End Joining Signatures

Charlie Soeder

4/4/2019

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

The dm6 reference genome was used for read alignment:

Table 1: Size and Consolidation of Reference Genomes

Reference Genome:	dm6
number_bases	144 M
number_contigs	1.87 k

2.2 Sequenced Reads

Two treatments have been sequenced: a control line, and an mcm5 mutant. The data available include two parent flies per treatment as well as 28 of their male offspring:

Table 2: Number of Sequenced Samples by Treatment

experimental	pedigree	sample_count
mutant	parent	2
mutant	child	28
control	parent	2
control	child	28

The samples from the mcm5 mutant cross were provided by Talia:

Table 3: Sequenced Samples from mcm5 Mutant Cross

name	source	pedigree	sex
Mcm5-A7	talia	parent	NA
DfMcm5	talia	parent	NA
mcm5-28	talia	child	M
mcm5-27	talia	child	M
mcm5-26	talia	child	M
mcm5-25	talia	child	M
mcm5-24	talia	child	M
mcm5-23	talia	child	M
mcm5-22	talia	child	M
mcm5-21	talia	child	M
mcm5-20	talia	child	M
mcm5-19	talia	child	M
mcm5-18	talia	child	M
mcm5-17	talia	child	M
mcm5-16	talia	child	M
mcm5-15	talia	child	M
mcm5-14	talia	child	M
mcm5-13	talia	child	M
mcm5-12	talia	child	M
mcm5-11	talia	child	M
mcm5-10	talia	child	M
mcm5-09	talia	child	M
mcm5-08	talia	child	M
mcm5-07	talia	child	M
mcm5-06	talia	child	M
mcm5-05	talia	child	M
mcm5-04	talia	child	M

name	source	pedigree	sex
mcm5-03	talia	child	M
mcm5-02	talia	child	M
mcm5-01	talia	child	M

The control samples were first published in (Miller et al. 2016); this cross consists of a male w1118 and a female Canton S, as well as their male offspring. 28 of the 196 sequenced offspring were selected at random and downloaded from NCBI, as were the sequenced parents:

> Males were numbered based on whether their father was homozygous w1118 or Canton-S and the number of their het- erozygous mother. For example, male cs12.3 had a Canton-S father, its mother was female number 12, and it was the third male selected for DNA extraction.

```
## Warning: Expected 2 pieces. Missing pieces filled with ‘NA‘ in 2 rows [1,
## 2].
```

```
## Warning: Expected 2 pieces. Missing pieces filled with ‘NA‘ in 1 rows [1].
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 1 rows [2].
```

```
## Warning: Expected 2 pieces. Missing pieces filled with ‘NA‘ in 1 rows [1].
```

Table 4: Sequenced Samples from Control Cross

name	source	pedigree	sex	offspring_id	father_type	mother_id
w1118	danny	parent	M	NA	NA	NA
CantonS	danny	parent	F	NA	NA	NA
w4_4	danny	child	M	4	w1118	4
w4_3	danny	child	M	3	w1118	4
w4_2	danny	child	M	2	w1118	4
w4_17	danny	child	M	17	w1118	4
w4_16	danny	child	M	16	w1118	4
w4_15	danny	child	M	15	w1118	4
w4_13	danny	child	M	13	w1118	4
w4_12	danny	child	M	12	w1118	4
w4_11	danny	child	M	11	w1118	4
w4_1	danny	child	M	1	w1118	4
w3_9	danny	child	M	9	w1118	3
w3_8	danny	child	M	8	w1118	3
w3_6	danny	child	M	6	w1118	3
w3_5	danny	child	M	5	w1118	3
w3_4	danny	child	M	4	w1118	3
w3_26	danny	child	M	26	w1118	3
w3_25	danny	child	M	25	w1118	3
w3_24	danny	child	M	24	w1118	3
w3_21	danny	child	M	21	w1118	3
w3_18	danny	child	M	18	w1118	3
w3_17	danny	child	M	17	w1118	3
w3_16	danny	child	M	16	w1118	3
w3_15	danny	child	M	15	w1118	3
w3_14	danny	child	M	14	w1118	3
w3_13	danny	child	M	13	w1118	3
w3_12	danny	child	M	12	w1118	3
w3_11	danny	child	M	11	w1118	3
w3_1	danny	child	M	1	w1118	3

name	source	pedigree	sex	offspring_id	father_type	mother_id

Table 5: Number of Male Offsping Sequenced, by Cross Type and Female ID

father_type	mother_id	count
w1118	3	18
w1118	4	10

2.2.1 Pre-Processing

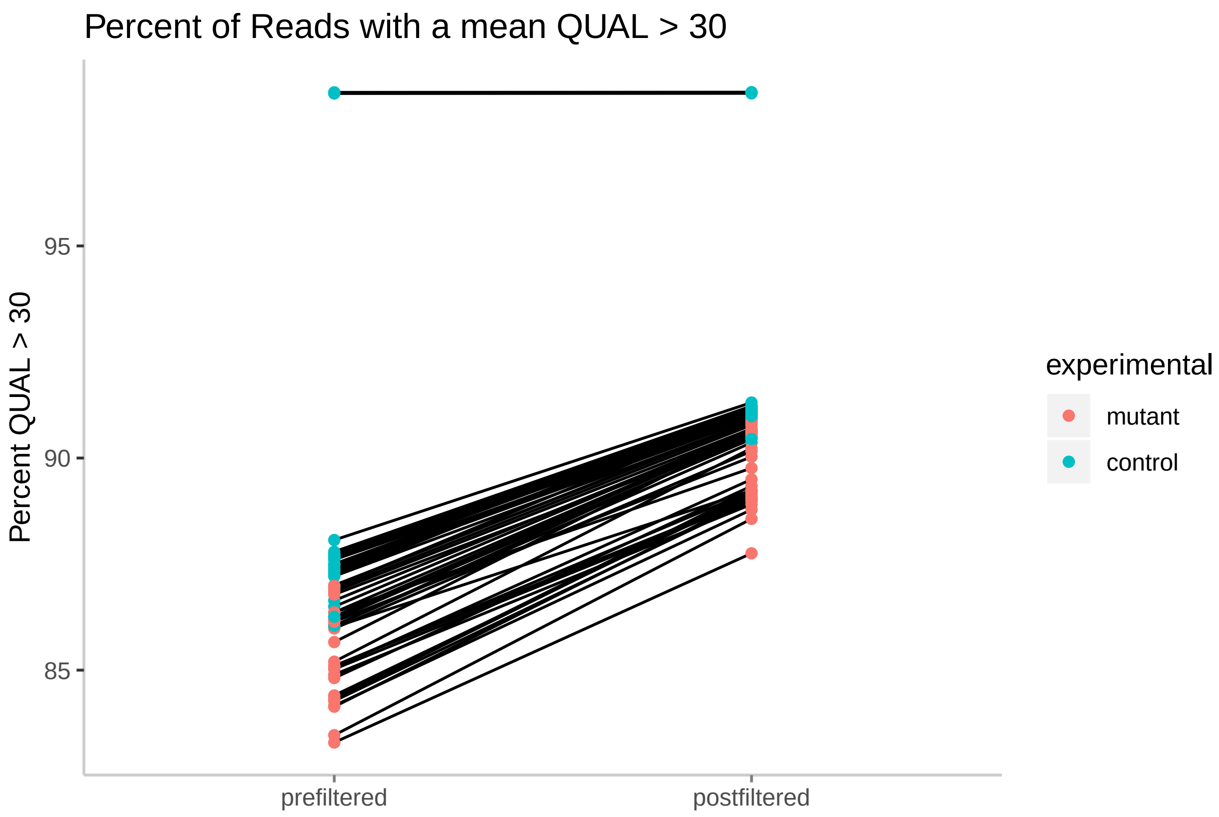
These reads were preprocessed with FASTP (S. Chen et al. 2018) for quality control and analytics.

Starting FASTQ files contained a total of $3.96G$ reads; after QC, this dropped to $3.7G$.

Table 6: Read Count and Percent Retention

type	minimum	average	maximum
prefiltered	42.3 M	66.1 M	93.6 M
postfiltered	39.4 M	61.7 M	87 M
percent retention	91.4	93.4	100

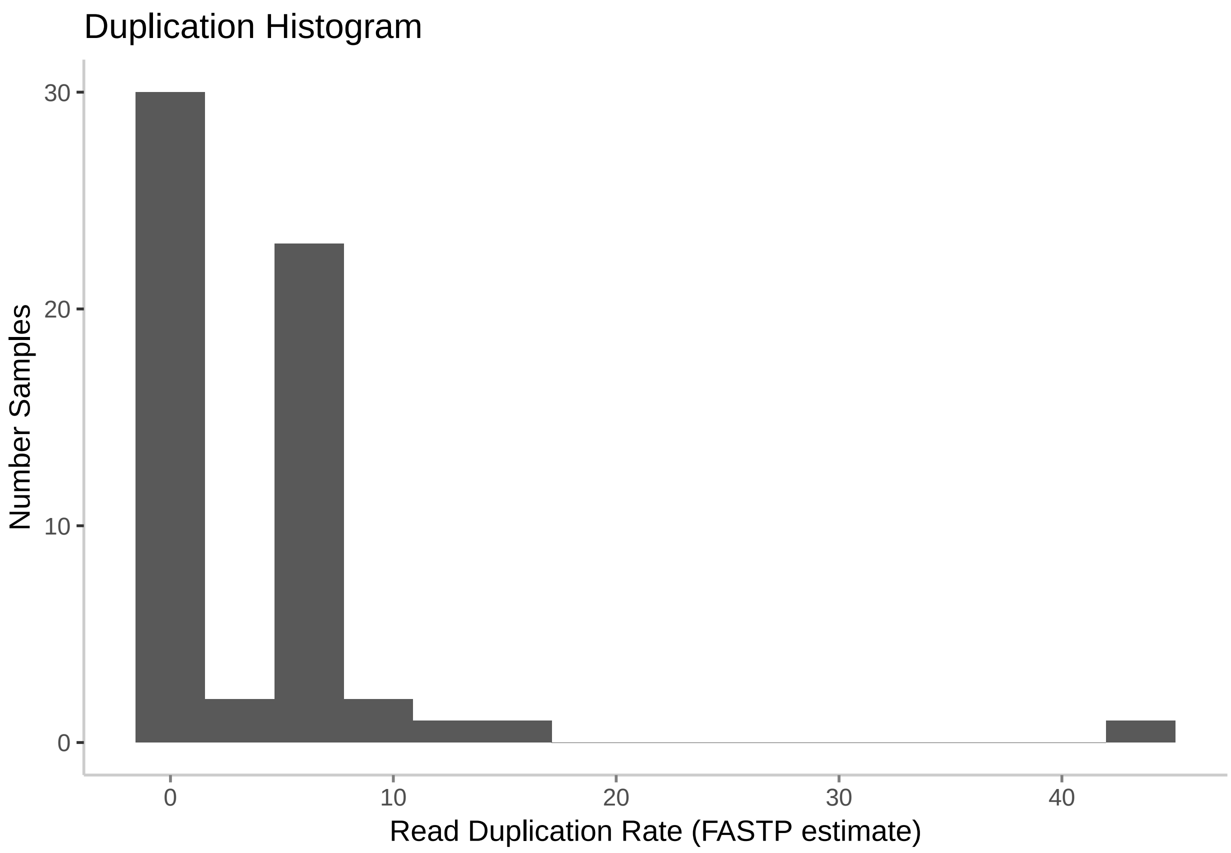
Filtration also increased the read quality, as seen in the increase in the fraction of reads with an average quality score > 30:



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

Table 7: Percentage Duplication

minimum	average	median	maximum
0	4	2.4	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:

```
## Warning: Column ‘sample‘/‘name‘ joining factors with different levels,
## coercing to character vector
```

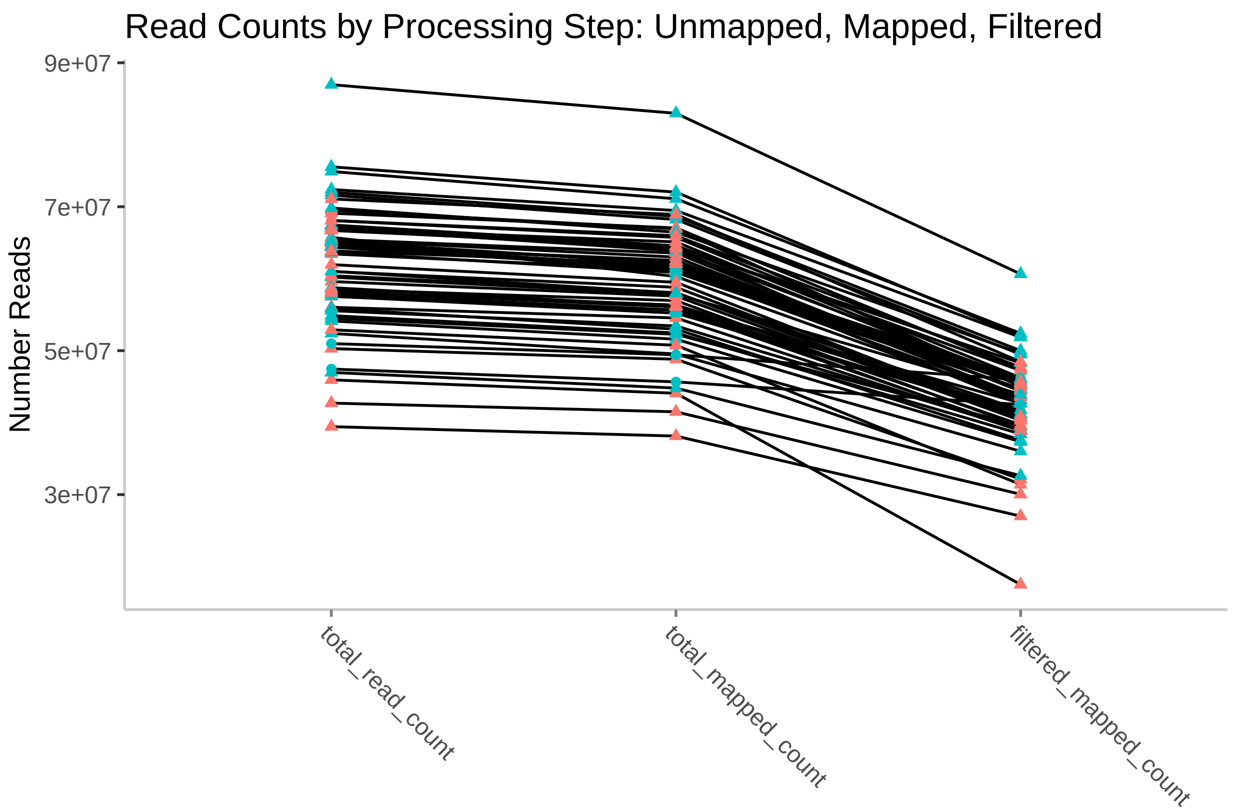


Table 8: Read Counts During Alignment & Filtration

measure	minimum	average	median	maximum
filtered_mapped_count	17.5 M	42.5 M	43.1 M	60.7 M
total_mapped_count	38.1 M	59.2 M	60.6 M	83 M
total_read_count	39.4 M	61.7 M	63.6 M	87 M

Table 9: Percentage of Reads Retained at Each Step

measure	minimum	average	median	maximum
filter_retention	39.7	71.7	72.1	93.1
mapping_retention	91.7	96	96	97.3

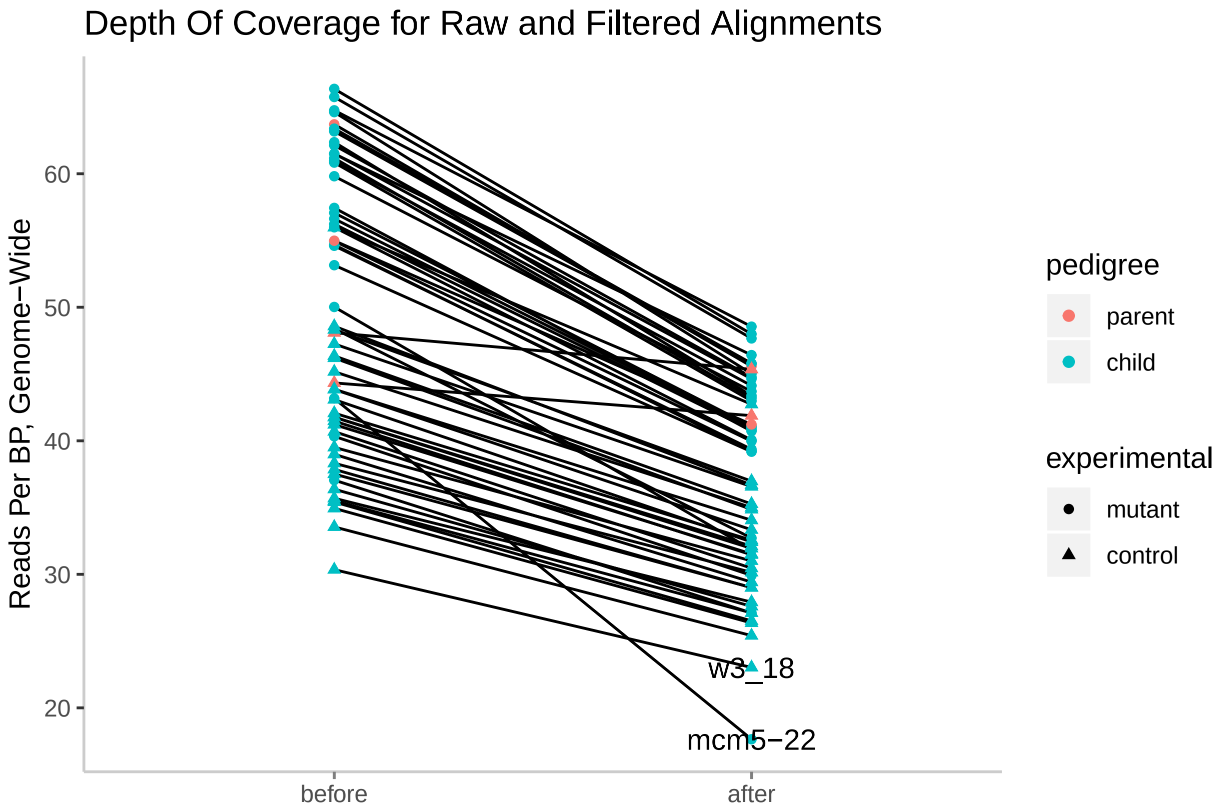
2.3.2 Depth & Breadth of Coverage

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

Table 10: Depth of Coverage Statistics for Raw and Filtered Alignments

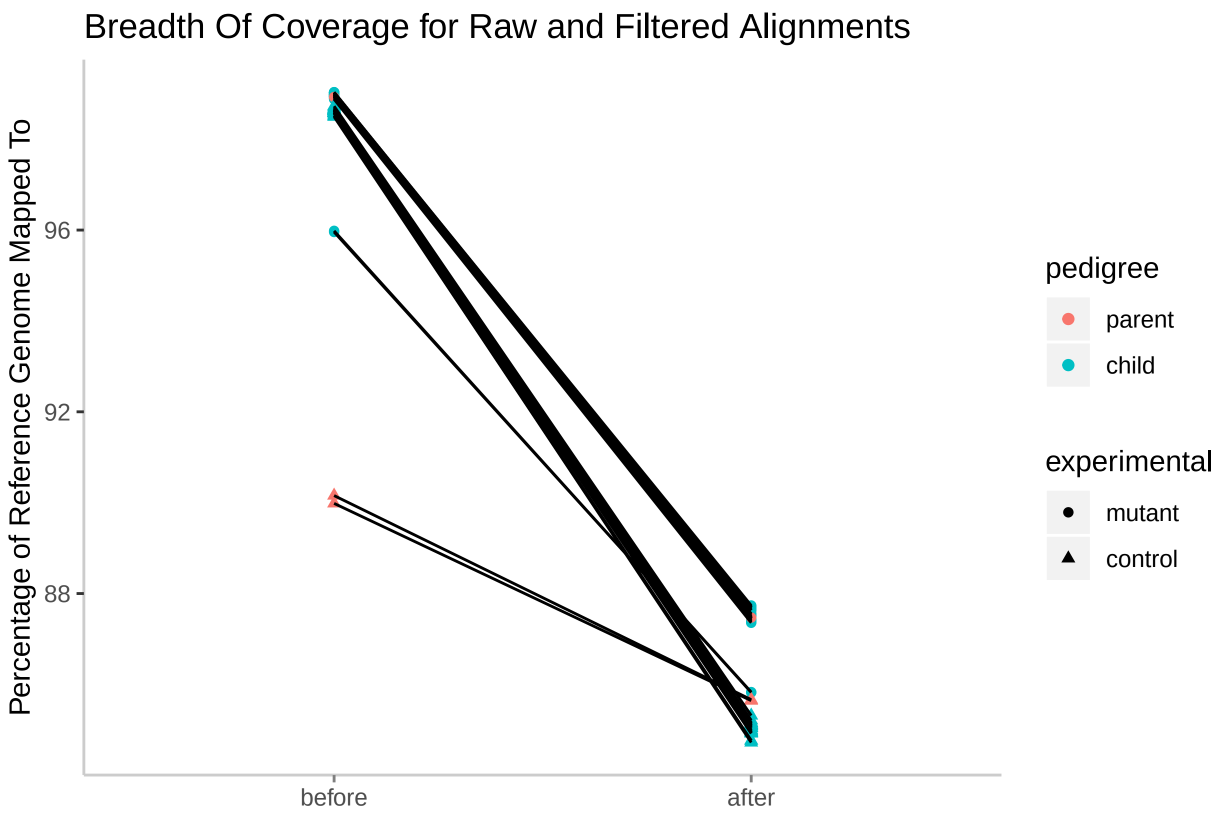
step	minimum	average	median	maximum
pre-filtration depth	30.3	49.3	48.2	66.4

step	minimum	average	median	maximum
post-filtration depth	17.6	36.5	36.6	48.5
depth retention percent	40.9	74.3	75.0	94.5



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

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



2.4 Called Variants

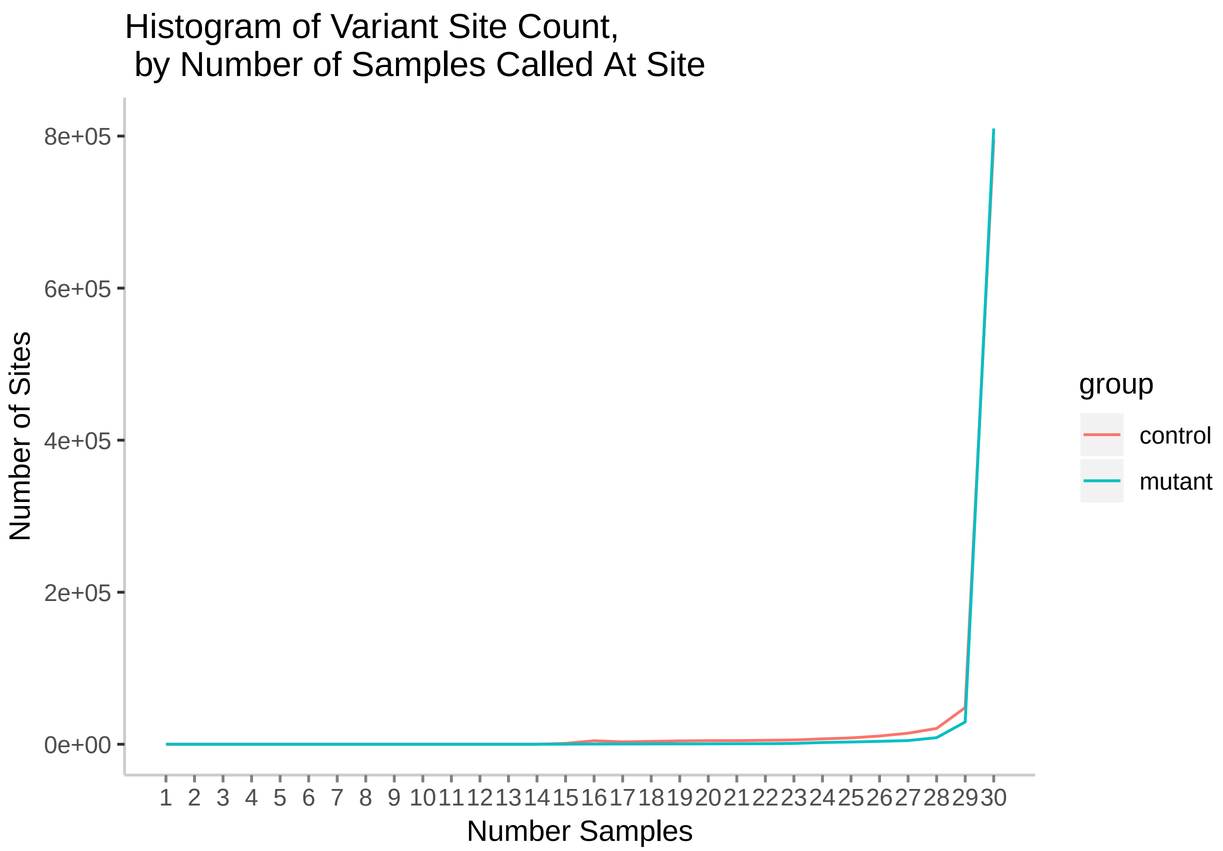
The BWA-Uniq alignments for the control and mutant flies were independently used to call variants in VCF format via Freebayes (Garrison and Marth 2012) using standard filters.

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   X2 = col_character(),
##   X3 = col_character(),
##   X4 = col_integer()
## )
```

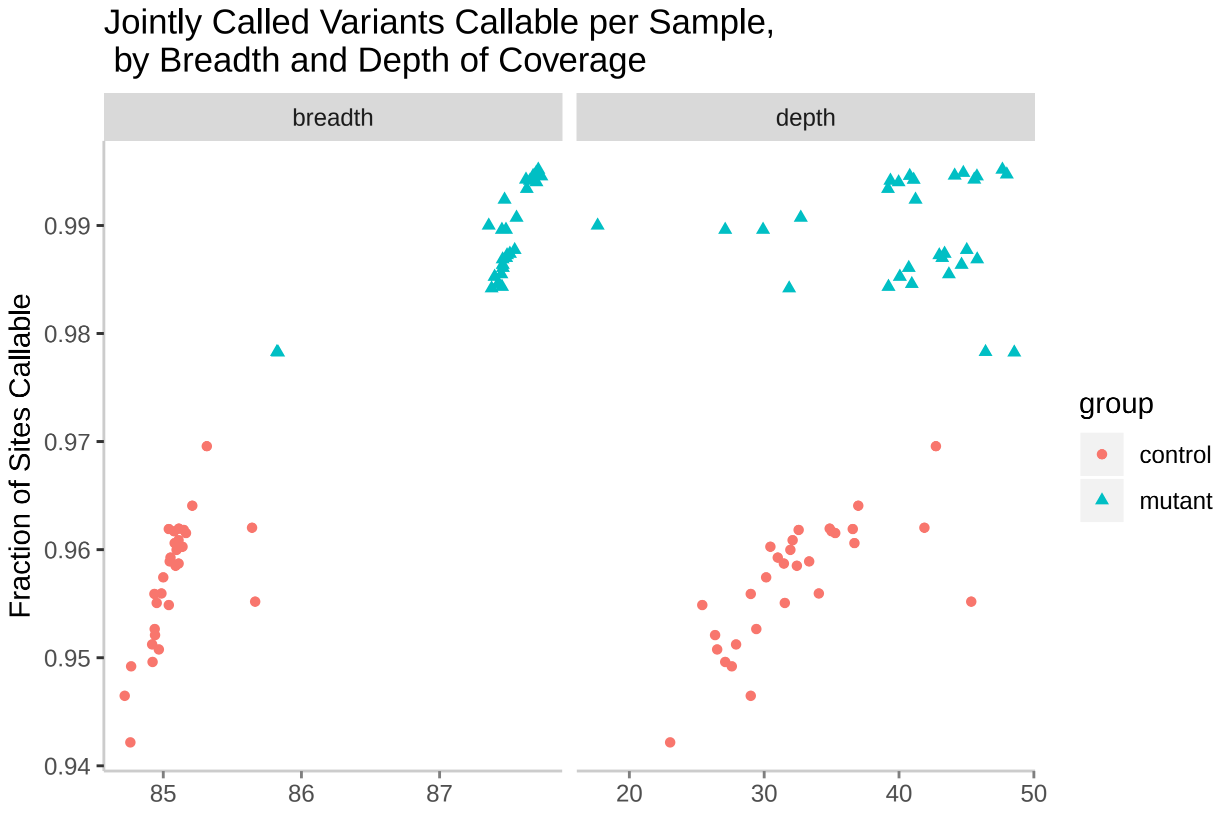
Table 11: Variant Counts and Frequency by Type and Group

aligner	INDELs	INDEL_per_kb	SNPs	SNP_per_kb
bwaUniq	715 k	1.6	715226	5.0
bwaUniq	700 k	1.2	699671	4.9

For each group VCF, 30 samples from the group 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:



The fraction of jointly called SNPs which are individually callable:



2.5 Variant Analysis

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.

Between the two treatments, this gave a total of 73700 sites:

Table 12: Quality Biallaeleic INDELs per Chromosome

chrom	control	mutant
chr2L	8134	10882
chr2R	7905	10085
chr3L	9200	9064
chr3R	9022	8891
chr4	228	289
total	34489	39211

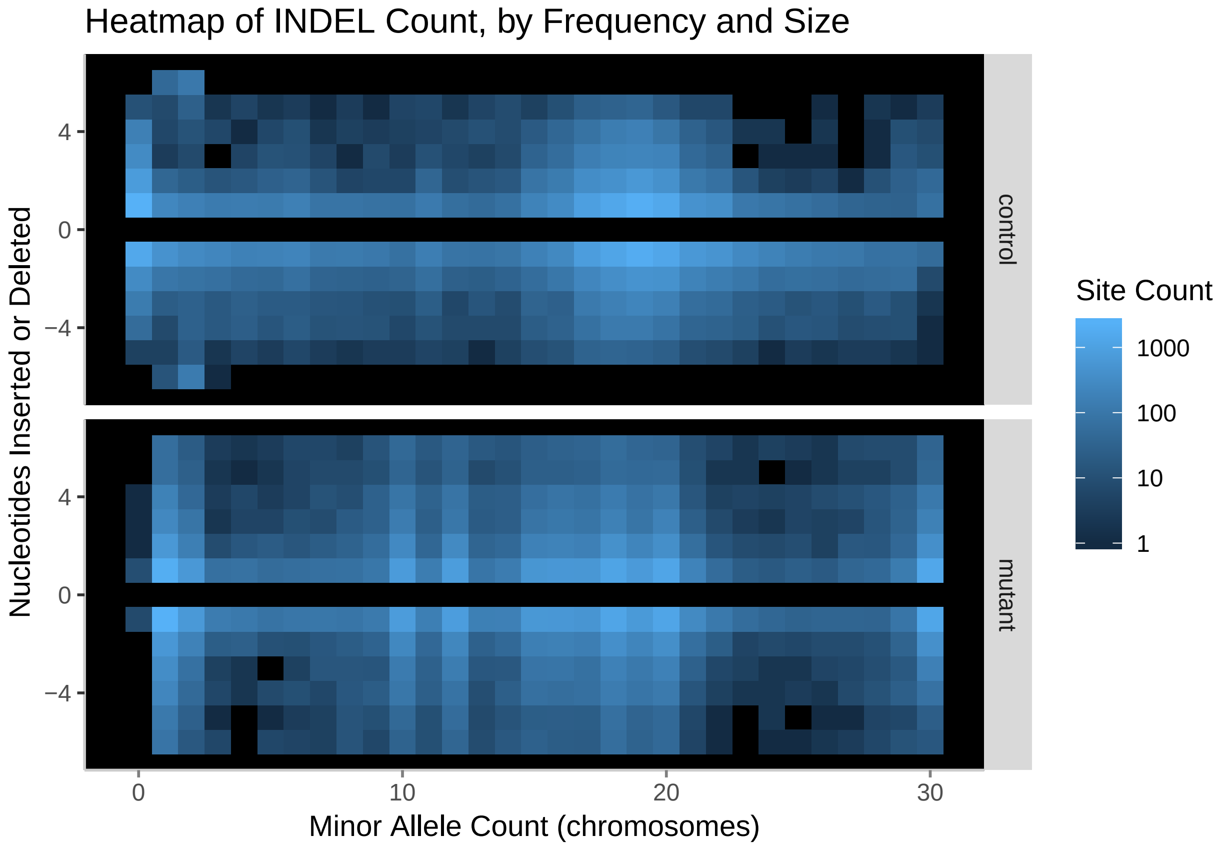
For these sites, minor and major alleles were assigned on the basis of which of the two variants had a smaller or larger allele count, respectively. INDEL type & size were determined relative to the major allele.

2.5.1 Novel Variant Identification

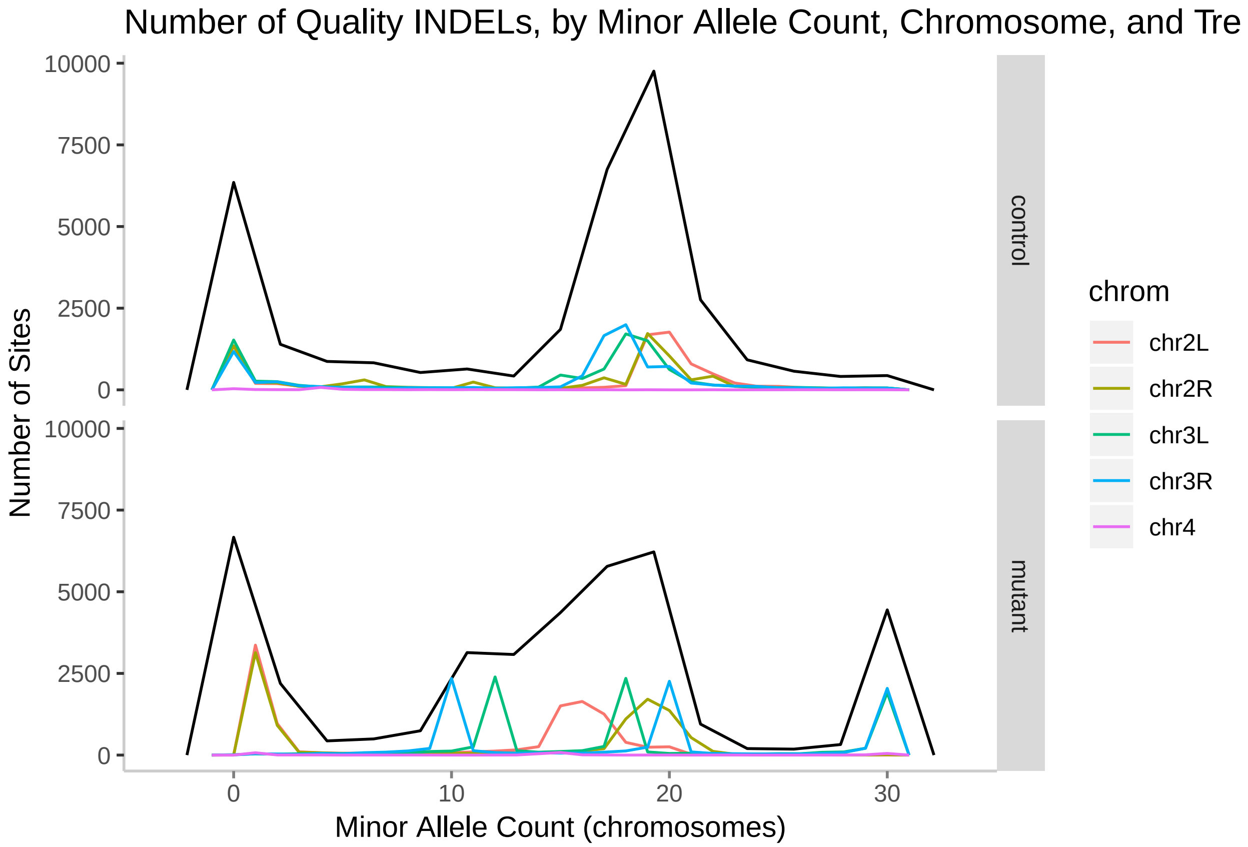
Double-strand breaks introduce new variants in the offspring which were not present in the parent flies. The VCFs were thus filtered to collect sites which have an allele in at least one offspring which was not detected in either parent. This can happen two ways: both parents may be homozygous for the reference, with an alternate allele sighted in the offspring, or both parents may be homozygous for an alternate allele, with the reference sighted in the offspring. These two cases have different connotations and were split in two with each case examined separately, with different results (see 3.1.2 below).

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.

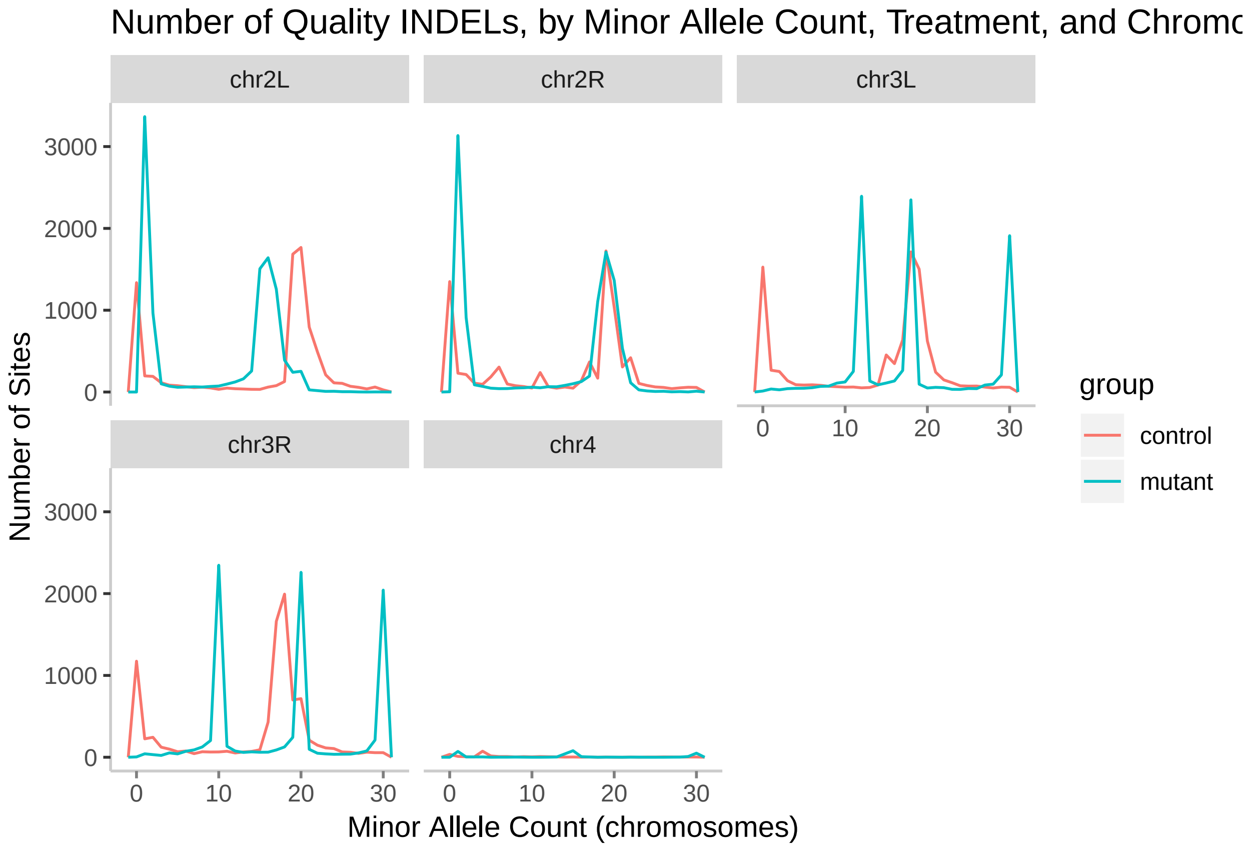


Both the control and mutant flies have peaks near 0, but on closer inspection these appear to be different phenomena. Whereas the control flies contain many sites with a minor allele count of zero (corresponding to fixed differences with the reference), the mutant flies had very few. The near-zero peak in the mutants consists of sites with one or two of the 60 chromosomes carrying the rarer allele:

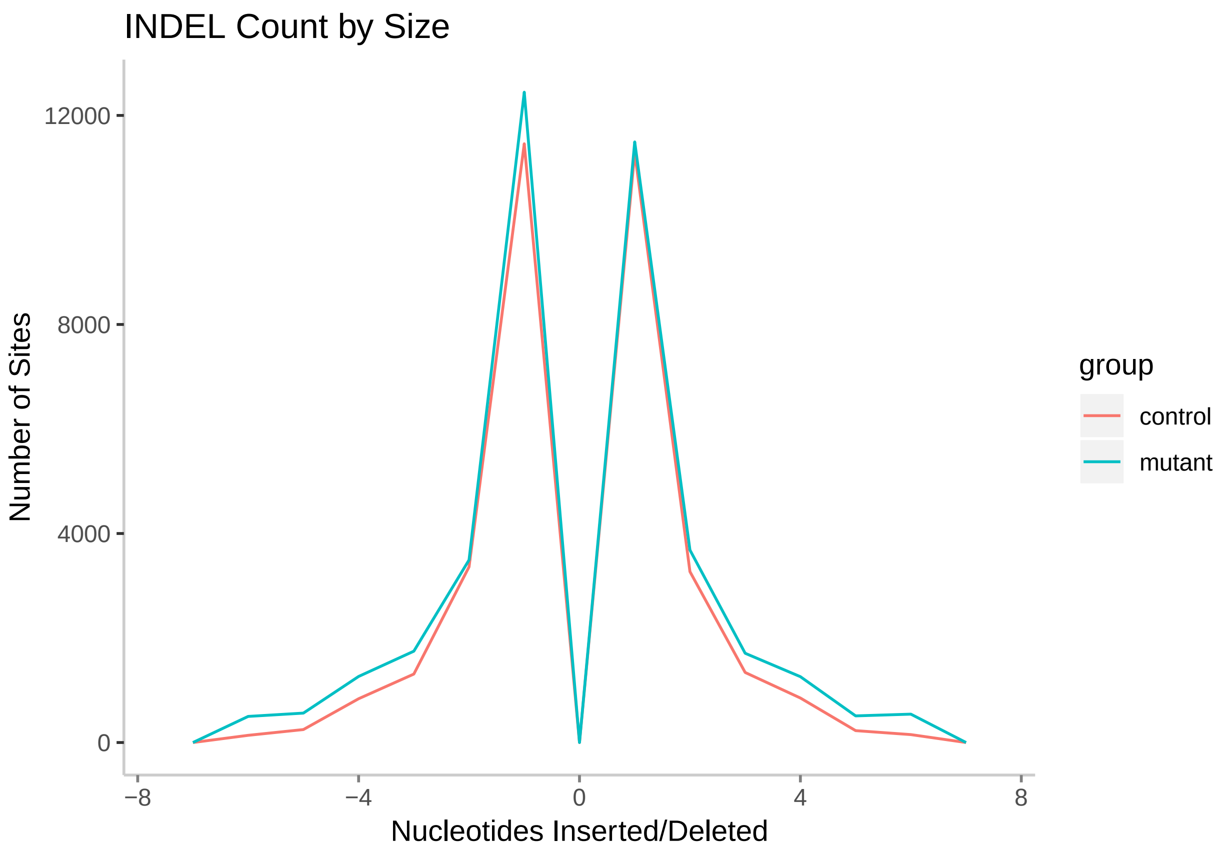
Table 13: Low-Frequency Site Details for Controls and Mutants

mac	control	mutant
0	5423	19
1	928	6651
2	906	1939
3	488	255

Both samples have slightly skewed but generally reasonable peaks in the 10-20 count range, corresponding to alleles with frequency ~25%. The control flies are depleted of alleles with a count near 30 (frequency ~50%), but the mutants are not. However, the mutants' variants in this range almost all on the arms of chromosome 3; conversely, almost all of its very rare variants are on the arms of chromosome 2. (*did we decide that this is because 2 is isogenized and 3 isn't?*)



Another unusual feature of both groups is the tendency of the mid-range variants (frequency ~25%) do not appear to be unimodes centered at MAC=15, but rather split into smaller peaks near MAC~10 and MAC~20. These would correspond to ratios of ~1/6 and 2/6.



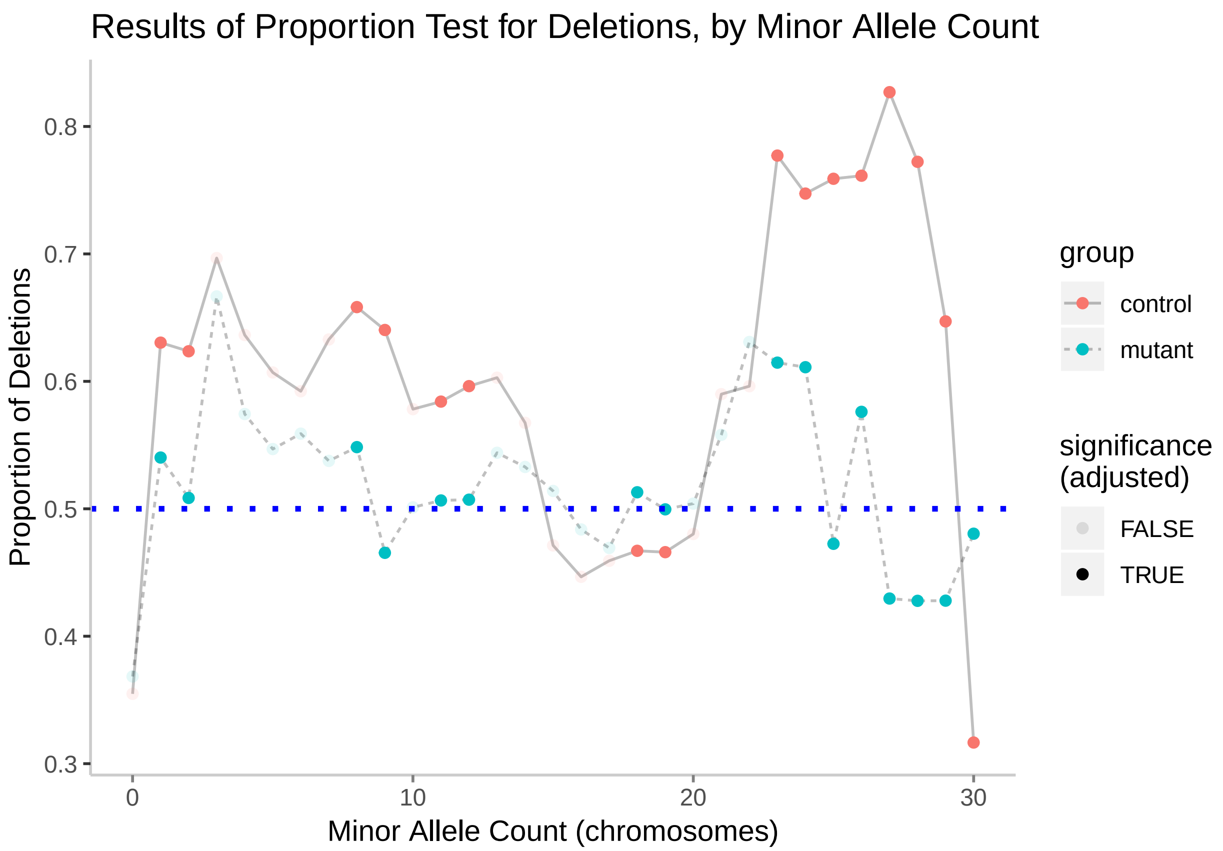
INDEL size ranged from 6bp deletions to 6bp insertions, with a small but significant bias towards deletion:

Table 14: Deletion Bias in mcm5 Mutants vs. Controls

group	del	ins	fraction_del	p.value
control	17353	17136	0.5031459	0.2447928
mutant	20013	19198	0.5103925	0.0000394

(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 degree of mutational bias varies, within and between the two treatments.



3.1.1 Novel INDELs (forward mutation)

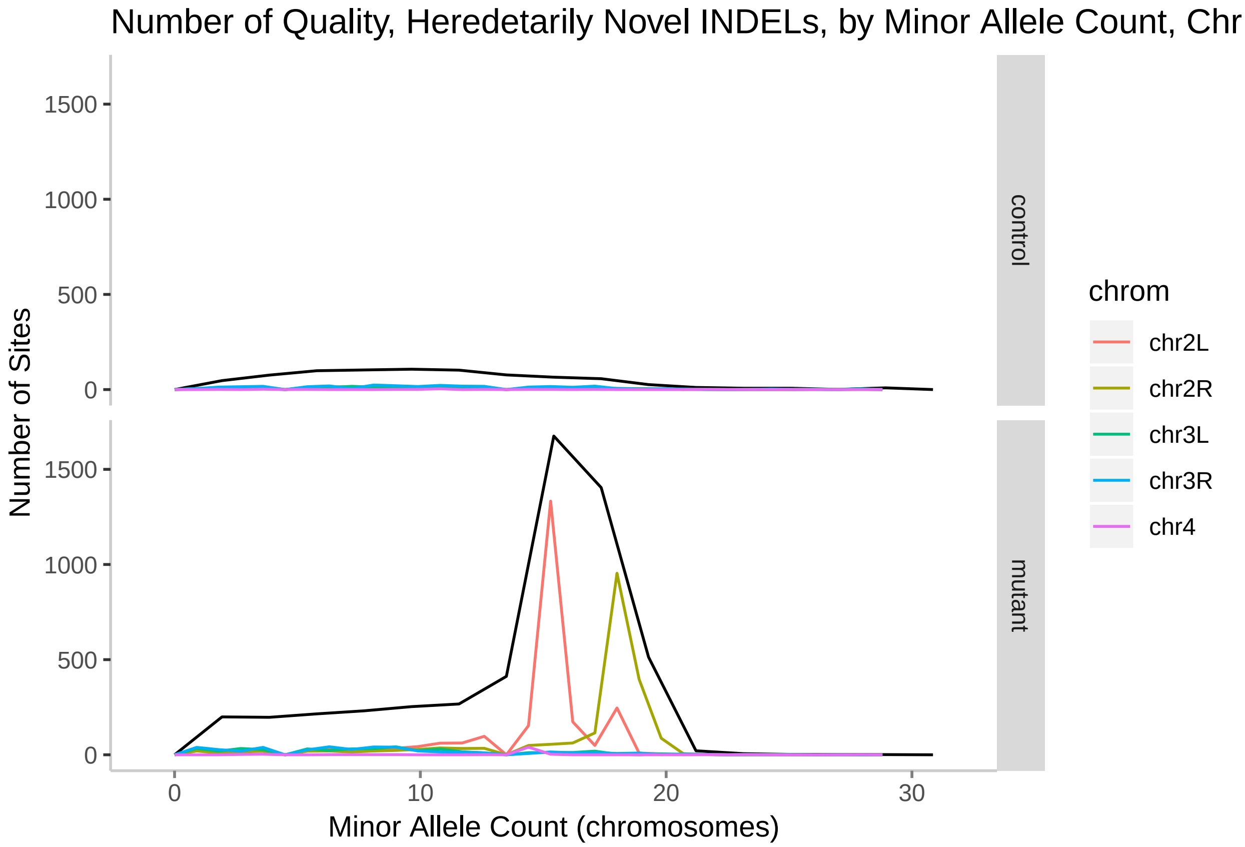
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

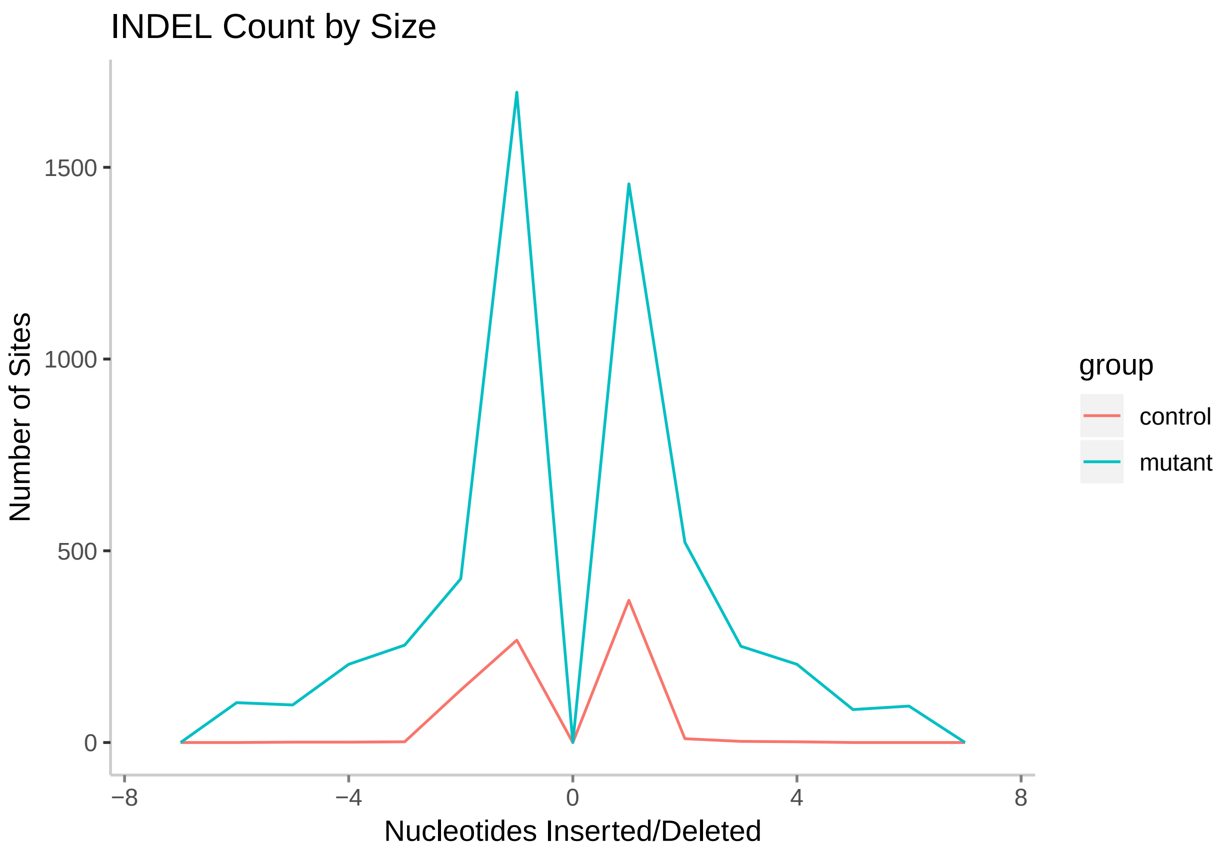
These sites were further filtered to collect those which are heredetarily novel, which is to say, non-reference alleles which appear in at least one offspring but neither parent. These are overall more common in the mutant flies than in the controls, especially on chromosome 2:

Table 15: Hereditarily Novel Biallaeleic INDELs, by Treatment and Chromosome

group	chrom	total	novel	percent_novel
control	chr2L	8134	137	1.68
control	chr2R	7905	180	2.28
control	chr3L	9200	176	1.91
control	chr3R	9022	286	3.17
control	chr4	228	15	6.58
mutant	chr2L	10882	2472	22.7
mutant	chr2R	10085	2024	20.1
mutant	chr3L	9064	418	4.61
mutant	chr3R	8891	428	4.81
mutant	chr4	289	56	19.4

In the mutant flies, the majority of the novel INDELs have an allele frequency ~25%, suggesting that they are cases in which one of the parents was actually a heterozygote but was misclassified:





The subset of heredetarily novel INDELs shows a significant deletion bias in mutants but not controls:

Table 16: Deletion bias in mcm5 Mutants (Hereditarily Novel)

group	del	ins	fraction_del	p.value
control	408	386	0.5138539	0.4561133
mutant	2783	2615	0.5155613	0.0230260

3.1.1.1 Novel, Singleton INDELs (forward mutation)

Assuming that DSBs can occur at many places in the genome with no preference, it is expected that a site will only experience a DSB once in a small population. Additionally, since the break occurs on one chromosome, the individual with the break is expected to be heterozygous for the resulting misrepair. Thus, the heredetarily novel INDELs were subsetted to retain only those with a minor allele count of 1.

Table 17: Hereditarily Novel, Singleton Biallaeleic INDELs, by Treatment and Chromosome

group	chrom	total	novel	percent_novel	novel_singleton	percent_novel_singleton
control	chr2L	8134	137	1.7	2	1.5
control	chr2R	7905	180	2.3	4	2.2
control	chr3L	9200	176	1.9	6	3.4
control	chr3R	9022	286	3.2	5	1.7
control	chr4	228	15	6.6	0	0.0
mutant	chr2L	10882	2472	22.7	27	1.1

group	chrom	total	novel	percent_novel	novel_singleton	percent_novel_singleton
mutant	chr2R	10085	2024	20.1	20	1.0
mutant	chr3L	9064	418	4.6	33	7.9
mutant	chr3R	8891	428	4.8	39	9.1
mutant	chr4	289	56	19.4	0	0.0

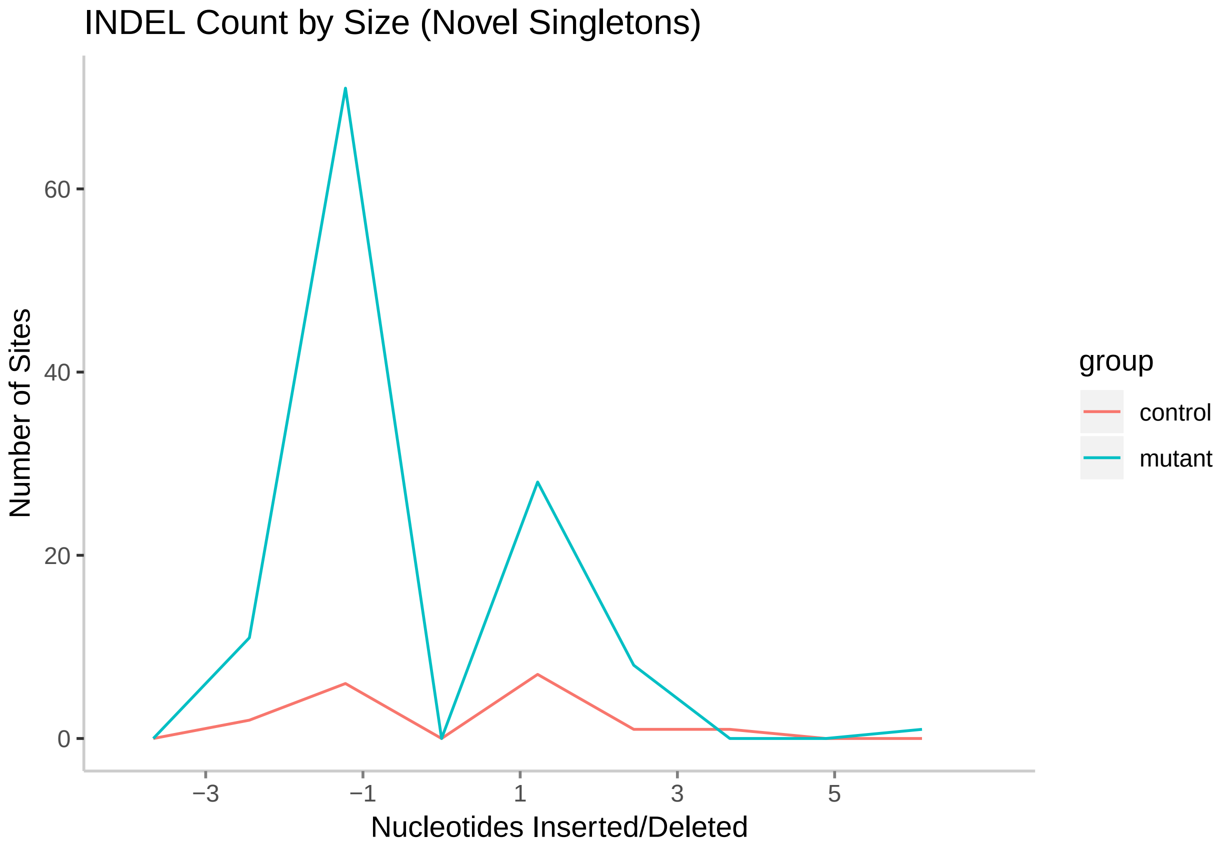


Table 18: Deletion bias in mcm5 Mutants (Hereditarily Novel Singletons)

group	del	ins	fraction_del	p.value
control	8	9	0.4705882	1.0e+00
mutant	82	37	0.6890756	5.5e-05

The mutant flies have more INDELs overall compared to the controls. The mutants also have a clear bias towards deletion, whereas none can be detected in the controls.

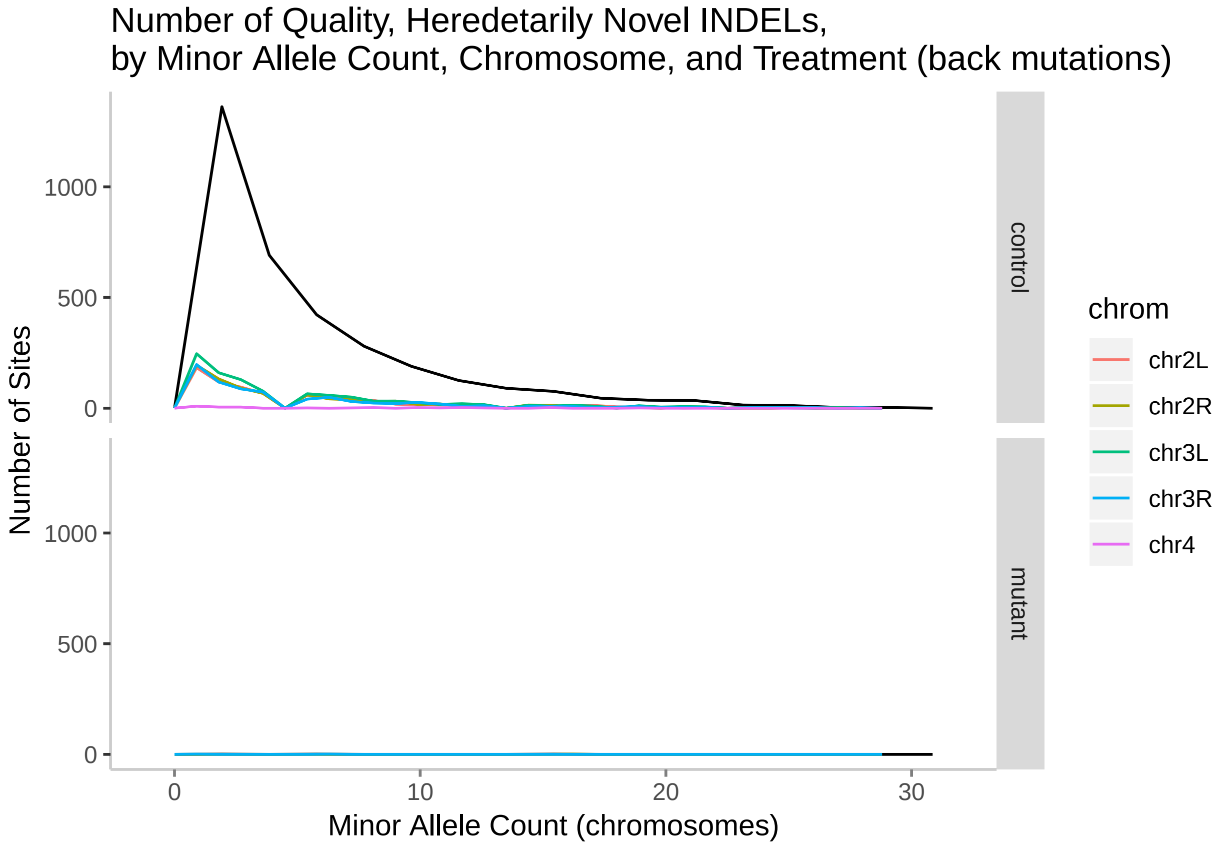
3.1.2 Novel INDELs (back mutation)

The above variants were classified as hereditarily novel if they went undetected in the parental alignments but appeared in the offspring. In particular, the parents were required to both be homozygous for the reference allele, whereas at least one offspring held an alternative allele. It is possible that the reverse could happen: both parents are homozygous for some alternate allele, and at least one of the offspring carried a reference allele. The biological interpretation would be that the site had a fixed difference with the reference, which mutated back to the reference in an offspring. This would probably be very rare. The computational interpretation would be that the variant caller has reverted to the reference allele as a default at a site which is difficult to resolve in some individuals. Here the same examination as above is applied to these apparent back-mutations.

There were very few such sites found in the mutants, compared to thousands in the control group:

Table 19: Hereditarily Novel Biallaeleic INDELs, by Treatment and Chromosome (Back Mutations)

group	chrom	total	novel	percent_novel
control	chr2L	8134	777	9.55
control	chr2R	7905	806	10.20
control	chr3L	9200	1008	10.96
control	chr3R	9022	759	8.41
control	chr4	228	32	14.04
mutant	chr2R	10085	1	0.01
mutant	chr3R	8891	2	0.02
mutant	chr2L	10882	0	0.00
mutant	chr3L	9064	0	0.00
mutant	chr4	289	0	0.00



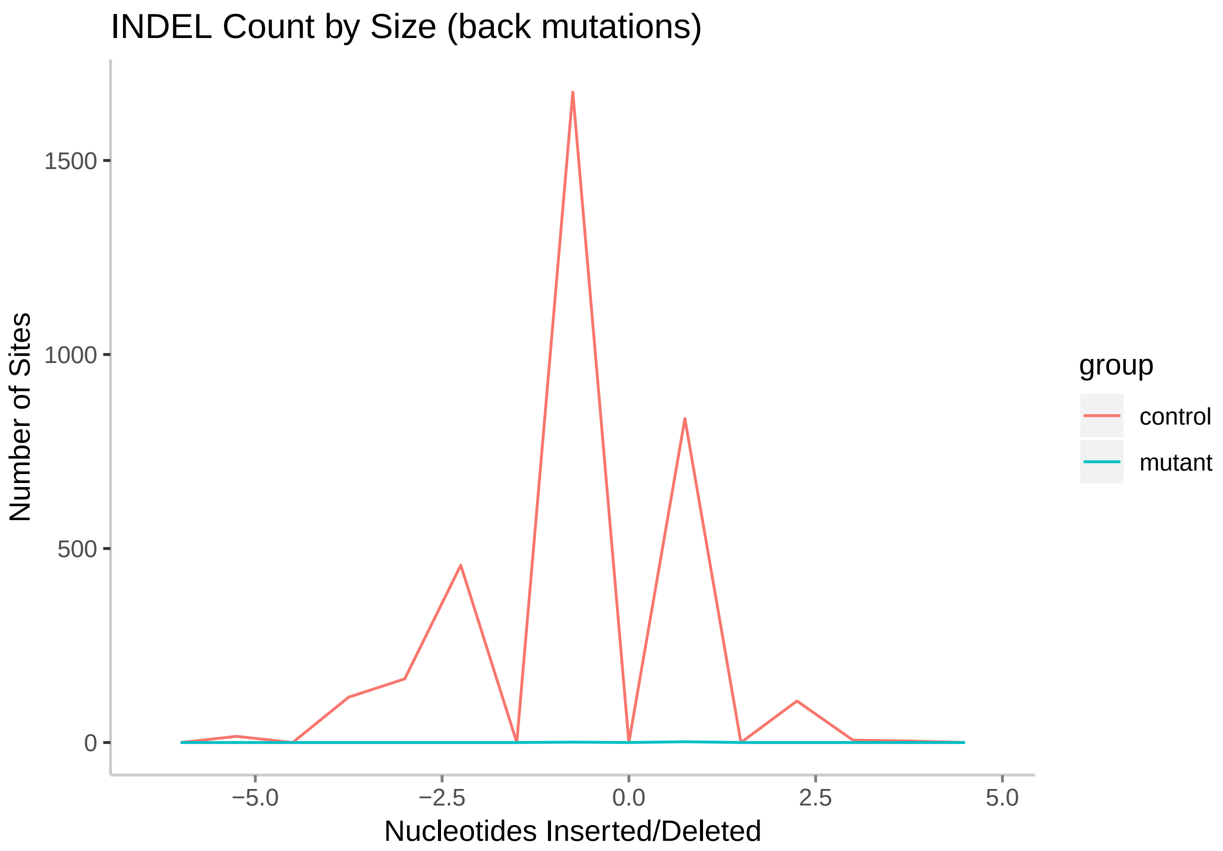


Table 20: Deletion bias in mcm5 Mutants (Hereditarily Novel; Back Mutations)

group	del	ins	fraction_del	p.value
control	2430	952	0.7185098	0
mutant	1	2	0.3333333	1

3.1.2.1 Novel, Singleton INDELs (back mutation)

The novel, apparent back mutations were then filtered to retain only those with a minor allele count of 1.

Table 21: Hereditarily Novel, Singleton Biallaeleic INDELs, by Treatment and Chromosome (Back Mutations)

group	chrom	total	novel	percent_novel	novel_singleton	percent_novel_singleton
control	chr2L	8134	777	9.5524957	183	23.55212
control	chr2R	7905	806	10.1960784	194	24.06948
control	chr3L	9200	1008	10.9565217	246	24.40476
control	chr3R	9022	759	8.4127688	197	25.95520
control	chr4	228	32	14.0350877	9	28.12500
mutant	chr2R	10085	1	0.0099157	0	0.00000
mutant	chr3R	8891	2	0.0224947	1	50.00000
mutant	chr2L	10882	0	0.0000000	0	NaN
mutant	chr3L	9064	0	0.0000000	0	NaN

group	chrom	total	novel	percent_novel	novel_singleton	percent_novel_singleton
mutant	chr4	289	0	0.0000000	0	NaN

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 6 rows containing missing values (geom_path).
```

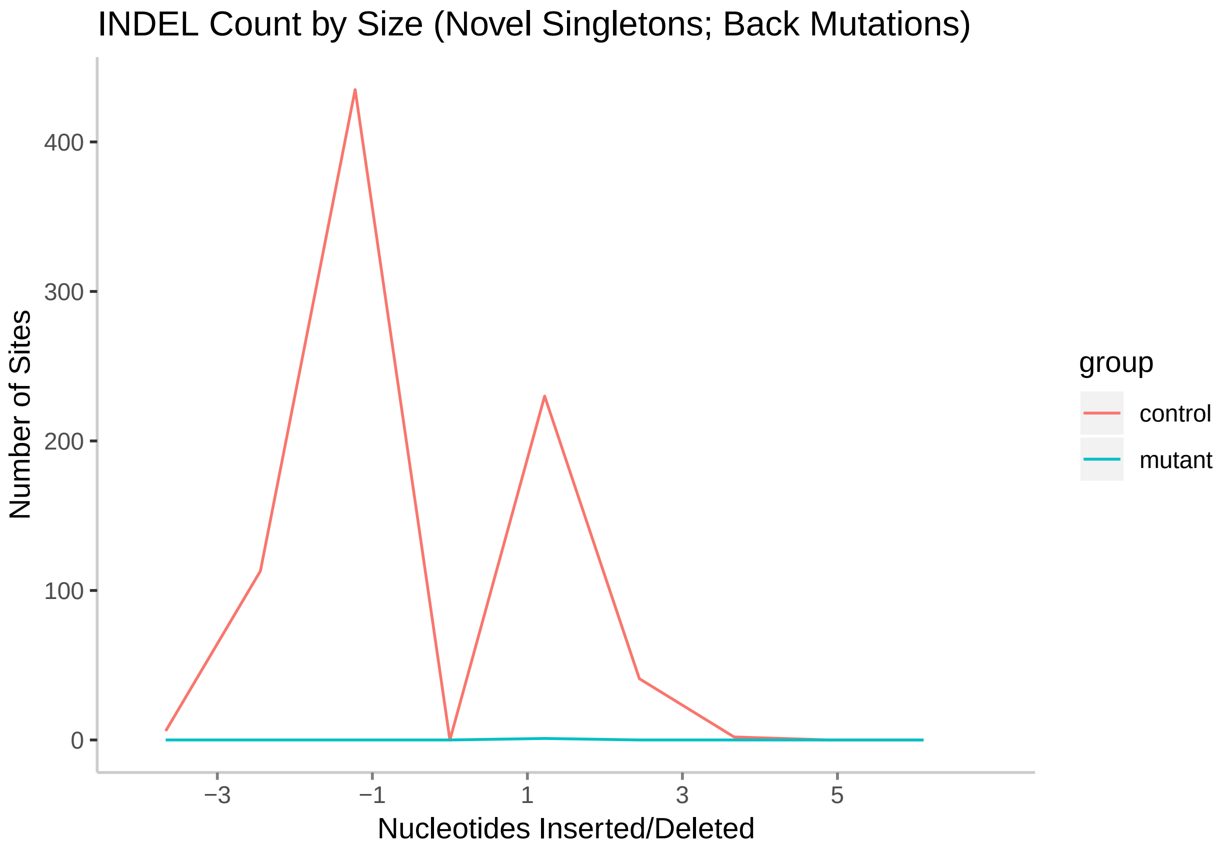


Table 22: Deletion bias in mcm5 Mutants (Hereditarily Novel Singletons; Back Mutations)

group	del	ins	fraction_del	p.value
control	556	273	0.6706876	0

The control flies have a clear excess of these apparent back mutations, many of which are listed as singeltons in the sample. These could be genuine back mutations or artefacts from the variant caller defaulting to the reference allele in some samples. Manual inspection of some of these sites suggested that they were glitchy in the effected samples; however, it is not obvious why either explanation would have such a lopsided impact on the controls but not the mutants.

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
- Resolve back-mutation mystery
-     - Local realignment?
- Include chromosome X - manage ploidy difference vs autosomes

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.