm5 and Mcm5-A7, respectively). Kind of a dodgy site even in the BWA-Uniq alignment??

The third, a T deletion at chr2R:8597548, appears in the VCF, heterozygous in a single individual. This is another site with 4 different insertion-deletion alleles called.

The fourth, a TATATA deletion at chr2L:21664793, doesn't appear in this VCF (filtered for depth, variants simplified, complex variants removed, indels only). However, it is picked up in the unfiltered BWA-Uniq VCF as a complex variant: TAC -> CAT. It is called as heterozygous in five individuals: mcm5-04,mcm5-03,mcm5-13,mcm5-18 mcm5-27,mcm5-19. mcm-24, instead of the TATATA deletion, is called as homozygous for the reference (the unfiltered BWA alignment for mcm5-24 has some 5 and 7bp indels but this still gets resolved as an MNP in the BWA-derived VCF. These reads are gone in BWA-Uniq.). The alignments give weak support for the existence of the TAC->CAT variant: the variant sites are there in the reads, but coverage is pretty low and the variation always seems to occur near the ends of the reads. In some cases the complex variant has been imputed from a single SNP near the end of a read.

Bibliography

McVey, Mitch, and Sang Eun Lee. 2008. "MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings." *Trends in Genetics* 24 (11): 529–38. doi:10.1016/j.tig.2008.08.007.

Miller, Danny E., Clarissa B. Smith, Nazanin Yeganeh Kazemi, Alexandria J. Cockrell, Alexandra V. Arvanitakis, Justin P. Blumenstiel, Sue L. Jaspersen, and R. Scott Hawley. 2016. "Whole-genome analysis of individual meiotic events in Drosophila melanogaster reveals that noncrossover gene conversions are insensitive to interference and the centromere effect." *Genetics* 203 (1): 159–71. doi:10.1534/genetics.115.186486.