End Joining Signatures

Charlie Soeder

3/18/2019

Contents

1 Introduction

(McVey and Lee 2008) (Miller et al. 2016)

2 Materials, Methods, Data, Software

High-thoughoutput sequences in FASTQ format were aligned to the reference genome using BWA(Li and Durbin 2009) and processed with SAMtools (Li et al. 2009) and BEDtools (Quinlan and Hall 2010). Variants were called from these alignments using Freebayes (Garrison and Marth 2012) and processed using VCFTools(Danecek et al. 2011).

2.1 Reference Genomes

Table 1: Size and Consolidation of Reference Genomes

Reference Genome:	dm6
number_bases	144 M
number_contigs	1.87 k

2.2 Sequenced Reads

Control sequences were provided by Danny Miller; these consist of the offspring of two individual flies' mating.

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
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

2.2.1 Pre-Processing

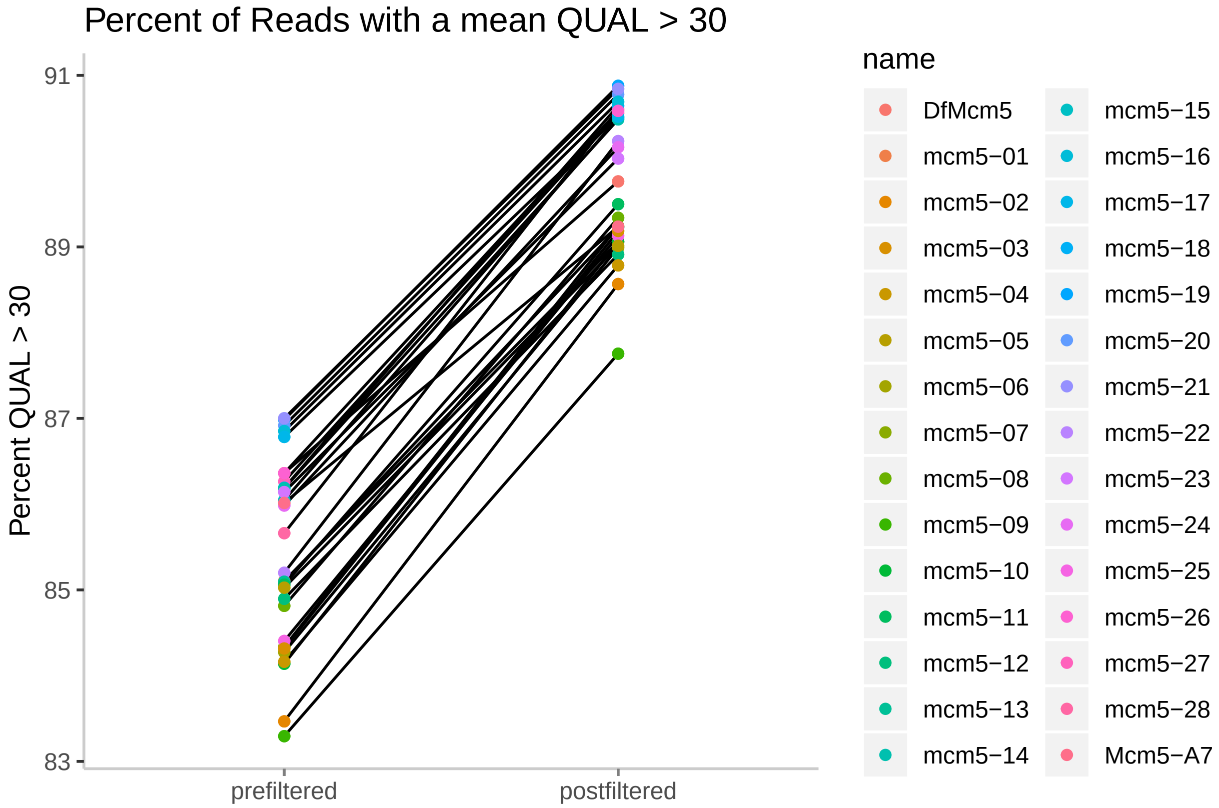
These reads were preprocessed with FASTP (S. Chen et al. 2018) for quality control and analytics.

Starting FASTQ files contained a total of $3.88G$ reads; after QC, this dropped to $3.64G$.

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

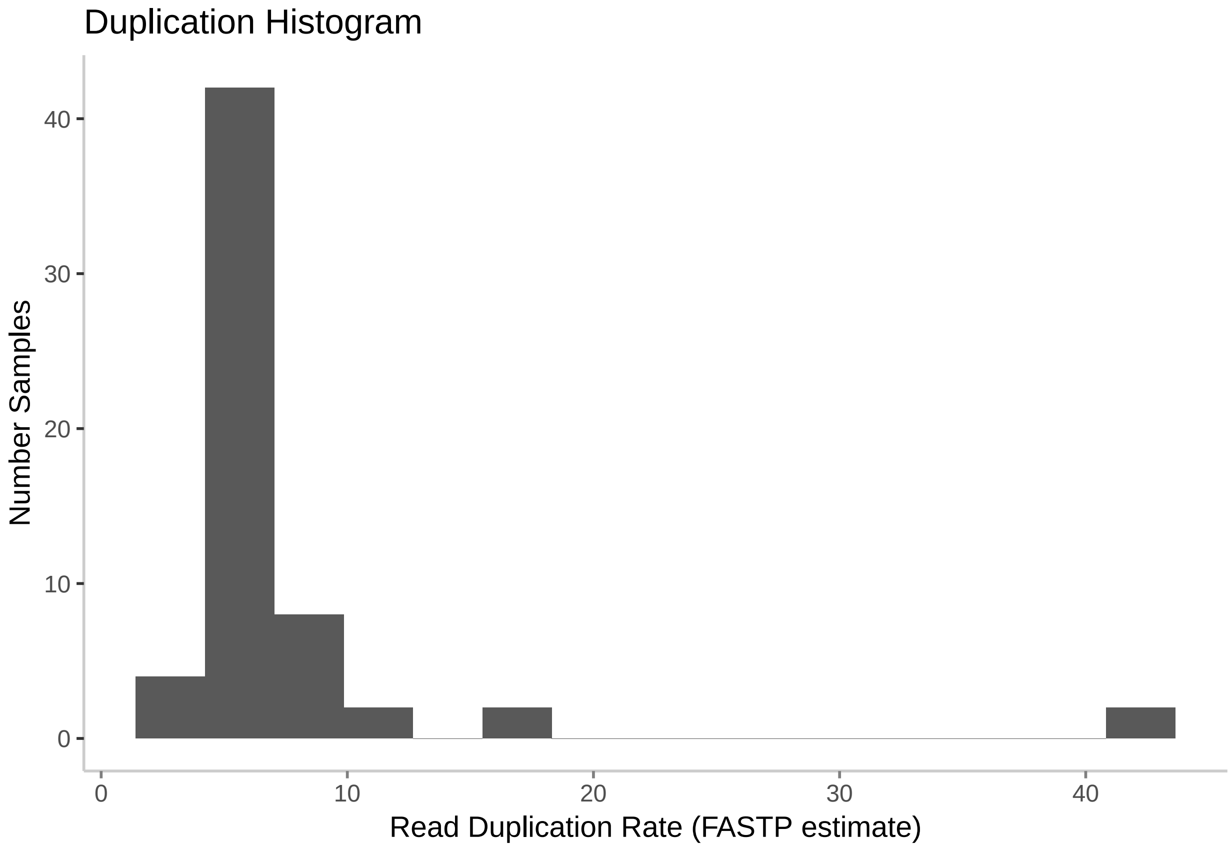
Filtration also increased the read quality, as seen in the increase in the fraction of reads with an average quality score > 30:



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

Table 5: Percentage Duplication

minimum	average	median	maximum
4.1	7.6	5.8	43.5



2.3 Mapped Reads

Reads were first mapped to the 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

The fraction of reads retained at each filtration step:

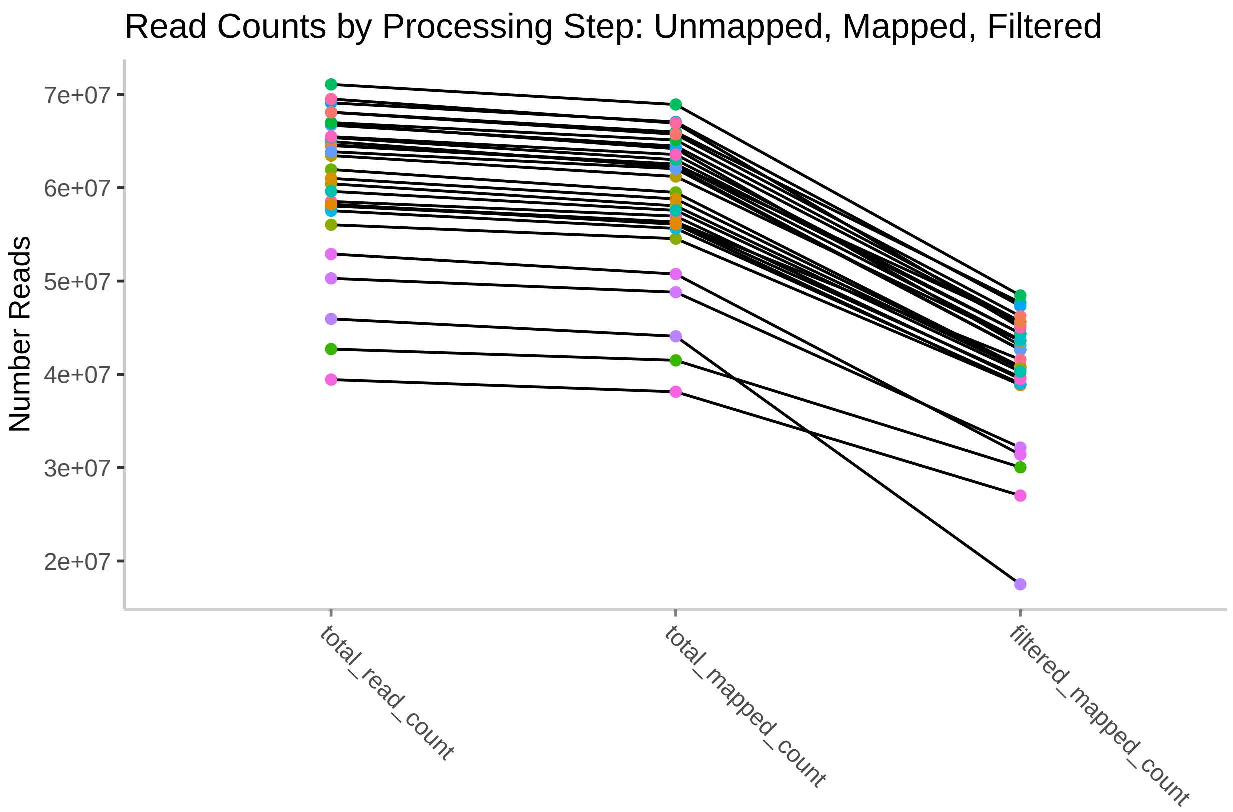


Table 6: Read Counts During Alignment & Filtration

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

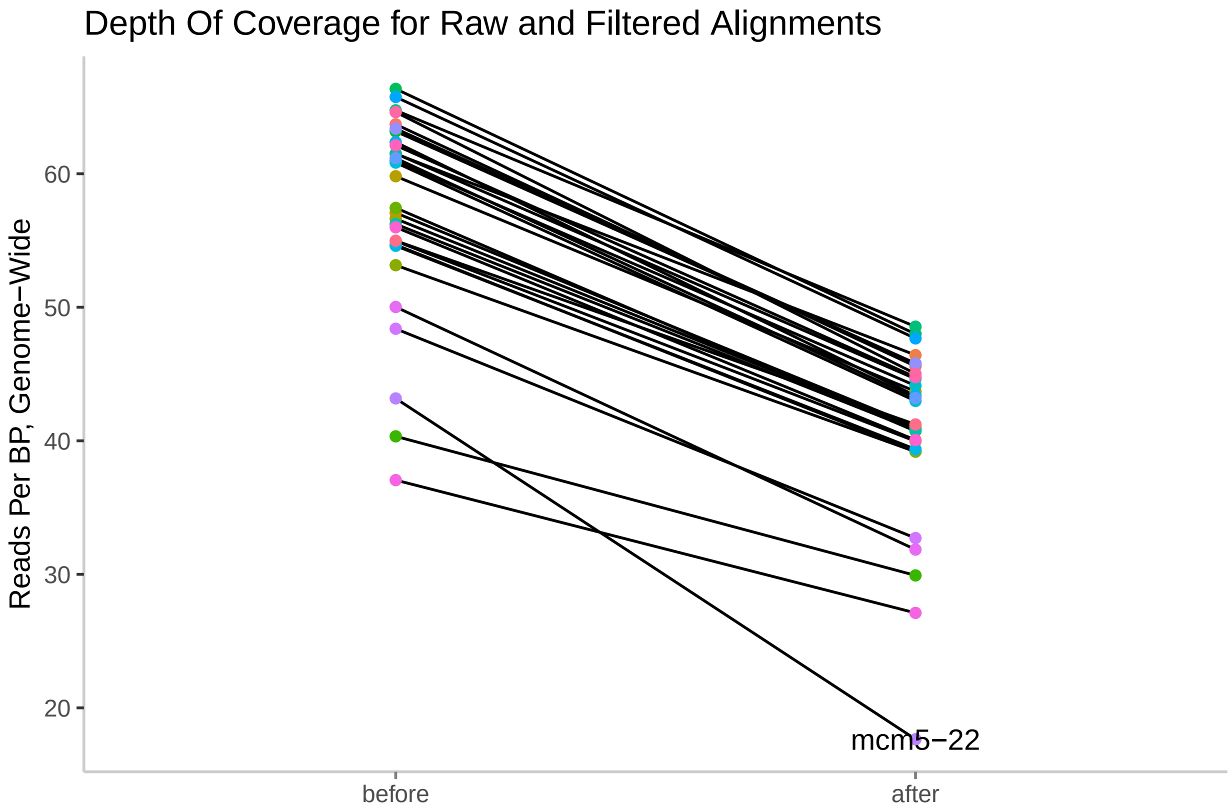
2.3.2 Depth & Breadth of Coverage

Depth of coverage, ie, the genome-wide average number of mapped reads per base pair:

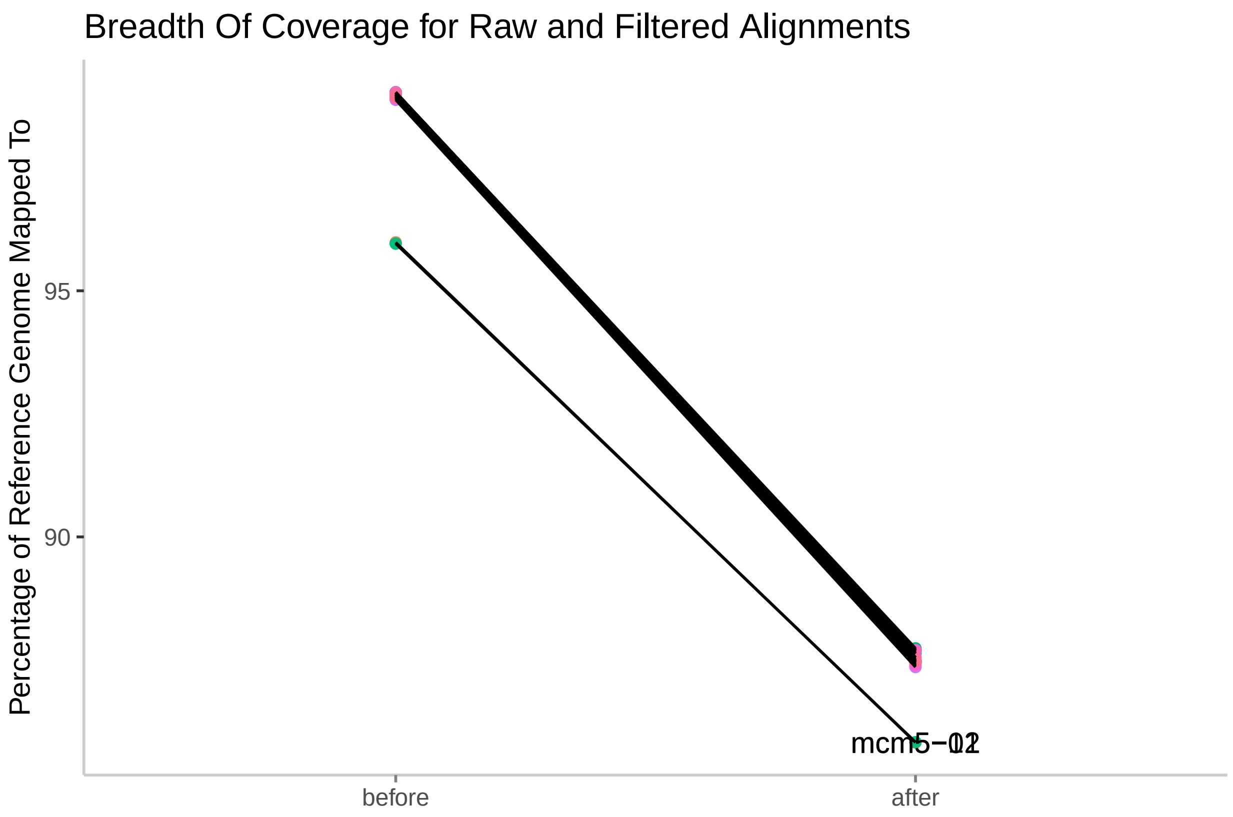
Table 8: Depth of Coverage Statistics for Raw and Filtered Alignments

step	minimum	average	median	maximum
pre-filtration depth	37.1	57.2	58.6	66.4

step	minimum	average	median	maximum
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, ie, the percentage of the genome covered by at least one read:



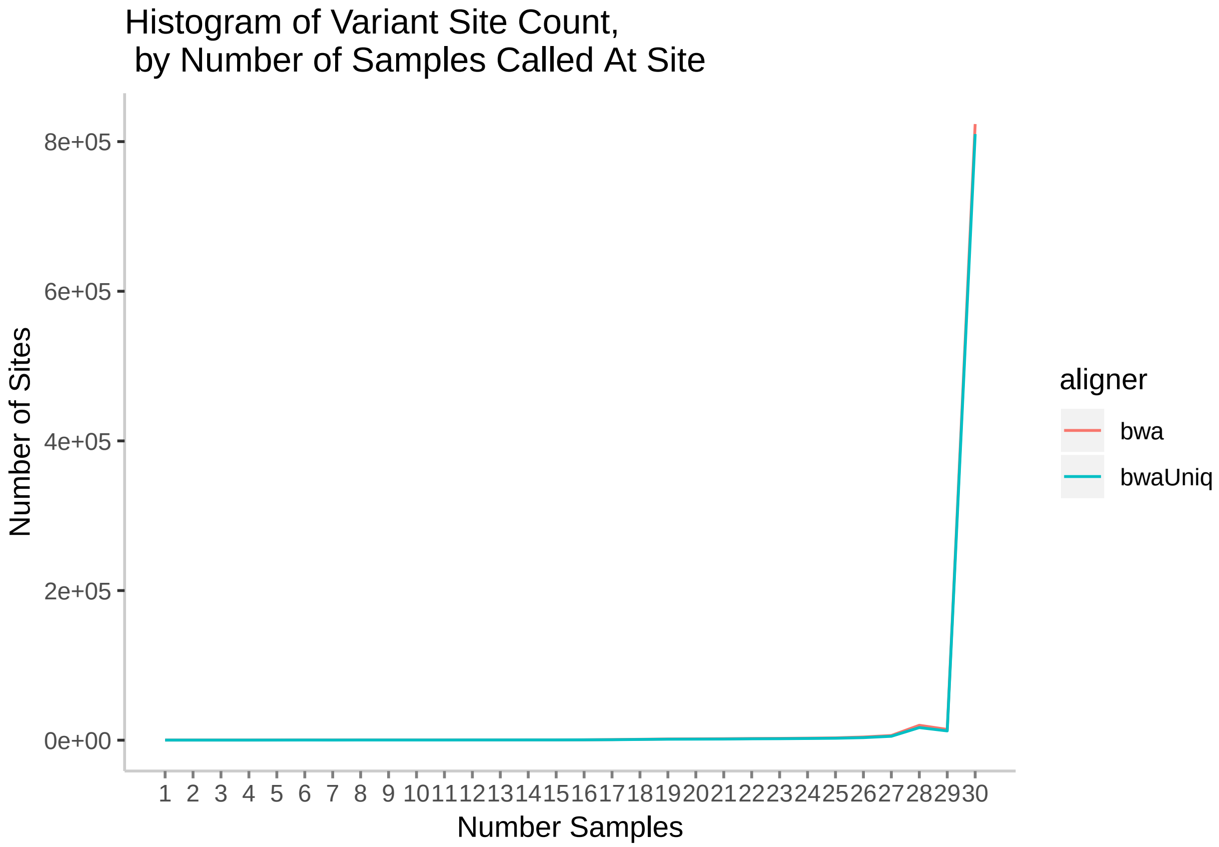
2.4 Called Variants

The BWA and BWA-Uniq alignments were independently used to call variants:

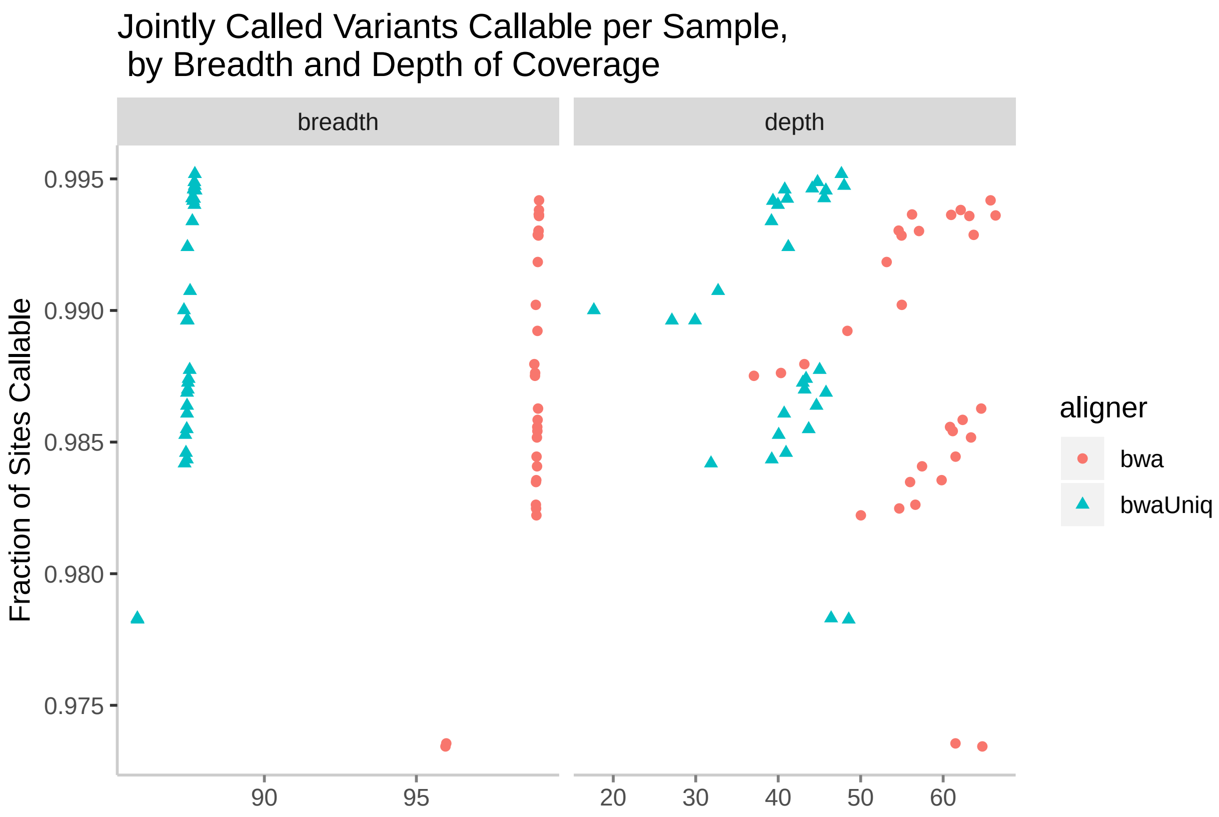
mappings were used to jointly call variants in VCF format via Freebayes (Garrison and Marth 2012) using standard filters.

To build this VCF, 30 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).
```



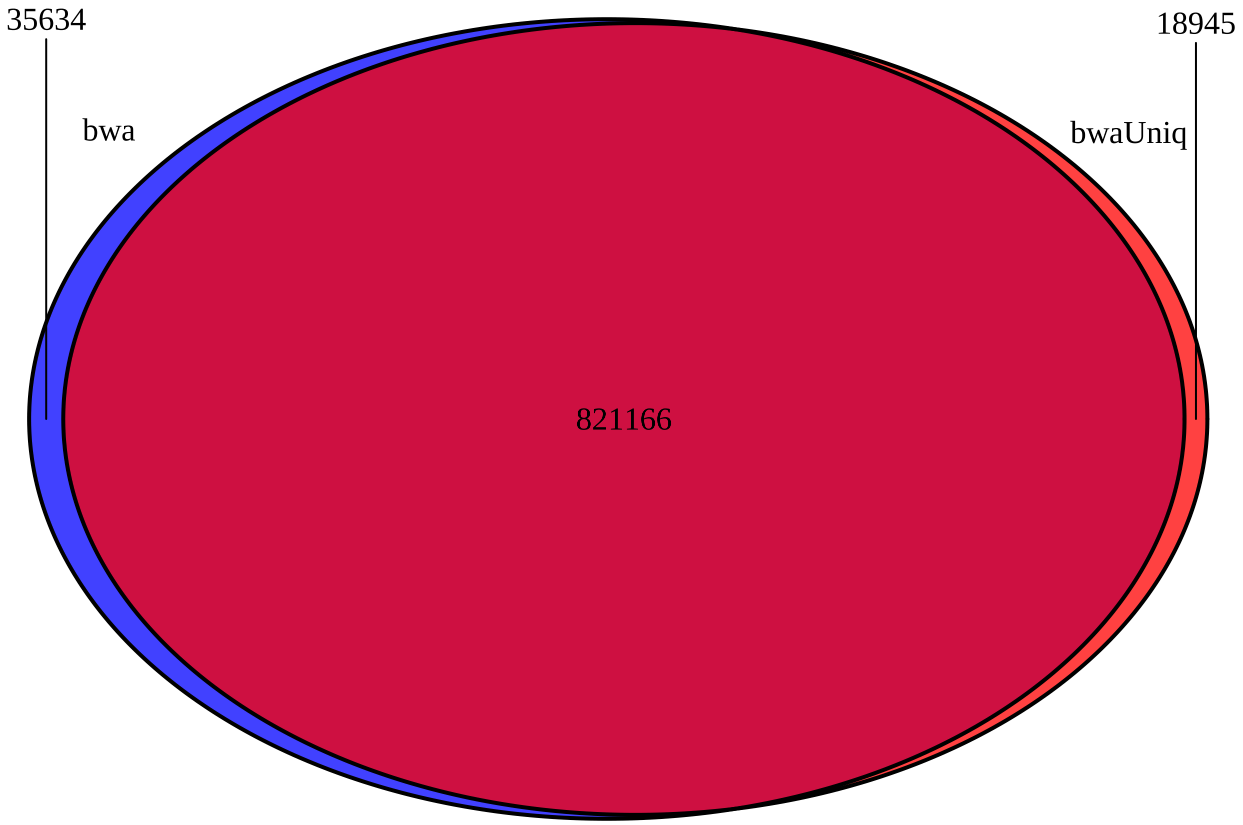
The fraction of jointly called SNPs which are individually callable:



2.4.1 VCF comparison & contrast

The two variant-calling methods are compared and contrasted over the main-line chromosomes (chr2L, chr2R, chr3L, chr3R, chr4, chrM, chrX, chrY).

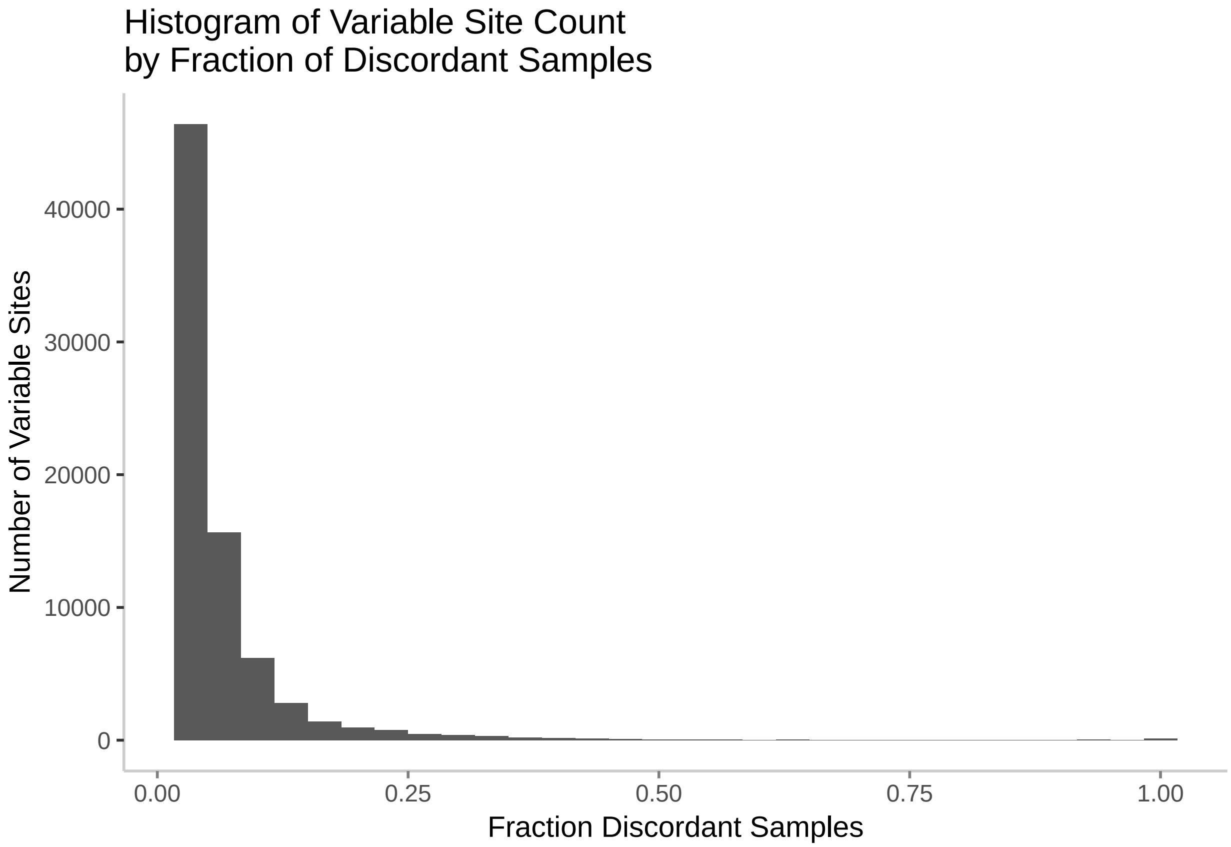
Of the $876k$ total variant sites called between the two VCFs, most ($821k$) were called in both files; $54.6k$ were unique to one of them:



```
## (polygon[GRID.polygon.610], polygon[GRID.polygon.611], polygon[GRID.polygon.612], polygon[GRID.polyg
```

Of the $821k$ variable sites shared by the two VCF files, $76.6k$ contained at least once sample which was called differently between the two VCF files.

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



Combining disjoint and discordant variable sites, there were a total $131k$ out of $876k$ disagreements between the two calling methods.

2.5 Variant Analysis

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

To search for traces of end-joining, the BWA-Uniq derived variants were further filtered, requiring a depth of 10 reads at the site for all flies sequenced: the per-site probability of sampling the same chromosome 10 times is $< 0.1\%$ given a fair draw, but this threshold is lower than the minimum average depth of coverage among the samples. Although it is possible that the join could occur near an already polymorphic site, or might manifest as a complex variant rather than a simple insertion/deletion, for the time being only proper biallelic indels were retained. Finally, only the consolidated autosomes were considered, and only sites which could be called in all 30 sequenced flies.

This gave a total of 41143 sites:

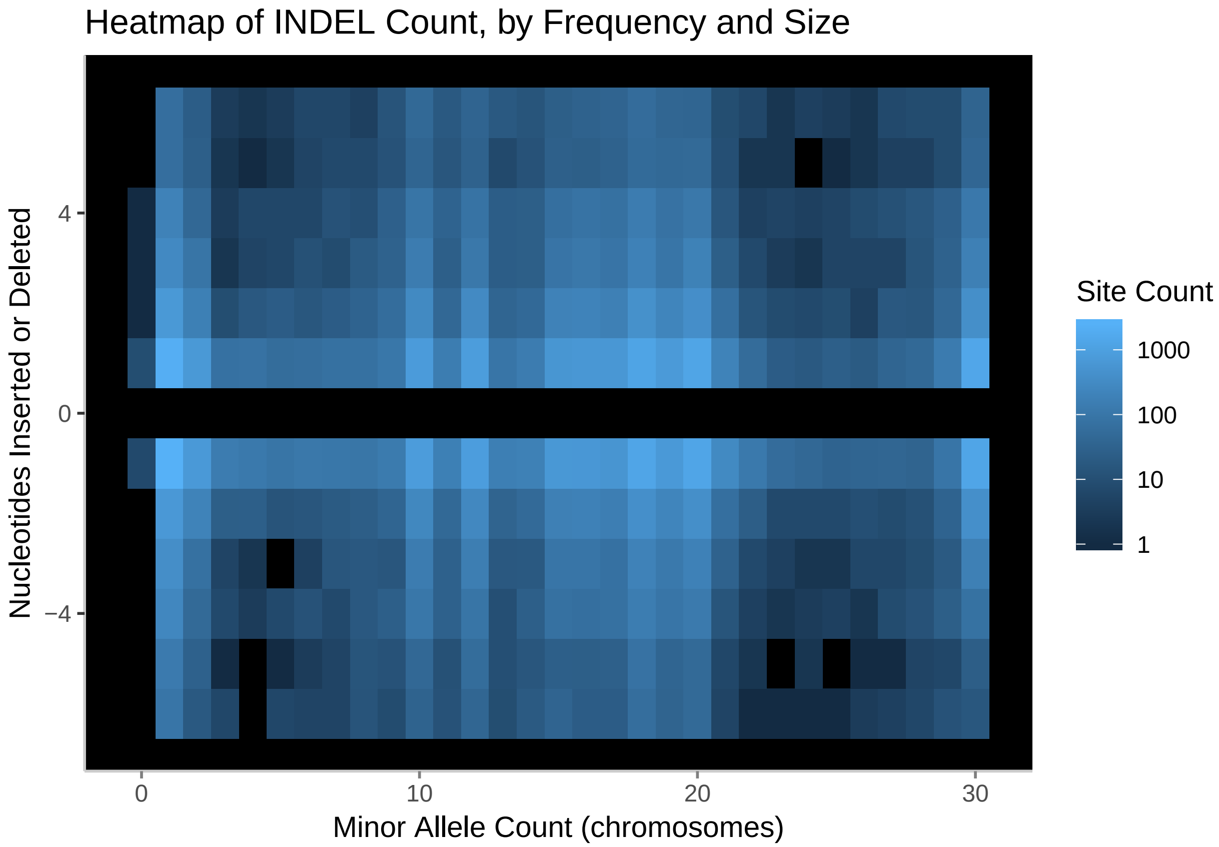
Table 9: Quality Biallaeleic INDELs per Chromosome

chrom	count
chr2L	11545
chr2R	10639
chr3L	9479
chr3R	9183
chr4	297

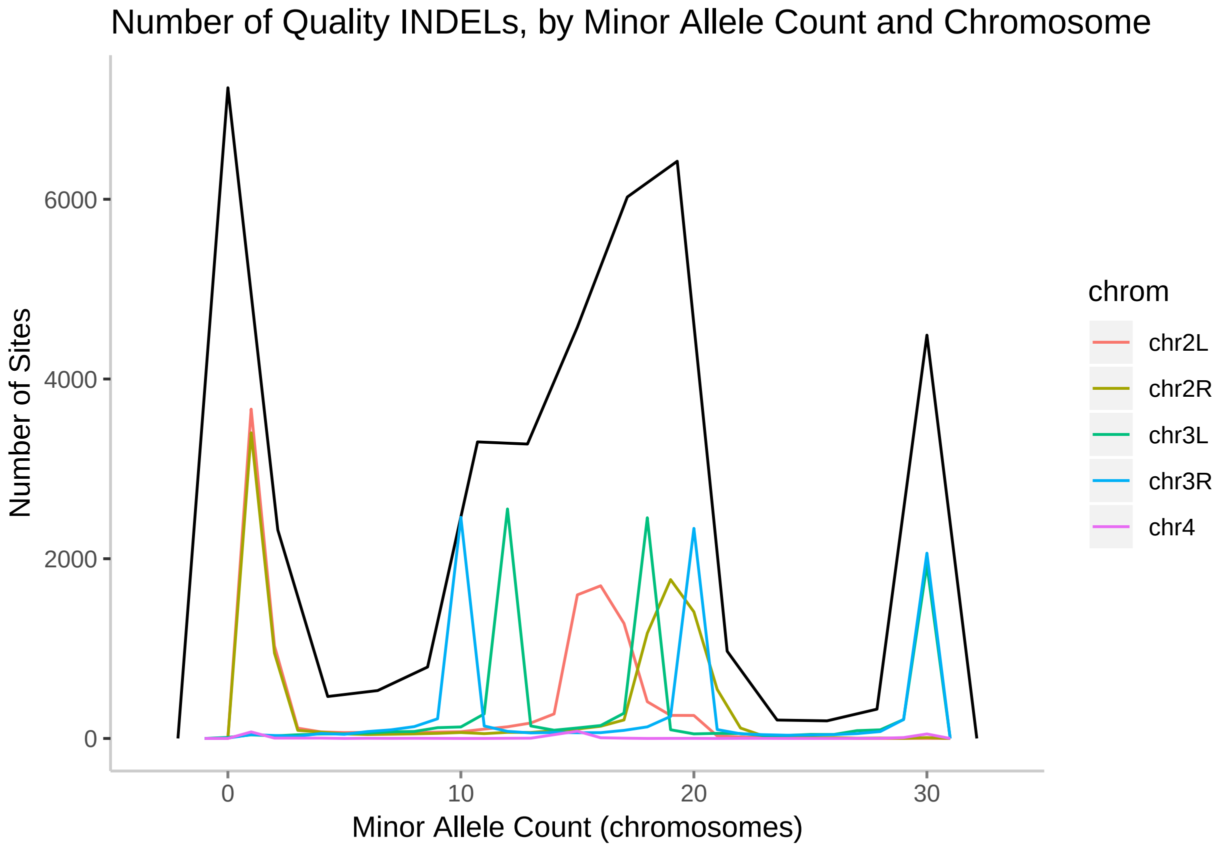
For these sites, minor and major alleles were assigned on the basis of which of the two variants had a smaller or larger allele count, respectively. INDEL type & size were determined relative to the major allele.

3 Results

3.1 INDEL Size & Frequency

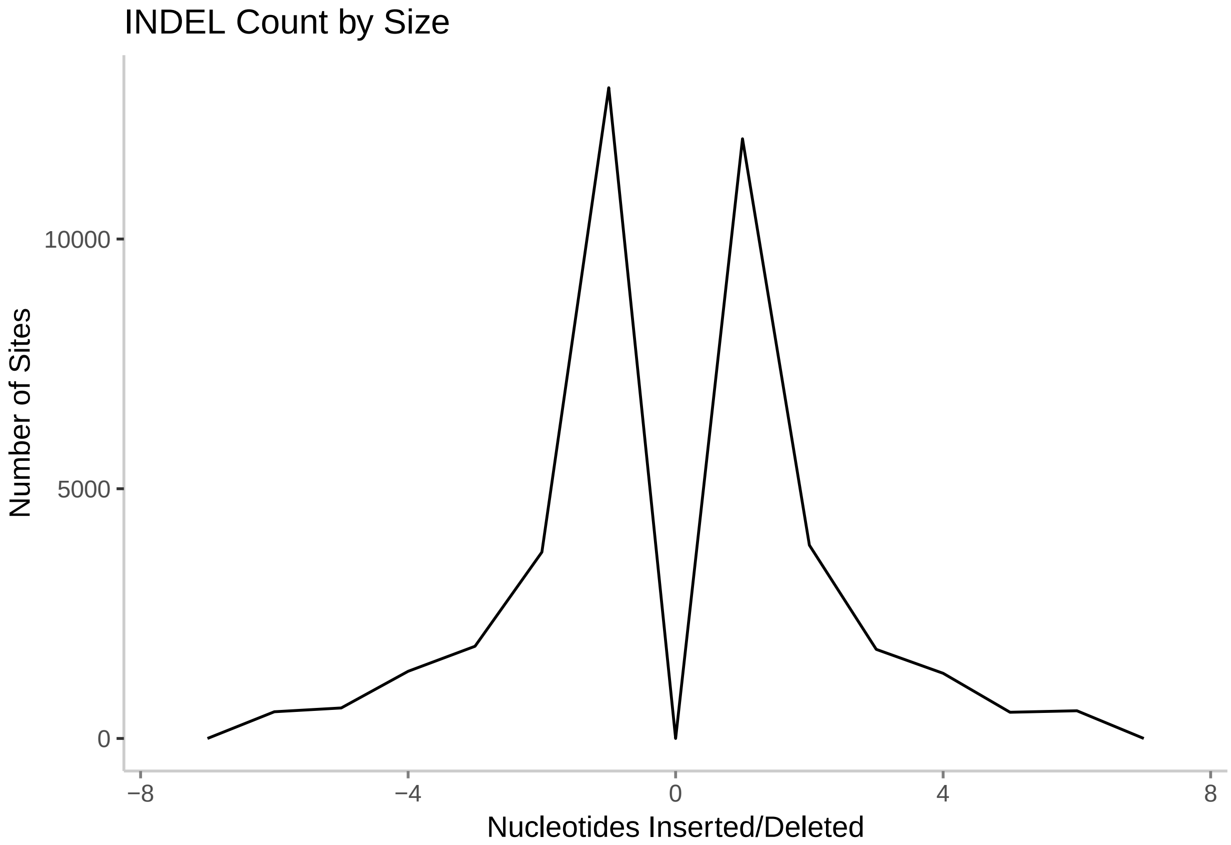


Since DSB is likely to create unique variants in the population, allele frequency is an important property to filter on. To get an idea of the expected allele frequency distribution, consider that for a given variable site, a diploid parent may be homozygous reference (a/a), homozygous alternate (A/A), or heterozygous (a/A). Thus, there are five relevant pairings of genotype: AA x AA, aa x aA, aa x AA, aA x AA, and aA x aA. (aa x aa corresponds to the case in which both parents have the reference allele, and the expectation would be the offspring would as well.) Under random assortment, the aa x aA and aA x AA crosses give a minor allele frequency of 1/4 and the aa x AA and aA x aA crosses give a minor allele frequency of 1/2. Finally, the AA x AA cross corresponds to the case where the parents are fixed for an alternate allele, and thus the offpsring would be expected to be fixed as well; the minor allele frequency would thus be zero. So, given sample of 30 diploid flies, the expected distribution of minor allele counts would be peaks centered on 0, 15, and 30.



At first glance, the distribution is as expected (black line above), but closer inspection reveals some anomalies. First, the peak near zero is not driven by fixed differences with the reference; there are in fact suprisingly few such fixed INDELs (19 out of 41143). Instead, it is almost entirely sites with a variant on one (7222) or two (2046) chromosomes across all samples. (All but 2 were alternate alleles, ie, disagreements with the reference genome.)

Under random assortment, there would be no reason to expect autosomes to have different allele-frequency distributions. However, this isn't the case when the variant sites are grouped by chromsome: the arms of chromosome 3 are enriched in sites with allele frequency ~50% and depleted of the very rare alleles. Conversely, the arms of chromsome 2 are depleted of the medium-frequency alleles and enriched for the rare ones. It also appears that the arms of chromosome 3 are split around the expected peak at MAC=15, into two smaller groups of peaks near MAC~10 and MAC~20. These would correspond to ratios of ~1/6 and 2/6.



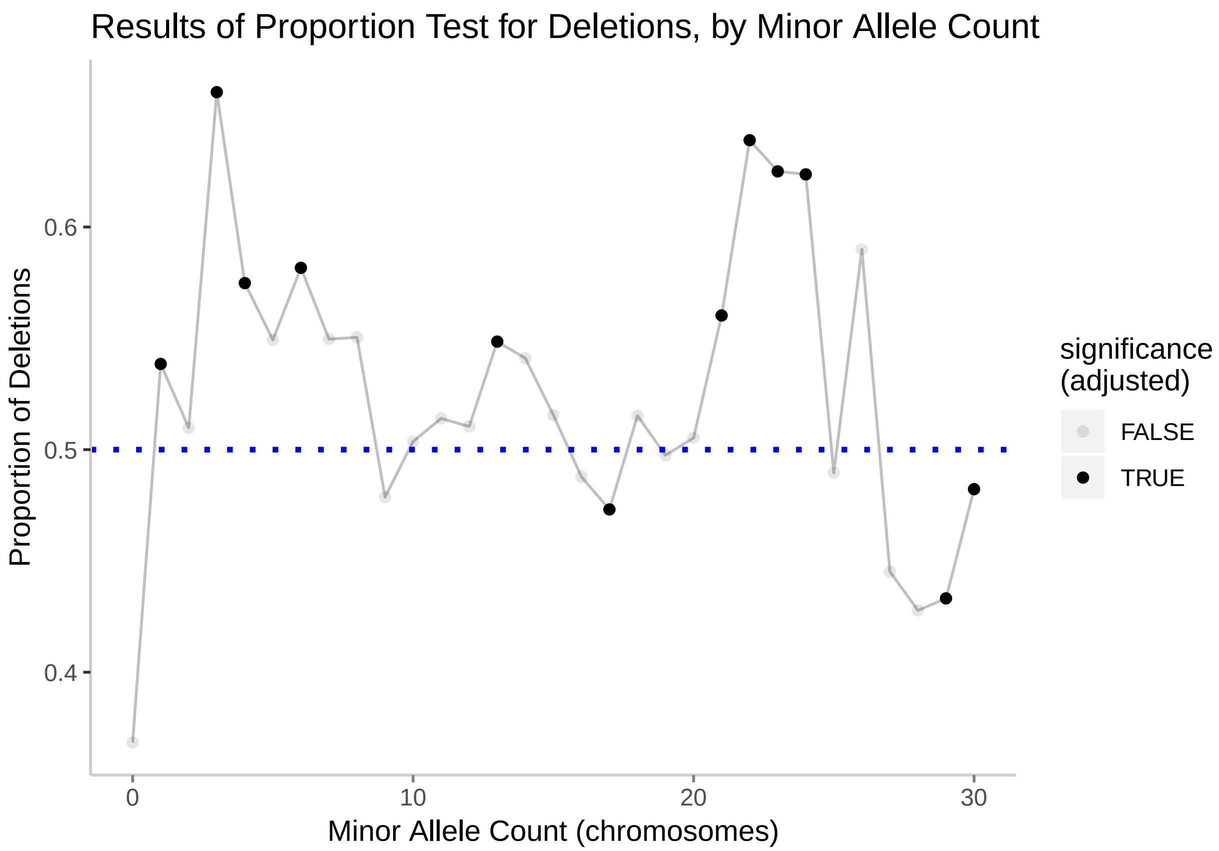
INDEL size ranged from 6bp deletions to 6bp insertions, with a small but significant bias towards deletion:

```
##
##  1-sample proportions test with continuity correction
##
## data:  insert_truth.Tbl$del out of insert_truth.Tbl$del + insert_truth.Tbl$ins, null probability 0.5
## X-squared = 26.797, df = 1, p-value = 2.26e-07
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5079296 0.5176130
## sample estimates:
##         p
## 0.5127725
```

(A similar result holds if the count by type is scaled by the change in length, for a ratio of bp removed to total bp change).

When constrained to a single MAC value, the basic form of the distribution stays the same but the degree of mutational bias varies, with the highest-MAC variants even having a significant excess of insertions.

```
## Warning: Using alpha for a discrete variable is not advised.
```



3.1.1 Singletons

Of particular interest in the context of DSB repair are INDELs which appear in exactly one sample as a heterozygote, ie, with an allele count of 1. These make up 7222 of 41143 total sites (17.6) %. As previously discussed, they disproportionately appear on the arms of chromosome 2:

Table 10: Distribution of Singleton INDELs Across Autosomes

chrom	count
chr2L	3665
chr2R	3401
chr3L	40
chr3R	43
chr4	73

3.2 Already-called variants

Danny Miller has already investigated these flies for likely candidates, identifying the following:

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

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).

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.

4 Next Steps

- Explore PINDEL (Ye et al. 2009) for calling larger variants
- Incorporate Talia's data

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.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. "The variant call format and VCFtools." *Bioinformatics* 27 (15): 2156–8. doi:10.1093/bioinformatics/btr330.

Garrison, Erik, and Gabor Marth. 2012. "Haplotype-based variant detection from short-read sequencing," July. http://arxiv.org/abs/1207.3907.

Li, Heng, and Richard Durbin. 2009. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics (Oxford, England)* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25 (16): 2078–9. doi:10.1093/bioinformatics/btp352.

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.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A flexible suite of utilities for comparing genomic features." *Bioinformatics* 26 (6): 841–42. doi:10.1093/bioinformatics/btq033.

Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. 2009. "Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads." *Bioinformatics* 25 (21): 2865–71. doi:10.1093/bioinformatics/btp394.