Bioinformatic Workflow for Deanhardt et al. 2021

Charlie Soeder

2023-05-05

Contents

1 Introduction

The primary question being investigated here is: how does gene expression change in the fruit fly nervous system, under a) various genome mutations (in OR47b, OR67d, and Fruitless) and b) different developmental environments (raised in groups or in social isolation). Towards this antennal RNASeq has been collected from flies under various conditions.

The basic analysis outline is:

I. Sequenced reads are filtered for quality, mapped to a reference genome, and counted against a reference gene annotation.

II. Read counts are used to estimate differential gene expression.
   A. A specific 1-factor model (~ housing) is implemented
   B. This is expanded to generic 1-factor models (~ condition, condition = housing, genotype...)
   C. This implementation is generalized further to an arbitrary multi-factor model, which is applied to a 2-factor model (~ housing + genotype)

III. Estimates for expression were compared across experimental treatments to identify genes of interest with similar behaviors.

Fruitless was examined in particular to check for differential exon use.

Along the way, two exterior concerns arose: that one of the Fru mutant samples were problematic, and that we were not certain of the sex composition of the samples. The former was examined by rerunning the models without it. This produced remarkably little change in the 1-factor models and remarkably large change in the 2-factor models. The latter concern was examined by comparing coverage on the sex chromosomes using published reads from NCBI as controls; this approach was inconclusive.

Late in analysis the base model was reverted from the 2-variable model to a collection of single variable models; the justification for this reversion was unclear:

> i had no idea that you had to merge two different experiments to generate the base mean and keep it like that. I also thought that since we have all the count data each experiment should be represented separately. Otherwise what is the purpose of doing the experiment or having the sh being an experimental condition? it makes no biological sense. ... the questions one asks about the data is what constantly changes. to understand and make sense of the biology not the stats
> (P. Volkan, Slack 24 Nov 2020)

2 Materials, Methods, Data, Software

At a top level, the workflow compares sequenced reads to bioinformatic databases, then uses specialized statistical software to analyze the results.

2.1 Reference Genomes

The dm6.13 reference genome was used for read alignment:

Table 1. Size and Consolidation of Reference Genomes
Drosophila Melanogaster

number bases	$138M$
number contigs	8

2.2 Reference Annotations

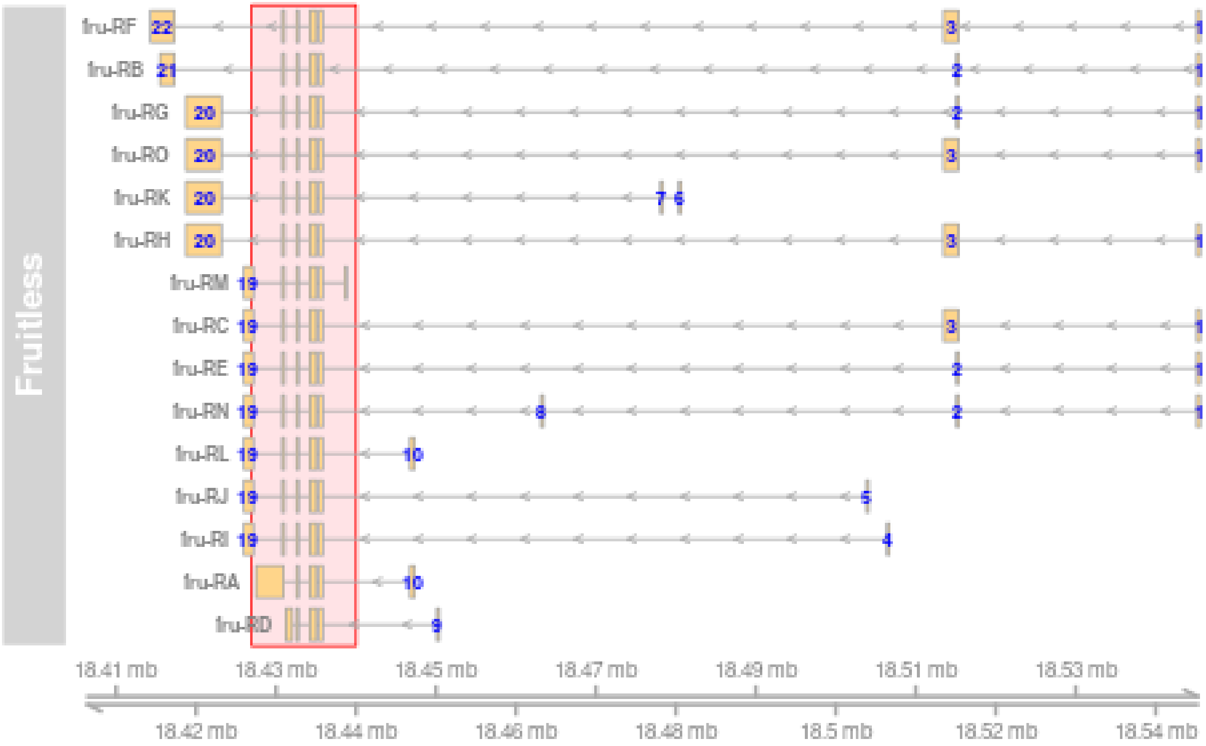
The dm6 reference annotations were used to define gene locii for differential expression analysis:

Table 2. Reference Annotations and their Sizes

	size (bp)		
annot	average	total	total count
dm6_genes	$5.8K$	$102.2M$	$17.7K$
dm6_repeats	197.1	$25.5M$	$129.4K$
fru_exons	939.3	$20.7K$	22
fru_intron	939.3	$20.7K$	22
fru_isoid	939.3	$20.7K$	22
fru_junct	939.3	$20.7K$	22

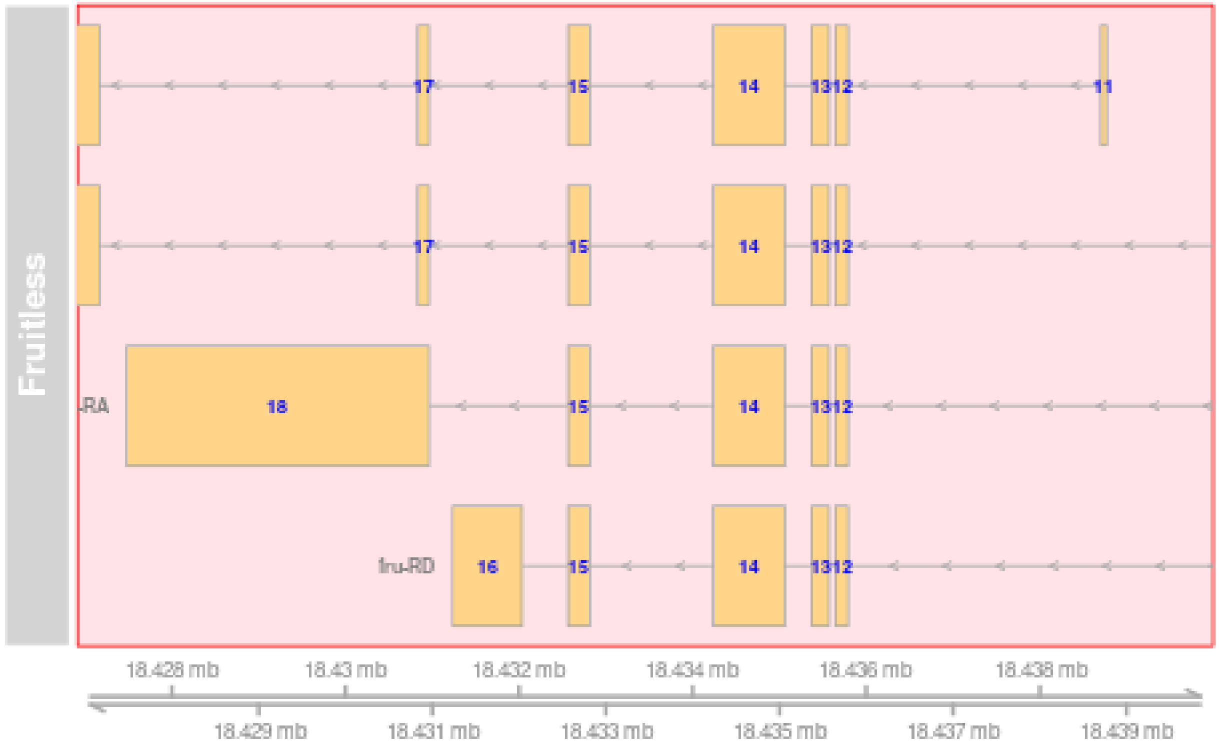
In addition to the genome as a whole, the gene Fruitless was given particular attention.

Figure 1. Fruitless gene model: exons and transcripts



```
## png
##   2
```

Figure 1 a. Fruitless gene model: exons and transcripts (detail)



```
## png
##   2
```

In order to focus on exon usage in Fru, the GTF entry was selected and decomposed into individual records per exon:

Table 3. Fru exons by Name
(chromosome 3R)

	start	stop
exon_22	18414273	18417301
exon_21	18415473	18417301
exon_20	18418716	18423183
exon_19	18425959	18427167
exon_18	18427480	18430965
exon_17	18430832	18430965
exon_16	18431233	18432035
exon_15	18432564	18432819
exon_14	18434235	18435063
exon_13	18435370	18435571
exon_12	18435643	18435791
exon_11	18438700	18438772
exon_10	18446701	18447330
exon_9	18450235	18450255
exon_8	18463267	18463282

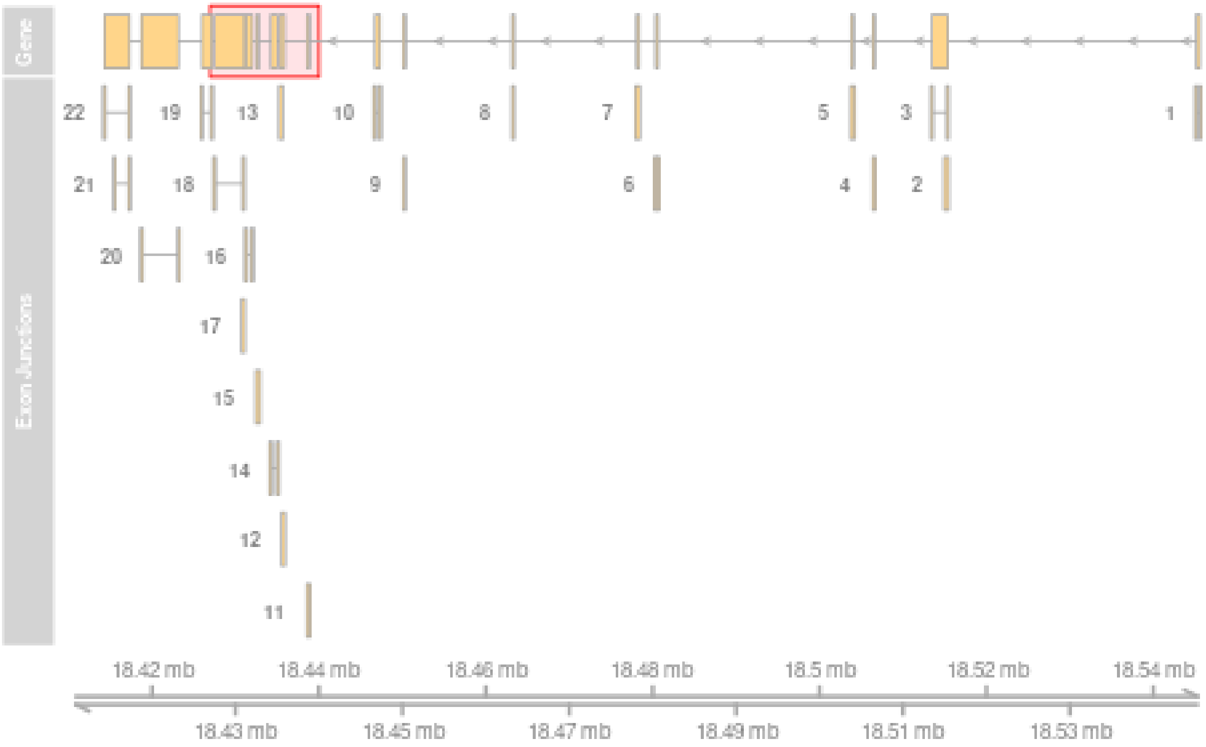
exon_7	18478064	18478333
exon_6	18480328	18480677
exon_5	18503846	18504067
exon_4	18506494	18506563
exon_3	18513451	18515344
exon_2	18515052	18515344
exon_1	18545113	18545587

```
cat /proj/cdjones_lab/Genomics_Data_Commons/annotations/drosophila_melanogaster/dmel-all-r6.13.gtf | gre
cat /proj/cdjones_lab/Genomics_Data_Commons/annotations/drosophila_melanogaster/dmel-all-r6.13.gtf | gre
cat fru.test.gtf.exon fru.test.gtf.gene | bedtools sort > utils/annotations/fru_ex.gtf
```

```
cat fru.test.gtf.exon | cut -f 1,4,5,7,9 | tr -d '"' | tr -d ";" | sed -e 's/gene_id //g' | awk '{print$

cat utils/annotations/fru_ex.bed.tmp | tr "_" "\t" | awk '{print$1,$2,$3,$4"_"$5,$6,$7}' | tr " " "\t" >
```

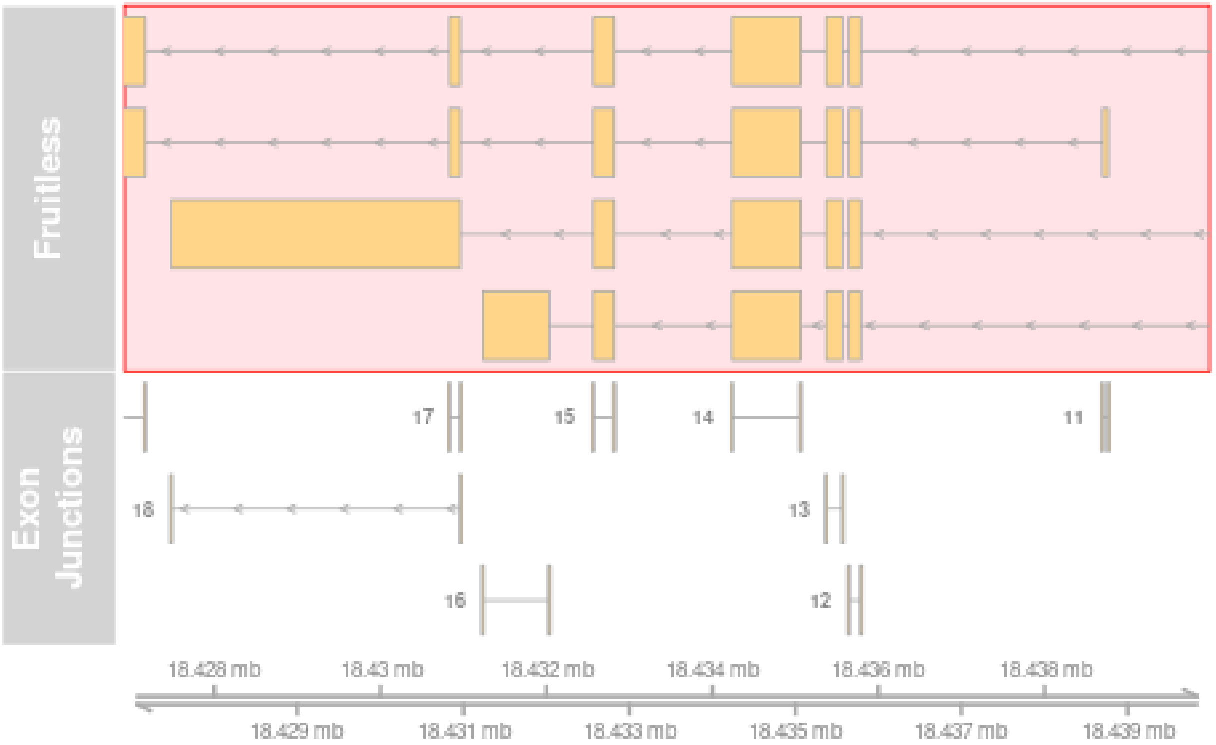
This gave the "fru_exons" annotation, to use for by-exon read counting. A further annotation, "fru_junct", was constructed by removing all of each exon except for splice junctions, ie, the 1bp boundaries of each exon which isn't a transcription start or stop site:

Figure 2. Fruitless gene model: junctions



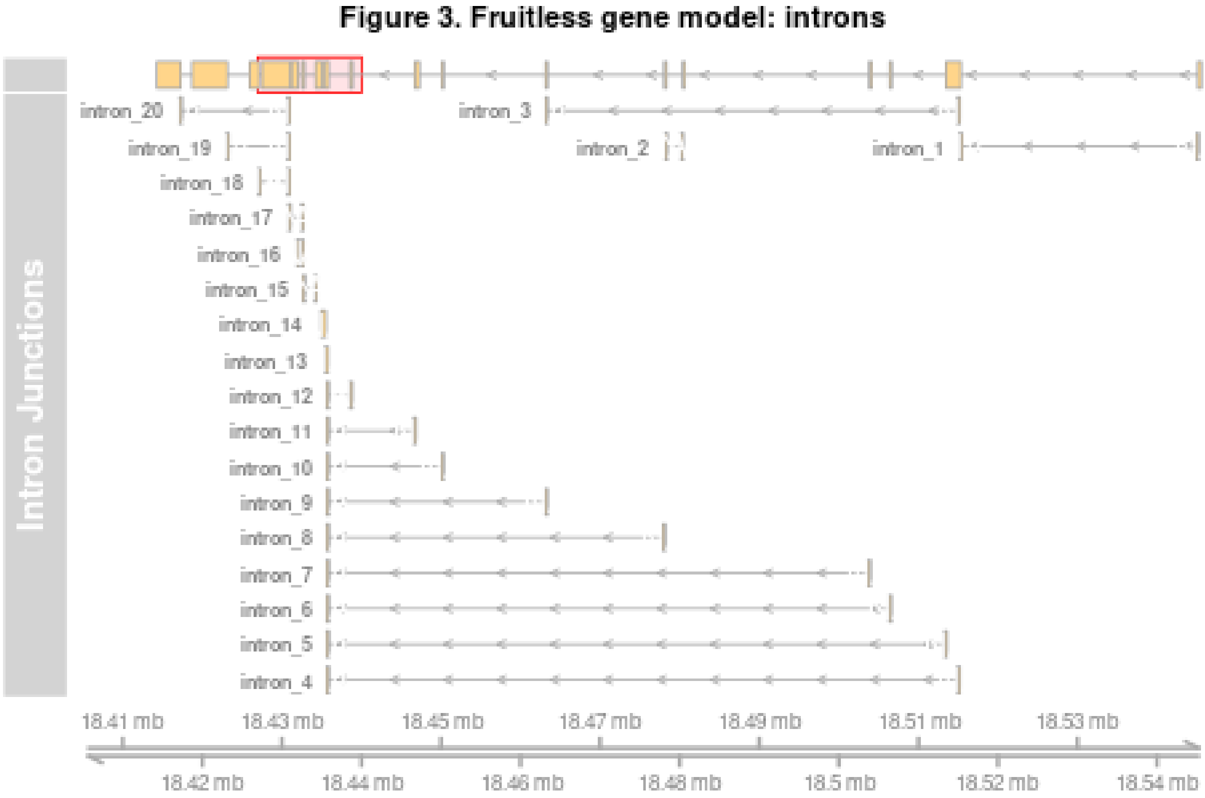
```
## png
##   2
```

Figure 2 a. Fruitless gene model: junctions (detail)



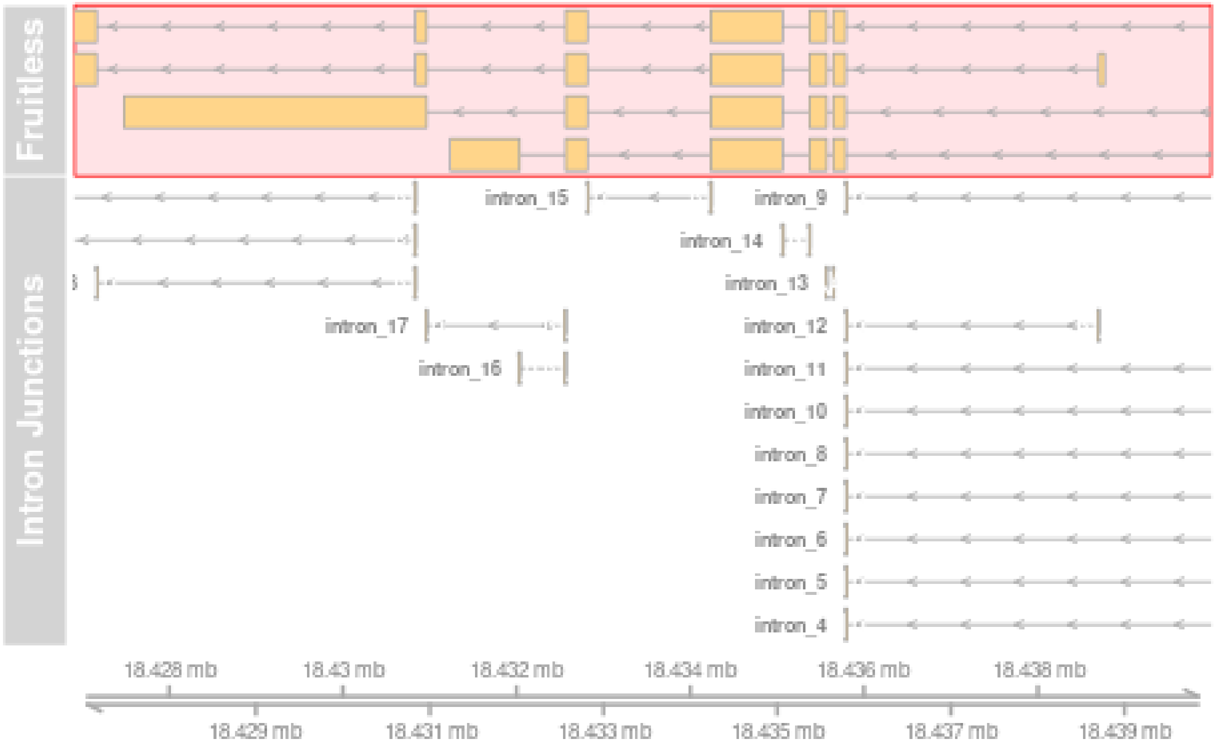
```
## png
##   2

cat utils/annotations/fru_ex.gtf  | grep -w gene > utils/annotations/fru_exonEdges.gtf
cat utils/annotations/fru_ex.gtf  | grep -w gene | cut -f 1,2 > utils/annotations/fru_exonEdges.gtf.fron
paste <( cat utils/annotations/fru_ex.gtf  | grep -w gene | cut -f 6- ) <( cat utils/annotations/fru_ex

paste utils/annotations/fru_exonEdges.gtf.front <(cat utils/annotations/fru_ex.gtf  | grep -w gene | awk

paste utils/annotations/fru_exonEdges.gtf.front <(cat utils/annotations/fru_ex.gtf  | grep -w gene | awk

cp utils/annotations/fru_exonEdges.gtf utils/annotations/fru_exonJunctions.gtf

cat utils/annotations/fru_introns.gtf | grep -w "exon" | awk '{print"chr"$0}' | cut -f 1,4,5 | sort | un
cat utils/annotations/fru_exonEdges.gtf | grep -w "exon" | cut -f 1,4,5 | sort | uniq > edges.bed
bedtools intersect -v -a <( cat edges.bed | awk '{print"chr"$0}' ) -b introns.bed > TSS_startStop.bed

bedtools subtract -a utils/annotations/fru_exonEdges.gtf -b  <( cat TSS_startStop.bed | cut -f 2- -d r

#cat utils/annotations/fru_exonJunctions.gtf.tmp | grep -v transcript_id | sed -e 's/exon_/exon_\t/g' |

#cat utils/annotations/fru_exonJunctions.gtf.tmp | grep transcript_id | sed -e 's/exon_/exon_\t/g' | sed

#cat utils/annotations/fru_exonJunctions.wrongStrand.gtf | sed -e 's/exon_/exon~/g' |sed -e 's/intron~2
```

Because a splice site represents two semi-independent exons but one intron, another annotation, “fru_intron”, was constructed consisting of the introns in Fruitless. The same 1-bp subintervals were used as in “fru_junct”, but in this case they were organized by the intron they bounded rather than by the exon:



```
## png
##   2
```

Figure 3 a. Fruitless gene model: introns (detail)



```
## png
##   2
```

```
rm -f coords.all
for transcript in $(cat /proj/cdjones_lab/Genomics_Data_Commons/annotations/drosophila_melanogaster/dmel
    echo $transcript;

    cat /proj/cdjones_lab/Genomics_Data_Commons/annotations/drosophila_melanogaster/dmel-all-r6.13.gtf

    head -n 1 coords.tmp | cut -f 2 | awk '{print "0\t"$0}' >> coords.all
    tail -n 1 coords.tmp | cut -f 2 | awk '{print $0"\t0"}' >> coords.all

    paste   <(cut -f 2 coords.tmp |tail -n +2 ) <(cut -f 1 coords.tmp | head -n -1 ) >> coords.all

done

cat coords.all | sort | uniq | grep -v -w 0 |awk -F'\t' 'NR>0{$0=$0"\tintron_"NR} 1'> coords.unq


cat coords.unq  | awk '{print"3R\tFlybase\tgene\t"$1"\t"$2"\t.\t-\t.\tgene_id ~"$3"~;"}' | tr '~' '"' |
cat coords.unq  | awk '{print"3R\tFlybase\tgene\t"$1"\t"$2"\t.\t-\t.\tgene_id ~"$3"~;"}' | tr '~' '"' |

cat <(cat coords.unq | grep -v -w 0 | awk '{print"3R\tFlybase\tgene\t"$1"\t"$2"\t.\t-\t.\tgene_id ~"$3"
cat utils/annotations/fru_introns.gtf.tmp | sed -e 's/intron_/intron~/g' |sed -e 's/intron~20/intron_1/g
```

(pull these into an annotation-builder rule?)

fru_junct and fru_intron annotations were used with the *_SplicedOnly aligments (section ~~)

2.3 Gene Lists

In addition to the full annotations, subsets containing prespecified genes of interest will also be used.

Here are those subsets and their sizes:

Table 4. Predefined Subsets of Gene Annotation

measure	brysonPriority	brysonsList	histoneMod	ionChannel	ionotropic	mating	nervSysDev
total count	25	35	8	250	246	3	93
annotated count	25	34	8	250	246	3	90
percent of annotations	0.1%	0.2%	0.0%	1.4%	1.4%	0.0%	0.5%
total size	554.5K	3.1M	46.9K	4.0M	3.7M	5.0K	1.8M
avg size	22.2K	91.1K	5.9K	16.2K	15.2K	1.7K	19.8K
percent genome size	0.4%	2.3%	0.0%	2.9%	2.7%	0.0%	1.3%
percent annotation size	0.5%	3.0%	0.0%	4.0%	3.7%	0.0%	1.7%

2.3.1 Ionotropic

A list of ionotropic receptors supplied by Corbin via Flybase & George et al 2019 (email 28 May 2019). This contained 335 entries, some with mutiple genes, some not unique. Once merged & uniqued : 246 Annotation symbols (CGxxxxx) converted to FlyBase gene games (FBgnxxxx) using flybase ID converter (http://flybase.org/convert/id)

239 converted cleanly; 5 had duplicate conversions and were corrected by hand:

```
CG11430 is FBgn0041585, not FBgn0050323
CG43368 is FBgn0263111, not FBgn0041188
CG8885 is FBgn0262467, not FBgn0081377
CG9090 is FBgn0034497, not FBgn0082745
CG9126 is FBgn0045073, not FBgn0053180
```

Two were corrected to be consistent with the dm6_genes annotation:

```
CG9907 (para), is listed as FBgn0264255 not FBgn0285944
CG42345 (straw) is listed as FBgn0259247 (laccase2)
```

2.3.2 Derived from GO terms

```
Sub Pull out by particular GO terms?
o Nervous system development - http://flybase.org/cgi-bin/cvreport.pl?rel=is_a&id=GO:0007399
o Mating - http://flybase.org/cgi-bin/cvreport.pl?rel=is_a&id=GO:0007618
o Histone modification - http://flybase.org/cgi-bin/cvreport.pl?rel=is_a&id=GO:0016570
```

```
o Dna-binding transcription factor - http://flybase.org/cgi-bin/cvreport.pl?id=GO%3A0003700
o Synaptic signaling - http://flybase.org/cgi-bin/cvreport.pl?rel=is_a&id=GO:0099536
o Synapse organization - http://flybase.org/cgi-bin/cvreport.pl?id=GO%3A0050808
```

(Bryson, email 24 July 2019)

```
o Ion Channel Activity - http://flybase.org/cgi-bin/cvreport.pl?rel=is_a&id=GO:0005216
```

(Bryson, email 12 May 2020)

melanogaster-specific genes with these GO terms were retrieved using the FlyBase QueryBuilder.

Nervous System Development:

```
nrd, FBgn0002967, no annotated gene model
l(2)23Ab, FBgn0014978, same
aloof, FBgn0020609, same
Imp, FBgn0285926, is FBgn0262735
```

Mating:

```
Only three, but all good
```

synapse signalling

```
1 gene
```

Histone modification, DNA trans factor act, synapse org

```
MT
```

Ion Channel Activity

```
251, all good
```

2.3.3 Bryson's Lists

Interest: (email, 29 Oct 2019)

```
Neverland: annotated as FBgn0259697, not FBgn0287185
```

Priority: (email, 5 Nov 2019; 7 Nov 2019)

2.4 Sequenced Reads

The sequenced reads covered three replicates each of 5 experimental conditions. The conditions included varying genotype, housing, and age (all RNA was collected from antenna tissue).

Table 5. Experimental Conditions and Replicates

&nbsp

genotype	housing	age (days)	tissue	# replicates
47b1	group	7	antennae	3
67d	group	7	antennae	3
FruLexaFru440	group	7	antennae	3
wt	group	7	antennae	3
wt	isolated	7	antennae	3

In addition to the novel reads, RNA-Seq from drosophila melanogaster antennae were downloaded from NCBI (PRJNA388757; Shiao et al. (2015)), one annotated as male and the other as female. These will be compared to the unpublished samples to try to confirm the sex of the flies they came from. This analysis was computationally problematic and ultimately inconclusive, and has been deactivated in this version.

2.4.1 Pre-Processing

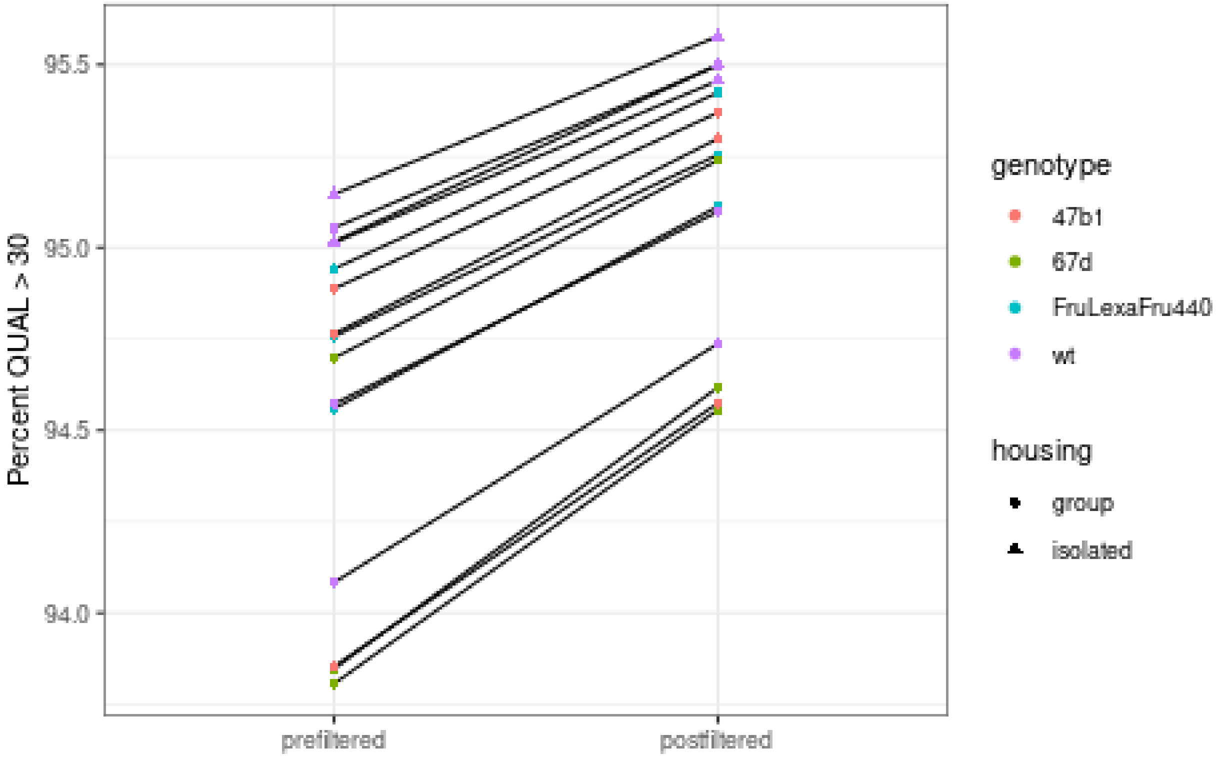
These reads were preprocessed with FASTP (Chen et al. 2018) for quality control and analytics.

Starting FASTQ files contained a total of $452M$ reads; after QC, this dropped to $445M$.

Table 6. Read Retention Rate during Preprocessing

	minimum	average	maximum
prefiltered	$22M$	$30M$	$43M$
postfiltered	$22M$	$30M$	$43M$
percent retention	98	98	99

Figure 4. Percent of Reads with a mean QUAL > 30



```
## png
##   2
```

Duplicate reads were also detected

Table 7. Percentage Duplication
FASTP estimate

minimum	average	median	maximum
5.5	7.4	6.8	17.9

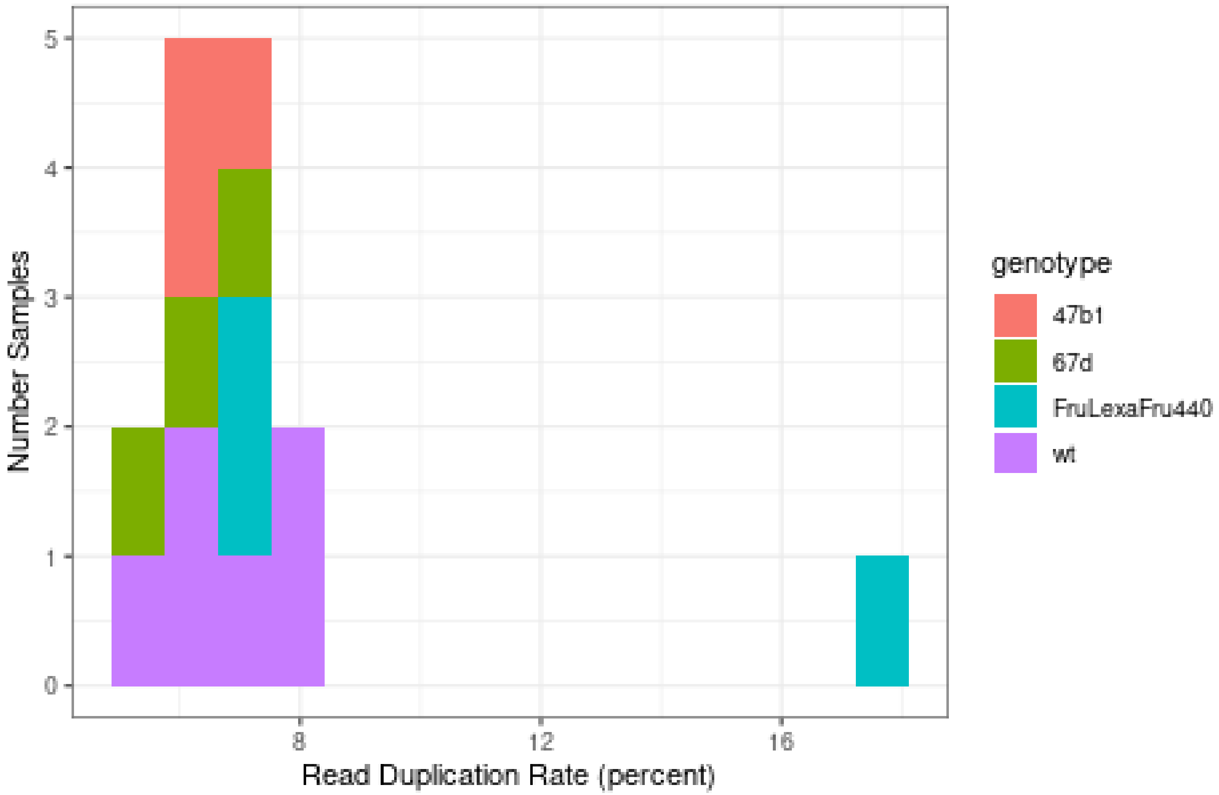


Figure 5. Duplication Histogram (FASTP estimate)

```
## png
##   2
```

2.5 Mapped Reads

Reads were mapped to the reference genome using MapSplice2 (Wang et al. 2010). Because MapSplice is written in python2, the code was downloaded and automatically refactored using the 2to3 python utility so that it would run in the python3 snakemake environment: https://docs.python.org/2/library/2to3.html

2.5.1 Raw Mapsplice

Of the 445M reads, MapSplice was able to align 442M of them, for an overall mapping rate of 99.2251335 %.

Individual mapping rates were generally more than 98%.

Table 8. Percent of Reads Mapping
raw mapsplice output

maximum	mean	median	minimum
99.7%	99.2%	99.1%	98.5%

Table 9. Individual Mapping Rates

raw mapsplice output

rep	day	total reads	reads mapped	percent mapped
group - 47b1				
1	7	32.1M	31.9M	99.5%
2	7	28.9M	28.8M	99.7%
3	7	24.3M	24.3M	99.6%
group - 67d				
1	7	25.1M	25.0M	99.6%
2	7	31.2M	31.0M	99.5%
3	7	24.1M	24.0M	99.6%
group - wt				
1	7	42.6M	42.2M	99.0%
2	7	31.5M	31.0M	98.5%
3	7	30.2M	29.9M	99.0%
isolated - wt				
1	7	30.7M	30.4M	99.1%
2	7	27.2M	27.1M	99.5%
3	7	33.8M	33.5M	99.0%
group - FruLexaFru440				
1	7	22.0M	21.7M	98.9%
2	7	30.7M	30.4M	99.1%
3	7	30.7M	30.4M	99.1%

Table 10. Percent of Duplicate Reads

raw mapsplice output

maximum	mean	median	minimum
122.3%	104.7%	109.3%	53.5%

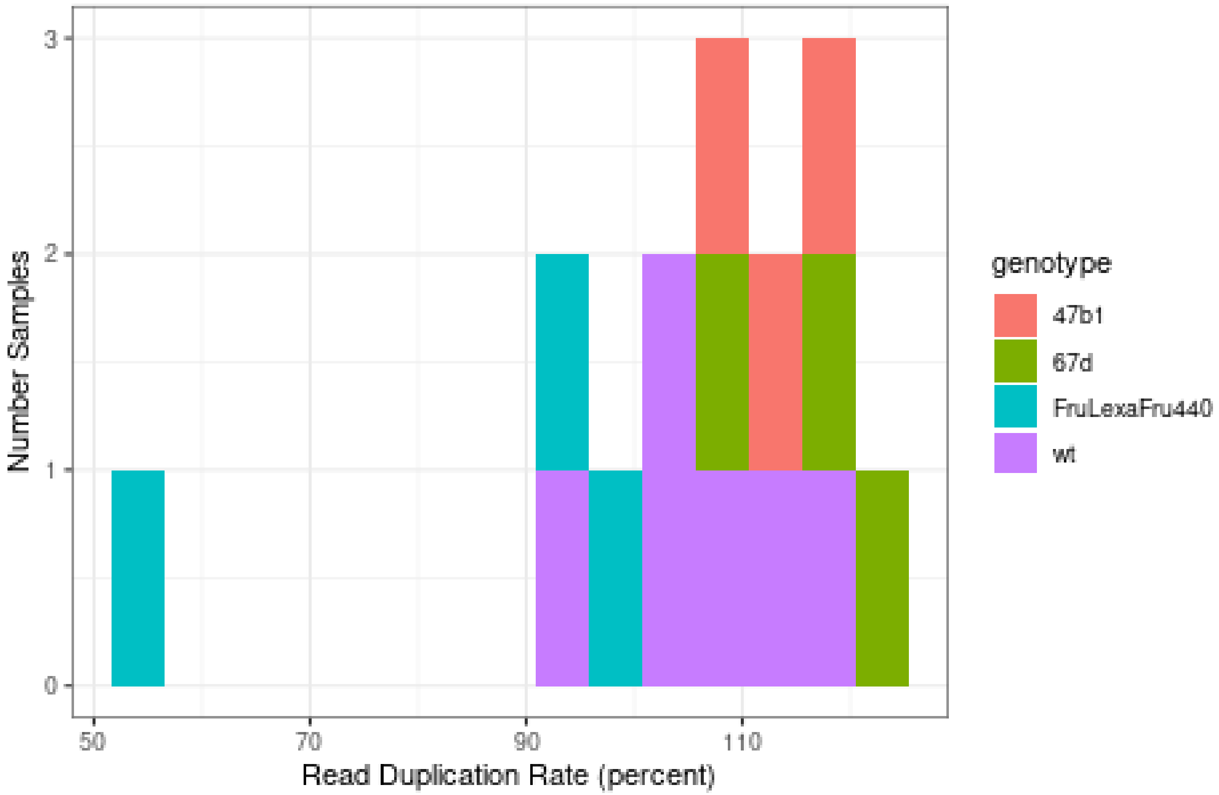
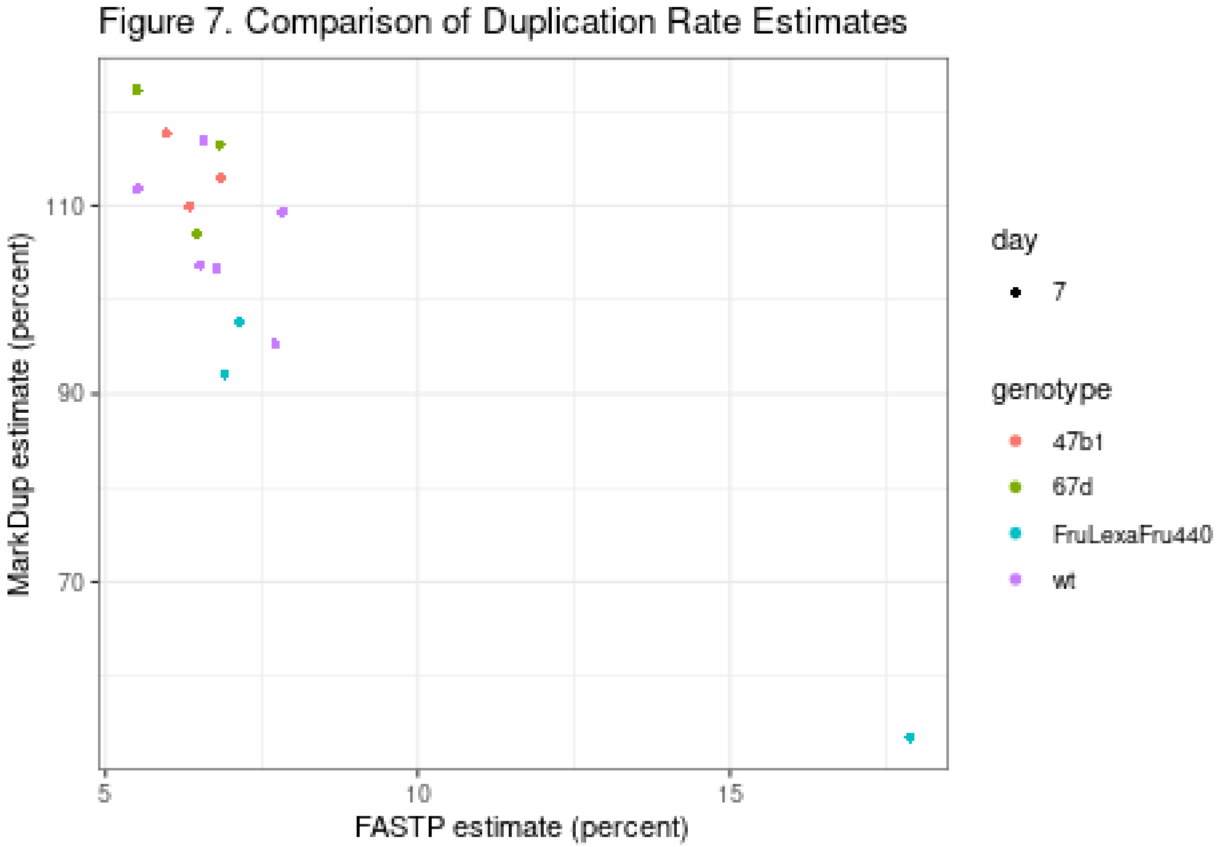


Figure 6. Duplication Histogram (Raw Mapsplice Alignment)

```
## png
##   2
```

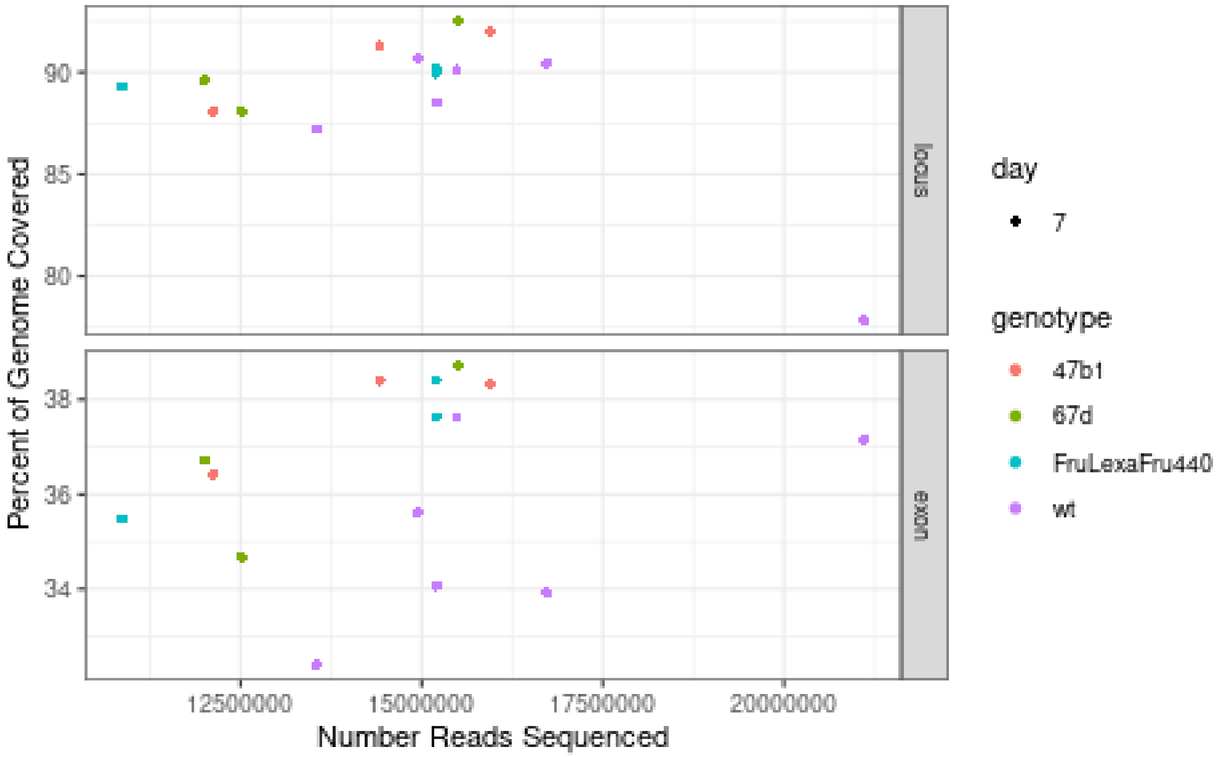
Although Samtools marks duplicates at a higher rate than FASTP, the estimates are correlated; in particular, both agree that FruLexa/Fru440 day 7 replicate 1 is a highly duplicated outlier. The NCBI reads are anomalous.



```
## png
##   2
```

Genome-wide depth of coverage is not very meaningful here, in the case of RNA-Seq. Breadth of coverage (the fraction of the genome which is covered by at least one read) is, but the ideal case is not 100% coverage like in a DNA-Seq; rather, we'd expect breadth to approximate the fraction of the genome which is under active transcription. Another complication is whether the reads which fall on splice junctions are treated as covering the intronic region or not (this corresponds to the distinction between the percent of the genome which is a transcribed locus vs the percent which is a transcribed exon).

Figure 8. Breadth of Coverage of Raw Mapsplice Alignment Compared to Read Count

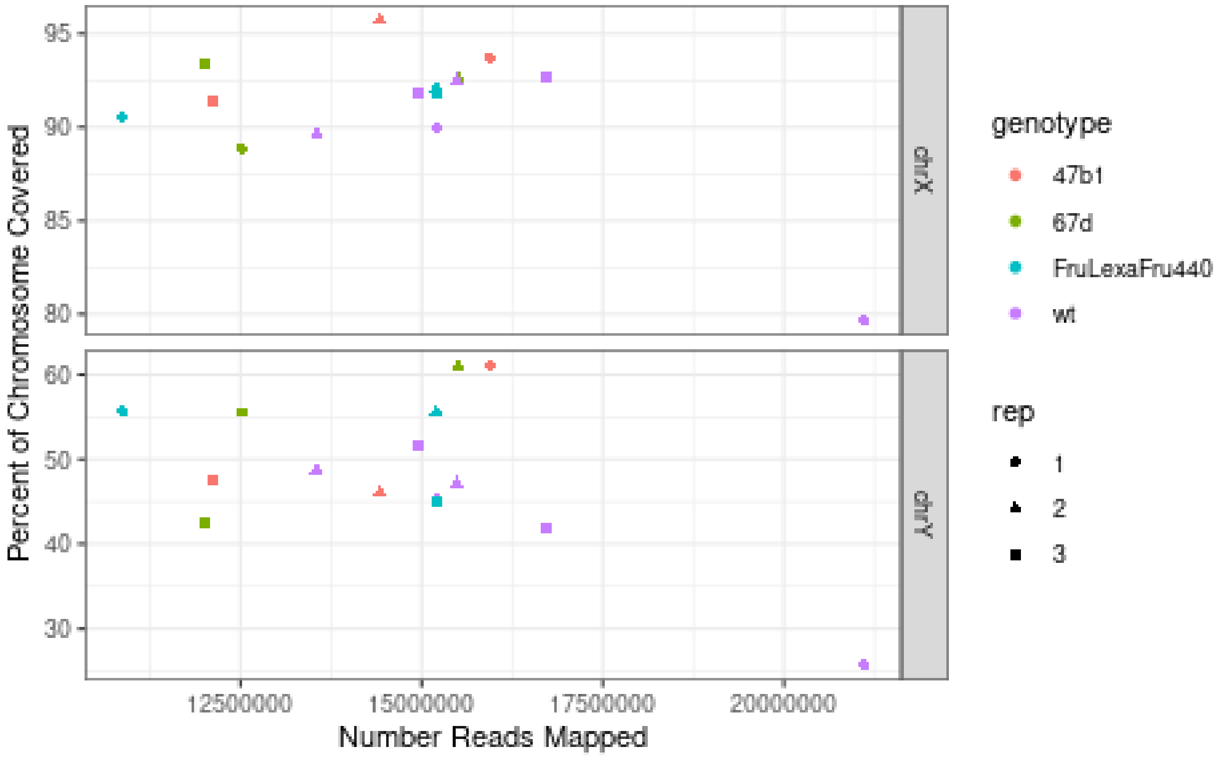


```
## png
##   2
```

There appears to be a slight dependence of breadth upon sequencing depth (ie, the number of reads sequenced), meaning that sequencing depth of these samples is not so great that the breadth covered is saturated. The breadth of the CantonS flies is unusually low for their depth of mapping.

We can also compare the breadth of coverage on the X and Y chromosomes to confirm that the flies sampled are all the same sex. The only outlier is the group-housed wildtype replicate 1, which is also anomalous genome-wide. The two samples from (Shiao et al. 2015) (not shown) agree well on the X chromosome, which is not unexpected, and the female-annotated sample has lower coverage on the Y, as expected. However, the difference between the NCBI controls is well within the variation of the new sequences, so this doesn't work as a decisive diagnostic.

Figure 9. Fraction of Sex Chromosome Covered
in Raw Mapsplice Alignments Compared to Read Count



```
## png
##   2
```

2.5.2 Filtered Multimap

From the raw MapSplice output, three filtered alignments were produced. The first, mapspliceMulti, has had duplicates marked and removed, and has been filtered to require proper pairing and a minimum mapping quality (SAM flags “-q 20 -F 0x0200 -F 0x04 -f 0x0002”; markdup flags “-rS”). Thus, mapspliceMulti is a filtered alignment that retains all locii for multimapped reads.

The filtration process removed a total -2.56 of $445M$ mapped reads, an overall mapped retention rate of 39.0986331 %.

Table 11. Sample Read Retention Rate
percent of reads retained when filtering raw alignment

	maximum	mean	median	minimum
mapped retention	80.9%	78.1%	79.2%	64.5%

Table 12. Sample Coverage Retention Rate
percent of coverage retained when filtering raw alignment

	maximum	mean	median	minimum

spanned breadth retention	99.7%	99.6%	99.6%	99.4%
split breadth retention	97.2%	97.0%	97.1%	96.7%

Although filtration removed some (45.921069 %) of the multimapping reads, 6.21M remain ambiguously mapped. A given read mapped, on average, to 1.10158077929945 locations These will be kept as-is in mapspliceMulti, but will be further filtered in other alignments.

Table 13. Mapping Uniqueness & Multiplicity
effect of filtering on multimapping reads

	percent of reads uniquely mapping		average per-read mapping multiplicity	
rep	raw	multi	raw	multi
47b1 - group - 7				
1	96.0%	96.0%	1.19	1.13
2	95.5%	95.7%	1.20	1.14
3	95.6%	95.6%	1.19	1.12
67d - group - 7				
1	96.7%	97.0%	1.15	1.10
2	95.8%	96.0%	1.23	1.15
3	96.0%	96.3%	1.21	1.14
wt - group - 7				
1	97.6%	97.8%	1.09	1.06
2	95.8%	95.9%	1.11	1.07
3	97.4%	97.8%	1.10	1.06
wt - isolated - 7				
1	97.7%	98.0%	1.08	1.06
2	97.7%	98.1%	1.08	1.05
3	97.7%	98.0%	1.08	1.06
FruLexaFru440 - group - 7				
1	95.2%	95.0%	1.22	1.17
2	96.7%	96.8%	1.15	1.11
3	95.7%	95.6%	1.15	1.10

2.5.3 Downsampled Multimapped

mapspliceRando is a downsampled alignment constructed by selecting at random a single location for each multimapped read, then merging the unambiguously located reads with mapspliceUniq.

Table 14. Downsampling Retention Rate
percent of alignment retained when multimappers are downsampled

	maximum	mean	median	minimum
mapped retention	99.2%	98.3%	98.3%	97.0%
spanned breadth retention	99.4%	99.1%	99.1%	98.0%
split breadth retention	90.1%	89.7%	89.7%	89.1%

2.5.4 Uniquely Mapped

mapspliceUniq is derived from mapspliceMulti by further filtering out the multimapped reads and keeping only those which map uniquely.

Table 15. Uniquely Mapped Retention Rate
percent of alignment retained when multimappers are excluded

	maximum	mean	median	minimum
mapped retention	98.1%	96.6%	96.3%	95.0%
spanned breadth retention	99.1%	98.8%	98.9%	97.6%
split breadth retention	87.6%	87.2%	87.3%	86.5%

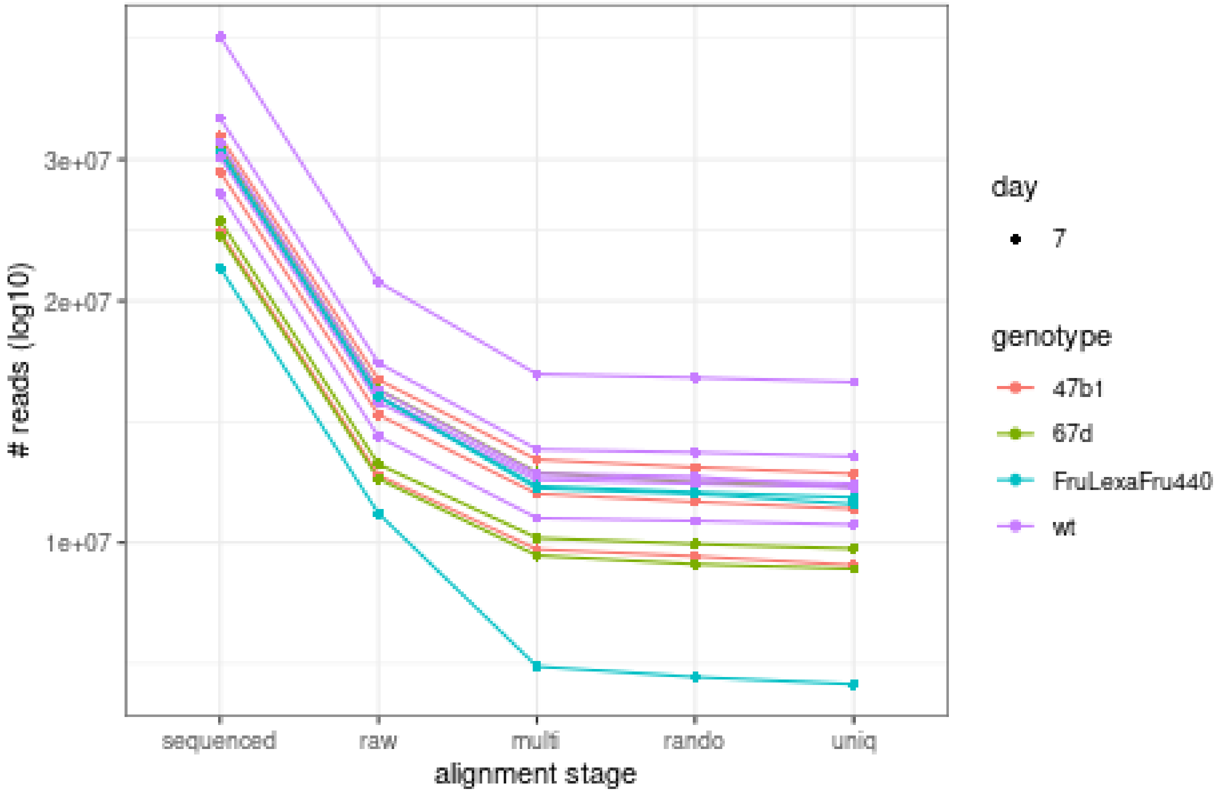
2.5.5 Spliced-Only

For each of Multi, Rando, and Uniq, a _SpliceOnly alignment was constructed by first filtering to only include spliced reads (awk '($6 ~ /N/)'), then reducing the reads to 1 bp on either side of the splice site. These are used with the fru_junct and fru_intron annotations .

2.5.6 Alignment Process Overview

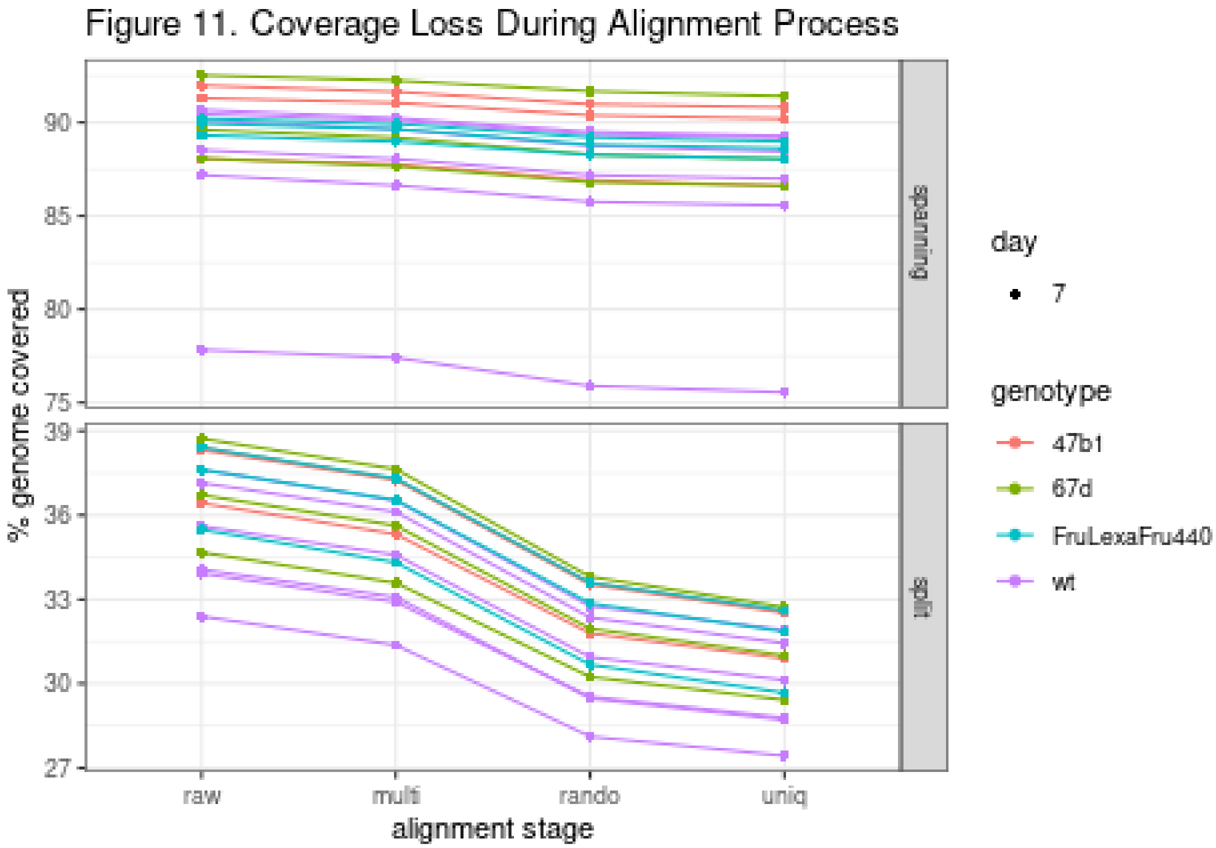
Here are the number of reads per sample, from the intial sequencing to the most heavily filtered alignment:

Figure 10. Read-count Dropout During Alignment Process



```
## png
##   2
```

The coverage dropout during the alignment filtration can be similarly tracked:



```
## png
##   2
```

When restricted to the sex chromosomes, the NCBI controls were almost indistinguishable, with the difference between them much smaller than the difference between experimental samples. So, accounting for multimapping reads also doesn't make this a useful diagnostic:

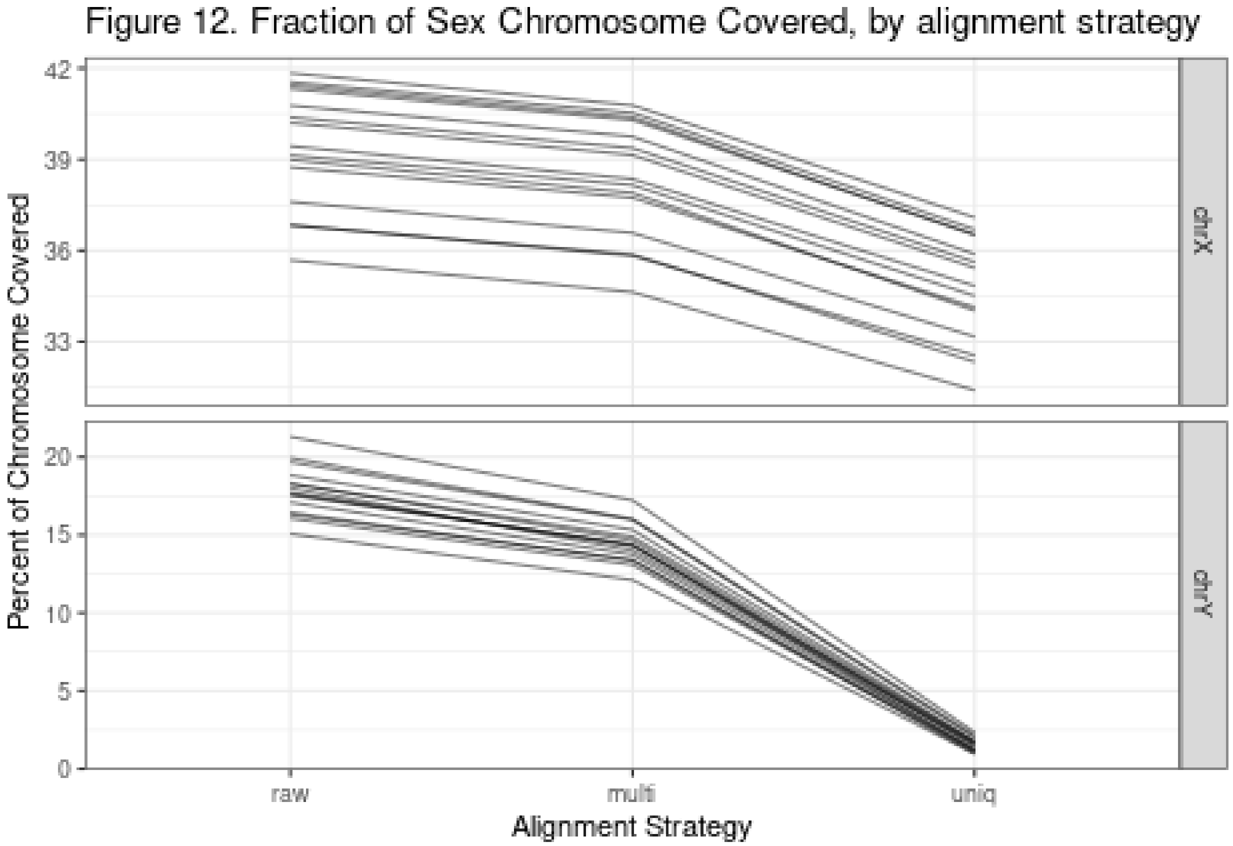


Figure 12. Fraction of Sex Chromosome Covered, by alignment strategy

```
## png
##   2
```

2.6 Assigning Reads to Annotated Features

Mapped reads were assigned and counted using the featureCounts function from the SubRead package. (Liao, Smyth, and Shi 2014). In particular, the reads were assigned to exons in the dm6_genes GTF annotation, and these were counted towards the genes containing the exons. The two ends of paired reads were counted as separate fragments. To be counted, both ends of the paired reads must map, and map to the same chromosome. Any multimapped reads are counted at all of their mapped locations. (Command line options: "-t exon -g gene_id -M -J -p -B -C").

By default, a read overlapping multiple genes is considered ambiguous and not counted. This makes sense when the feature being counted is a gene, but becomes problematic when counting by exon, since:

- reads which span splice junctions necessarily overlap multiple features, and thus aren't counted
- exons which are small compared to read size will have few or no reads unspliced
- some exons are completely contained within other exons, and are precluded from having reads assigned.

Thus, some counts (filenames containing "MpBCO") have reads assigned to all overlapping features, instead of none (filenames containing "MpBC"). featureCounts offers a third option, to assign 1/nth of a read to each of n features it overlaps; however, DESeq2 requires integer counts so this is not appropriate here.

Table 16. Percentage of Reads Assignable to Features in dm6_genes

fraction of the reads which can be unambiguously counted under different alignment strategies

	mapping strategy		
rep	multi	rando	uniq
47b1 - group - 7			
1	176.6%	178.7%	181.2%
2	174.7%	177.1%	179.9%
3	176.2%	177.7%	181.2%
67d - group - 7			
1	178.9%	180.7%	182.4%
2	173.9%	177.7%	179.4%
3	174.9%	177.9%	179.7%
FruLexaFru440 - group - 7			
1	165.5%	169.2%	172.0%
2	176.7%	178.5%	180.5%
3	175.5%	176.1%	180.1%
wt - group - 7			
1	181.3%	181.4%	183.0%
2	177.2%	176.7%	181.2%
3	180.3%	180.7%	182.2%
wt - isolated - 7			
1	181.5%	181.7%	183.2%
2	182.1%	182.3%	183.6%
3	181.7%	181.9%	183.4%

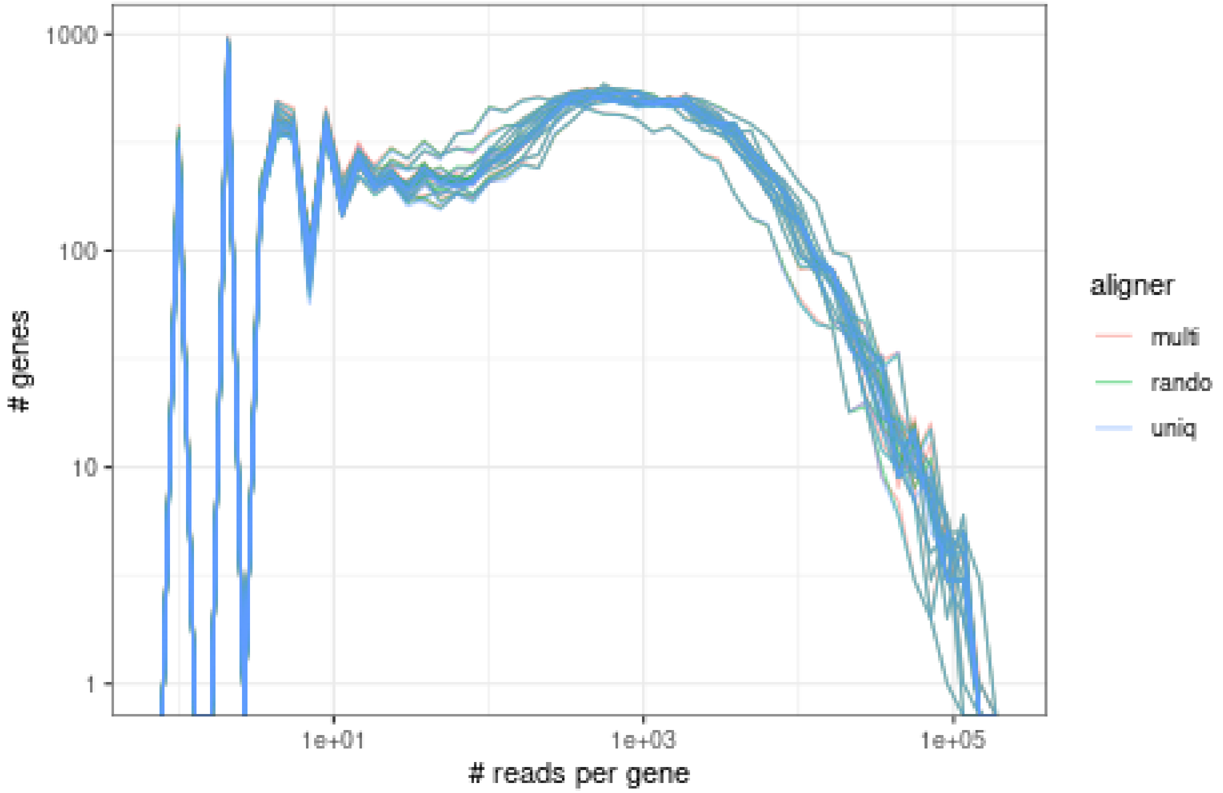
Table 17. Averaged Percentage of Reads Not Assignable to Features in dm6_genes

average fraction of mapped reads which were unassigned

	mapping strategy		
	multi	rando	uniq
Ambiguous	7.9%	7.8%	7.8%
No Overlap	31.9%	11.9%	11.3%

The values for “multi” are inflated because each appearance of a multi-mapped read is counted, whereas the denominator is the actual read count (FIX THIS)

Figure 13. Per-Gene Read Count Histogram (by aligner and sample)



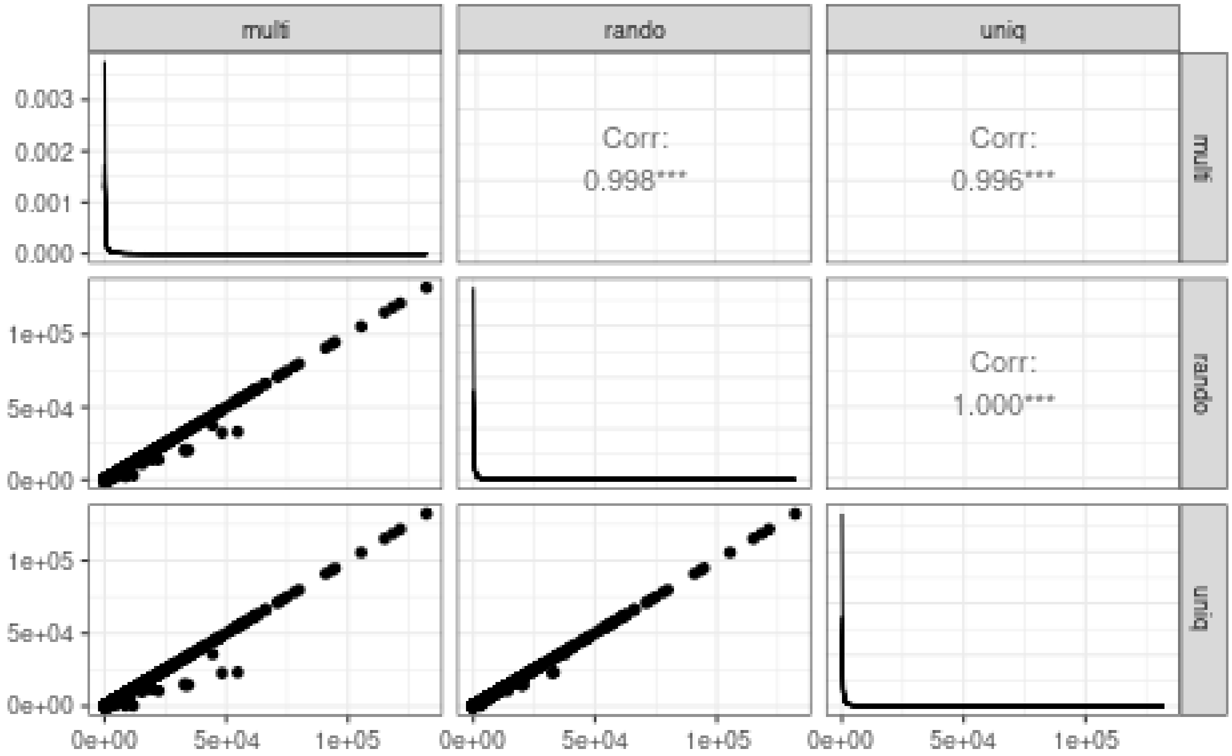
```
## png
##   2
```

One average, a gene had 1160.39477542723 reads assigned to it, but most genes had relatively fewer, with more than a quarter having no reads assigned at all, almost half having fewer than 10 reads, and almost two thirds having fewer than 100.

Table 18. Averaged Percentage of Genes by Threshold Read Counts
average fraction of genes with low number of reads

	read count threshold		
aligner	< 1	< 10	< 100
multi	28.0%	42.3%	55.0%
rando	28.8%	42.6%	55.2%
uniq	29.2%	43.0%	55.4%

Figure 14. Correlations between Read Count Assigned to Gene Across Alignment Strategy (downsampled to 10%)

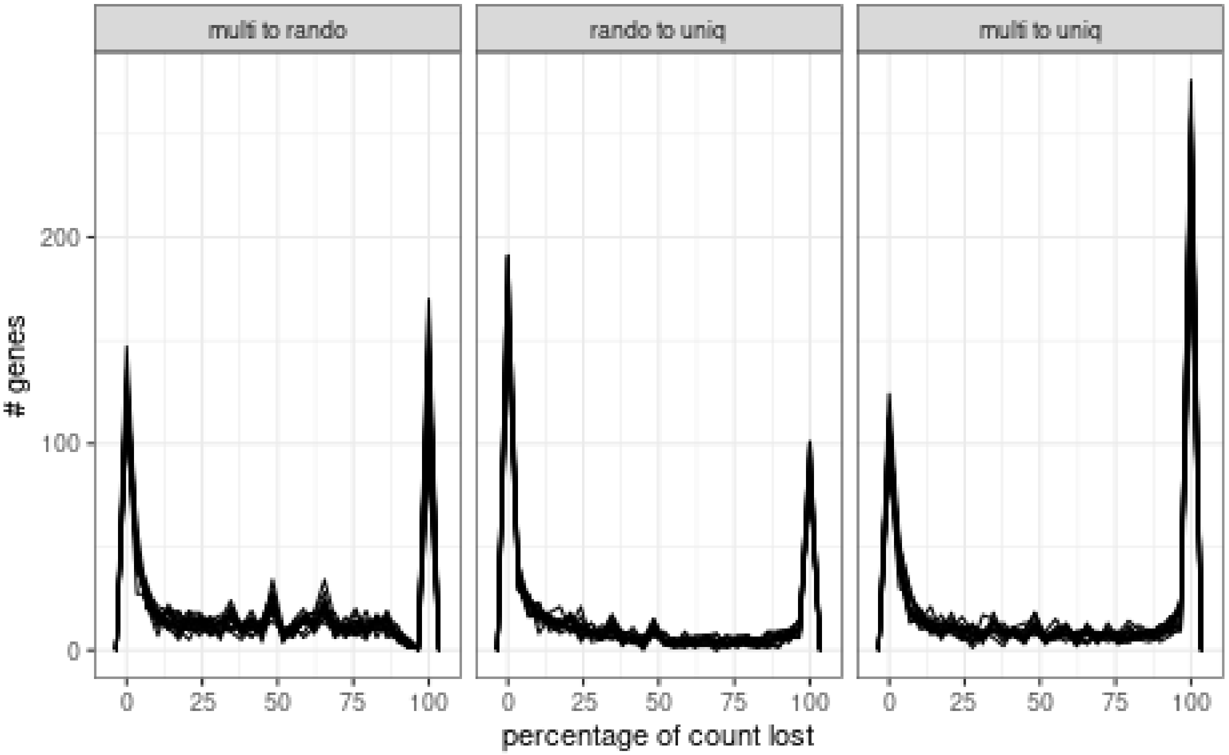


```
## png
##   2
```

The three mapping strategies generally agreed well; for 93.8364779874214 % of genes, the same number of reads were assigned by all three strategies in all samples. (Restricted to genes with at least one nonzero count, the proportion was 93.1525122276567 %)

By construction, the read count assigned to a gene is supposed to decrease across strategy: multi >= rando >= uniq. It's not clear why but for a very small number of cases (0; 0 %), rando > multi.

Figure 15. Percent Loss in Assigned Read Count
Between Mapping Strategies (Discrepancies Only)



```
## png
##   2
```

2.6.1 Fruitless by exon

To study Fru on an exon-by-exon case, the existing GTF annotation was subsetted to isoforms of only this gene, and reformatted such that each exon was an individual feature to be counted. featureCounts was then run as usual on this new annotation. With many genes to study on a per-exon basis, the featureCounts -f flag might be more useful.

(Counts are so small compared to total that percentages aren't informative here)

Table 19. Number of Reads Assignable to Features in fru_exons
number of the reads which can be counted by alignment/assignment strategy

	all			none		
	multi	rando	uniq	multi	rando	uniq
47b1 - group - 7						
1	1567	1567	1567	1148	1148	1148
2	1759	1759	1759	1328	1328	1328
3	1349	1349	1349	1034	1034	1034
67d - group - 7						
1	1287	1287	1287	958	958	958

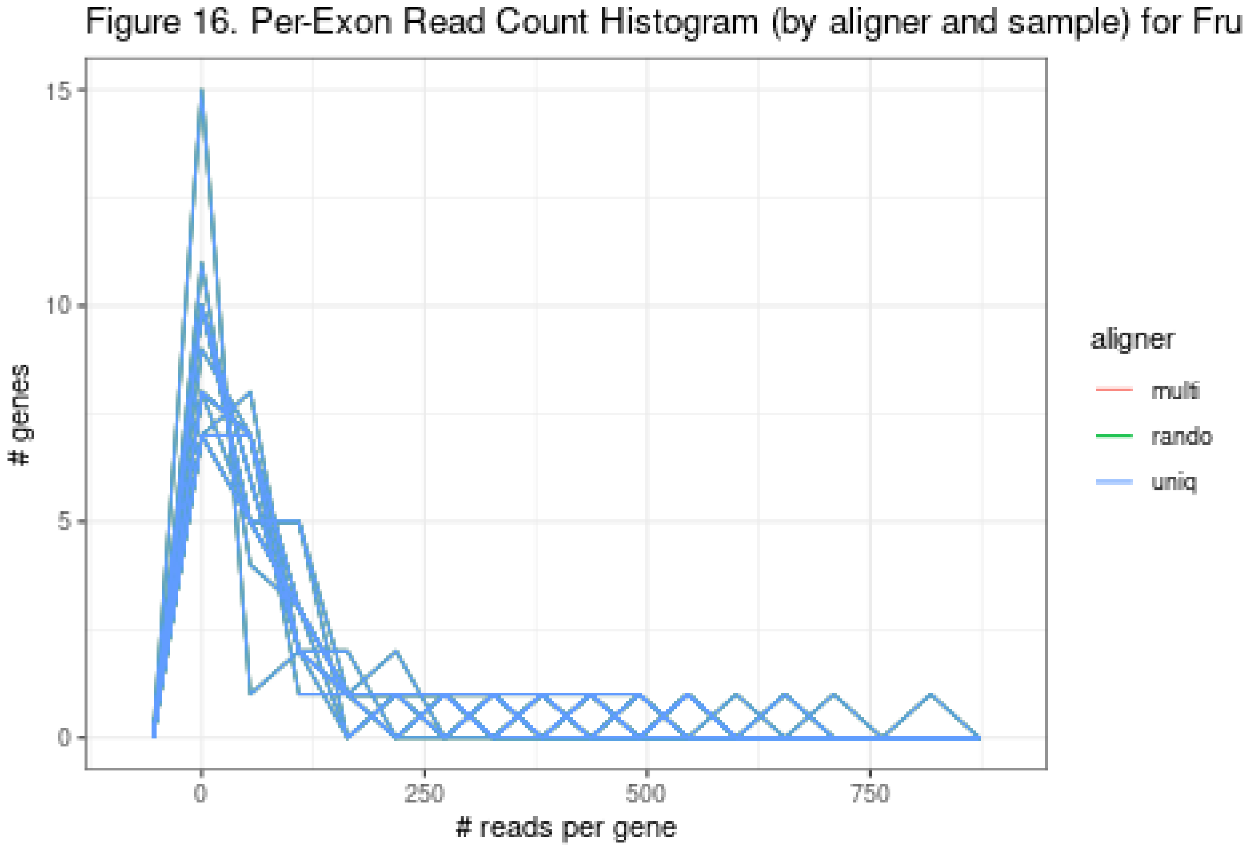
2	1912	1912	1912	1492	1492	1492
3	1481	1481	1481	1123	1123	1123
FruLexaFru440 - group - 7						
1	1914	1914	1914	1613	1613	1613
2	1420	1420	1420	1119	1119	1119
3	1456	1456	1456	1102	1102	1102
wt - group - 7						
1	2227	2227	2227	1594	1594	1594
2	1836	1836	1836	1336	1336	1336
3	1659	1659	1659	1244	1244	1244
wt - isolated - 7						
1	1625	1625	1625	1162	1162	1162
2	1919	1919	1919	1409	1409	1409
3	2161	2161	2161	1608	1608	1608

One average, a exon had 98.4272727272727 reads assigned to it, but most exons had relatively fewer, with almost a quarter having no reads assigned at all, more than a third having fewer than 10 reads, and almost two thirds having fewer than 50. These figures are for the “All” assignment strategy, and are necessarily lower for the “None”.

Table 20. Averaged Percentage of Exons by Threshold Read Counts (Fruitless)
average fraction of genes with low number of reads

	read count threshold		
aligner	< 1	< 10	< 50
multi	22.7%	33.9%	54.2%
rando	22.7%	33.9%	54.2%
uniq	22.7%	33.9%	54.2%

There is VERY little difference between alignment strategies when it comes to read count here. total overplot in count histogram; no point in showing other comparisons…



```
## png
##   2
```

2.6.2 Fruitless by splice junction

The fru_junct annotation is only counted under the “All” assignment strategy, since the reads being counted are spliced and thus necessarily overlap multiple exons. As well, the “SpliceOnly” version of each alignment will be used (ie, only spliced reads and only the 1bp subintervals which correspond to splice junctions)

Table 21. Number of Reads Assignable to Features in fru_junct
number of the reads which can be counted by alignment/assignment strategy

	all		
	multi	rando	uniq
47b1 - group - 7			
1	127	127	127
2	117	117	117
3	91	91	91
67d - group - 7			
1	96	96	96
2	110	110	110
3	74	74	74

FruLexaFru440 - group - 7			
1	29	29	29
2	90	90	90
3	80	80	80
wt - group - 7			
1	158	158	158
2	157	157	157
3	117	117	117
wt - isolated - 7			
1	125	125	125
2	151	151	151
3	158	158	158

2.6.3 Fruitless by intron

The fru_intron annotation is counted under the “All” assignment strategy, using the “SpliceOnly” alignments. Because the same reads are being counted against the same intervals, the number of reads countable are identical to those in fru_junct

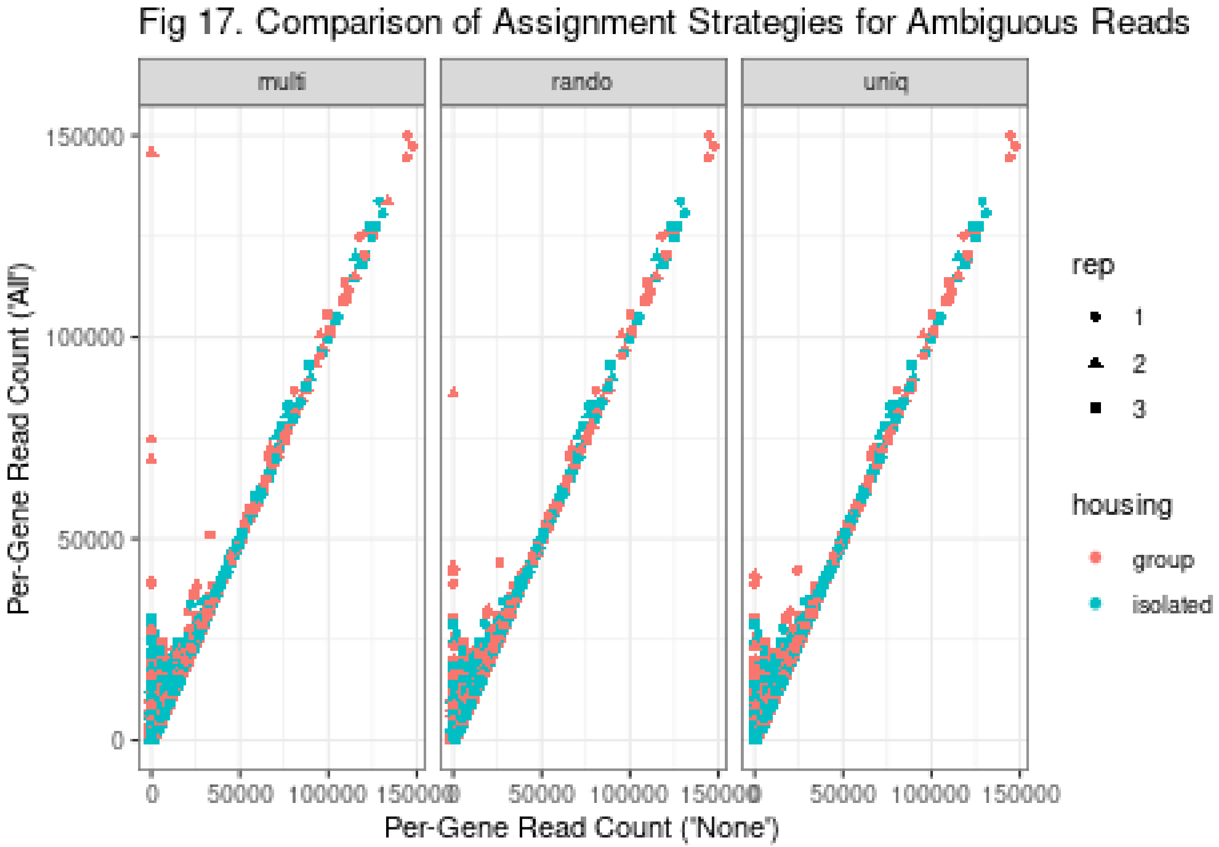
Table 22. Number of Reads Assignable to Features in fru_intron
number of the reads which can be counted by alignment/assignment strategy

	all		
	multi	rando	uniq
47b1 - group - 7			
1	127	127	127
2	117	117	117
3	91	91	91
67d - group - 7			
1	96	96	96
2	110	110	110
3	74	74	74
FruLexaFru440 - group - 7			
1	29	29	29
2	90	90	90
3	80	80	80
wt - group - 7			
1	158	158	158
2	157	157	157
3	117	117	117
wt - isolated - 7			
1	125	125	125
2	151	151	151
3	158	158	158

2.6.4 Ambiguous Assignment Strategy Comparison

The whole gene annotation and the Fruitless exons are currently having readcounts assigned with slightly different strategies. When all genes are considered, ambiguously assigned reads (those which overlap multiple features) are simply discarded; we will call this the "None" strategy. When the exons of Fru are considered, ambiguously assigned reads count towards the tally of every exon they overylap we'll call this "all").

There is a big difference between these strategies at the exon level; how well do they agree on the enitre dm6_genes annotation?



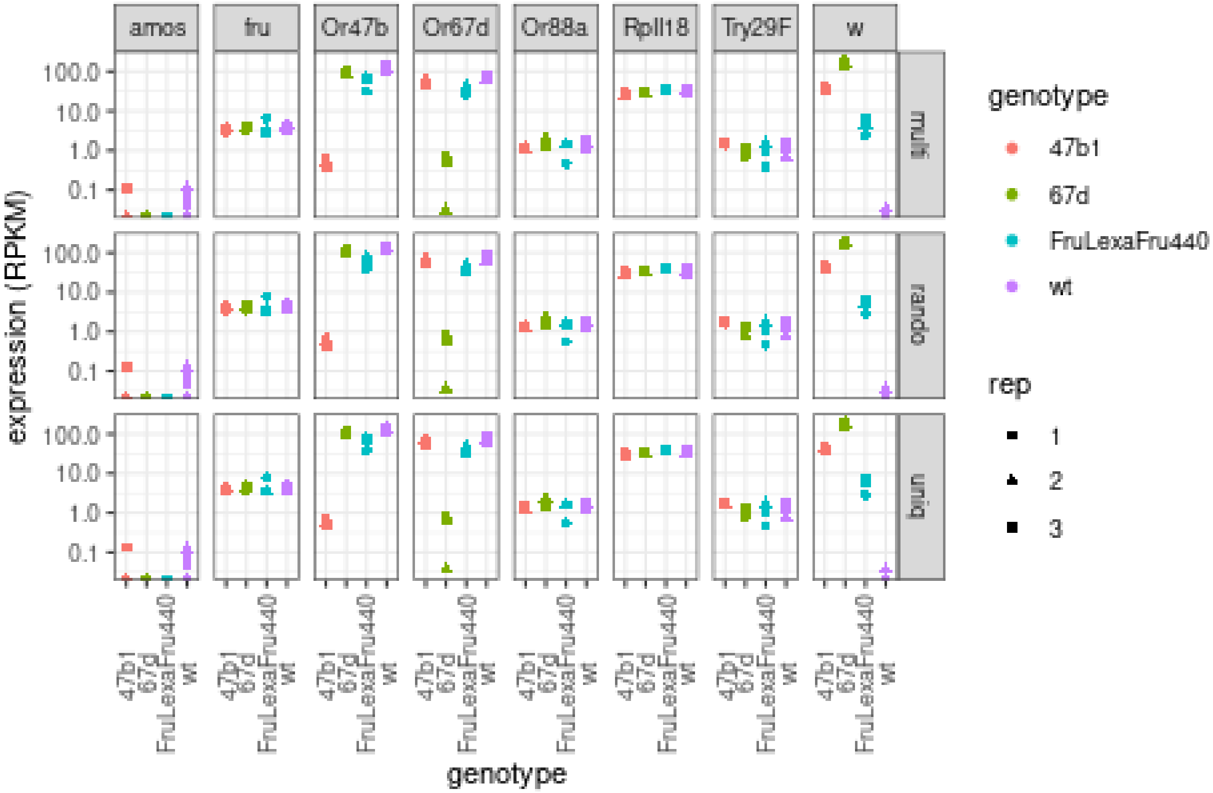
```
## png
##   2
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0      0.0      0.0    118.4      4.0 145767.0
```

2.7 Expression

Using the per-gene read counts, the per-alignment total mapped counts, and the gene lengths, the gene expression was calculated as reads per kilobase per million mapped (RPKM). In particular, these can confirm that the knock-outs are not being expressed. This appears to be the case in the 47b and 67d mutants. The Fru440FruLexa mutants do not show any obvious reduction in expression of Fruitless (not knockouts - is expression expected though?) CantonS appears similar to wt in all cases. CatnonS-Amos mutants have very low expression of OR47b, OR67d, and OR88a but have typical expression values for fru and control genes.For context, a positive control (RNA polymerase) and a negative control (trypsin) have also been included.

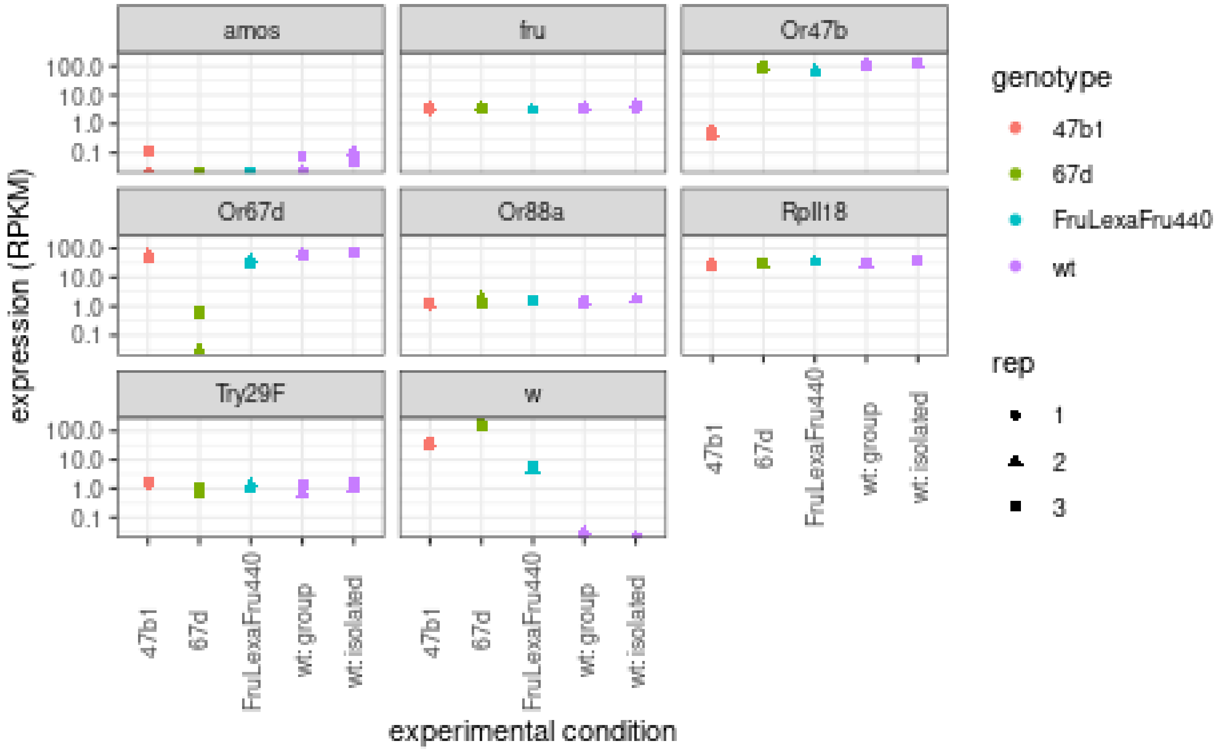
Fig 18. Expression Levels for Key Genes, by Genotype



```
## png
##   2
```

Fig 18a. Expression Levels for Key Genes, by Genotype

Subset of Key Genes, FruLexaFru440 replicate 1 excluded



```
## png
##   2
```

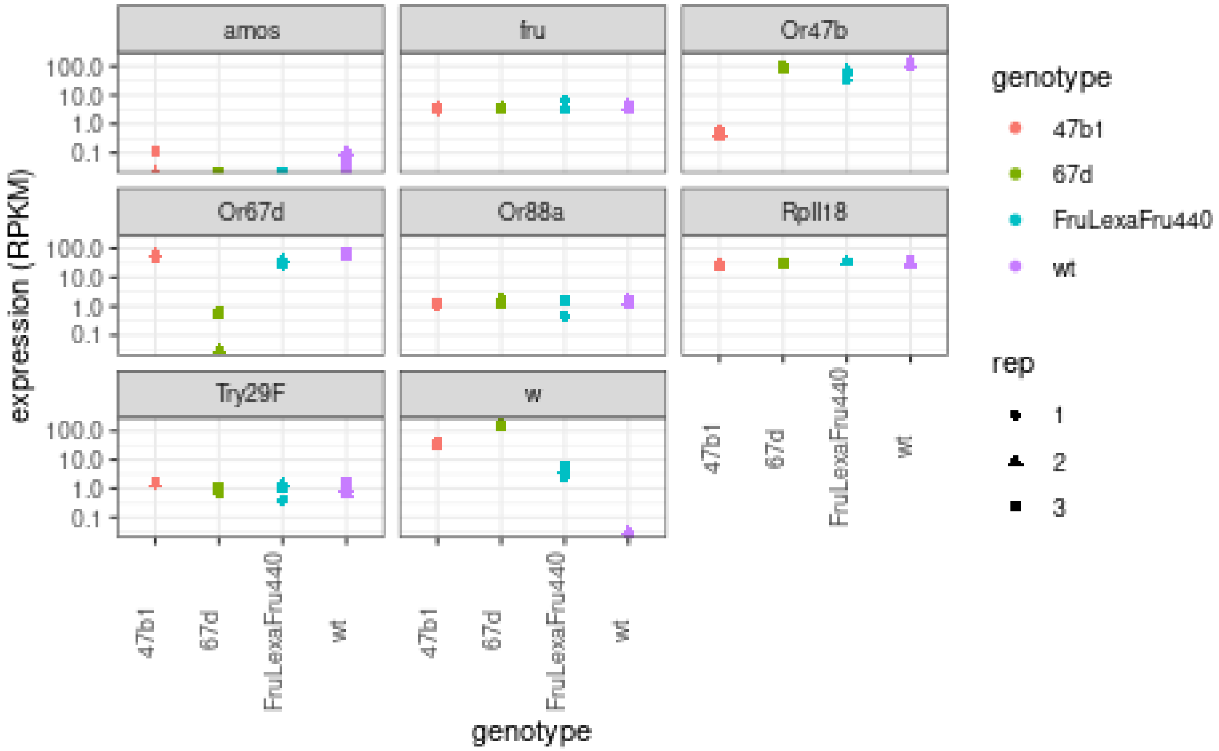
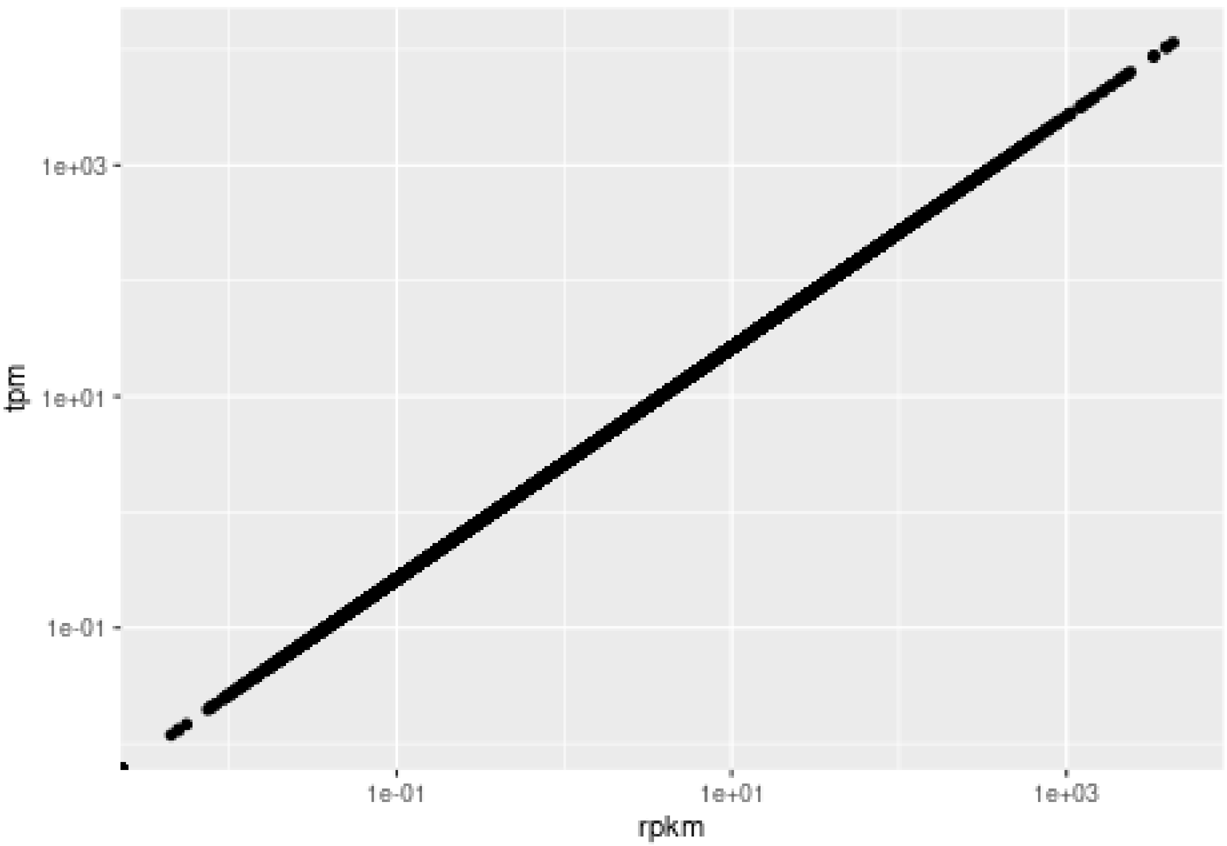
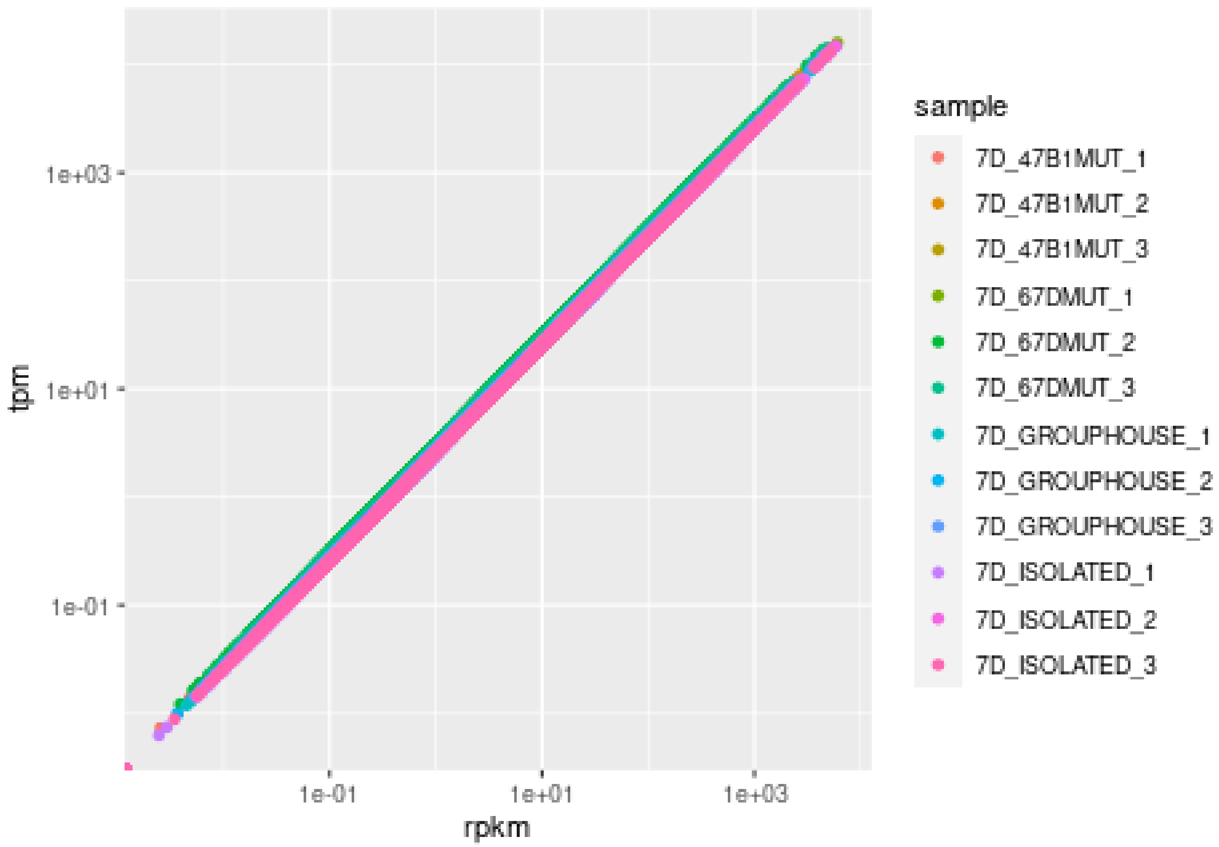


Fig 18b. Expression Levels for Key Genes, by Genotype

simplified - multi only

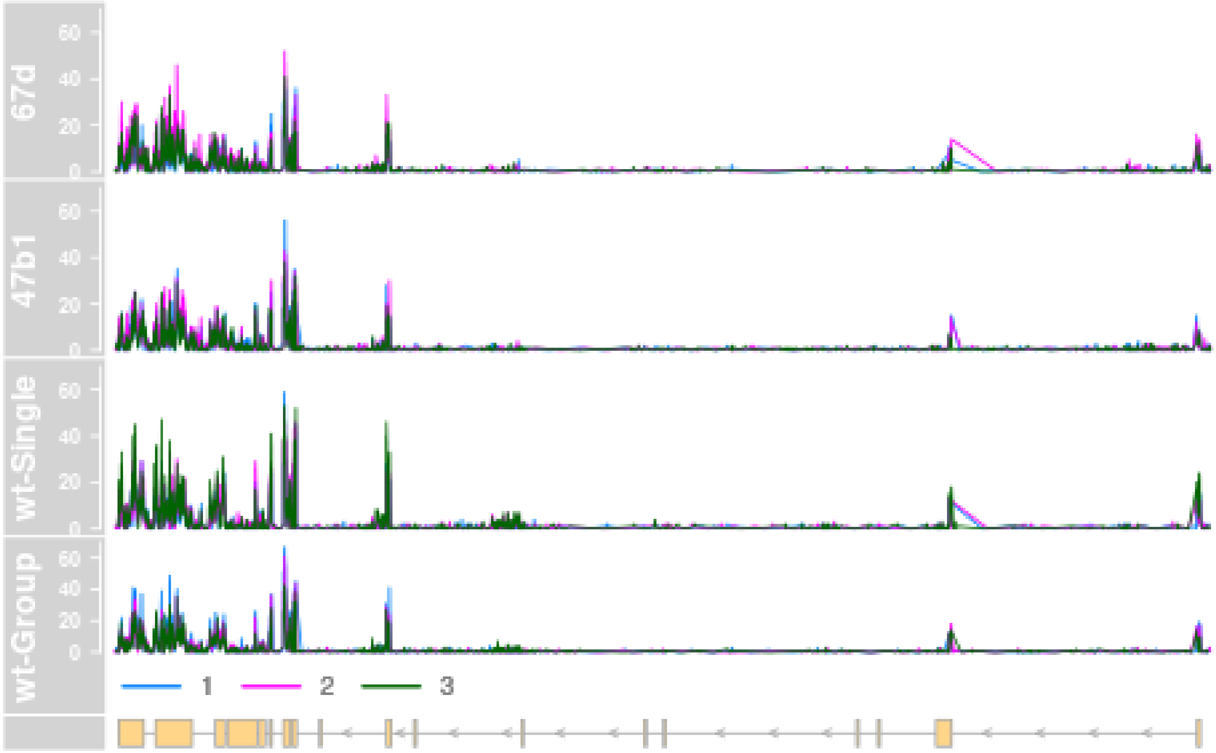
```
## png
##   2
```


1e+03
1e+01
1e-01
tpm
1e-01
1e+01
1e+03
rpkm


sample
7D_47B1MUT_1
7D_47B1MUT_2
7D_47B1MUT_3
7D_67DMUT_1
7D_67DMUT_2
7D_67DMUT_3
7D_GROUPHOUSE_1
7D_GROUPHOUSE_2
7D_GROUPHOUSE_3
7D_ISOLATED_1
7D_ISOLATED_2
7D_ISOLATED_3
tpm
1e+03
1e+01
1e-01
1e-01
1e+01
1e+03
rpkm

2.7.1 Focus on Fruitless

Figure 19. Focus on Fruitless:
Gene-Scale Depth of Coverage (by treatment and replicate)



```
## png
##   2
```

2.8 Differential Expression Analysis.

DESeq2 (Love, Huber, and Anders 2014) was used to detect changes in expression from read-count data, following the official vignette as a guide (http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html ; see also http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html).

DESeq2 builds a statistical model in which the read counts are normalized and then fit to explanatory variables ("factors"). Each value a factor may take on is called a "level". For example, genotype is a factor, whereas the 47b mutation is a level of the genotype factor. The model fit to the counts may contain one or more factors.

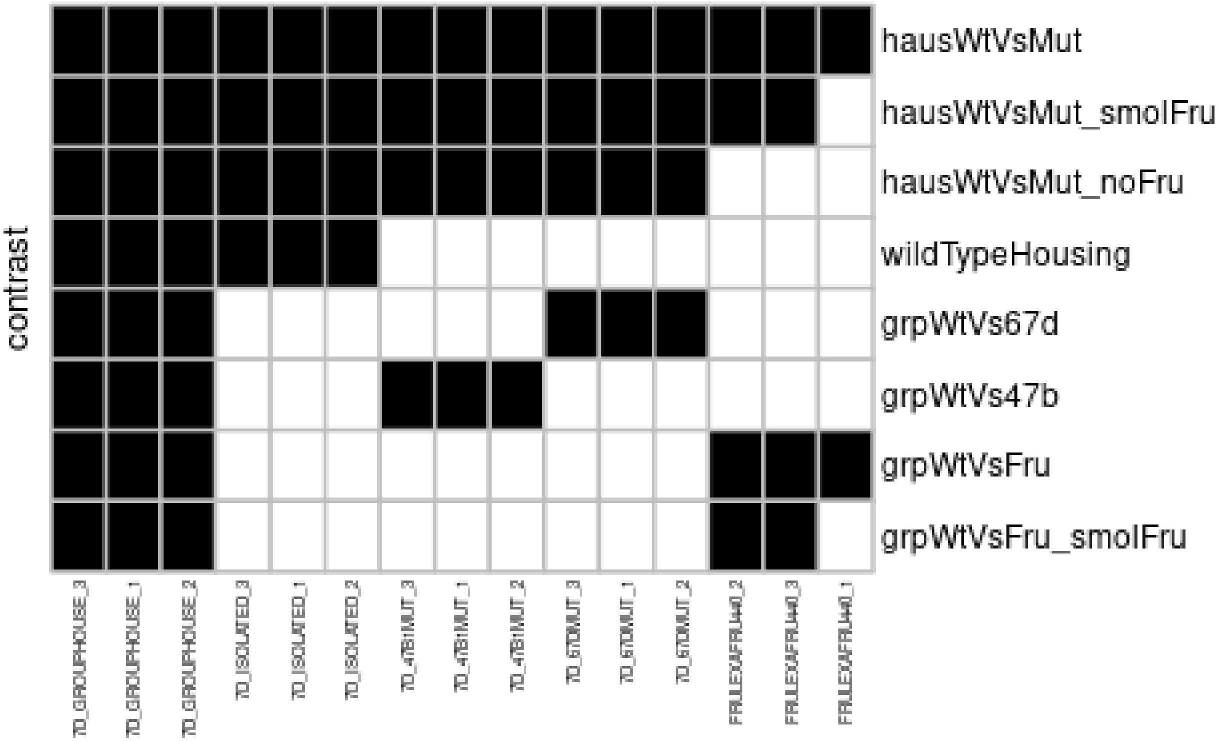
Single-factor models (wildTypeHousing, grpWtVs47b, grpWtVs67d, grpWtVsFru, grpWtVsMut) were built by specifying the axis of comparison (eg, housing) and subsetting samples to the relevant contrast (eg, wt group reps 1,2,3 and wt isolated reps 1,2,3).

Current results mostly come from a two-factor model in which both housing and genotype are considered simultaneously (hausWtVsMut).

Table 23. Differential Expression Contrasts
with model and reference levels

	fit model	reference levels
grpWtVs47b	~ genotype	genotype: wt
grpWtVs67d	~ genotype	genotype: wt
grpWtVsFru	~ genotype	genotype: wt
grpWtVsFru_smolFru	~ genotype	genotype: wt
hausWtVsMut	~ genotype + housing	genotype: wt, housing: group
hausWtVsMut_noFru	~ genotype + housing	genotype: wt, housing: group
hausWtVsMut_smolFru	~ genotype + housing	genotype: wt, housing: group
wildTypeHousing	~ housing	housing: group

Figure 20. RNASeq Samples Used in DESeq2 Contrast



contrast	7D_GROUPHOUSE_3	7D_GROUPHOUSE_1	7D_GROUPHOUSE_2	7D_ISOLATED_3	7D_ISOLATED_1	7D_ISOLATED_2	7D_47B1MUT_3	7D_47B1MUT_1	7D_47B1MUT_2	7D_67DMUT_3	7D_67DMUT_1	7D_67DMUT_2	FRULEXAFRU440_2	FRULEXAFRU440_3	FRULEXAFRU440_1
hausWtVsMut	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hausWtVsMut_smolFru	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□
hausWtVsMut_noFru	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□
wildTypeHousing	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□
grpWtVs67d	■	■	■	□	□	□	□	□	□	■	■	■	□	□	□
grpWtVs47b	■	■	■	□	□	□	■	■	■	□	□	□	□	□	□
grpWtVsFru	■	■	■	□	□	□	□	□	□	□	□	□	■	■	■
grpWtVsFru_smolFru	■	■	■	□	□	□	□	□	□	□	□	□	■	■	□

```
## png
##   2
```

For each factor and level, DESeq2 returns two key pieces of information: an effect size and an adjusted p-value.

The effect size is reported as the base-2 logarithm of fold-change in expression between the reference level and some alternate level. Thus, if the 47b contrast for some gene G has a log2FoldChange of 1, it means that the 47b mutants express G at 2^1 = 2 times

as much as the wildtype flies. A log2FoldChange of -1 means that the 47b mutants express G at 2^-1 = 0.5 times as much as the wildtype flies. No change at all would be a foldchange of 1, and a log2 fold change of 0.

The p-value gives the odds that an effect size as large would be observed if there were no change in expression, just random noise. Since a p-value is estimated for each gene in the annotation, a correction for multiple comparisons (Benjamini-Hochberg) is applied.

DESeq2 reports the normalized mean counts for each level; an expression level was derived from it by scaling by feature length. (More on interpretation & use of the "baseMean": https://support.bioconductor.org/p/75244; https://support.bioconductor.org/p/63567/; https://www.biostars.org/p/219093/; https://www.biostars.org/p/248486/)

Counts filtered to remove genes with less than 10 reads combined across all samples. Effect-size shrinkage is currently done using apeglm; other shrinkage estimators have not yet been explored.

2.8.1 Differential Exon Use

For a number of reasons, estimating changes in transcript or exon usage is more challenging than estimating coarse-scaled gene expression (eg, https://www.biostars.org/p/424242/#424343). Some of approaches here have been to modify the reads/annotations and then analyze within a standard DESeq2 framework. Two approaches outside this schema were also explored.

There is an important distinction between exon USE, which reflects the proportion of expressed transcripts containing a given exon, and exon EXPRESSION, which is the total amount of exon RNA being transcribed (a function of both exon use and gene expression.) (Anders, Reyes, and Huber 2012) In the DESeq2-based approaches, the counts have been subset to the Fruitless gene, on the effects of gene expression change will be absorbed into the calculation of sizeFactors, thus any residual changes left represent changes in exon use.

2.8.1.1 The edgeHog

The Fru annotation was divided up into intervals corresponding to exons. The intervals were then subdivided further at exon boundaries, eg if two exons A and B share a 5' edge but B is longer, the interval would be divided into two adjacent intervals, AB and B. Each interval then corresponds to an element of the power set of transcripts, that is, the interval AB corresponds to the set of transcripts which include the AB interval. An interval only has one set of transcript associated, but a set of transcripts can have multiple intervals associated. For example, the set of all transcripts would be associated with the intervals of all constitutive exons.

The intervals were partitioned according to the set of transcripts associated. A new annotation of "isoids" is generated by stitching together the intervals in each partition, and the reads counted. The isoid counts for Fru were provided to DESeq; presumably differences in gene-level expression will be rolled into the size factor estimation and leave behind differences coming from exon use.

For each transcript, the p-value was collected from each isoid associated with the transcript and converted to Z-scores. Stouffer's test was applied to estimate a significance value for an overall change in relative transcript use.

2.8.1.2 DEXSeq

To compare these results to an established tool, DEXSeq (Anders, Reyes, and Huber 2012) was used. DEXSeq repurposes the DESeq2 statistical methods but modifies the underlying annotation (in a way that doesn't necessarily respect the exon naming/grouping I've used) and counts them in a somewhat idiosyncratic way:

```
The central data structure for our method is a table that, in the simplest case,
contains for each exon of each gene the number of reads in each sample
that overlap with the exon. Special attention is needed, however, if an exon's
boundary is not the same in all transcripts. In such cases, we cut the exon
in two or more parts (Fig. 1). We use the term "counting bin" to refer to
```

```
exons or parts of exons derived in this manner. Note that a read that overlaps
with several counting bins of the same gene is counted for each of these.
```

(Anders, Reyes, and Huber 2012)

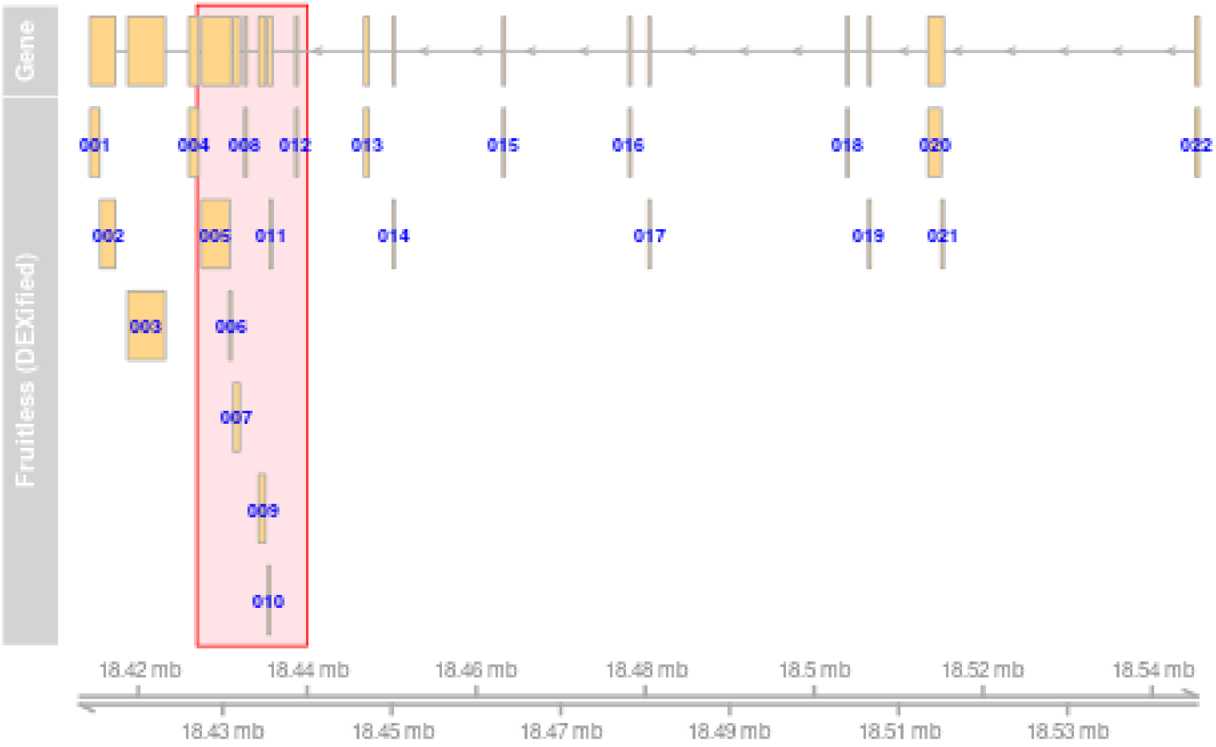
The suggested featureCounts implementation was used: https://github.com/vivekbhr/Subread_to_DEXSeq

In order for the dexseq_prepare_annotation.py script to run correctly, it had to be modified in order to account for transcriptional edge cases (eg, the polycistronic pre-mod(mdg4)-*)

```
for f in HTSeq.GFF_Reader( gtf_file ):
   if f.type != "exon":
      continue
   f.attr['gene_id'] = f.iv.chrom + '_' + f.attr['gene_id'].replace( ":", "_" ) + f.iv.strand # THIS WO
   #f.attr['gene_id'] = f.attr['gene_id'].replace( ":", "_" ) # THIS DOESN'T
   exons[f.iv] += ( f.attr['gene_id'], f.attr['transcript_id'] )
```

(source: https://stat.ethz.ch/pipermail/bioconductor/2012-June/046494.html)

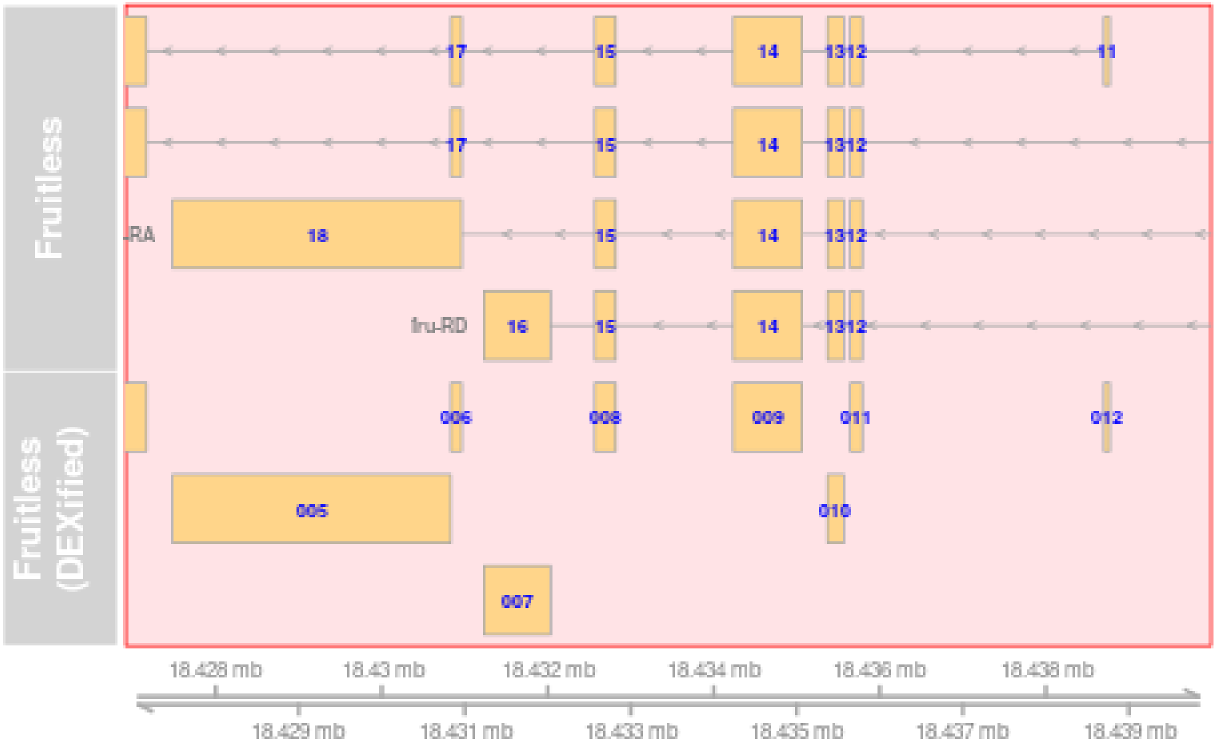
Figure 21. Fruitless gene model: DEXSeq Intervals Derived From Exons



```
## png
##   2
```

Take a careful look at the relationship between exon_18 and intervals 005 and 006.

Figure 21 a. Fruitless gene model: DEXSeq Intervals Derived From Exons (detail)



```
## png
##   2
```

In order to estimate differential exon usage, DEXSeq fits the (size & dispersion-corrected) counts with two models, one containing an interaction term and then other not, and compares the two:

```
Having the dispersion estimates and the size factors, we can now test for differential exon usage.
For each gene, DEXSeq fits a generalized linear model with the formula
~sample + exon + condition:exon
and compare it to the smaller model (the null model)
~ sample + exon
```

(Official vignette: https://bioconductor.org/packages/devel/bioc/vignettes/DEXSeq/inst/doc/DEXSeq.html)

Models using more than one explanatory variable are possible but more involved & haven't been explored yet

Table 24. Differential Exon Use Contrasts
all modeled as '~ sample + exon + condition:exon'

	sample subset	reference levels
dex_grpWtVs47b	grpWtVs47b	genotype: wt
dex_grpWtVs67d	grpWtVs67d	genotype: wt

dex_grpWtVsFru	grpWtVsFru	genotype: wt
dex_grpWtVsFru_smolFru	grpWtVsFru_smolFru	genotype: wt
dex_wtHousing	wtHousing	housing: group

2.9 Gene Ontology Enrichment

Gene Ontology Enrichment was studied using topGO. https://bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf

For each set of DESeq data studied, the genes and their expression differences were subsetted by factor and level. Two tests were used: Fisher's Exact, which uses counts from a discrete subset of genes (here, those with adjusted $p < 0.01$), and Kolmogorov-Smirnov, which uses the p-values as a quantitative score. The "classic" algorithm was used, and the top 50 nodes were collected and saved for each GO type: Molecular Function, Biological Component, Cellular Process.

topGO appears to still be plagued by an intermittent error, "There are no adj nodes for node: GO:xxxxxxx Error in switch", for which there is not yet a clear solution or explanation. (eg, https://support.bioconductor.org/p/116048/; https://support.bioconductor.org/p/103640/; https://www.biostars.org/p/311104/)

From experience, I can prevent it by masking ~30 genes. Some of these are significantly differentially expressed, however!

```
    flybase_gene_id external_gene_name
1       FBgn0261268               Cul3
33      FBgn0032470             CG5142
45      FBgn0031450                Hrs
84      FBgn0051999            CG31999
88      FBgn0011828                Pxn
108     FBgn0014388                sty
142     FBgn0038358             CG4525
152     FBgn0016075                vkg
164     FBgn0050046            CG30046
169     FBgn0028573                prc
180     FBgn0033710            CG17739
184     FBgn0261800              LanB1
204     FBgn0005695                gcl
229     FBgn0020269               mspo
233     FBgn0039257                tnc
242     FBgn0262733             Src64B
296     FBgn0263930              dally
343     FBgn0040206                krz
363     FBgn0026562              SPARC
378     FBgn0041604                dlp
411     FBgn0004907         14-3-3zeta
442     FBgn0032252                loh
448     FBgn0035049               Mmp1
477     FBgn0050203            CG30203
482     FBgn0026721        fat-spondin
487     FBgn0003969                vap
505     FBgn0004390            RasGAP1
531     FBgn0031850                Tsp
```

Additionally, BioMart does not appear to have descriptions listed for some GO IDs; these currently need to be looked up on a case-by-case basis at http://geneontology.org/

Multiple comparison adjustment isn't done (see topGO vignette section 6.2)

Currently applied to the simultaneous model only.

2.10 Variant Calling & Genetic Distance

> The authors indicate there are "modest differences in genetic backgrounds", without providing data to support this statement. This is not addressed in the methods that I could find, though the authors indicate that more description is in the methods. Relatedness is something that can be calculated, so the authors should use scientific language that is supported by the data. For example, for measurements of relatedness see: PMID: 24714809. I am not suggesting the authors do these calculations, but I am pointing out that they have no basis to say there are modest differences, without presenting data to support the idea. It appears that the strains used are all different laboratory stocks that were not outcrossed into a common background. Why do the authors indicate there are modest genetic differences?

- Reviewer #1

The PMID in question is: Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines Huang et al. (2014)

18 DGRP lines were picked at random and DNA sequences from Huang et al. (2014) were downloaded and mapped to the reference genome. These mapspliceUniq alignments, and those from this study, were used to jointly call variants in VCF format via Freebayes (Garrison and Marth 2012) using standard filters. To improve time economy, bedtools (Quinlan and Hall 2010) was used to restrict variant calling to those regions which have nonzero coverage in all samples. vcftools (Danecek et al. 2011) was used to filter the called variants, retaining only biallelic SNPs with no missing calls. The problematic FruM replicate was excluded from variant calling.

ALSO RESTRICT TO AUTOSOMES!!

```
--remove-indv FRULEXAFRU440_1 --min-alleles 2 --max-alleles 2  --max-missing-count 0
```

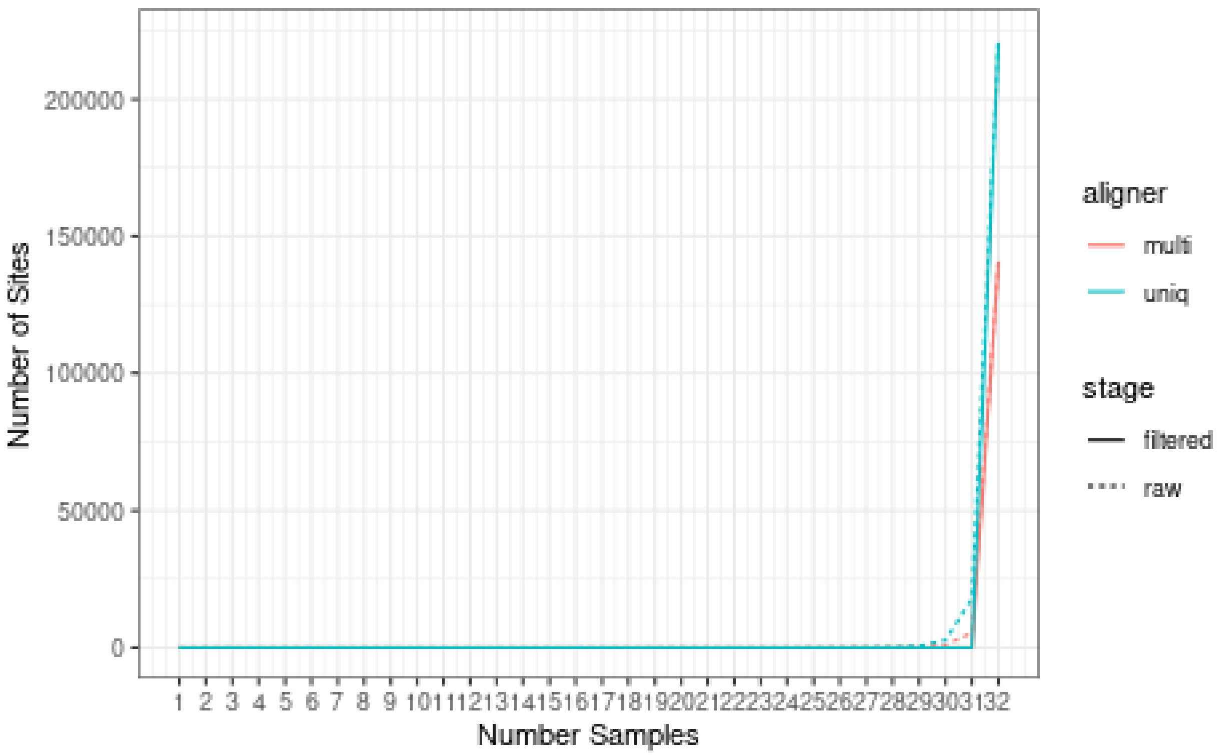
```
## Rows: 1 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (4): X1, X2, X3, X4
## dbl (1): X5
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Table 25. SNP count and per-KB SNP rate across all samples

stage	refGenome	# SNPs	SNP rate (per kb)
mapspliceUniq filtered	dm6main	$236.7K$	1.72

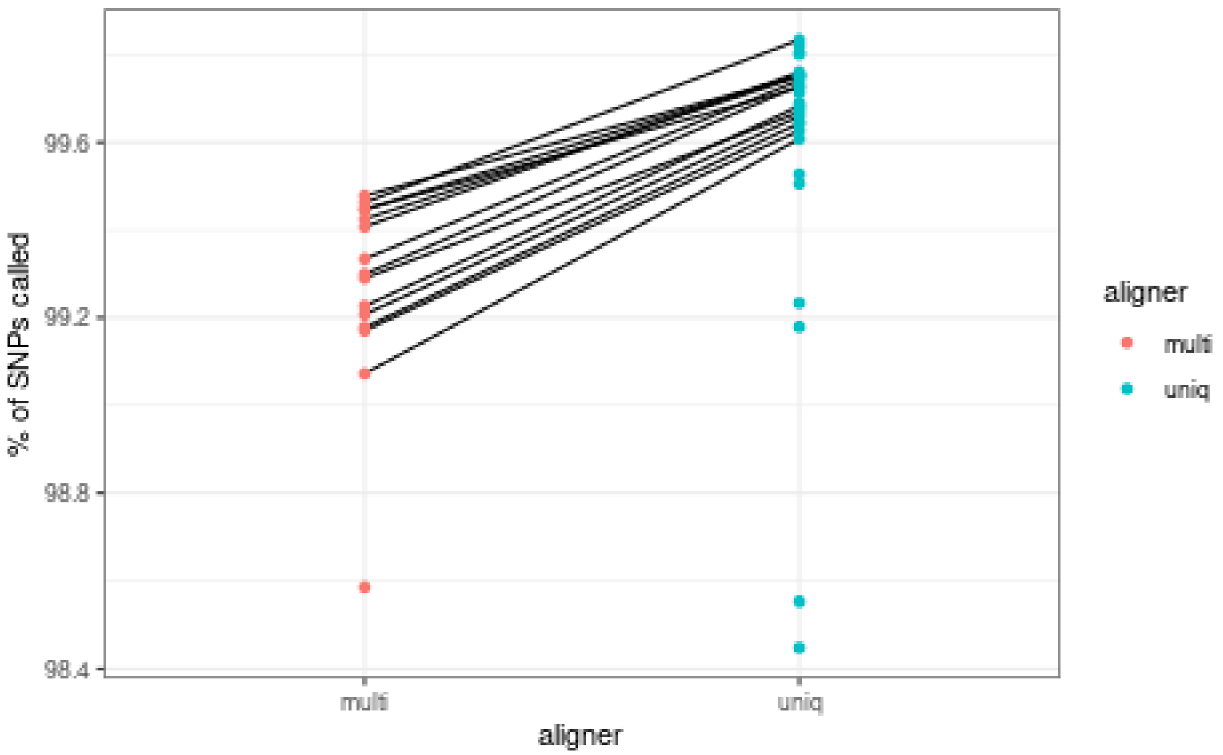
To build this VCF, 32 samples called jointly. However, not all sites were called in all samples (eg, due to coverage differences). The sites had the following group-wide call rate:

Figure 22. Histogram of SNPs by Number of Samples Called At Site



The fraction of jointly called SNPs which are individually callable:

Figure 23. Jointly Called SNPs Callable per Sample



For each sample in a given VCF, a pseudogene was created by extracting and concatenating the variable alleles. Each site contains two alleles, one for each strand, and these were simply concatenated, ie each site contributes 2 nucleotides. This was done using bcftools(Danecek et al. 2021):

```
bcftools query -s {samp} -f '[%TGT]' {input.vcf_in} | tr -d "/"
```

These pseudogenes were combined in a single fasta file and treated as a multiple sequence alignment. Genetic distance was then computed for the MSA using clustalW2 (Larkin et al. 2007):

```
clustalw2 -infile={msa_fa} -tree -outputtree=phylip -clustering=Neighbour-joining
```

The resulting tree was visualized with treeio (Wang et al. 2020) and ggtree (Yu 2020)

3 Results

Earlier results were based upon the 1-factor models; these results are largely hidden in the */supp/ folders

3.1 Wildtype: Group-housed vs. Isolated

In the first contrast, wildtype flies with group-housed and isolated life histories are compared (experimental design: ~ housing). Group-housing was used as a reference level; fold changes are reported relative to it.

After filtering to remove genes with too few reads for analysis, about 12.6k of 17.7k annotated genes (71.0674105 %) remain available for testing:

3.1.1 preshrunk comparison across alignment strategies

The differential expression data were examined before shrinkage. The most discrepancy appeared between the mapspliceUniq alignement and the two which included multimappers, and in genes with small effect sizes.

```
## png
##   2
```

3.1.2 effect size: preshrunk vs shrunk

The shrinkage step attempts to correct for the large apparent effect sizes in genes with small read counts. As expected, the shrinkage narrows the distribution around zero.

```
## png
##   2
```

3.1.3 shrunk comparison across alignment strategies

The shrunk effect sizes agree well between alignment strategies; the "cloud" around unshrunk data at low effect size has disappeared.

```
## png
##   2
```

??what's up with the outliers??

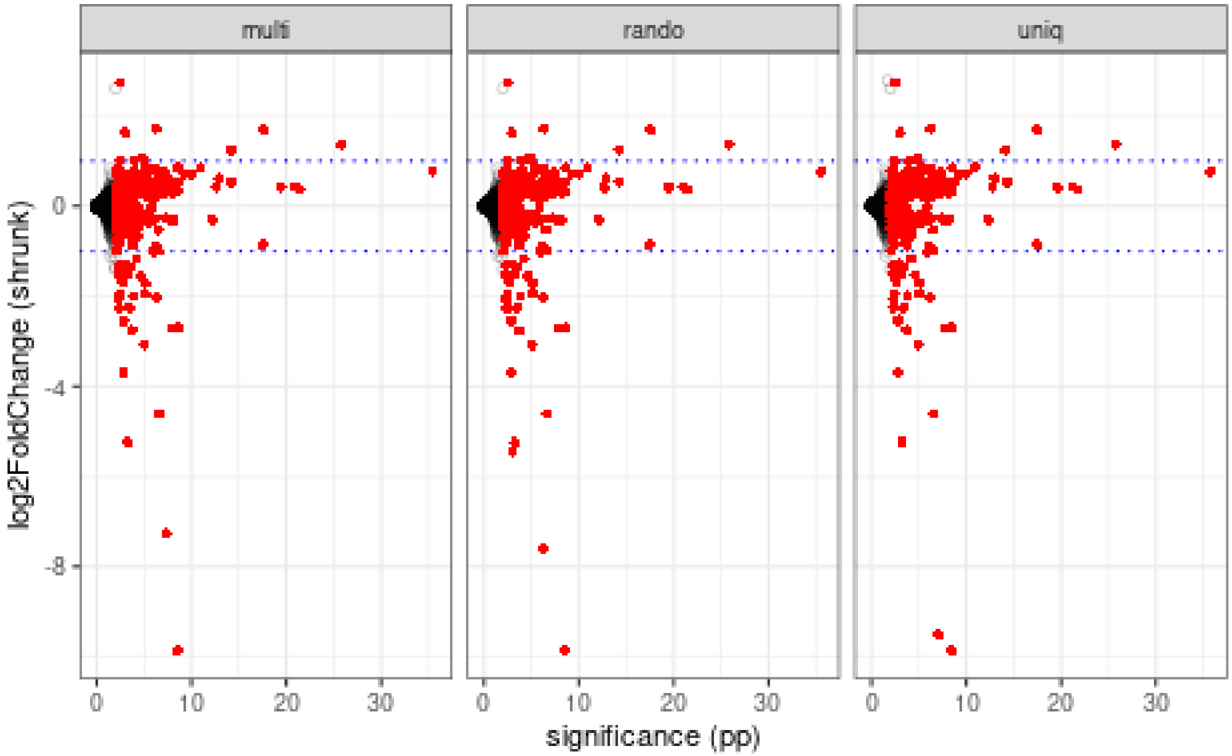
The alignment strategies also agree well when it comes to significance (shrinkage doesn't impact significance so this is the same before and after.)

```
## png
##   2
```

3.1.4 differential expression overview

Here is a volcano plot for the three alignment strategies, with significance on the horizontal axis and log2 fold change on the vertical. Significant (padj<0.01) differences are highlighted in red. Dashed blue guidelines mark a log2 fold change of +/-1 (ie, a difference in expression of a factor of 2). Genes with negative log2 fold changes are depleted relative to the group-housed condition; positive fold changes are enriched

Figure 28. Volcano Plot: Fold Change vs. Significance
(between isolated and group-housed wildtypes)



```
## png
##   2
```

From the volcano plots, we can pull out genes with large (ie, a fold change greater than 2 or less than 1/2), significant (ie, padj < 0.01) changes. There were 39 such genes, mostly shared across alignment strategy:

Table 27. Genes with Large (2< fold change), Significant (padj < 0.01) Changes
between isolated and group-housed wildtypes

	multi	rando	uniq
MtnB	yes	yes	yes
CG11852	yes	yes	yes
TotC	yes	yes	yes
Amy-p	yes	yes	yes
amd	yes	yes	yes
CG15144	yes	yes	yes
Prat2	yes	yes	yes
CG7470	yes	yes	yes
CG10799	yes	yes	yes
Amy-d	no	yes	no
CG42369	yes	yes	yes
CG2736	yes	yes	yes
CG15822	yes	yes	yes
LUBEL	yes	yes	yes

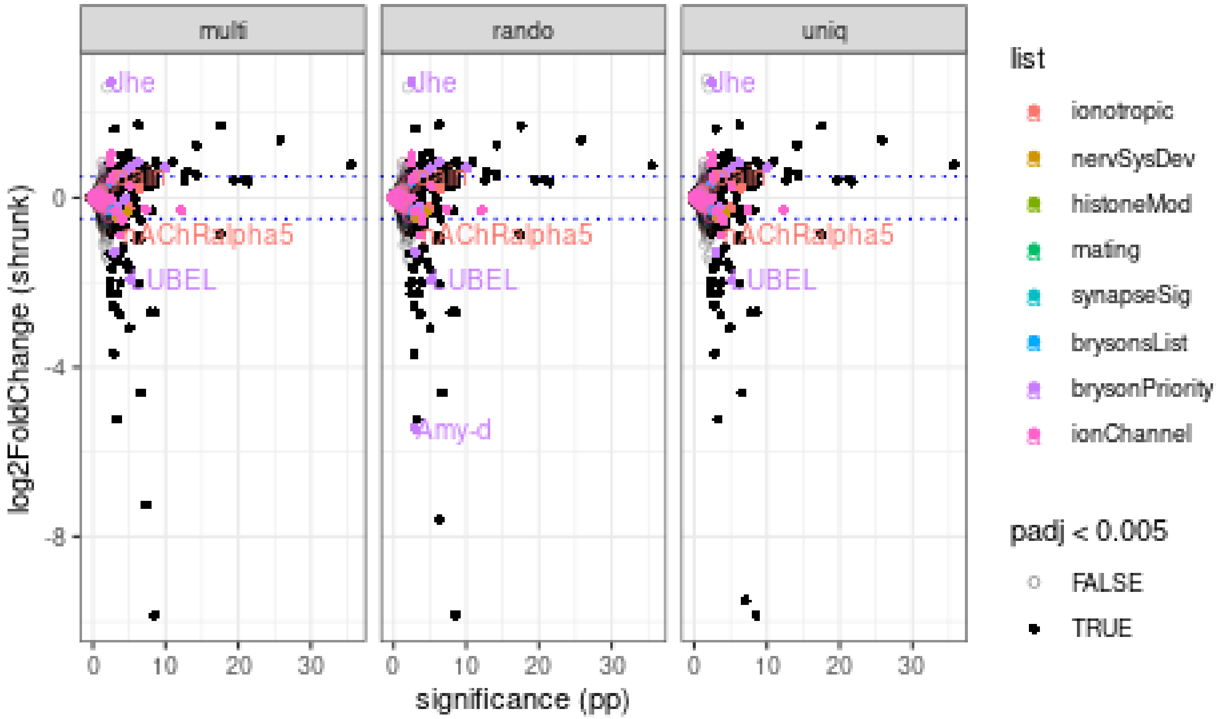
Mal-B2	yes	yes	yes
hgo	yes	yes	yes
CG14838	yes	yes	yes
phu	yes	yes	yes
BomBc2	yes	yes	yes
Cpr64Ac	yes	yes	yes
CG8745	yes	yes	yes
Lst	yes	yes	yes
CG5435	yes	yes	yes
CG11400	yes	yes	yes
CG18003	yes	yes	yes
Jhe	yes	yes	yes
CG5171	yes	yes	yes
CG9572	yes	yes	yes
lectin-28C	yes	yes	yes
Spag1	yes	yes	yes
CG31324	yes	yes	yes
CG14105	yes	yes	yes
CG33233	yes	yes	yes
Srr	yes	yes	yes
Mal-A5	yes	yes	yes
Gbp2	yes	yes	yes
CG11842	yes	yes	yes
Apoltp	yes	yes	yes
bib	yes	yes	yes

3.1.5 In relation to gene lists

```
## png
##   2
```

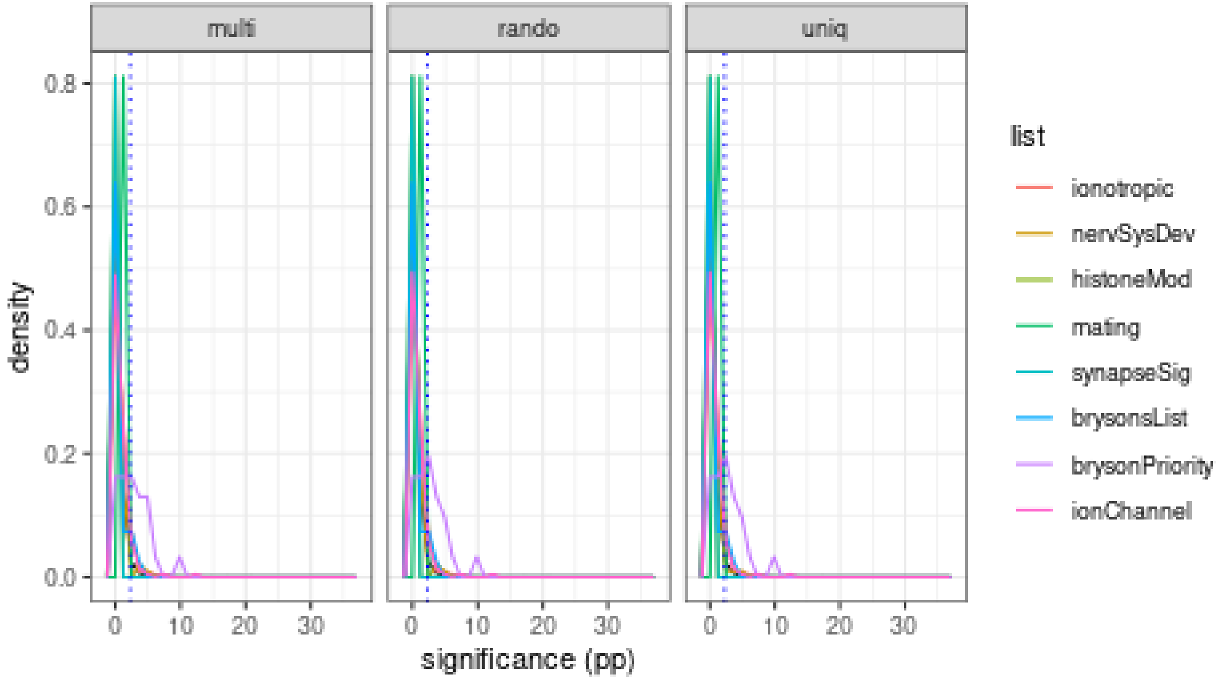
```
## png
##   2
```

Figure 29. Volcano Plot: Fold Change vs. Significance with Gene Lists (between isolated and group-housed wildtypes)
abs(lfc)<0.5 & adjusted p < 0.005 highlighted



```
## png
##   2
```

Figure 30. p-value Distribution: Distribution of Fold-Change Significance
in General and in Genes of Greatest inTerest
(between isolated and group-housed wildtypes)
adjusted p < 0.005 highlighted



```
## png
##   2
```

3.1.6 Genes with top 10 most significant changes

Ordered in decreasing significance, the alignemnt strategies agree on the top 10 most significant changes:

Table 29. Top Ten Most Significantly (padj<0.01) Differentially Expres
between isolated and group-housed wildtypes

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldChang
1	CG10050	0.82	3.43×10^{-36}	0.768	CG10050	0.82	2.89×10^{-36}	0.76
2	MtnB	1.67	1.55×10^{-26}	1.366	MtnB	1.67	1.44×10^{-26}	1.36
3	CG14687	6.77	3.48×10^{-22}	0.369	CG14687	6.77	3.41×10^{-22}	0.37
4	CG31663	0.94	1.06×10^{-21}	0.426	CG31663	0.94	1.14×10^{-21}	0.42
5	Cln3	1.73	3.56×10^{-20}	0.416	Cln3	1.73	3.41×10^{-20}	0.41
6	CG11852	0.35	2.42×10^{-18}	1.695	CG11852	0.35	2.83×10^{-18}	1.69
7	Dhc36C	0.05	3.08×10^{-18}	−0.861	Dhc36C	0.05	3.57×10^{-18}	−0.86
8	Obp84a	1.12	5.98×10^{-15}	0.536	Obp84a	1.12	5.77×10^{-15}	0.53
9	amd	3.50	6.13×10^{-15}	1.246	amd	3.50	5.77×10^{-15}	1.24
10	CG13659	0.47	1.21×10^{-13}	0.613	CG13659	0.47	1.44×10^{-13}	0.61

3.1.7 Top 10 genes with biggest (significant) effect sizes

The alignment strategies agree on the top 10 largest fold changes (though not completely on their order):

Table 30. Top Ten Largest Magnitude Fold Changes which were Significa
between isolated and group-housed wildtypes

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldChange
1	TotC	0.09	3.01×10^{-9}	-9.856	TotC	0.09	2.99×10^{-9}	-9.848
2	Amy-p	0.04	4.46×10^{-8}	-7.256	Amy-p	0.03	5.15×10^{-7}	-7.593
3	CG10799	0.03	5.53×10^{-4}	-5.240	Amy-d	0.02	9.94×10^{-4}	-5.438
4	Prat2	0.02	2.28×10^{-7}	-4.606	CG10799	0.03	5.65×10^{-4}	-5.241
5	CG14838	0.00	1.46×10^{-3}	-3.687	Prat2	0.02	2.27×10^{-7}	-4.607
6	CG2736	0.03	9.46×10^{-6}	-3.080	CG14838	0.00	1.47×10^{-3}	-3.687
7	phu	0.01	1.85×10^{-4}	-2.763	CG2736	0.03	9.25×10^{-6}	-3.081
8	Jhe	1.23	3.27×10^{-3}	2.733	phu	0.01	1.84×10^{-4}	-2.764
9	CG7470	0.03	1.16×10^{-8}	-2.708	Jhe	1.23	3.31×10^{-3}	2.733
10	CG15144	0.02	2.64×10^{-9}	-2.699	CG7470	0.03	1.21×10^{-8}	-2.708

3.1.8 Top 10 highest expressed genes with significant change

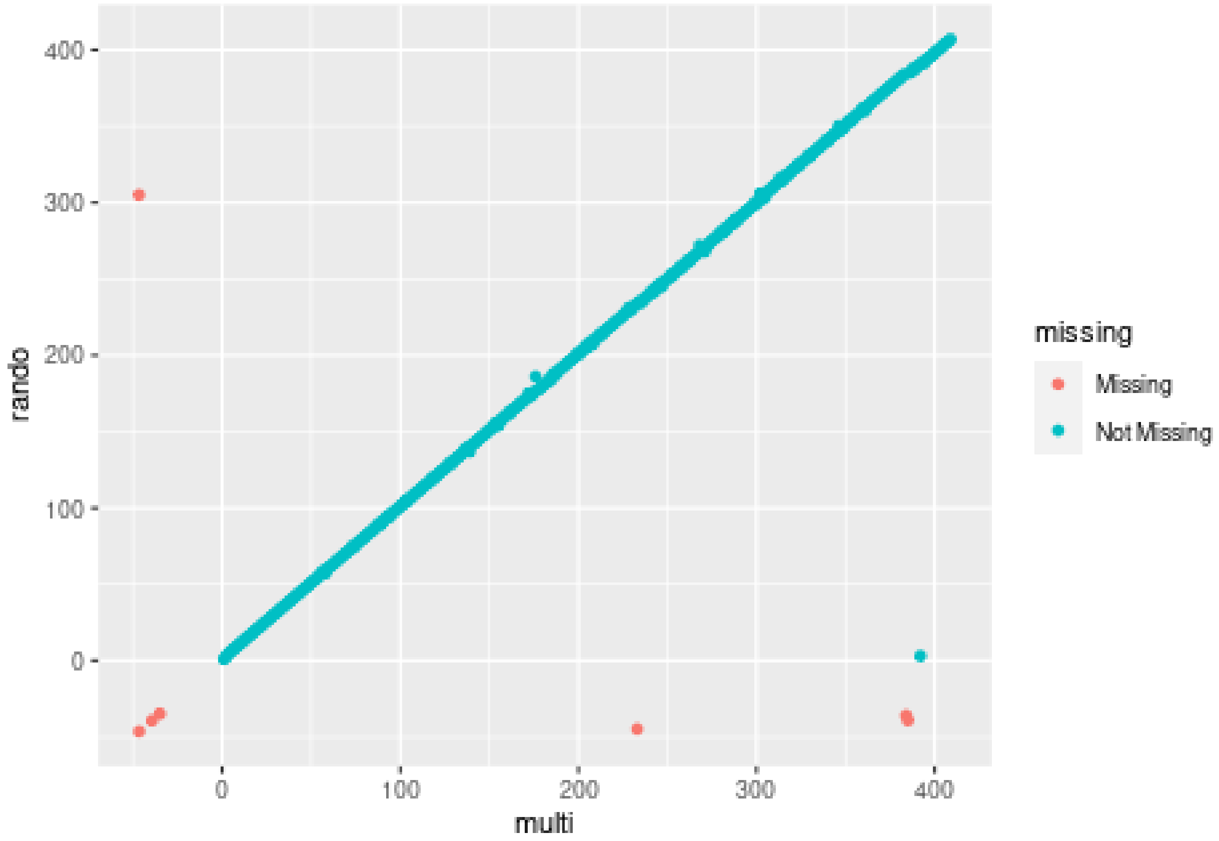
Ranking by DESeq2-based expression (ie, basemean scaled by gene length, in units of standard reads per base)

The “multi” and “rando” alignment strategies agree completely on the top 10 most expressed genes with significant changes. The “uniq” strategy differs in rank order and includes Gs2 and Msp300 instead of Calr and bun:

Table 31. Top Ten Highest Expressed Genes with Significant (padj <
Difference
between isolated and group-housed wildtypes

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldChange
1	Obp28a	64.92	4.55×10^{-4}	0.309	Obp28a	64.93	4.27×10^{-4}	0.310
2	a5	55.86	2.33×10^{-4}	0.260	a5	55.86	2.11×10^{-4}	0.262
3	CG9691	29.67	8.16×10^{-3}	0.178	CG9691	29.67	8.60×10^{-3}	0.175
4	CG11550	22.91	1.51×10^{-3}	0.305	Gs2	29.01	9.63×10^{-3}	0.221
5	RpL41	19.08	8.99×10^{-4}	0.219	CG11550	22.91	1.47×10^{-3}	0.306
6	Obp59a	17.76	8.33×10^{-4}	0.266	RpL41	19.08	8.80×10^{-4}	0.218
7	Cyt-b5	14.40	2.04×10^{-3}	0.190	Obp59a	17.76	8.12×10^{-4}	0.267
8	RpL36	13.79	8.04×10^{-3}	0.121	Cyt-b5	14.40	2.02×10^{-3}	0.190
9	vir-1	13.60	8.88×10^{-7}	0.227	RpL36	13.79	8.08×10^{-3}	0.121
10	Ldsdh1	13.59	3.71×10^{-3}	0.246	vir-1	13.60	7.75×10^{-7}	0.227

3.1.9 rank-correllation between alignment strategies



3.2 Group Housed: Wildtype vs Mutants

3.2.1 wt vs OR47b

After filtering to remove genes with too few reads for analysis, about 12.9k of 17.7k annotated genes (72.6226029 %) remain available for testing:

3.2.1.1 preshrunk comparison across alignment strategies

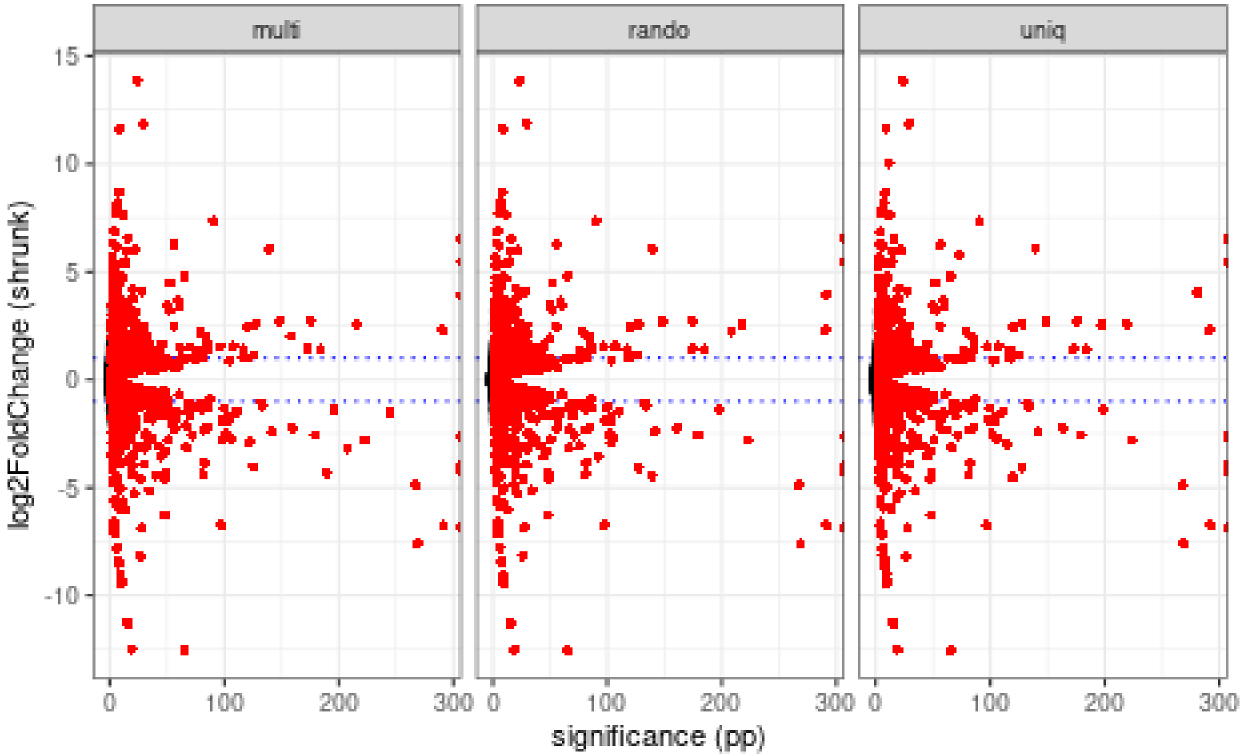
```
## png
##   2
```

```
## png
##   2
```

3.2.1.2 differential expression overview

Here is a volcano plot for the three alignment strategies, with significance on the horizontal axis and log2 fold change on the vertical. Significant (padj<0.01) differences are highlighted in red. Dashed blue guidelines mark a log2 fold change of +/-1 (ie, a difference in expression of a factor of 2). Genes with negative log2 fold changes are depletion relative to the group-housed condition; positive fold changes are enriched

Figure 33. Volcano Plot: Fold Change vs. Significance
(between group-housed wildtypes and 47b mutants)



```
## png
##   2
```

Some of the effect sizes and p values are outrageous!!

From the volcano plots, we can pull out genes with large (ie, a fold change greater than 2 or less than 1/2), significant (ie, padj < 0.01) changes. There were 602 such genes, mostly shared across alignment strategy: (see supplementary tables folder, $results/tables/supp/grpWtVs47b_chonky.html$)

3.2.1.3 Genes with top 10 most significant changes

Ordered in decreasing significance, the alignemnt strategies agree on the top 10 most significant changes:

Table 34. Top Ten Most Significantly (padj<0.01) Differentially Exp
between group-housed wildtypes and 47b mutants

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldCh
1	DIP-alpha	0.18	0.00	−4.256	DIP-alpha	0.18	0.00	−
2	CG6912	1.32	0.00	6.538	CG6912	1.32	0.00	
3	CG7900	1.93	0.00	5.462	CG7900	1.93	0.00	
4	Idgf2	0.89	0.00	−3.980	Idgf2	0.89	0.00	−
5	Drip	2.30	0.00	−2.615	Drip	2.30	0.00	−
6	Cyp6a17	0.92	0.00	−6.856	Cyp6a17	0.92	0.00	−

7	phr	0.38	0.00	3.886	phr	0.34	3.81×10^{-292}	
8	5-HT2A	0.26	6.89×10^{-292}	−6.739	5-HT2A	0.26	7.17×10^{-292}	−
9	Cyp9b2	5.46	4.49×10^{-291}	2.305	Cyp9b2	5.46	2.16×10^{-291}	
10	Or47b	1.41	1.59×10^{-269}	−7.608	Or47b	1.41	1.66×10^{-269}	−

rando and uniq alignment strategies agree very well; in multi, the gene “Unc-115a” has moved from off the chart to the #1 spot, bumping off “Ugt86Dd”.

3.2.1.4 Top 10 genes with biggest (significant) effect sizes

The alignment strategies agree well for the top 4, and disagree on order and content lower:

Table 35. Top Ten Largest Magnitude Fold Changes which v

between group-housed wildtypes and 47b mutants

	multi				rando		
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p
1	mthl8	0.14	1.48×10^{-24}	13.842	mthl8	0.14	1.42×10^{-24}
2	CG40486	4.88	3.44×10^{-66}	−12.538	CG40486	4.87	3.56×10^{-66}
3	CG30428	0.20	8.46×10^{-20}	−12.468	CG30428	0.20	8.41×10^{-20}
4	w	0.52	4.21×10^{-30}	11.860	w	0.52	4.08×10^{-30}
5	CG43149	0.16	1.26×10^{-9}	11.618	CG43149	0.16	1.24×10^{-9}
6	ppk19	0.08	3.44×10^{-16}	−11.319	ppk19	0.08	3.44×10^{-16}
7	lncRNA:CR45502	0.15	6.46×10^{-16}	−11.224	lncRNA:CR45502	0.15	6.43×10^{-16}
8	lncRNA:CR44377	0.02	1.23×10^{-10}	−9.447	lncRNA:CR44377	0.02	1.20×10^{-10}
9	CG14563	0.07	1.26×10^{-10}	−9.319	CG14563	0.07	1.21×10^{-10}
10	asRNA:CR44030	0.04	5.10×10^{-10}	−9.053	asRNA:CR44030	0.04	4.85×10^{-10}

3.2.1.5 Top 10 highest expressed genes with significant change

Ranking by DESeq2-based expression (ie, basemean scaled by gene length, in units of standard reads per base)

The three alignment strategies agree well on the top 10 highest expressed genes with significant change:

Table 35. Top Ten Highest Expressed Genes with Significant (padj
Difference

between group-housed wildtypes and 47b mutants

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldChang
1	Obp83b	119.29	1.33×10^{-4}	0.305	Obp83b	119.29	8.49×10^{-5}	0.30
2	Obp19d	104.32	9.71×10^{-4}	0.352	Obp19d	104.31	5.90×10^{-4}	0.35
3	Obp83a	100.03	4.58×10^{-9}	0.409	Obp83a	100.03	8.97×10^{-10}	0.40
4	Obp28a	65.56	4.72×10^{-20}	0.510	Obp28a	65.56	8.86×10^{-21}	0.50
5	OS9	65.21	4.44×10^{-5}	0.276	OS9	65.20	2.71×10^{-5}	0.27
6	Obp19a	58.27	1.97×10^{-4}	0.204	Obp19a	58.26	1.33×10^{-4}	0.20
7	GstE4	44.81	7.66×10^{-4}	0.194	GstE4	44.80	5.35×10^{-4}	0.19
8	Ugt35B1	40.66	1.62×10^{-7}	0.573	Ugt35B1	40.66	8.77×10^{-8}	0.57
9	Obp69a	40.08	1.42×10^{-4}	0.304	Obp69a	40.07	1.15×10^{-4}	0.30
10	CG11391	38.64	3.34×10^{-3}	0.301	CG11391	38.64	2.53×10^{-3}	0.30

3.2.2 wt vs 67d

After filtering to remove genes with too few reads for analysis, about 12.8k of 17.7k annotated genes (72.3915779 %) remain available for testing:

3.2.2.1 preshrunk comparison across alignment strategies

```
## png
##   2
```

```
## png
##   2
```

3.2.2.2 differential expression overview

Here is a volcano plot for the three alignment strategies, with significance on the horizontal axis and log2 fold change on the vertical. Significant (padj<0.01) differences are highlighted in red. Dashed blue guidelines mark a log2 fold change of +/-1 (ie, a difference in expression of a factor of 2). Genes with negative log2 fold changes are depleted relative to the group-housed condition; positive fold changes are enriched

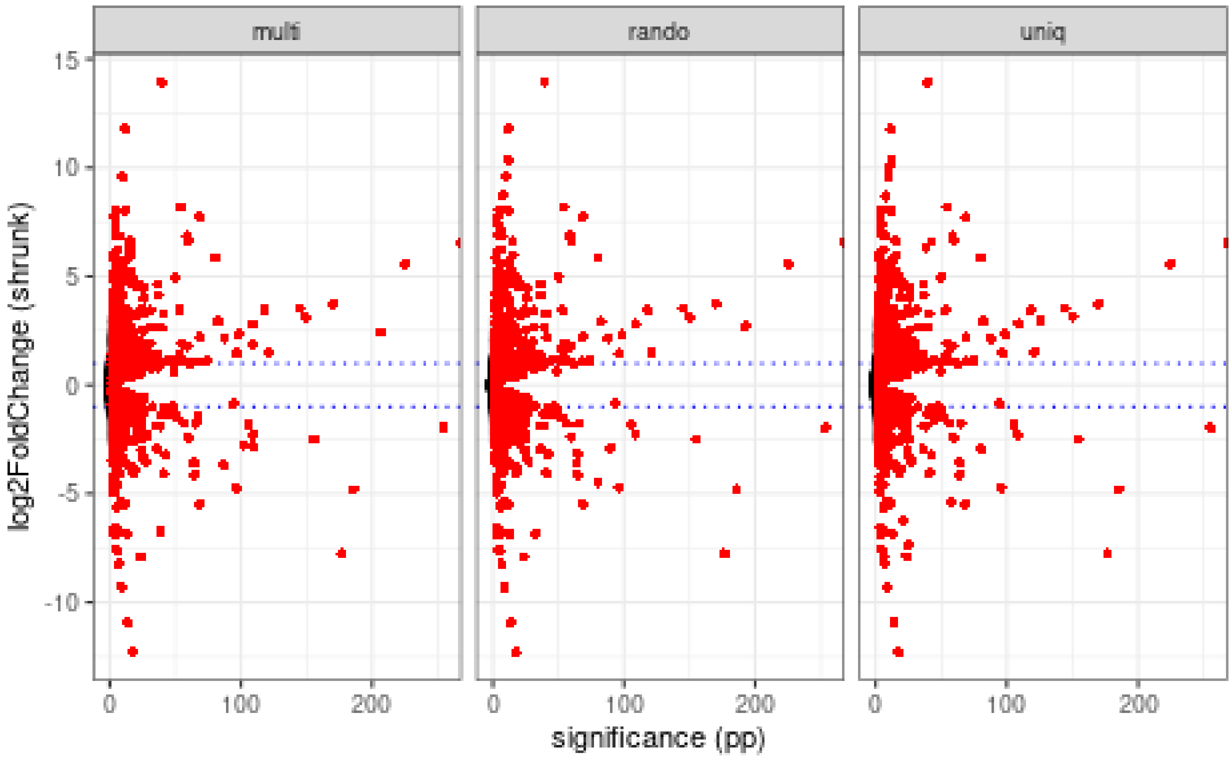


Figure 36. Volcano Plot: Fold Change vs. Significance (between group-housed wildtypes and 67d mutants)

```
## png
##   2
```

From the volcano plots, we can pull out genes with large (ie, a fold change greater than 2 or less than 1/2), significant (ie, padj < 0.01) changes. There were 591 such genes, mostly shared across alignment strategy: (see tables folder, $results/tables/supp/grpWtVs67d_{c}honky.html$)

3.2.2.3 Genes with top 10 most significant changes

Ordered in decreasing significance, the alignemnt strategies agree on the top 4 most significant changes, but disagree on the order & content after that.

Table 39. Top Ten Most Significantly (padj<0.01) Differentially Exp
between group-housed wildtypes and 67d mutants

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldCh
1	CG7900	4.04	0.00	6.575	CG7900	4.04	0.00	
2	l(2)03659	0.86	0.00	6.534	l(2)03659	0.86	0.00	
3	NijC	2.51	4.19×10^{-255}	−1.960	NijC	2.51	1.87×10^{-254}	−
4	CG6912	0.67	1.01×10^{-225}	5.569	CG6912	0.67	6.86×10^{-226}	
5	CG32641	4.74	2.25×10^{-207}	2.430	CG32641	3.19	1.68×10^{-193}	
6	DIP-alpha	0.17	1.74×10^{-186}	−4.825	DIP-alpha	0.17	1.11×10^{-186}	−
7	5-HT2A	0.25	1.95×10^{-177}	−7.768	5-HT2A	0.25	2.70×10^{-177}	−
8	ppk25	0.61	3.94×10^{-171}	3.745	ppk25	0.61	7.07×10^{-171}	
9	CG9447	2.75	1.72×10^{-156}	−2.500	CG9447	2.75	9.10×10^{-156}	−
10	Cyp9b1	1.71	6.02×10^{-150}	3.123	Cyp9b1	1.71	6.83×10^{-151}	

3.2.2.4 Top 10 genes with biggest (significant) effect sizes

The alignment strategies agree relatively well on the genes with the top 10 largest (significant) fold changes (though not on their order):

Table 40. Top Ten Largest Magnitude Fold Changes which w
between group-housed wildtypes and 47b mutants

	multi				rando		
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p
1	w	2.12	2.06×10^{-40}	13.942	w	2.12	1.99×10^{-40}
2	CG32437	0.10	3.21×10^{-18}	−12.284	CG32437	0.10	3.24×10^{-18}
3	CG43149	0.16	2.37×10^{-12}	11.788	CG43149	0.16	2.27×10^{-12}
4	lncRNA:CR44111	0.14	1.67×10^{-14}	−10.916	lncRNA:CR44111	0.14	1.71×10^{-14}
5	ppk9	0.02	1.10×10^{-10}	9.595	CG43291	0.03	1.76×10^{-12}
6	lncRNA:CR44377	0.02	1.15×10^{-9}	−9.309	ppk9	0.02	1.13×10^{-10}
7	CG43919	0.05	1.52×10^{-7}	−8.232	lncRNA:CR44377	0.02	1.18×10^{-9}
8	lncRNA:dntRL	0.29	7.39×10^{-55}	8.195	His-Psi:CR31614	0.03	1.95×10^{-8}
9	CheB42a	0.02	1.62×10^{-5}	8.072	CG43919	0.05	1.56×10^{-7}
10	Obp83g	0.16	1.41×10^{-12}	8.028	lncRNA:dntRL	0.29	7.64×10^{-55}

3.2.2.5 Top 10 highest expressed genes with significant change

Ranking by DESeq2-based expression (ie, basemean scaled by gene length, in units of standard reads per base)

The alignment strategies agree well on the top 10 highest expressed genes with significant changes (though not on their order):

Table 41. Top Ten Highest Expressed Genes with Significant (pa
Difference

between group-housed wildtypes and 67d mutants

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 Fold
1	Obp83b	120.39	1.01×10^{-3}	0.398	Obp83b	120.40	9.80×10^{-4}	
2	Obp83a	100.96	6.45×10^{-6}	0.492	Obp83a	100.97	6.19×10^{-6}	
3	Obp69a	45.23	1.20×10^{-16}	0.675	Obp69a	45.24	1.17×10^{-16}	
4	lncRNA:noe	37.67	1.04×10^{-3}	0.371	lncRNA:noe	37.67	1.01×10^{-3}	
5	Drsl5	34.00	4.54×10^{-5}	−0.404	Drsl5	34.00	4.59×10^{-5}	
6	GstE4	32.92	4.51×10^{-6}	−0.641	GstE4	32.92	4.58×10^{-6}	
7	EbpIII	29.12	6.99×10^{-3}	0.358	EbpIII	29.12	6.82×10^{-3}	
8	Cyp6w1	26.24	8.69×10^{-9}	0.606	Cyp6w1	26.24	8.12×10^{-9}	
9	lush	26.17	1.97×10^{-14}	0.799	lush	26.17	1.84×10^{-14}	
10	Snmp1	21.44	3.00×10^{-5}	−0.381	Snmp1	21.44	3.15×10^{-5}	

3.2.3 wt vs FruLexaFru440

After filtering to remove genes with too few reads for analysis, about 13*k* of 17.7*k* annotated genes (73.0132792 %) remain available for testing:

3.2.3.1 preshrunk comparison across alignment strategies

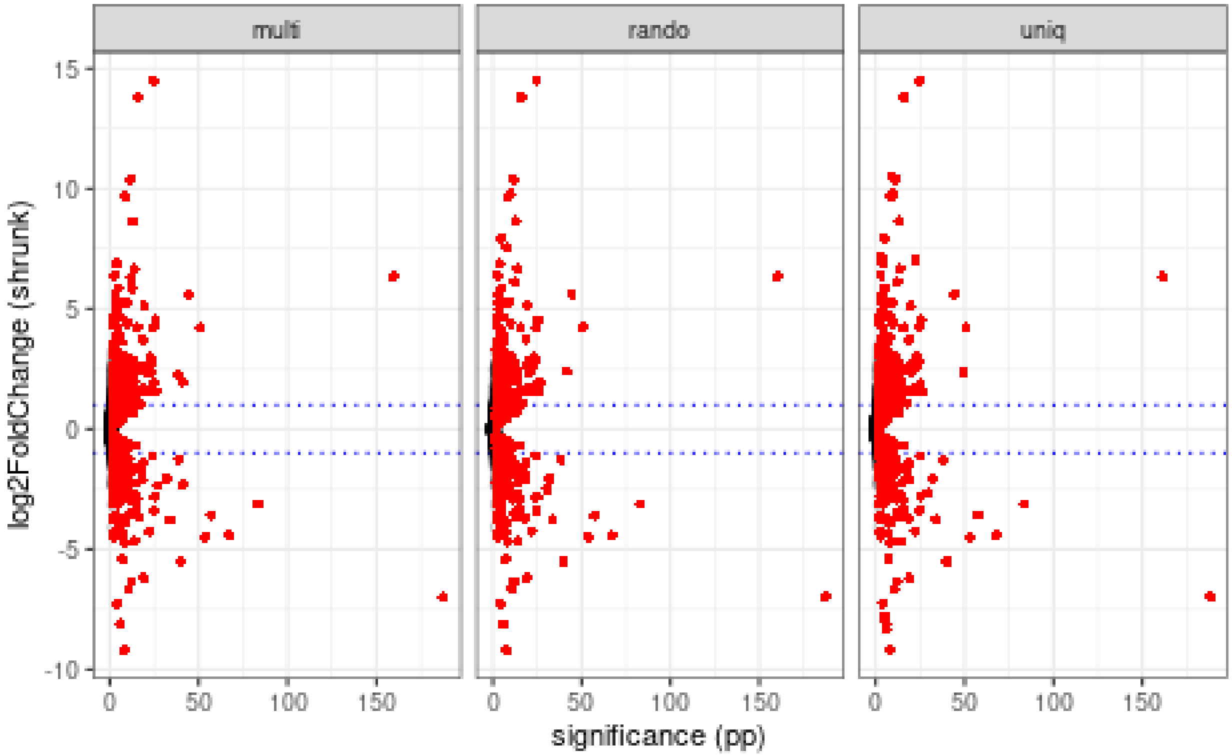
```
## png
##   2
```

```
## png
##   2
```

3.2.3.2 differential expression overview

Here is a volcano plot for the three alignment strategies, with significance on the horizontal axis and log2 fold change on the vertical. Significant (padj<0.01) differences are highlighted in red. Dashed blue guidelines mark a log2 fold change of +/-1 (ie, a difference in expression of a factor of 2). Genes with negative log2 fold changes are depleted relative to the group-housed condition; positive fold changes are enriched

Figure 39. Volcano Plot: Fold Change vs. Significance
(between group-housed wildtypes and 67d mutants)



```
## png
##   2
```

From the volcano plots, we can pull out genes with large (ie, a fold change greater than 2 or less than 1/2), significant (ie, padj < 0.01) changes. There were 413 such genes, mostly shared across alignment strategy: (see tables folder, $results/tables/supp/grpWtVsFru_chonky.html$)

3.2.3.3 Genes with top 10 most significant changes

Ordered in decreasing significance, the alignment strategies agree very well on the top 10 most significant changes:

Table 43. Top Ten Most Significantly (padj<0.01) Differentially Exp
between group-housed wildtypes and Fru mutants

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldC
1	5-HT2A	0.24	3.00×10^{-188}	−6.986	5-HT2A	0.24	4.68×10^{-188}	−
2	CG7900	3.27	2.15×10^{-160}	6.342	CG7900	3.28	3.73×10^{-161}	
3	Ets21C	0.12	6.06×10^{-84}	−3.128	Ets21C	0.12	1.08×10^{-83}	−
4	BomBc1	0.89	1.01×10^{-67}	−4.438	BomBc1	0.89	9.25×10^{-68}	−
5	BomS1	1.83	9.58×10^{-58}	−3.585	BomS1	1.83	7.37×10^{-58}	−
6	DIP-alpha	0.16	2.09×10^{-54}	−4.511	DIP-alpha	0.16	2.71×10^{-54}	−
7	CG42526	0.16	1.07×10^{-51}	4.228	CG42526	0.16	1.18×10^{-51}	

8	CG11893	0.64	5.34×10^{-45}	5.606	CG11893	0.64	4.18×10^{-45}	
9	CG32640	5.39	7.27×10^{-42}	1.966	CG32641	2.57	1.72×10^{-42}	
10	Cyp12d1-p	0.15	7.27×10^{-42}	−2.329	CG9010	0.16	1.37×10^{-40}	−

3.2.3.4 Top 10 genes with biggest (significant) effect sizes

The alignment strategies agree on the genes with the top 5 largest fold changes, less so for the next 5:

Table 44. Top Ten Largest Magnitude Fold Changes which

between group-housed wildtypes and Fru mutants

	multi				rando		
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p
1	mthl8	0.20	1.79×10^{-25}	14.513	mthl8	0.20	1.83×10^{-25}
2	CG43149	0.52	1.13×10^{-16}	13.809	CG43149	0.52	1.09×10^{-16}
3	CG9287	0.03	2.98×10^{-12}	10.390	CG9287	0.03	2.94×10^{-12}
4	ppk27	0.03	3.53×10^{-9}	9.678	CG43291	0.02	1.31×10^{-10}
5	lncRNA:CR44377	0.02	1.12×10^{-8}	−9.198	ppk27	0.03	3.64×10^{-9}
6	w	0.06	9.19×10^{-14}	8.655	lncRNA:CR44377	0.02	1.18×10^{-8}
7	CG43919	0.04	1.15×10^{-6}	−8.116	w	0.06	8.80×10^{-14}
8	CG18577	0.02	7.97×10^{-5}	−7.269	CG43919	0.04	1.18×10^{-6}
9	5-HT2A	0.24	3.00×10^{-188}	−6.986	lncRNA:CR46123	0.04	2.26×10^{-5}
10	lncRNA:CR44285	0.12	1.23×10^{-4}	6.889	CR45496	0.11	1.65×10^{-8}

3.2.3.5 Top 10 highest expressed genes with significant change

Ranking by DESeq2-based expression (ie, basemean scaled by gene length, in units of standard reads per base)

The alignment strategies agree on the top 10 highest expressed genes with significant changes.

Table 45. Top Ten Highest Expressed Genes with Significant (pa Difference

between group-housed wildtypes and Fru mutants

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldCh
1	Obp19d	97.15	4.38×10^{-3}	0.389	Obp19d	97.19	4.21×10^{-3}	0
2	Obp56d	20.04	6.84×10^{-4}	1.129	Obp56d	20.05	6.57×10^{-4}	
3	Obp59a	15.84	1.51×10^{-3}	0.367	Obp59a	15.84	1.32×10^{-3}	0
4	CG6908	15.05	1.42×10^{-3}	0.507	CG6908	15.06	1.32×10^{-3}	0
5	CG9449	12.97	5.14×10^{-6}	0.276	CG9449	12.97	3.80×10^{-6}	0
6	Cyp6a20	11.71	1.29×10^{-39}	−1.281	Cyp6a20	11.71	4.29×10^{-39}	−
7	CG5973	9.78	2.99×10^{-5}	0.562	CG5973	9.78	2.91×10^{-5}	0
8	CG8369	8.84	8.88×10^{-3}	0.721	CG8369	8.85	8.59×10^{-3}	0
9	Vha55	8.64	5.16×10^{-3}	−0.281	Vha55	8.64	5.42×10^{-3}	−0
10	ND-MLRQ	7.75	2.32×10^{-4}	0.430	ND-MLRQ	7.75	2.26×10^{-4}	0

3.3 Comparing Expression Changes from Housing with Expression Changes from Genotype (Single-Factor Comparison Summary)

We want to see if the difference in life history creates similar changes in expression as various mutations. To do this, the differenctial expression data from DESeq2 are joined across pairs of contrasts. For example, the statistics from the wt-group vs wt-isolation contrast are joined by gene with the statistics from the wt-group vs 67d-group contrast. The p-values were readjusted with a Bonferroni correction using n=2 to reflect this new comparison. Candidate genes of interest are then collected by filtering this joint comparison for genes which show a significant change in both contrasts. These candidates are further classified as to whether the expression changes are in the same direction (ie, both enriched or both depleted) or not (ie, one enriched and the other depleted).

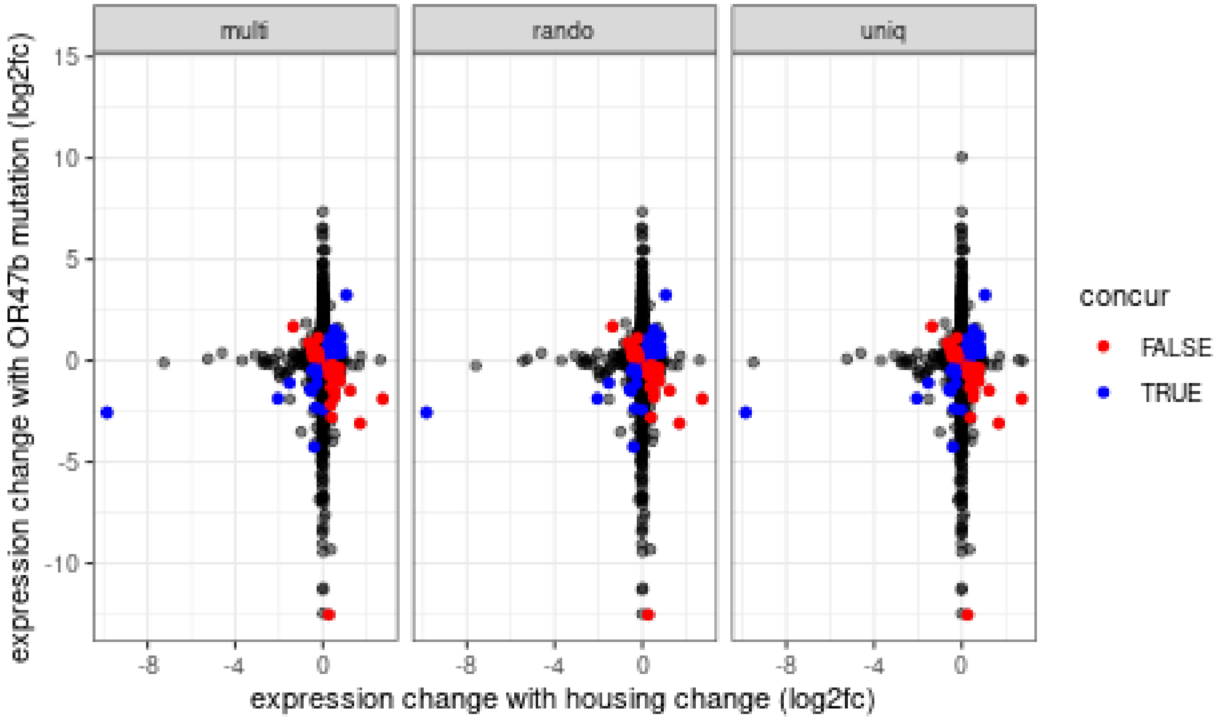
Average significance for gene is currently computed as exp((ln(p1)+ln(p2))/2). (Better to apply stouffer's?)

look at NAs in fulljoin (gene dropout may be interesting…)

3.3.1 Housing & OR47b

Here is a scatterplot of the log2 fold change of the 47b & wt contast vs the housing contrast (wt group & wt isolated). The upper right quadrant contains genes which are enriched in both cases; the lower left contains genes which are depleted in both cases. The other two quadrants contain mismatches between expression patterns. Significant changes are highlighted accordingly.

Figure 40. Scatterplot of Expression Changes in OR47b mutants vs Expression Changes in Housing
(Significant Similarities and Differences Highlighted)



```
## png
##   2
```

Of the mututally significant genes, slightly more have the same direction of change than not:

Table 46. Number of Genes with Significant Changes in Both Contrasts, by Shared Direction of Change
change in housing vs OR47b

	multi	rando	uniq
Agree	63	62	62
Disagree	62	61	60

Of those mutually significant genes with the same direction of change, the top 10 most significant agree well across alignment strategy:

Table 47. Top Ten Most Significant Gene
in difference expression between housing and OR47b

	multi						
rank	name	mean expression	mean readusted p	housing l2fc	mutation l2fc	name	mean expression
1	DIP-alpha	0.25	0.00	−0.393	−4.256	DIP-alpha	0.25
2	CG7272	3.53	1.45×10^{-60}	0.233	1.065	CG7272	3.53
3	CG9717	1.78	1.67×10^{-58}	0.531	1.532	CG9717	1.78
4	jv	0.16	7.59×10^{-42}	0.494	1.223	jv	0.16
5	Obp59a	18.22	2.35×10^{-29}	0.266	0.569	Obp59a	18.22
6	Cpr64Ac	0.28	5.54×10^{-29}	1.069	3.231	Cpr64Ac	0.28
7	NA	0.45	2.92×10^{-27}	0.413	0.844	NA	0.45
8	CAH7	1.08	1.24×10^{-22}	0.363	0.715	CAH7	1.08
9	CG31313	2.62	3.27×10^{-21}	0.275	0.907	Lgr1	0.22
10	Lgr1	0.22	9.17×10^{-20}	−0.287	−1.134	CG9338	2.11

When mutually significant genes with the same direction of change are ranked by the magnitude of their mean log2FoldChange, the top 10 agree well across alignment strategy:

Table 48. Top Ten Largest Magnitude Changes In Significant G
in difference expression between housing and OR47b contrants

	multi				rando			
rank	name	mean l2fc	mean expression	mean readusted p	name	mean l2fc	mean expression	mean
1	TotC	−6.210	0.10	1.56×10^{-7}	TotC	−6.205	0.10	
2	DIP-alpha	−2.324	0.25	0.00	DIP-alpha	−2.325	0.25	
3	Cpr64Ac	2.150	0.28	5.54×10^{-29}	Cpr64Ac	2.151	0.28	
4	Srr	−1.964	0.01	6.05×10^{-3}	Srr	−1.965	0.01	
5	Dscam4	−1.329	0.12	2.58×10^{-6}	Dscam4	−1.352	0.12	
6	CG9572	−1.312	0.12	1.37×10^{-3}	CG9572	−1.310	0.12	
7	TotA	−1.226	0.27	8.09×10^{-5}	TotA	−1.225	0.27	
8	CG9717	1.031	1.78	1.67×10^{-58}	CG9717	1.032	1.78	
9	CG12986	1.015	0.14	7.14×10^{-5}	CG12986	1.016	0.14	
10	Idgf1	−1.010	0.17	3.93×10^{-6}	Idgf1	−1.008	0.17	

Of those mutually significant genes with different directions of change, the top 10 most significant agree well across alignment strategy. (“NA” is trol, “terribly reduced optic lobes”, FBgn0267911/FBgn0284408)

Table 49. Top Ten Most Significant Genes of
in difference expression between housing and OR47b cont

	multi							
rank	name	mean expression	mean readusted p	housing l2fc	OR47b l2fc	name	mean expression	mea
1	CG14400	2.64	2.94×10^{-114}	0.396	−2.804	CG14400	2.64	
2	amd	2.40	6.88×10^{-63}	1.246	−1.477	amd	2.40	
3	SPARC	6.54	1.96×10^{-53}	0.367	−1.325	SPARC	6.54	
4	didum	0.25	4.41×10^{-42}	−0.222	1.125	didum	0.25	
5	CG10050	0.62	1.55×10^{-41}	0.768	−1.019	CG10050	0.62	
6	Obp84a	0.85	2.80×10^{-37}	0.536	−1.537	Obp84a	0.85	
7	CG40486	8.19	1.35×10^{-35}	0.270	−12.538	CG40486	8.19	
8	Prx2540-2	0.98	3.62×10^{-35}	0.353	−2.173	Loxl2	0.96	
9	Loxl2	0.96	7.50×10^{-22}	0.450	−0.977	Jheh3	2.53	
10	Jheh3	2.53	2.36×10^{-21}	0.293	−0.600	CG11852	0.22	

When mutually significant genes with different directions of change are ranked by the magnitude of their difference in log2FoldChange, the top 10 genes agree well across alignment strategy, with minor disagreements about their order:

Table 50. Top Ten Most Serious Significant Differences betw
in difference expression between housing and OR47b contrants

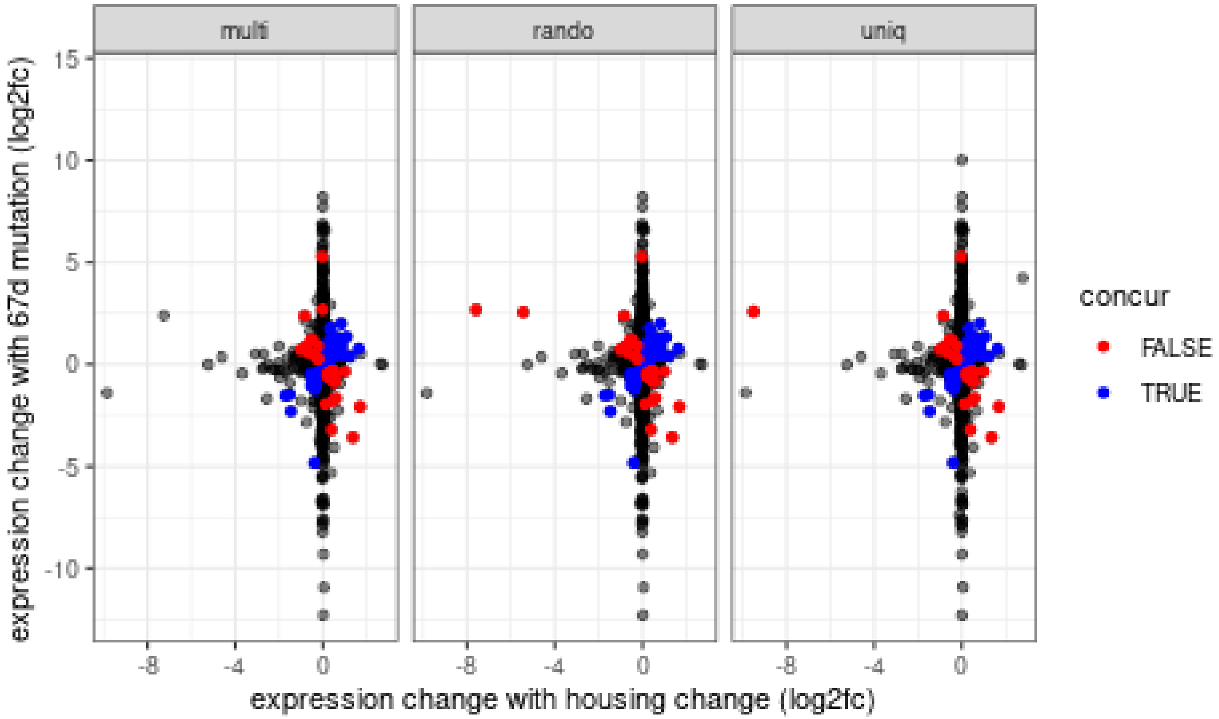
	multi				rando		
rank	name	l2fc difference	mean expression	mean readusted p	name	l2fc difference	mean expression
1	CG40486	12.808	8.19	1.35×10^{-35}	CG40486	12.812	8.19
2	CG11852	4.788	0.22	4.01×10^{-21}	CG11852	4.786	0.22
3	Jhe	4.622	0.68	9.57×10^{-4}	Jhe	4.619	0.68
4	CG14400	3.200	2.64	2.94×10^{-114}	CG14400	3.201	2.64
5	CG5171	−3.039	0.07	1.17×10^{-9}	CG5171	−3.041	0.07
6	amd	2.723	2.40	6.88×10^{-63}	amd	2.718	2.40
7	Prx2540-2	2.526	0.98	3.62×10^{-35}	MtnA	2.323	1.68
8	MtnA	2.326	1.68	2.99×10^{-21}	Obp84a	2.071	0.85
9	Obp84a	2.073	0.85	2.80×10^{-37}	CG10050	1.785	0.62
10	CG10050	1.787	0.62	1.55×10^{-41}	SPARC	1.691	6.54

The full joined comparisons can be found in the tables folder: $results/tables/supp/housingContrast_and_47bContrast.multi.ts$
$results/tables/supp/housingContrast_and_47bContrast.rando.tsv$ $results/tables/supp/housingContrast_and_47bContrast.un$

3.3.2 Housing & 67d

Here is a scatterplot of the log2 fold change of the 67d & wt contast vs the housing contrast (wt group & wt isolated). The upper right quadrant contains genes which are enriched in both cases; the lower left contains genes which are depleted in both cases. The other two quadrants contain mismatches between expression patterns. Significant changes are highlighted accordingly.

Figure 41. Scatterplot of Expression Changes in 67d mutants vs Expression Changes in Housing
(Significant Similarities and Differences Highlighted)



```
## png
##   2
```

Of the mutually significant genes, slightly fewer have the same direction of change as not:

Table 51. Number of Genes with Significant Changes in Both Contrasts, by Shared Direction of Change
change in housing vs 67d

	multi	rando	uniq
Agree	53	52	52
Disagree	38	39	37

Of those mutually significant genes with the same direction of change, the top 10 most significant agree well across alignment strategy:

Table 52. Top Ten Most Significant Genes of Ag
in difference expression between housing and 67d contrants

	multi					ran		
rank	name	mean expression	mean readusted p	housing l2fc	67d l2fc	name	mean expression	mean
1	DIP-alpha	0.25	4.12×10^{-95}	-0.393	-4.825	DIP-alpha	0.25	
2	Pop2	1.87	9.13×10^{-38}	0.259	1.078	Pop2	1.87	

3	CG1227	0.64	3.09×10^{-31}	0.323	1.778	CG1227	0.64
4	CG14400	4.34	5.81×10^{-20}	0.396	1.359	CG14400	4.34
5	CG9717	1.60	6.67×10^{-17}	0.531	1.206	CG9717	1.60
6	dmGlut	0.68	1.56×10^{-15}	0.838	1.130	dmGlut	0.68
7	CG13659	0.54	3.92×10^{-15}	0.613	1.352	CG13659	0.54
8	Cda5	0.17	1.26×10^{-14}	−0.526	−1.053	Cda5	0.17
9	jv	0.15	2.08×10^{-14}	0.494	1.002	jv	0.15
10	CG31288	2.72	2.72×10^{-14}	0.855	0.986	CG31288	2.73

When mutually significant genes with the same direction of change are ranked by the magnitude of their mean log2FoldChange, the top 10 agree relatively well across alignment strategy, with differences in the placement of Amy-d and Amy-p and the inclusion of CG13332.

Table 53. Top Ten Largest Magnitude Changes In Significant Ge

in difference expression between housing and 67d contrants

	multi				rando			
rank	name	mean l2fc	mean expression	mean readusted p	name	mean l2fc	mean expression	mean
1	DIP-alpha	−2.609	0.25	4.12×10^{-95}	DIP-alpha	−2.610	0.25	3
2	lectin-28C	−1.887	0.05	1.22×10^{-4}	lectin-28C	−1.886	0.05	
3	hgo	−1.623	0.10	1.40×10^{-4}	hgo	−1.622	0.10	
4	CG9572	−1.507	0.12	1.92×10^{-3}	CG9572	−1.507	0.12	
5	CG12986	1.412	0.18	5.12×10^{-10}	CG12986	1.412	0.18	5
6	Cpr64Ac	1.212	0.15	1.79×10^{-4}	Cpr64Ac	1.212	0.15	
7	CG31324	1.194	0.20	2.05×10^{-3}	CG31324	1.194	0.20	
8	CG1227	1.050	0.64	3.09×10^{-31}	CG1227	1.050	0.64	3
9	dmGlut	0.984	0.68	1.56×10^{-15}	dmGlut	0.985	0.68	1
10	CG13659	0.982	0.54	3.92×10^{-15}	CG13659	0.983	0.54	4

Of those mutually significant genes with different directions of change, the top 10 most significant agree well across alignment strategy.

Table 54. Top Ten Most Significant Genes of Disa

in difference expression between housing and OR47b contrants

	multi					rand		
rank	name	mean expression	mean readusted p	housing l2fc	67d l2fc	name	mean expression	mean re
1	NijC	3.54	8.87×10^{-129}	0.127	−1.960	NijC	3.54	1.88
2	MtnB	1.06	2.14×10^{-45}	1.366	−3.581	MtnB	1.06	2.6
3	Loxl2	0.90	2.21×10^{-36}	0.450	−1.774	Loxl2	0.90	3.4
4	Tsp	0.13	7.17×10^{-23}	0.390	−3.207	Tsp	0.13	7.4
5	didum	0.23	4.75×10^{-20}	−0.222	0.919	didum	0.23	5.6
6	CG11852	0.22	1.04×10^{-12}	1.695	−2.076	CG11852	0.22	1.1
7	CG5895	1.42	2.41×10^{-11}	0.344	−0.668	CG5895	1.42	2.6
8	CG11425	0.41	1.61×10^{-10}	0.596	−1.663	CG11425	0.41	1.7
9	CG13937	1.67	4.32×10^{-9}	0.246	−0.587	CG13937	1.67	4
10	pug	0.21	1.05×10^{-8}	−0.491	1.224	pug	0.21	1

When mutually significant genes with different directions of change are ranked by the magnitude of their difference in log2FoldChange, the top 10 genes agree well across alignment strategy, with minor disagreements about their order:

Table 55. Top Ten Most Serious Significant Differences betw
in difference expression between housing and 67d contrants

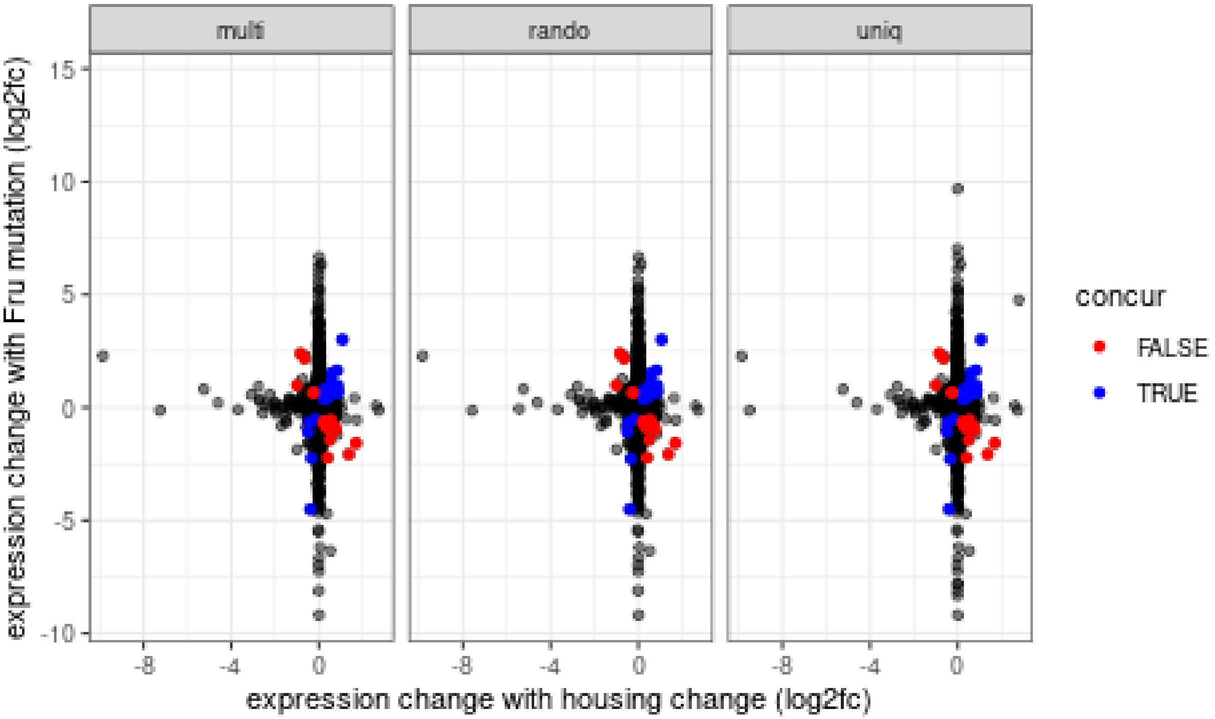
	multi				rando		
rank	name	l2fc difference	mean expression	mean readusted p	name	l2fc difference	mean expression
1	Muc68D	-5.321	0.39	1.97×10^{-4}	Amy-p	-10.269	0.14
2	MtnB	4.947	1.06	2.14×10^{-45}	Amy-d	-8.000	0.07
3	CG11852	3.771	0.22	1.04×10^{-12}	Muc68D	-5.316	0.37
4	Tsp	3.597	0.13	7.17×10^{-23}	MtnB	4.946	1.06
5	CG9812	-3.214	0.55	1.05×10^{-8}	CG11852	3.770	0.22
6	Amy-d	-2.712	0.11	2.93×10^{-4}	Tsp	3.596	0.13
7	CG11425	2.259	0.41	1.61×10^{-10}	CG9812	-3.214	0.55
8	Loxl2	2.224	0.90	2.21×10^{-36}	CG11425	2.259	0.41
9	NijC	2.087	3.54	8.87×10^{-129}	Loxl2	2.223	0.90
10	CG31769	-1.742	0.52	1.27×10^{-8}	NijC	2.086	3.54

The full joined comparisons can be found in the tables folder: $results/tables/supp/housingContrast_and_67dContrast.multi.ts$
$results/tables/supp/housingContrast_and_67dContrast.rando.tsv$ $results/tables/supp/housingContrast_and_67dContrast.un$

3.3.3 Housing & Fru

Here is a scatterplot of the log2 fold change of the Fru & wt contast vs the housing contrast (wt group & wt isolated). The upper right quadrant contains genes which are enriched in both cases; the lower left contains genes which are depleted in both cases. The other two quadrants contain mismatches between expression patterns. Significant changes are highlighted accordingly.

Figure 42. Scatterplot of Expression Changes in Fru mutants vs Expression Changes in Housing
(Significant Similarities and Differences Highlighted)



```
## png
##   2
```

Of the mutually significant genes, slightly more have the same direction of change as not:

Table 56. Number of Genes with Significant Changes in Both Contrasts, by Shared Direction of Change
change in housing vs Fru

	multi	rando	uniq
Agree	20	19	19
Disagree	19	19	18

Of those mutually significant genes with the same direction of change, the top 10 most significant agree well across alignment strategy:

Table 57. Top Ten Most Significant Genes of Ag
in difference expression between housing and Fru contrants

	multi					ran		
rank	name	mean expression	mean readusted p	housing l2fc	Fru l2fc	name	mean expression	mean
1	DIP-alpha	0.24	4.52×10^{-29}	−0.393	−4.511	DIP-alpha	0.24	
2	CG13659	0.55	1.10×10^{-18}	0.613	1.532	CG13659	0.55	

3	Cpr64Ac	0.24	2.64×10^{-14}	1.069	3.004	Cpr64Ac	0.24
4	CG31288	2.64	2.44×10^{-11}	0.855	0.972	CG31288	2.64
5	Pop2	1.68	2.33×10^{-9}	0.259	0.737	Pop2	1.68
6	CG31272	0.19	2.48×10^{-7}	0.853	0.641	CG31272	0.19
7	CG5835	0.40	2.07×10^{-6}	0.570	0.906	CG5835	0.40
8	Dscam4	0.12	3.33×10^{-6}	−0.341	−2.247	Dscam4	0.11
9	jv	0.13	5.24×10^{-6}	0.494	0.507	jv	0.13
10	Ugt301D1	2.58	5.49×10^{-6}	0.416	0.843	Ugt301D1	2.58

When mutually significant genes with the same direction of change are ranked by the magnitude of their mean log2FoldChange, the top 10 agree well across alignment strategy.

Table 58. Top Ten Largest Magnitude Changes In Significant Ge

in difference expression between housing and Fru contrants

	multi				rando			
rank	name	mean l2fc	mean expression	mean readusted p	name	mean l2fc	mean expression	mean
1	DIP-alpha	−2.452	0.24	4.52×10^{-29}	DIP-alpha	−2.452	0.24	5
2	Cpr64Ac	2.037	0.24	2.64×10^{-14}	Cpr64Ac	2.037	0.24	2
3	Dscam4	−1.294	0.12	3.33×10^{-6}	Dscam4	−1.313	0.11	
4	CG12986	1.233	0.15	8.50×10^{-6}	CG12986	1.234	0.15	
5	CG13659	1.072	0.55	1.10×10^{-18}	CG13659	1.073	0.55	1
6	CG31288	0.913	2.64	2.44×10^{-11}	CG31288	0.914	2.64	2
7	Cpr49Ae	0.856	0.37	9.15×10^{-6}	Cpr49Ae	0.856	0.37	
8	CG42806	0.811	1.17	1.81×10^{-5}	CG42806	0.812	1.17	
9	Cpr62Bb	−0.791	0.24	4.90×10^{-3}	Cpr62Bb	−0.789	0.24	
10	CG31272	0.747	0.19	2.48×10^{-7}	CG31272	0.748	0.19	

Of those mutually significant genes with different directions of change, the top 10 most significant agree well across alignment strategy.

Table 59. Top Ten Most Significant Genes of Disg

in difference expression between housing and Fru contrants

	multi					rand		
rank	name	mean expression	mean readusted p	housing l2fc	Fru l2fc	name	mean expression	mean re
1	MtnB	1.08	2.37×10^{-29}	1.366	−2.071	MtnB	1.08	2.8
2	CG10050	0.60	1.14×10^{-19}	0.768	−0.993	CG10050	0.60	1.0
3	CG11852	0.22	5.17×10^{-11}	1.695	−1.580	CG11852	0.22	5.6
4	CG14400	2.61	2.89×10^{-10}	0.396	−2.228	CG14400	2.61	2.8
5	Spn47C	0.14	5.71×10^{-9}	−0.654	2.175	Spn47C	0.14	6.
6	Or92a	3.77	7.05×10^{-8}	0.438	−0.745	Or92a	3.77	7.
7	CG9812	0.53	8.35×10^{-7}	−0.836	2.390	CG9812	0.54	8.
8	CG14275	0.98	2.85×10^{-6}	0.499	−1.433	CG14275	0.98	2.
9	didum	0.21	5.24×10^{-6}	−0.222	0.653	didum	0.21	4.
10	axo	0.58	7.41×10^{-6}	0.207	−0.659	axo	0.58	7.

When mutually significant genes with different directions of change are ranked by the magnitude of their difference in log2FoldChange, the top 10 genes agree well across alignment strategy.

Table 60. Top Ten Most Serious Significant Differences betw
in difference expression between housing and Fru contrasts

	multi				rando		
rank	name	l2fc difference	mean expression	mean readusted p	name	l2fc difference	mean expression
1	MtnB	3.437	1.08	2.37×10^{-29}	MtnB	3.435	1.08
2	CG11852	3.275	0.22	5.17×10^{-11}	CG11852	3.273	0.22
3	CG9812	−3.226	0.53	8.35×10^{-7}	CG9812	−3.228	0.54
4	Spn47C	−2.829	0.14	5.71×10^{-9}	Spn47C	−2.830	0.14
5	CG14400	2.625	2.61	2.89×10^{-10}	CG14400	2.621	2.61
6	Gbs-70E	−1.987	0.20	6.27×10^{-5}	Gbs-70E	−1.988	0.20
7	CG14275	1.932	0.98	2.85×10^{-6}	CG14275	1.931	0.98
8	CG10050	1.760	0.60	1.14×10^{-19}	CG10050	1.759	0.60
9	Dh44-R2	1.545	0.06	3.31×10^{-5}	Dh44-R2	1.542	0.06
10	SPARC	1.269	6.47	1.16×10^{-4}	SPARC	1.264	6.47

Notably, many genes which have significant, high-ranking similarities in both the housing contrast and the Fru contrast … are points of significant, high-ranking differences between the housing contrast and the 47b or 67d contrasts. In particular:

```
DIP-alpha
CG13659
Cpr49Ae
Dscam4
CG31288
Pop2
CG7272
```

As well, many genes which have significant, high-ranking difference in both the housing contrast and the Fru contrast … are points of significant, high-ranking similarity between the housing contrast and the 47b or 67d contrasts. In particular:

```
MtnB
CG10050
CG14400
CG5895
CG11852
Spn47C
```

Full data are in the tables folder:

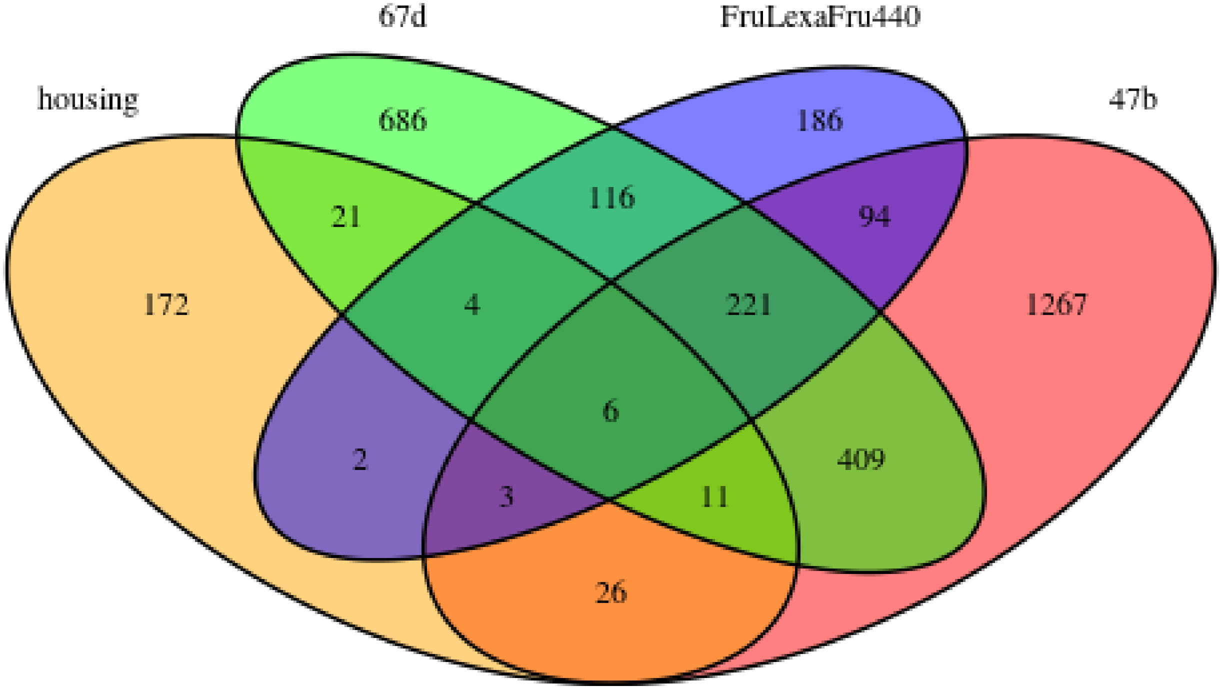
$results/tables/supp/housingContrast_and_FruContrast.multi.tsv$ $results/tables/supp/housingContrast_and_FruContrast.r$
$results/tables/supp/housingContrast_and_FruContrast.uniq.tsv$

3.3.4 Overview

How to perform multiple comparisons adjustment on a all-contrasts venn/upset plot????

Let's bonferroni-correct for n=4 comparisons, one for each single-factor model.

Figure 43 . Venn Diagram: # genes with shared significant change,
by experimental contrast intersection
(multi-only, inner-joined single-factor, bf=4,)



```
## null device
##           1

## null device
##           1
```

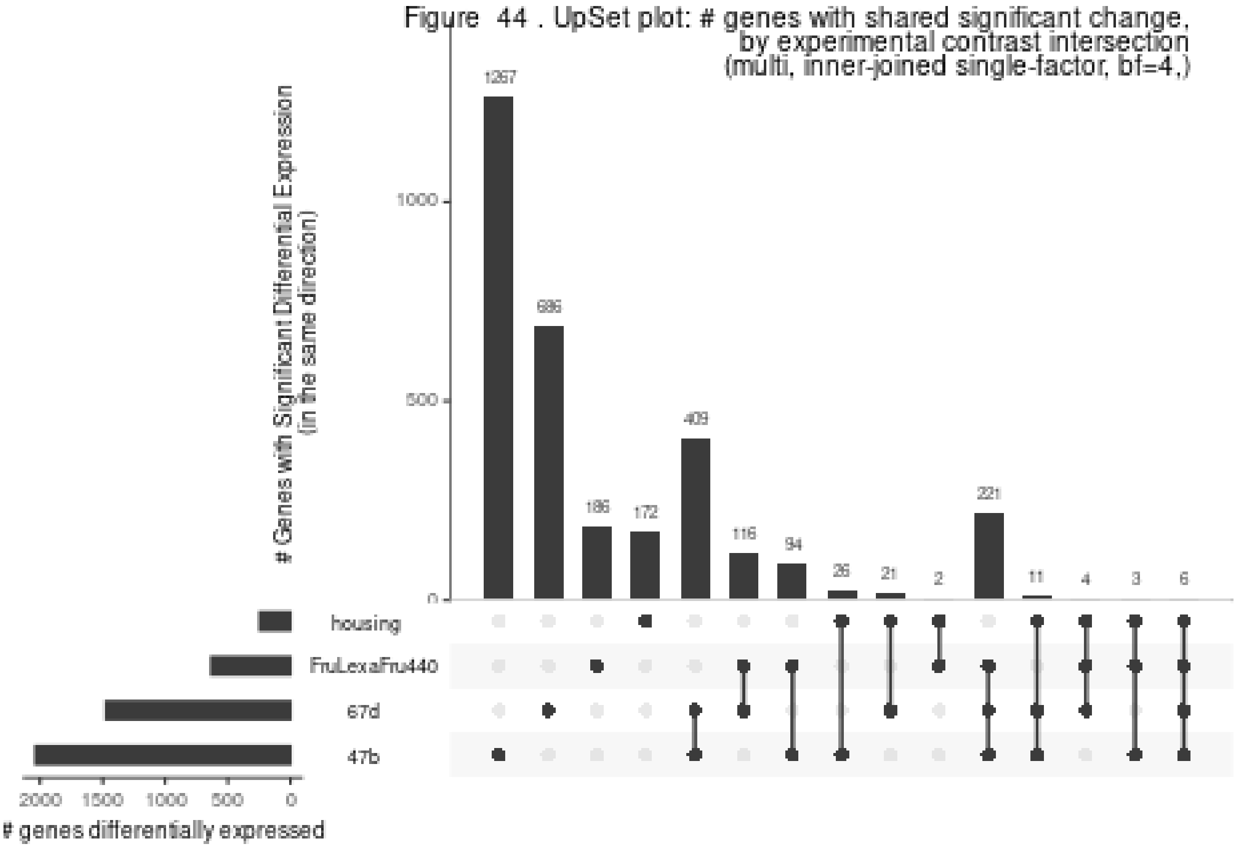


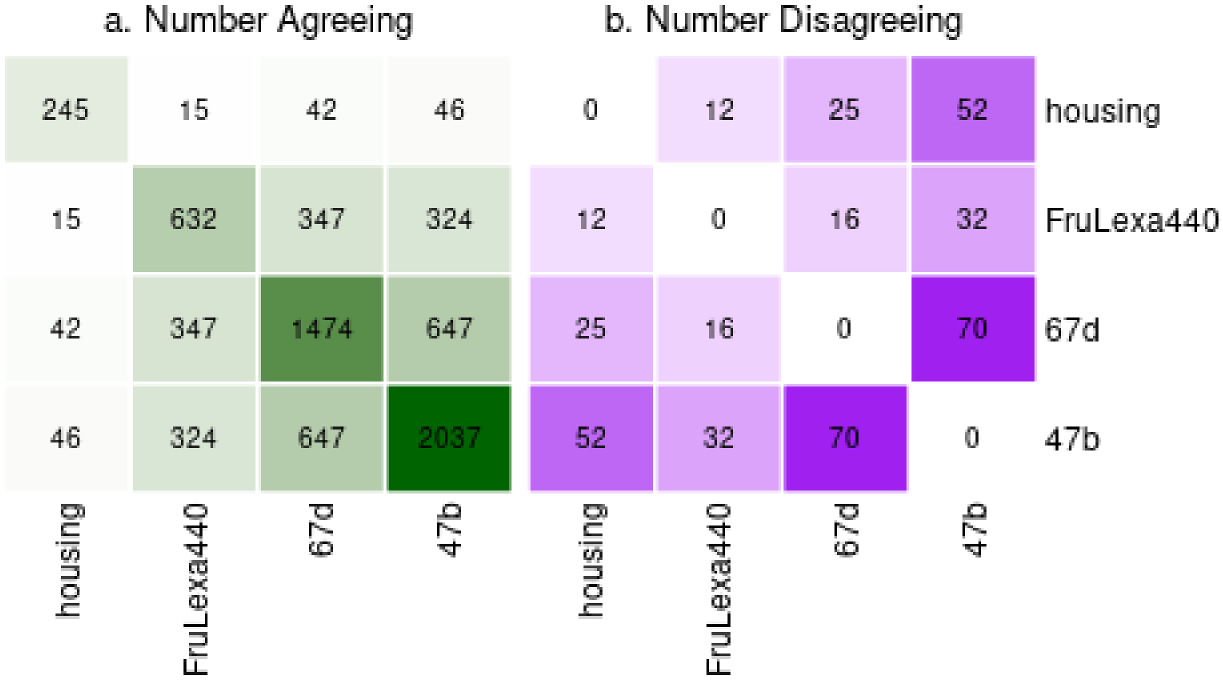
Figure 44 . UpSet plot: # genes with shared significant change, by experimental contrast intersection (multi, inner-joined single-factor, bf=4,)

```
## null device
##           1
```

```
## null device
##           1
```

The results using full join and no correction are qualitatively similar (indeed the more sets being intersected the closer an inner join will approximate a fulljoin)

Figure 45 Heatmap of Pairwise Comparisons between Contrasts:
# significant genes with the same (left)
or different (right) directions of change
(single factor models; multi only; bf=4)



a. Number Agreeing

	housing	FruLexa440	67d	47b
housing	245	15	42	46
FruLexa440	15	632	347	324
67d	42	347	1474	647
47b	46	324	647	2037

b. Number Disagreeing

	housing	FruLexa440	67d	47b
housing	0	12	25	52
FruLexa440	12	0	16	32
67d	25	16	0	70
47b	52	32	70	0

```
## png
##   2
```

3.4 Simultaneously Modeling Housing & Genotype.

gives us eye-to-eye results for all treatments

These data are in the file “results/tables/supp/hausWtVsMut.allAligners.DESeq2.MpBC.reformatted.tsv”; columns are defined as follows:

```
external_gene_name :
    human-readable gene symbol

geneid :
    flybase gene ID

baseMean.(factor).(level) :
    the normalized mean read count for all samples in (level) of contrast (factor).
    Example: baseMean.genotype.wt is the normalized mean read count for wild types
    (of any housing status).

expression.(factor).(level) :
    expression level, calculated as baseMean.(factor).(level)/gene length in bp

baseMean.(factor).vs_(level).apeglm
```

```
log2FoldChange.(factor).vs_(level).apeglm
lfcSE.(factor).vs_(level).apeglm
pvalue.(factor).vs_(level).apeglm
padj.(factor).vs_(level).apeglm
expression.(factor).vs_(level).apeglm :
    equivalent to the "shrunk" data in the single-factor contrast
    for (level) compared to reference
```

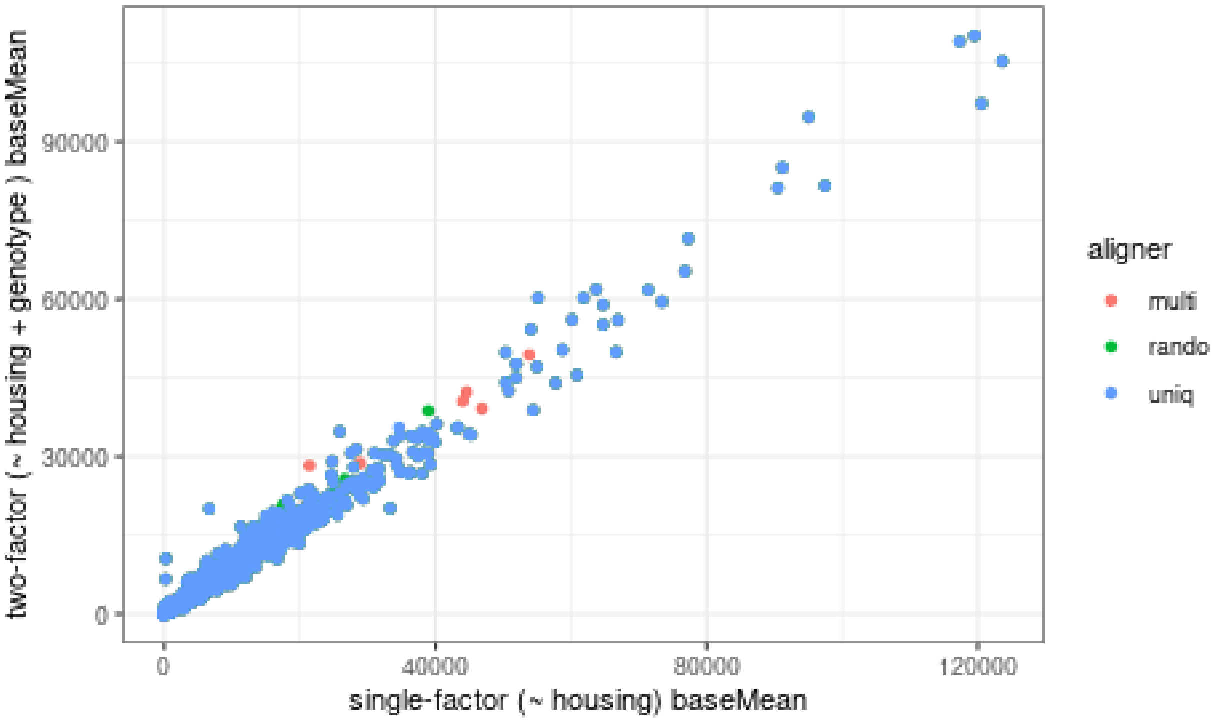
-> do % of genes available for analysis

3.4.1 compare wildTypeHousing results to housing results from hausWtVsMut

To examine consistency with single-factor models, the two-factor model results are subsetted to the housing comparison.

Normalized mean read counts are different between the two models (which is not unexpected) but are correlated:

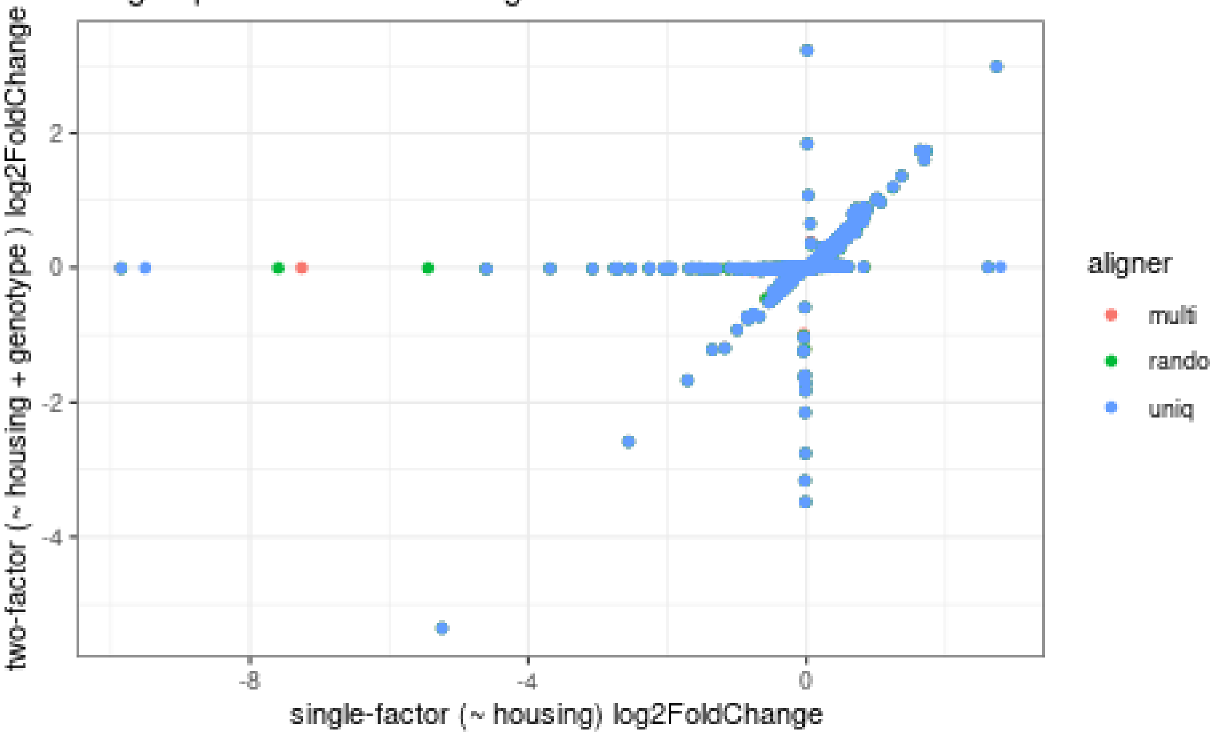
Figure 46. Scatterplot of per-gene normalized mean counts
in full- vs. single-factor models
for group vs isolated housing treatment



```
## png
##   2
```

Effect-size estimates from the two models either agree very well, or not at all. I have not had an opportunity to investigate this discrepancy.

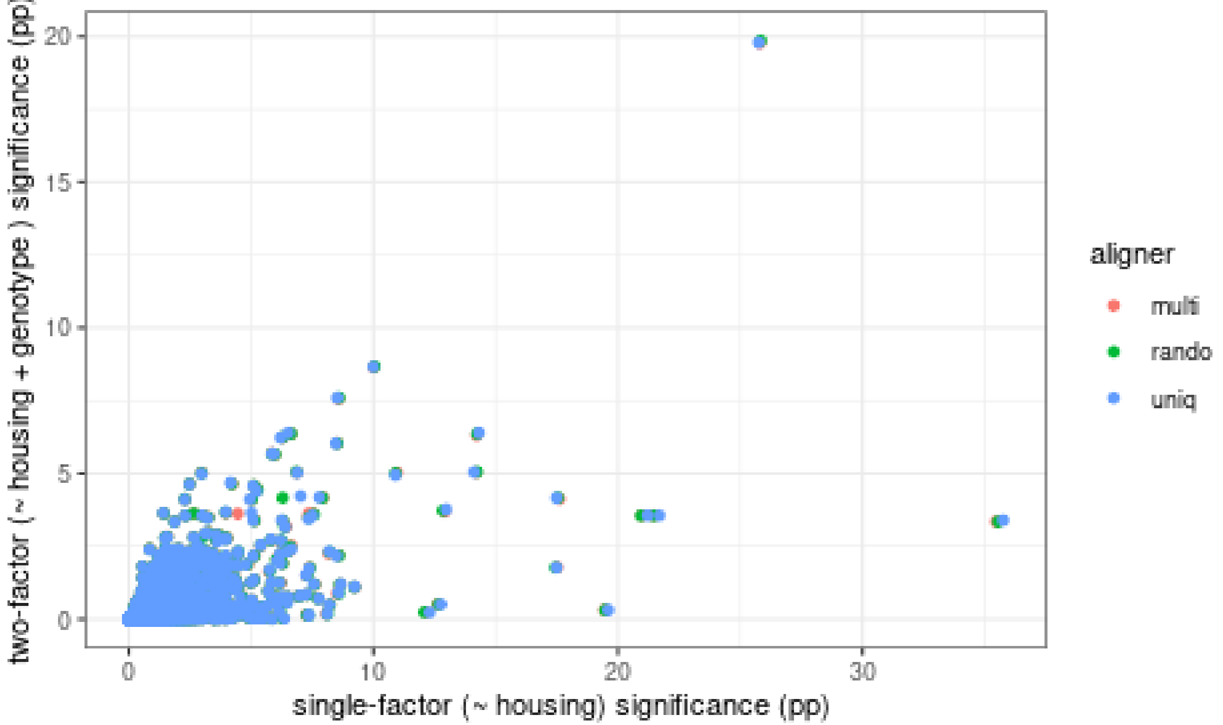
Figure 47. Scatterplot of per-gene expression difference
in full- vs. single-factor models
for group vs isolated housing treatment



```
## png
##   2
```

Significance of differenctial expression estimates agree well enough, I guess:

Figure 48. Scatterplot of per-gene DE significance
in full- vs. single-factor models
for group vs isolated housing treatment

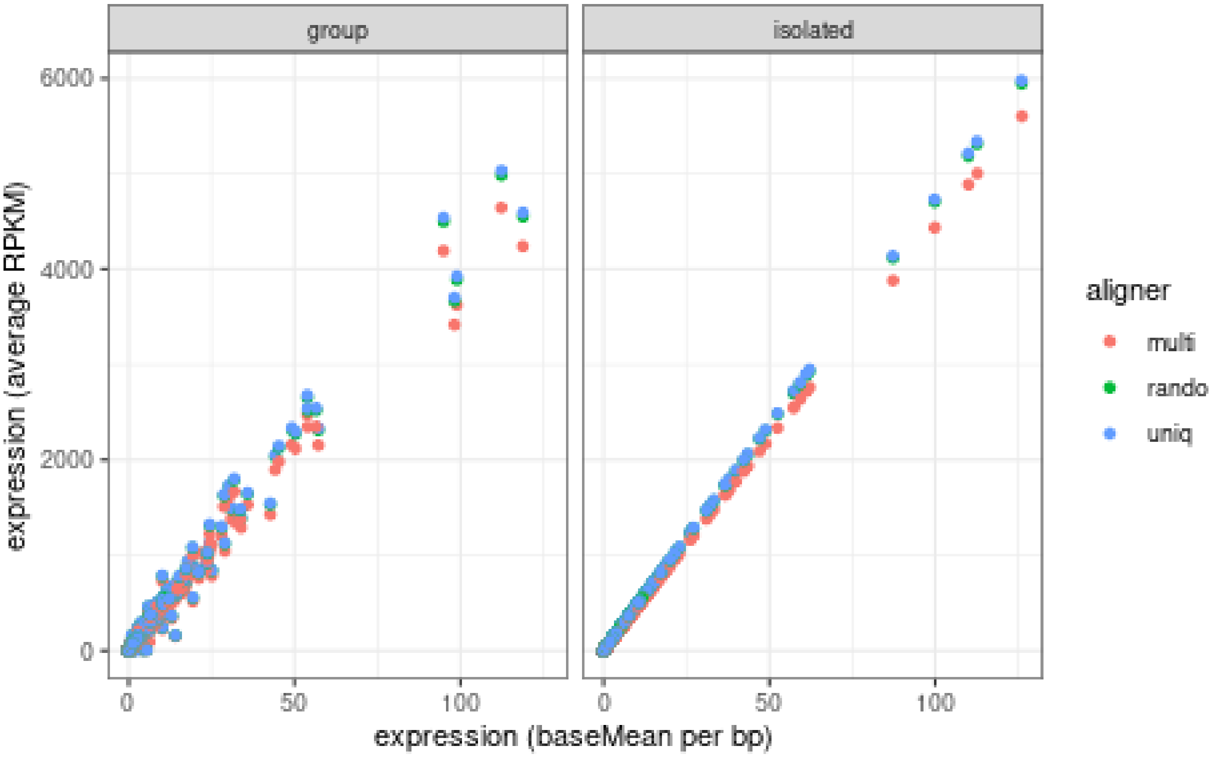


```
## png
##   2
```

3.4.2 compare RPKM to baseMean expression

Two estimates of gene expression have been made: one is based upon normalized mean read count from DESeq2, and the other is an RPKM value calculated from the raw counts. Let's see how they agree

Figure 49. Comparison of Expressionas Inferred by Direct Read Counting vs. DESeq2 Normalization

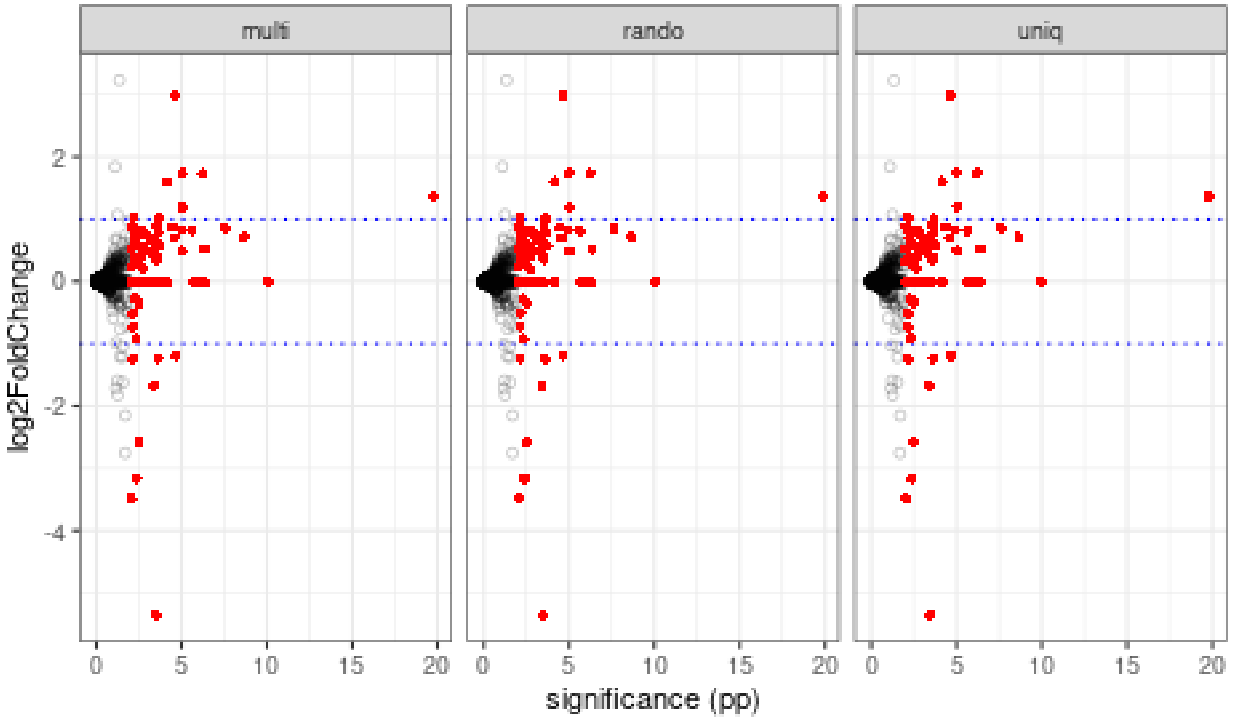


```
## png
##   2
```

3.4.3 Housing Contrast (Simultaneous model)

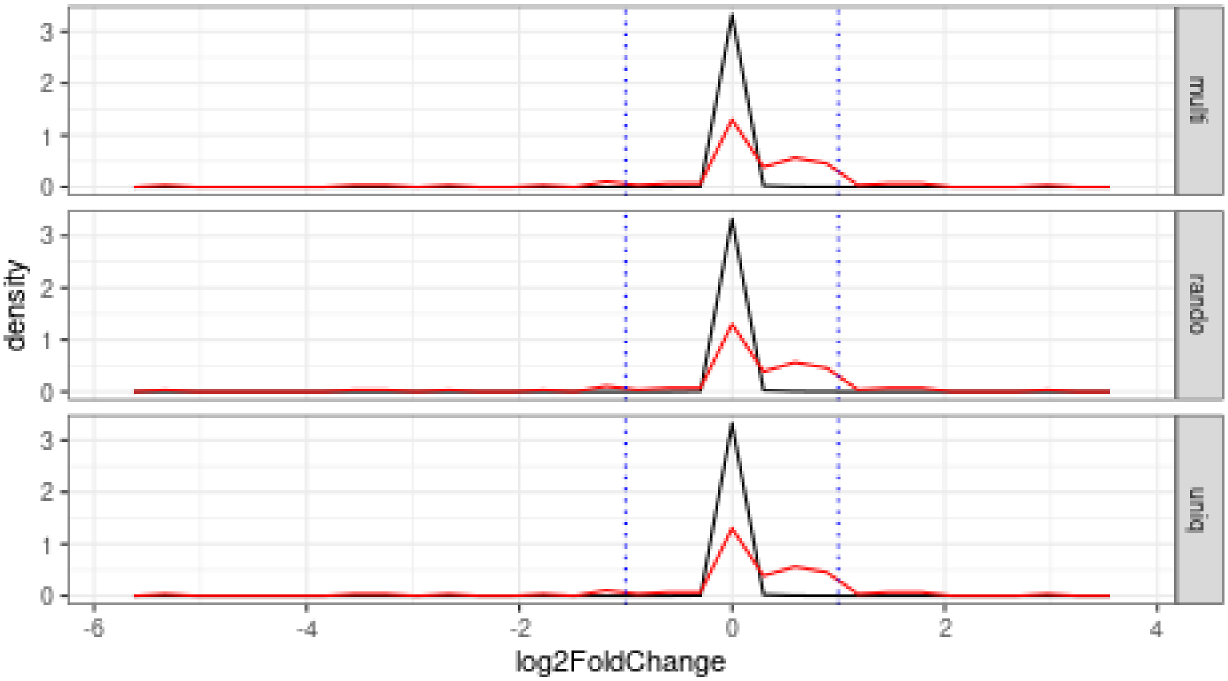
Here is a volcano plot for the three alignment strategies, with significance on the horizontal axis and log2 fold change on the vertical. Significant (padj<0.01) differences are highlighted in red. Dashed blue guidelines mark a log2 fold change of +/-1 (ie, a difference in expression of a factor of 2). Genes with negative log2 fold changes are depleted relative to the group-housed condition; positive fold changes are enriched

Figure 50. Volcano Plot: Fold Change vs. Significance
(between isolated and group-housed wildtypes)
(Simultaneous Model)



```
## png
##   2
```

Figure 51. histogram of fold change
withsignificant(padj<0.01) changes highlighted in red
(between isolated and group-housed wildtypes)
(Simultaneous Model)

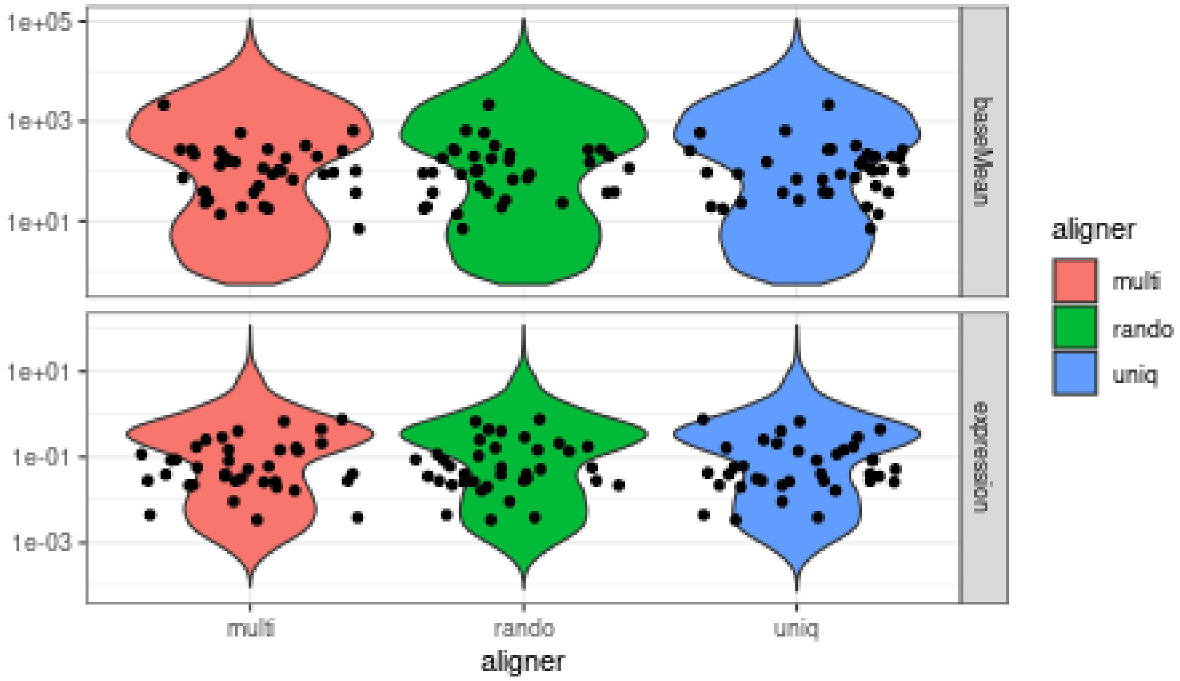


```
## png
##   2
```

I’m concerned about the “tail” of genes with very small effect sizes but high significance. . . .

They do not appear to be unusual in terms of read count or in terms of expression. Here background distributions are shown as violin plots, with anomalous points overplotted:

Figure 52. High Significance, Low Effect Size Genes Do Not Have Unsual Read Counts or Expression Levels
(isolated and group-housed contrast)
(Simultaneous Model)



```
## png
##   2
```

The gene content is not obviously skewed (eg, no tRNA genes, no rRNA genes, not overwhelmed with sketchy CGs. . . .)

Table 61. Genes with Low Effect Size and High Significance Are Mainstream
isolated and group-housed contrast; Simultaneous Model)

gene	where anomalous
Amy-d	multi,rando,uniq
Amy-p	multi,rando,uniq
CG12239	multi,rando,uniq
CG1468	multi,rando,uniq
CG14838	multi,rando,uniq
CG15144	multi,rando,uniq
CG15293	multi,rando,uniq
CG16826	multi,rando,uniq
CG18003	multi,rando,uniq
CG2736	multi,rando,uniq
CG31178	multi,rando,uniq
CG43147	multi,rando,uniq
CG4461	multi,rando,uniq
CG45076	multi,rando,uniq

CG45078	multi,rando,uniq
CG4716	multi,rando,uniq
CG6503	multi,rando,uniq
CG7470	multi,rando,uniq
Hasp	multi,rando,uniq
LUBEL	multi,rando,uniq
Mf	multi,rando,uniq
Mlp60A	multi,rando,uniq
Npc2g	multi,rando,uniq
Npc2h	multi,rando,uniq
PPO1	multi,rando,uniq
PPO2	multi,rando,uniq
Prat2	multi,rando,uniq
TotA	multi,rando,uniq
TotC	multi,rando,uniq
Ubx	multi,rando,uniq
Zasp66	multi,rando,uniq
l(2)efl	multi,rando,uniq
lcs	multi,rando,uniq
lncRNA:CR32652	multi,rando,uniq
nAChRalpha1	multi,rando,uniq
phu	multi,rando,uniq
wupA	multi,rando,uniq

Here's what Mike Love has to say: (email, 13 July 2020):

```
So for one thing, the shrinkage tends to be more conservative than the p-value w/o shrinkage.
If you do svalue=TRUE you will get s-values that correspond to this conservativeness
that you see on the y-axis

The other thing is that, you probably would also lose these genes if you used lfcThreshold=x,
for some x that's higher than 0.

We talk about this in the DESeq2 paper, that rejection of LFC=0 doesn't necessarily mean
that that fold changes are practically meaningful, just that we have evidence that they are
not equal to 0. Typically with more samples we can reject nulls when LFC is quite close to 0...
```

Here's Love, Huber, and Anders (2014):

```
Most approaches to testing for differential expression, including the
default approach of DESeq2,test against the null hypothesis of zero LFC.
However, if any biological processes are genuinely affected by the
difference in experimental treatment, this null hypothesis implies that
the gene under consideration is perfectly decoupled from these processes.
Due to the high interconnected- ness of cells' regulatory networks, this
hypothesis is, in fact, implausible, and arguably wrong for many if not
most genes. Consequently, with sufficient sample size, even genes with a
very small but non-zero LFC will eventually be detected as differentially
expressed. A change should therefore be of sufficient magnitude to be
consid- ered biologically significant. For small-scale experiments,
statistical significance is often a much stricter requirement than
biological significance, thereby relieving the researcher from the
need to decide on a threshold for biological significance.
```

```
For well-powered experiments, however, a statistical
test against the conventional null hypothesis of zero LFC may report
genes with statistically significant changes that are so weak in
effect strength that they could be consid- ered irrelevant or
distracting.
```

Of the 11759 genes with significance scores available, 97 have an adjusted p < 0.01 (0.8249001 %)

From the volcano plots, we can pull out genes with large (ie, a fold change greater than 2 or less than 1/2), significant (ie, padj < 0.01) changes. There were 16 such genes, mostly shared across alignment strategy:

Table 62. Genes with Large (2< fold change), Significant (padj < 0.01) Changes
between isolated and group-housed wildtypes, simultaneous model

	multi	rando	uniq
MtnB	yes	yes	yes
CG10799	yes	yes	yes
CG15822	yes	yes	yes
Jhe	yes	yes	yes
CG31324	yes	yes	yes
amd	yes	yes	yes
CG11400	yes	yes	yes
CG11852	yes	yes	yes
CG13912	yes	yes	yes
CG5819	yes	yes	yes
CG5435	yes	yes	yes
hgo	yes	yes	yes
CG6912	yes	yes	yes
CG33056	yes	yes	yes
CG1146	yes	yes	yes
bib	yes	yes	yes

3.4.3.1 Top Tens

Genes with top 10 most significant changes

Ordered in decreasing significance, the alignemnt strategies agree on the top 10 most significant changes:

Table 63. Top Ten Most Significantly (padj<0.01) Differenti
between isolated and group-housed wildtypes; simultaneous model

	multi				rando		
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p
1	MtnB	0.76	1.67×10^{-20}	1.361	MtnB	0.76	1.37×10^{-20}
2	lncRNA:CR32652	0.02	9.13×10^{-11}	-0.001	lncRNA:CR32652	0.02	8.80×10^{-11}
3	magu	0.40	2.28×10^{-9}	0.713	magu	0.40	2.11×10^{-9}
4	CG31288	2.72	2.77×10^{-8}	0.852	CG31288	2.72	2.52×10^{-8}
5	Prat2	0.04	4.42×10^{-7}	-0.019	Prat2	0.04	4.33×10^{-7}
6	Obp84a	0.75	4.90×10^{-7}	0.524	Obp84a	0.75	4.33×10^{-7}
7	CG15822	0.01	5.92×10^{-7}	1.734	CG15822	0.01	5.75×10^{-7}
8	TotC	0.29	9.11×10^{-7}	-0.009	TotC	0.29	9.28×10^{-7}
9	TotA	0.44	2.07×10^{-6}	-0.007	TotA	0.44	2.15×10^{-6}

10	CG10512	0.44	2.30×10^{-6}	0.814	CG10512	0.44	2.22×10^{-6}

Top 10 genes with biggest (significant) effect sizes

Table 64. Top Ten Largest Magnitude Fold Changes which were Signifi
between isolated and group-housed wildtypes; simultaneous model

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldChang
1	CG10799	0.07	3.20×10^{-4}	−5.353	CG10799	0.07	3.32×10^{-4}	−5.35
2	CG5819	0.01	8.81×10^{-3}	−3.479	CG5819	0.01	8.94×10^{-3}	−3.48
3	CG13912	0.02	4.26×10^{-3}	−3.162	CG13912	0.02	4.21×10^{-3}	−3.16
4	Jhe	0.50	2.29×10^{-5}	2.989	Jhe	0.50	2.31×10^{-5}	2.98
5	CG5435	0.01	3.19×10^{-3}	−2.584	CG5435	0.02	3.22×10^{-3}	−2.58
6	CG31324	0.17	9.49×10^{-6}	1.740	CG31324	0.17	9.89×10^{-6}	1.73
7	CG15822	0.01	5.92×10^{-7}	1.734	CG15822	0.01	5.75×10^{-7}	1.73
8	hgo	0.07	4.10×10^{-4}	−1.674	hgo	0.07	4.19×10^{-4}	−1.67
9	CG11852	0.14	7.55×10^{-5}	1.600	CG11852	0.14	6.93×10^{-5}	1.60
10	MtnB	0.76	1.67×10^{-20}	1.361	MtnB	0.76	1.37×10^{-20}	1.36

Top 10 highest expressed genes with significant change

Ranking by DESeq2-based expression (ie, basemean scaled by gene length, in units of standard reads per base)

Table 65. Top Ten Highest Expressed Genes with Significant (padj <
Difference
between isolated and group-housed wildtypes; simultaneous model

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldChange
1	Fer2LCH	8.34	1.86×10^{-3}	0.202	Fer2LCH	8.34	1.95×10^{-3}	0.202
2	CG14687	5.43	2.77×10^{-4}	0.345	CG14687	5.43	2.82×10^{-4}	0.344
3	CG32276	4.00	3.03×10^{-3}	0.264	CG32276	4.00	3.32×10^{-3}	0.262
4	Or92a	3.09	4.25×10^{-4}	0.422	Or92a	3.09	4.19×10^{-4}	0.422
5	CG31288	2.72	2.77×10^{-8}	0.852	CG31288	2.72	2.52×10^{-8}	0.851
6	CG18135	2.20	3.90×10^{-3}	0.826	CG18135	2.20	3.96×10^{-3}	0.824
7	amd	2.00	9.28×10^{-6}	1.190	amd	2.00	8.90×10^{-6}	1.187
8	Pop2	1.78	8.46×10^{-3}	0.238	Pop2	1.78	9.30×10^{-3}	0.240
9	Nep4	1.54	2.40×10^{-4}	0.785	Nep4	1.54	2.37×10^{-4}	0.782
10	CG33056	1.46	2.31×10^{-4}	1.005	CG33056	1.46	2.19×10^{-4}	1.005

3.4.3.2 Gene Ontology Enrichment

Genes were analyzed for GO Term Enrichment using topGO, using Fisher's test applied to those whose expression difference passed a significance threshold ($p < 0.01$), and applying the Kolmogorov-Smirnov test using p-values as scores.

-> check consistency between GO terms between alignment strategies -> filter out very broad/very specific terms?

Correlation between significance values for the two tests are…. well, folks,

```
## png
##   2
```

Table 66. Enriched GO Terms among Significantly Differentially Expressed Genes
simultaneous housing contrast; multi only

		p-value		
GO Term	Description	Fisher	K-S	ontology
GO:0015116	sulfate transmembrane transporter activity	7.59×10^{-3}	4.70×10^{-3}	MF
GO:0015291	NA	8.13×10^{-3}	2.40×10^{-3}	MF
GO:0009408	response to heat	1.40×10^{-4}	5.50×10^{-5}	BP
GO:0009607	NA	4.60×10^{-4}	5.10×10^{-4}	BP
GO:0043207	NA	4.60×10^{-4}	5.10×10^{-4}	BP
GO:0051707	NA	4.60×10^{-4}	5.10×10^{-4}	BP
GO:0044419	NA	7.30×10^{-4}	3.30×10^{-4}	BP
GO:0006576	NA	7.60×10^{-4}	3.81×10^{-3}	BP
GO:0007599	hemostasis	9.20×10^{-4}	2.60×10^{-4}	BP
GO:0042381	hemolymph coagulation	9.20×10^{-4}	2.60×10^{-4}	BP
GO:0050817	NA	9.20×10^{-4}	2.60×10^{-4}	BP
GO:0009308	NA	9.40×10^{-4}	1.55×10^{-3}	BP
GO:0044106	NA	9.40×10^{-4}	1.55×10^{-3}	BP
GO:0009266	response to temperature stimulus	1.35×10^{-3}	1.20×10^{-4}	BP
GO:0007498	mesoderm development	2.87×10^{-3}	4.76×10^{-3}	BP
GO:0030239	myofibril assembly	3.61×10^{-3}	4.00×10^{-5}	BP
GO:0055002	NA	3.61×10^{-3}	4.00×10^{-5}	BP
GO:0050878	NA	3.86×10^{-3}	2.99×10^{-3}	BP
GO:0042692	muscle cell differentiation	4.55×10^{-3}	3.70×10^{-7}	BP
GO:0061077	chaperone-mediated protein folding	4.57×10^{-3}	1.90×10^{-4}	BP
GO:0055001	muscle cell development	6.92×10^{-3}	2.50×10^{-6}	BP
GO:0008272	sulfate transport	7.10×10^{-3}	4.65×10^{-3}	BP
GO:0005576	extracellular region	3.80×10^{-5}	8.40×10^{-4}	CC
GO:0030017	sarcomere	7.00×10^{-5}	2.00×10^{-6}	CC
GO:0030016	myofibril	9.50×10^{-5}	4.50×10^{-6}	CC
GO:0043292	NA	1.30×10^{-4}	4.60×10^{-6}	CC
GO:0036379	NA	3.07×10^{-3}	2.40×10^{-4}	CC
GO:0030018	Z disc	3.36×10^{-3}	5.00×10^{-5}	CC
GO:0031674	I band	3.58×10^{-3}	8.70×10^{-5}	CC
GO:0015629	actin cytoskeleton	8.36×10^{-3}	8.80×10^{-4}	CC

```
catechol-containing compound metabolic process (GO:0009712)
coagulation (GO:0050817)
response to biotic stimulus (GO:0009607)
response to external biotic stimulus (GO:0043207)
response to other organism (GO:0051707)
alpha-amino acid metabolic process (GO:1901605)
regulation of body fluid levels (GO:0050878)
response to inorganic substance (GO:0010035)
response to chemical (GO:0042221)
obsolete contractile fiber part (GO:0044449)
contractile fiber (GO:0043292)
```

“Bare” GO terms are mostly response (stimulus, chemical) and metabolic (amino acid, catechol) processes, and contractile fiber components.

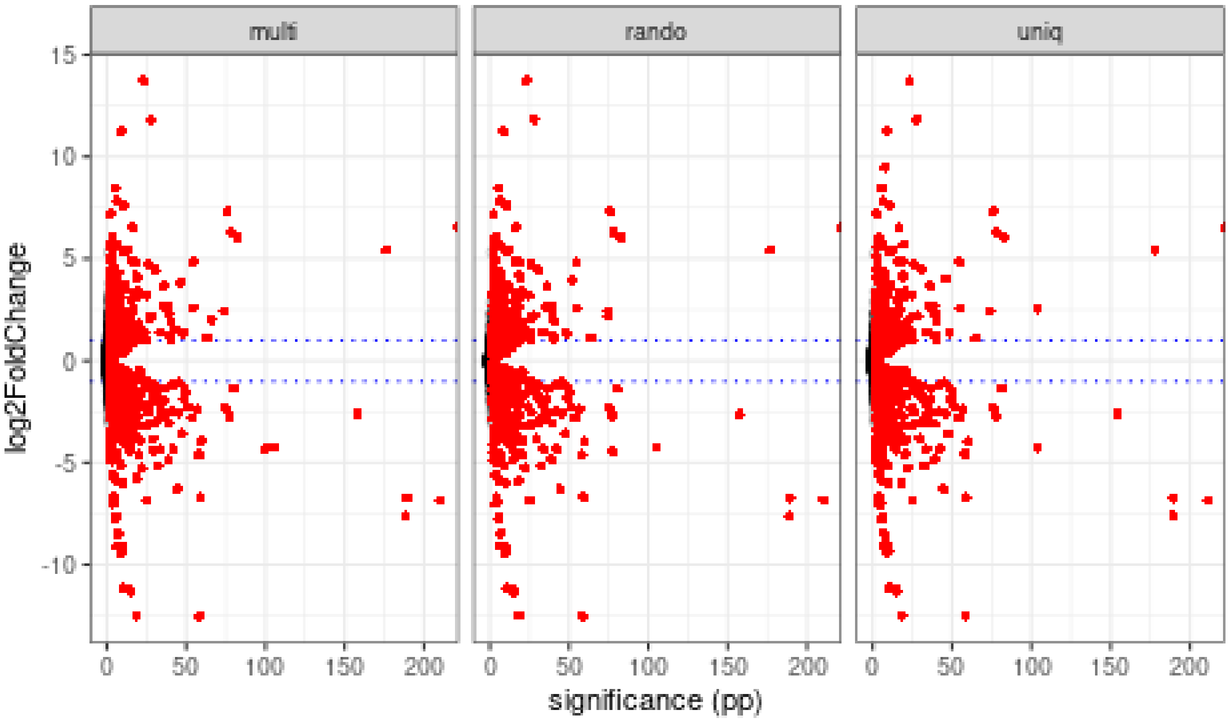
Minor differences w/aligner; see tables/supp/

3.4.4 47b vs wt (Simultaneous model)

Of the 12588 genes with significance scores available, 1099 have an adjusted p < 0.01 (8.730537 %)

Here is a volcano plot for the three alignment strategies, with significance on the horizontal axis and log2 fold change on the vertical. Significant (padj<0.01) differences are highlighted in red. Dashed blue guidelines mark a log2 fold change of +/-1 (ie, a difference in expression of a factor of 2). Genes with negative log2 fold changes are depleted relative to the group-housed condition; positive fold changes are enriched.

Figure 54. Volcano Plot: Fold Change vs. Significance
(between group-housed 47b mutants and wildtypes)
(simultaneous model)



```
## png
##   2
```

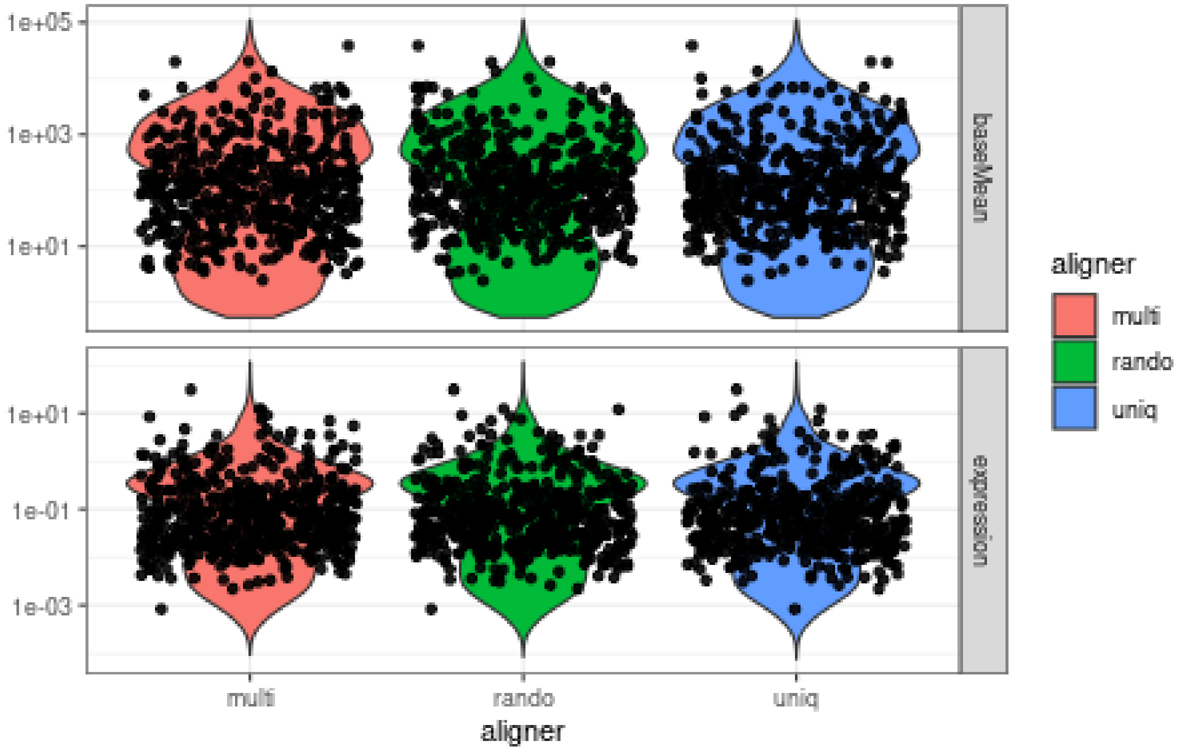
From the volcano plots, we can pull out genes with large (ie, a fold change greater than 2 or less than 1/2), significant (ie, padj < 0.01) changes. There were 553 such genes, mostly shared across alignment strategy:

(Table available at $results/tables/tbl67_{h}ausWtVsMut_{g}enotype47b_{c}honky.html$

Are “chonky” genes prone to unusually low expression?

Figure 55. 'Chonky' Gene Expression Changes Are Not Prone to Low-E

47b1 simultaneous contrast, adjusted p < 0.01, abs(l2fc) > 1



```
## png
##   2
```

3.4.4.1 Top Tens

Genes with top 10 most significant changes

Ordered in decreasing significance, the alignemnt strategies agree on the top 10 most significant changes:

Table 67. Top Ten Most Significantly (padj<0.01) Differentially Exp
between group-housed 47b mutants and wildtypes (simultaneous model)

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldC
1	CG6912	0.73	0.00	6.522	CG6912	0.73	0.00	
2	Cyp6a17	1.75	7.56×10^{-211}	−6.855	Cyp6a17	1.75	4.18×10^{-211}	−
3	5-HT2A	0.19	7.10×10^{-190}	−6.734	5-HT2A	0.19	6.22×10^{-190}	−
4	Or47b	1.83	3.90×10^{-189}	−7.608	Or47b	1.83	1.28×10^{-189}	−
5	CG7900	3.51	7.23×10^{-177}	5.442	CG7900	3.51	1.43×10^{-177}	
6	Drip	2.79	7.61×10^{-159}	−2.612	Drip	2.79	4.23×10^{-158}	−
7	DIP-alpha	0.12	7.00×10^{-106}	−4.242	DIP-alpha	0.12	1.08×10^{-105}	−
8	Cyp12d1-p	0.11	2.00×10^{-100}	−4.337	Cpr62Bc	0.14	1.24×10^{-83}	
9	Cpr62Bc	0.14	2.38×10^{-83}	6.031	PICK1	0.47	4.13×10^{-81}	−
10	PICK1	0.47	1.45×10^{-80}	−1.366	CG8665	0.13	5.04×10^{-79}	

Top 10 genes with biggest (significant) effect sizes

Table 68. Top Ten Largest Magnitude Fold Changes which w

group-housed 47b mutants and wildtypes (simultaneous model)

	multi				rando		
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p
1	mthl8	0.13	6.90×10^{-24}	13.724	mthl8	0.13	6.47×10^{-24}
2	CG40486	7.44	3.30×10^{-59}	−12.525	CG40486	7.44	3.01×10^{-59}
3	CG30428	0.30	3.10×10^{-19}	−12.497	CG30428	0.30	3.10×10^{-19}
4	w	1.01	1.39×10^{-28}	11.814	w	1.01	1.32×10^{-28}
5	ppk19	0.08	9.14×10^{-16}	−11.322	ppk19	0.08	8.77×10^{-16}
6	CG43149	0.32	1.43×10^{-9}	11.261	CG43149	0.32	1.37×10^{-9}
7	lncRNA:CR45502	0.29	6.75×10^{-15}	−11.216	lncRNA:CR45502	0.29	6.55×10^{-15}
8	CheA7a	0.15	2.22×10^{-11}	−11.154	CheA7a	0.15	2.17×10^{-11}
9	lncRNA:CR44377	0.01	1.04×10^{-9}	−9.419	lncRNA:CR44377	0.01	1.00×10^{-9}
10	CG14563	0.07	1.21×10^{-10}	−9.345	CG14563	0.07	1.16×10^{-10}

Top 10 highest expressed genes with significant change

Ranking by DESeq2-based expression (ie, basemean scaled by gene length, in units of standard reads per base)

Table 69. Top Ten Highest Expressed Genes with Significant (pa
Difference

group-housed 47b mutants and wildtypes (simultaneous model)

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 Fold
1	Obp28a	58.19	3.93×10^{-3}	0.485	Obp28a	58.20	3.76×10^{-3}	
2	Drsl5	32.91	2.38×10^{-6}	−0.923	Drsl5	32.91	2.29×10^{-6}	
3	lncRNA:noe	32.02	4.79×10^{-7}	−0.863	lncRNA:noe	32.03	5.07×10^{-7}	
4	to	30.97	1.93×10^{-34}	−2.188	to	30.97	1.32×10^{-34}	
5	Est-6	24.00	5.12×10^{-3}	0.401	Est-6	24.00	4.81×10^{-3}	
6	lush	23.82	6.22×10^{-3}	0.516	lush	23.83	5.77×10^{-3}	
7	CG11550	21.31	2.01×10^{-4}	0.598	CG11550	21.31	1.76×10^{-4}	
8	Obp59a	16.26	1.44×10^{-11}	0.559	Obp59a	16.26	8.52×10^{-12}	
9	CG30197	12.24	9.03×10^{-11}	1.101	CG30197	12.24	7.41×10^{-11}	
10	CG43093	12.13	8.00×10^{-12}	−0.871	CG43093	12.13	6.95×10^{-12}	

3.4.4.2 Gene Ontology Enrichment

Genes were analyzed for GO Term Enrichment using topGO, using Fisher's test applied to those whose expression difference passed a significance threshold ($p < 0.01$), and applying the Kolmogorov-Smirnov test using p-values as scores.

-> check consistency between GO terms between alignment strategies -> filter out very broad/very specific terms?

Correlation between significance values for the two tests

```
## png
##   2
```

```
molecular transducer activity (GO:0060089)
sensory perception (GO:0007600)
system process (GO:0003008)
DNA packaging complex (GO:0044815)
obsolete membrane part (GO:0044425)
```

Table 71. Enriched GO Terms among Significantly Differentially Expressed Genes
simultaneous 47b contrast; multi only; top 10 most significant per category

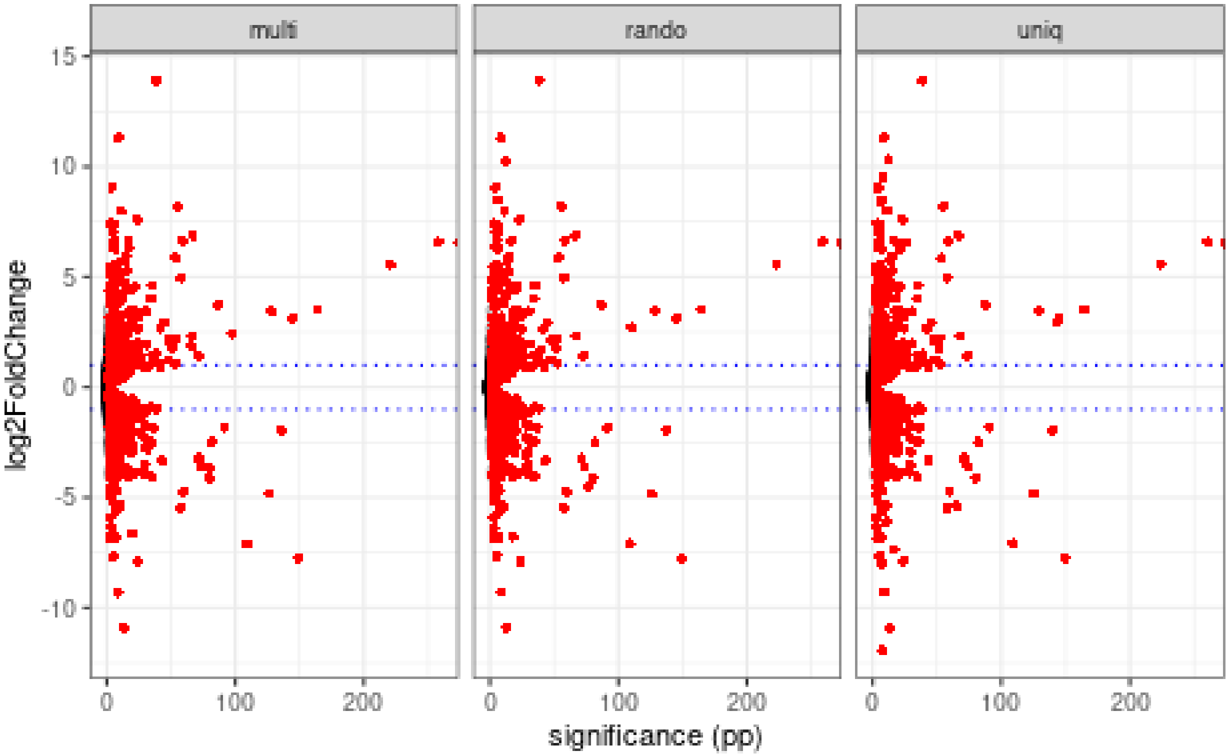
		p-value	
GO Term	Description	Fisher	K-S
MF			
GO:0005549	odorant binding	6.20×10^{-9}	6.10×10^{-8}
GO:0046982	protein heterodimerization activity	1.80×10^{-8}	3.30×10^{-7}
GO:0004888	transmembrane signaling receptor activity	1.40×10^{-7}	2.10×10^{-5}
GO:0004984	olfactory receptor activity	8.00×10^{-7}	1.30×10^{-5}
GO:0005506	iron ion binding	2.80×10^{-6}	7.90×10^{-4}
GO:0038023	signaling receptor activity	4.80×10^{-6}	8.90×10^{-7}
GO:0060089	NA	4.80×10^{-6}	8.90×10^{-7}
GO:0046983	protein dimerization activity	9.60×10^{-6}	1.80×10^{-4}
GO:0031492	nucleosomal DNA binding	3.60×10^{-5}	2.90×10^{-7}
GO:0046873	metal ion transmembrane transporter activity	9.90×10^{-5}	1.80×10^{-4}
BP			
GO:0007606	sensory perception of chemical stimulus	4.60×10^{-12}	1.80×10^{-12}
GO:0007600	NA	1.00×10^{-11}	1.70×10^{-12}
GO:0007608	sensory perception of smell	3.80×10^{-9}	7.70×10^{-9}
GO:0050907	detection of chemical stimulus involved in sensory perception	1.30×10^{-7}	2.10×10^{-6}
GO:0050906	detection of stimulus involved in sensory perception	1.80×10^{-7}	3.20×10^{-6}
GO:0050896	response to stimulus	2.10×10^{-7}	5.00×10^{-7}
GO:0009593	detection of chemical stimulus	3.10×10^{-7}	8.80×10^{-6}
GO:0050877	nervous system process	4.90×10^{-7}	2.20×10^{-7}
GO:0003008	NA	6.40×10^{-7}	1.70×10^{-8}
GO:0042221	response to chemical	2.50×10^{-6}	5.10×10^{-4}
CC			
GO:0071944	cell periphery	4.60×10^{-14}	9.00×10^{-11}
GO:0031224	intrinsic component of membrane	2.20×10^{-12}	1.60×10^{-7}
GO:0016021	integral component of membrane	3.20×10^{-12}	2.60×10^{-7}
GO:0000786	nucleosome	6.90×10^{-12}	1.20×10^{-10}
GO:0044815	NA	3.30×10^{-11}	8.10×10^{-10}
GO:0005886	plasma membrane	3.50×10^{-11}	3.10×10^{-8}
GO:0032993	protein-DNA complex	3.50×10^{-9}	2.50×10^{-8}
GO:0005576	extracellular region	3.90×10^{-9}	2.00×10^{-5}
GO:0016020	membrane	1.20×10^{-8}	1.90×10^{-6}
GO:0031226	intrinsic component of plasma membrane	4.50×10^{-8}	2.00×10^{-6}

3.4.5 67d vs wt (Simultaneous model)

Of the 11778 genes with significance scores available, 1355 have an adjusted $p < 0.01$ (11.5044999 %)

Here is a volcano plot for the three alignment strategies, with significance on the horizontal axis and log2 fold change on the vertical. Significant (padj<0.01) differences are highlighted in red. Dashed blue guidelines mark a log2 fold change of +/-1 (ie, a difference in expression of a factor of 2). Genes with negative log2 fold changes are depleted relative to the group-housed condition; positive fold changes are enriched

Figure 57. Volcano Plot: Fold Change vs. Significance
(between group-housed 67d mutants and wildtypes, simultaneous model)



```
## png
##   2
```

From the volcano plots, we can pull out genes with large (ie, a fold change greater than 2 or less than 1/2), significant (ie, padj < 0.01) changes. There were 600 such genes, mostly shared across alignment strategy:

(Table available at $results/tables/tbl72_hausWtVsMut_genotype67d_chonky.html$)

3.4.5.1 Top Tens

Genes with top 10 most significant changes

Ordered in decreasing significance, the alignemnt strategies agree on the top 10 most significant changes:

Table 73. Top Ten Most Significantly (padj<0.01) Differentially Exp
group-housed 67d mutants and wildtypes (simultaneous model)

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldCh
1	l(2)03659	0.33	0.00	6.537	l(2)03659	0.33	0.00	
2	CG7900	3.51	1.24×10^{-258}	6.571	CG7900	3.51	1.56×10^{-259}	
3	CG6912	0.73	2.30×10^{-221}	5.569	CG6912	0.73	1.42×10^{-223}	

4	CG10936	0.09	1.15×10^{-164}	3.522	CG10936	0.09	4.83×10^{-165}	
5	5-HT2A	0.19	5.64×10^{-150}	−7.763	5-HT2A	0.19	5.04×10^{-150}	−
6	Cyp9b1	0.89	1.45×10^{-145}	3.123	Cyp9b1	0.89	5.81×10^{-146}	
7	NijC	3.06	2.27×10^{-136}	−1.955	NijC	3.06	2.69×10^{-138}	−
8	CG32407	0.27	5.75×10^{-129}	3.453	CG32407	0.27	2.24×10^{-129}	
9	DIP-alpha	0.12	5.85×10^{-127}	−4.820	DIP-alpha	0.12	8.72×10^{-127}	−
10	Or67d	1.06	7.62×10^{-110}	−7.105	CG32641	2.99	5.01×10^{-111}	

Top 10 genes with biggest (significant) effect sizes

Table 74. Top Ten Largest Magnitude Fold Changes which v

group-housed 67d mutants and wildtypes (simultaneous model)

	multi				rando		
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p
1	w	1.01	2.92×10^{-39}	13.911	w	1.01	2.82×10^{-39}
2	CG43149	0.32	1.11×10^{-9}	11.303	CG43149	0.32	1.11×10^{-9}
3	lncRNA:CR44111	0.12	7.19×10^{-14}	−10.906	lncRNA:CR44111	0.12	7.05×10^{-14}
4	lncRNA:CR44377	0.01	1.50×10^{-9}	−9.305	CG43291	0.02	6.16×10^{-13}
5	ppk9	0.01	4.95×10^{-5}	9.052	lncRNA:CR44377	0.01	1.48×10^{-9}
6	lncRNA:dntRL	0.13	5.34×10^{-56}	8.199	ppk9	0.01	5.21×10^{-5}
7	Obp83g	0.07	5.63×10^{-12}	7.998	His-Psi:CR31614	0.01	8.18×10^{-7}
8	CG9010	0.12	9.29×10^{-25}	−7.907	lncRNA:dntRL	0.13	5.03×10^{-56}
9	5-HT2A	0.19	5.64×10^{-150}	−7.763	Obp83g	0.07	5.39×10^{-12}
10	c-cup	0.01	1.13×10^{-5}	−7.682	CG9010	0.12	9.26×10^{-25}

Top 10 highest expressed genes with significant change

Ranking by DESeq2-based expression (ie, basemean scaled by gene length, in units of standard reads per base)

Table 75. Top Ten Highest Expressed Genes with Significant (padj
Difference

group-housed 67d mutants and wildtypes (simultaneous model)

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldChang
1	Obp83b	117.12	1.57×10^{-3}	0.396	Obp83b	117.13	1.49×10^{-3}	0.39
2	Obp69a	41.49	1.11×10^{-6}	0.658	Obp69a	41.49	1.16×10^{-6}	0.65
3	lush	23.82	8.48×10^{-6}	0.778	lush	23.83	7.77×10^{-6}	0.77
4	Cyp6w1	22.67	1.78×10^{-4}	0.586	Cyp6w1	22.67	1.65×10^{-4}	0.58
5	Snmp1	19.98	1.66×10^{-3}	−0.376	Snmp1	19.98	1.53×10^{-3}	−0.37
6	Obp56d	18.04	2.52×10^{-3}	0.745	Obp56d	18.05	2.46×10^{-3}	0.74
7	CG1927	16.43	1.45×10^{-3}	0.333	CG1927	16.43	1.47×10^{-3}	0.33
8	Ldsdh1	12.31	1.77×10^{-3}	0.486	Ldsdh1	12.31	1.66×10^{-3}	0.48
9	CG30197	12.24	1.00×10^{-9}	1.047	CG30197	12.24	8.88×10^{-10}	1.04
10	Cyp6a2	11.95	3.99×10^{-21}	2.745	Cyp6a2	11.95	3.37×10^{-21}	2.74

3.4.5.2 Gene Ontology Enrichment

Genes were analyzed for GO Term Enrichment using topGO, using Fisher's test applied to those whose expression difference passed a significance threshold ($p < 0.01$), and applying the Kolmogorov-Smirnov test using p-values as scores.

-> check consistency between GO terms between alignment strategies -> filter out very broad/very specific terms?

Correlation between significance values for the two tests

```
## png
##   2
```

```
tetrapyrrole binding (GO:0046906)
sensory perception (GO:0007600)
detection of stimulus (GO:0051606)
cell projection membrane (GO:0031253)
obsolete plasma membrane part (GO:0044459)
leading edge membrane (GO:0031256)
obsolete membrane part (GO:0044425)
```

Table 75. Enriched GO Terms among Significantly Differentially Expressed Genes

simultaneous 67d contrast; multi only; top 10 most significant per category

GO Term	Description	
MF		
GO:0004984	olfactory receptor activity	7.20
GO:0005549	odorant binding	5.70
GO:0020037	heme binding	1.3
GO:0046906	NA	1.6
GO:0005506	iron ion binding	3.2
GO:0005215	transporter activity	5.1
GO:0016491	oxidoreductase activity	9.2
GO:0022857	transmembrane transporter activity	1.2
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	1.9
GO:0015157	NA	4.5
BP		
GO:0050907	detection of chemical stimulus involved in sensory perception	4.40
GO:0007608	sensory perception of smell	1.30
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	7.40
GO:0050906	detection of stimulus involved in sensory perception	1.60
GO:0009593	detection of chemical stimulus	2.90
GO:0007600	NA	6.60
GO:0007606	sensory perception of chemical stimulus	1.4
GO:0050896	response to stimulus	2.3
GO:0051606	NA	3.0
GO:0044782	cilium organization	7.1
CC		
GO:0071944	cell periphery	6.60
GO:0005886	plasma membrane	7.00

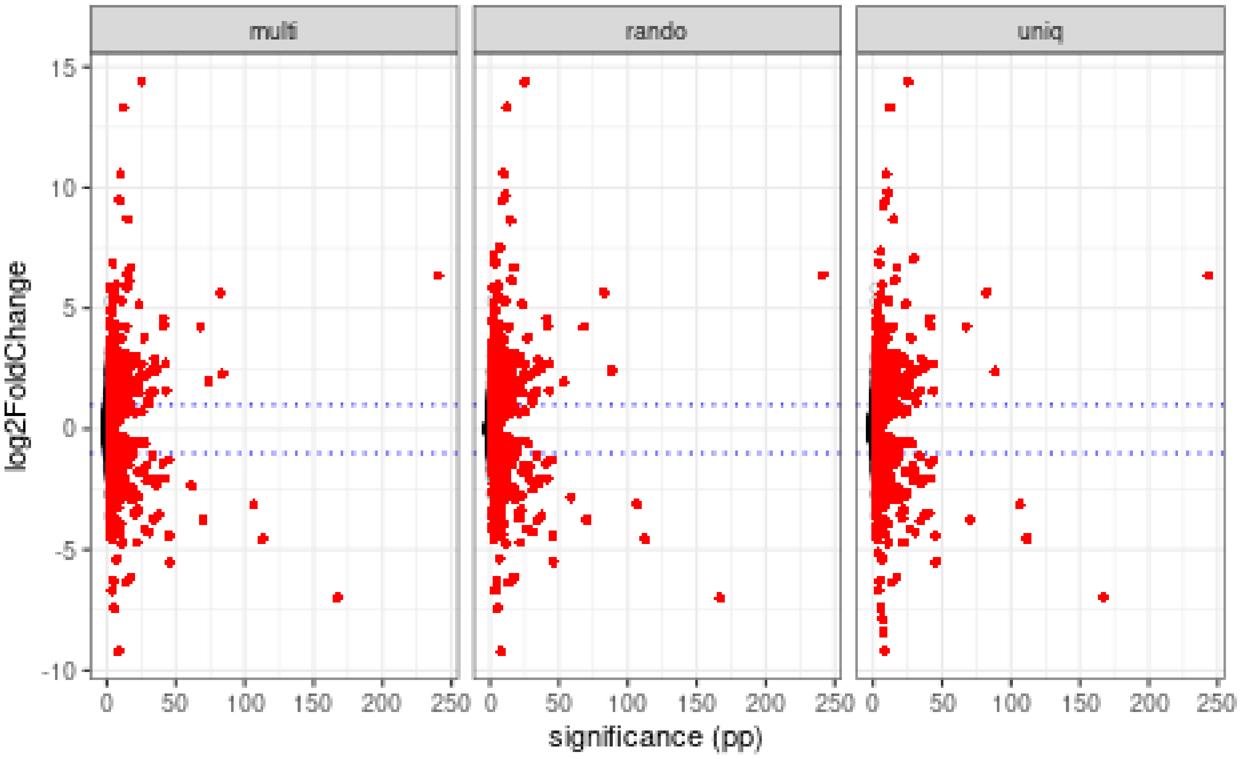
GO:0016020	membrane	2.0
GO:0016021	integral component of membrane	2.7
GO:0031224	intrinsic component of membrane	4.4
GO:0032590	dendrite membrane	2.6
GO:0042995	cell projection	4.2
GO:0005929	cilium	6.3
GO:0120025	plasma membrane bounded cell projection	8.4
GO:0031253	NA	1.9

3.4.6 FruLexA/Fru440 vs wt (Simultaneous model)

Of the 11779 genes with significance scores available, 1856 have an adjusted p $<$ 0.01 (15.7568554 %)

Here is a volcano plot for the three alignment strategies, with significance on the horizontal axis and log2 fold change on the vertical. Significant (padj<0.01) differences are highlighted in red. Dashed blue guidelines mark a log2 fold change of +/-1 (ie, a difference in expression of a factor of 2). Genes with negative log2 fold changes are depleted relative to the group-housed condition; positive fold changes are enriched

Figure 59. Volcano Plot: Fold Change vs. Significance
(between group-housed FruLexaFru440 mutants and wildtypes, simultaneo



```
## png
##   2
```

From the volcano plots, we can pull out genes with large (ie, a fold change greater than 2 or less than 1/2), significant (ie, padj $<$ 0.01) changes. There were 574 such genes, mostly shared across alignment strategy:

(Table available at $results/tables/tbl77_{h}ausWtVsMut_{g}enotypeFruLexa440_{c}honky.html$)

3.4.6.1 Top Tens

Genes with top 10 most significant changes

Ordered in decreasing significance, the alignemnt strategies agree on the top 10 most significant changes:

Table 78. Top Ten Most Significantly (padj<0.01) Differentially Exp
between isolated and group-housed wildtypes

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldCh
1	CG7900	3.51	3.37×10^{-241}	6.352	CG7900	3.51	3.68×10^{-242}	
2	5-HT2A	0.19	9.37×10^{-168}	−6.989	5-HT2A	0.19	9.30×10^{-168}	−
3	DIP-alpha	0.12	1.35×10^{-113}	−4.547	DIP-alpha	0.12	2.40×10^{-113}	−
4	Ets21C	0.13	3.78×10^{-107}	−3.130	Ets21C	0.13	2.03×10^{-107}	−
5	CG32641	4.55	1.12×10^{-84}	2.270	CG32641	2.99	5.00×10^{-89}	
6	CG11893	0.28	4.52×10^{-83}	5.643	CG11893	0.28	3.04×10^{-83}	
7	CG32640	5.44	3.06×10^{-74}	1.979	prom	0.07	6.18×10^{-71}	−
8	prom	0.07	1.50×10^{-70}	−3.772	CG42526	0.08	9.01×10^{-69}	
9	CG42526	0.08	1.03×10^{-68}	4.237	Or19b	0.55	1.86×10^{-59}	−
10	Or19b	0.74	4.91×10^{-62}	−2.349	CG32640	2.21	4.47×10^{-54}	

Top 10 genes with biggest (significant) effect sizes

Table 79. Top Ten Largest Magnitude
between group-housed FruLexaFru440 mutants ar

	multi						
rank	name	FB ID	expression	adjusted p	log2 FoldChange	name	FB ID
1	mthl8	FBgn0052475	0.13	4.52×10^{-26}	14.398	mthl8	FBgn00
2	CG43149	FBgn0262679	0.32	5.92×10^{-13}	13.321	CG43149	FBgn02
3	CG9287	FBgn0032057	0.01	1.99×10^{-10}	10.562	CG9287	FBgn00
4	ppk27	FBgn0035458	0.01	9.86×10^{-10}	9.460	CG43291	FBgn02
5	lncRNA:CR44377	FBgn0265527	0.01	3.89×10^{-9}	−9.202	ppk27	FBgn00
6	w	FBgn0003996	1.01	1.05×10^{-15}	8.679	lncRNA:CR44377	FBgn02
7	CG18577	FBgn0037870	0.01	3.86×10^{-6}	−7.426	w	FBgn00
8	5-HT2A	FBgn0087012	0.19	9.37×10^{-168}	−6.989	CR45496	FBgn02
9	lncRNA:CR44285	FBgn0265312	0.05	8.39×10^{-5}	6.864	CG18577	FBgn00
10	tRNA:Gly-GCC-1-8	FBgn0011867	0.13	1.48×10^{-4}	−6.718	lncRNA:CR46123	FBgn02

Table 79. Top Ten Largest Magnitude Fold Changes which
between group-housed FruLexaFru440 mutants and wildtypes, simultaneous

	multi				rando		
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted
1	mthl8	0.13	4.52×10^{-26}	14.398	mthl8	0.13	$4.39 \times 10^{-}$
2	CG43149	0.32	5.92×10^{-13}	13.321	CG43149	0.32	$6.09 \times 10^{-}$
3	CG9287	0.01	1.99×10^{-10}	10.562	CG9287	0.01	$1.83 \times 10^{-}$
4	ppk27	0.01	9.86×10^{-10}	9.460	CG43291	0.02	$1.47 \times 10^{-}$
5	lncRNA:CR44377	0.01	3.89×10^{-9}	−9.202	ppk27	0.01	$9.84 \times 10^{-}$

6	w	1.01	1.05×10^{-15}	8.679	lncRNA:CR44377	0.01	3.91×10
7	CG18577	0.01	3.86×10^{-6}	−7.426	w	1.01	$1.02 \times 10^{-}$
8	5-HT2A	0.19	9.37×10^{-168}	−6.989	CR45496	0.04	8.05×10
9	lncRNA:CR44285	0.05	8.39×10^{-5}	6.864	CG18577	0.01	3.75×10
10	tRNA:Gly-GCC-1-8	0.13	1.48×10^{-4}	−6.718	lncRNA:CR46123	0.02	2.47×10

Top 10 highest expressed genes with significant change

Ranking by DESeq2-based expression (ie, basemean scaled by gene length, in units of standard reads per base)

Table 80. Top Ten Highest Expressed Genes with Significant (padj <
Difference

between group-housed FruLexaFru440 mutants and wildtypes, simultaneous model

	multi				rando			
rank	name	expression	adjusted p	log2 FoldChange	name	expression	adjusted p	log2 FoldChange
1	Obp83b	117.12	7.76×10^{-4}	0.410	Obp83b	117.13	6.98×10^{-4}	0.412
2	Obp19d	99.12	2.72×10^{-3}	0.400	Obp19d	99.14	2.58×10^{-3}	0.401
3	Jhedup	55.34	4.81×10^{-3}	−0.457	Jhedup	55.34	4.78×10^{-3}	−0.469
4	Obp69a	41.49	7.67×10^{-4}	0.467	Obp69a	41.49	7.42×10^{-4}	0.469
5	Orco	25.74	7.43×10^{-4}	−0.594	Orco	25.74	7.18×10^{-4}	−0.597
6	Obp56d	18.04	5.86×10^{-7}	1.198	lush	23.83	9.57×10^{-3}	0.467
7	sesB	16.90	3.45×10^{-4}	0.379	Obp56d	18.05	5.45×10^{-7}	1.200
8	CG1927	16.43	9.40×10^{-3}	0.271	sesB	16.91	3.44×10^{-4}	0.375
9	Obp59a	16.26	1.11×10^{-5}	0.378	CG1927	16.43	8.91×10^{-3}	0.274
10	CG6908	13.64	6.54×10^{-3}	0.500	Obp59a	16.26	8.41×10^{-6}	0.380

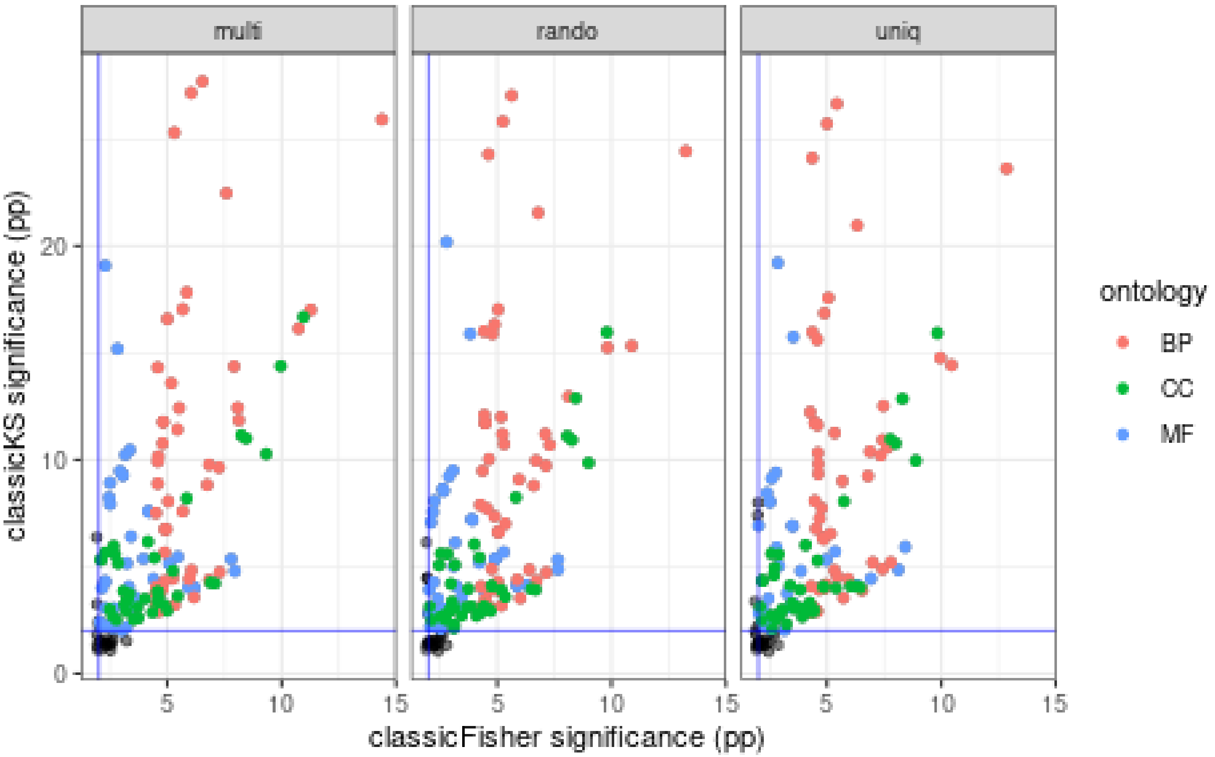
3.4.6.2 Gene Ontology Enrichment

Genes were analyzed for GO Term Enrichment using topGO, using Fisher's test applied to those whose expression difference passed a significance threshold ($p < 0.01$), and applying the Kolmogorov-Smirnov test using p-values as scores.

-> check consistency between GO terms between alignment strategies -> filter out very broad/very specific terms?

Correlation between significance values for the two tests

Figure 60. Scatterplot of GO Term Enrichment Significance for Two Tests (FruLexa/Fru440 contrast from simultaneous model)



```
## png
##   2
```

```
molecular transducer activity (GO:0060089)
mannosyl-oligosaccharide mannosidase activity (GO:0015924)
adenyl nucleotide binding (GO:0030554)
response to chemical (GO:0042221)
transport (GO:0006810)
establishment of localization (GO:0051234)
cellular response to stimulus (GO:0051716)
plasma membrane bounded cell projection (GO:0120025)
obsolete cell projection part (GO:0044463)
obsolete plasma membrane bounded cell projection part (GO:0120038)
obsolete plasma membrane part (GO:0044459)
```

Table 81. Enriched GO Terms among Significantly Differentially Expressed Genes
simultaneous FruLexa440 contrast; multi only; top 10 most significant per category

		p-value	
GO Term	Description	Fisher	K-S
MF			
GO:0004888	transmembrane signaling receptor activity	1.10×10^{-8}	1.60×10^{-5}
GO:0004984	olfactory receptor activity	1.60×10^{-8}	4.40×10^{-6}

GO:0038023	signaling receptor activity	6.70×10^{-7}	9.70×10^{-5}
GO:0060089	NA	6.70×10^{-7}	9.70×10^{-5}
GO:0005549	odorant binding	1.20×10^{-6}	9.20×10^{-5}
GO:0005096	GTPase activator activity	3.40×10^{-6}	3.50×10^{-6}
GO:0030695	GTPase regulator activity	8.30×10^{-6}	6.80×10^{-6}
GO:0060589	NA	8.30×10^{-6}	6.80×10^{-6}
GO:0008092	cytoskeletal protein binding	4.00×10^{-5}	3.80×10^{-5}
GO:0030554	NA	6.10×10^{-5}	2.80×10^{-8}
BP			
GO:0050896	response to stimulus	4.10×10^{-15}	1.20×10^{-26}
GO:0051179	localization	5.30×10^{-12}	9.90×10^{-18}
GO:0042221	response to chemical	1.80×10^{-11}	7.40×10^{-17}
GO:0051234	establishment of localization	7.50×10^{-9}	1.40×10^{-12}
GO:0006810	NA	8.30×10^{-9}	3.70×10^{-13}
GO:0010033	response to organic substance	1.20×10^{-8}	4.30×10^{-15}
GO:0051716	cellular response to stimulus	2.60×10^{-8}	3.40×10^{-23}
GO:0050907	detection of chemical stimulus involved in sensory perception	5.10×10^{-8}	1.80×10^{-5}
GO:0010970	transport along microtubule	5.40×10^{-8}	2.30×10^{-10}
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	1.50×10^{-7}	4.20×10^{-5}
CC			
GO:0005886	plasma membrane	1.10×10^{-11}	2.10×10^{-17}
GO:0071944	cell periphery	1.10×10^{-10}	4.10×10^{-15}
GO:0016020	membrane	4.70×10^{-10}	5.30×10^{-11}
GO:0042995	cell projection	3.60×10^{-9}	1.00×10^{-11}
GO:0120025	plasma membrane bounded cell projection	5.80×10^{-9}	7.10×10^{-12}
GO:0031224	intrinsic component of membrane	8.40×10^{-8}	6.10×10^{-5}
GO:0016021	integral component of membrane	1.30×10^{-7}	5.60×10^{-5}
GO:0043005	neuron projection	1.40×10^{-6}	6.50×10^{-9}
GO:0032590	dendrite membrane	3.50×10^{-6}	2.40×10^{-4}
GO:0005856	cytoskeleton	5.60×10^{-6}	1.60×10^{-5}

3.4.7 Transcriptional Profiles

Backing away from differential expression, we can also look at transcriptional profiles of treatment groups by gene.

Heatmaps were made representing expression as color intensity; for genes which did were not modeled due to low overall read count, the baseMean-derived expression was filled in as zero. Since these values range over several orders of magnitude, logarithmic scales were uesed; log10(0) was defined as -999

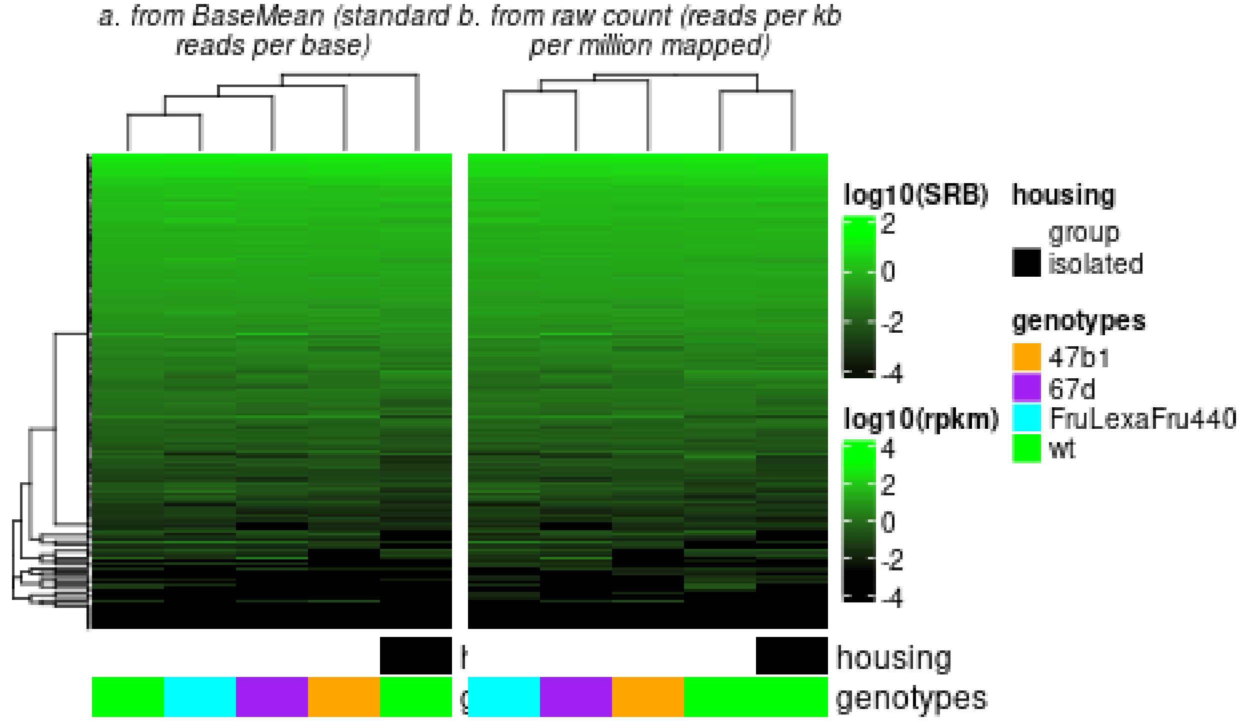
Since the RPKM-derived expression values have more clusterable genes (b/c fewer genes with all 0's), these were used to cluseter both heatmaps. (This doesn't necessarily mean that the finer clustering is meaningful!)

To try to put the expression on a common scale, absolute expression values were scaled on a by-gene basis, with each gene's expression values being divided by the sum of those expressions, to calculate an expression share. Genes with a sum-expression of zero were assigned an expression share of 0 for all treatments. The relative expressions were clustered and heatplots graphed. These are susceptible to low-level noise: a single stray read is enough to mean the difference between all samples having 0% of the reads and one sample having 100% of all reads.

3.4.7.1 ion channel activity genes

Here is a heatmap of transcriptional profiles for ion-channel activity genes (GO:0005216) from the samples in the housing and genotype comparison (Table 5a).

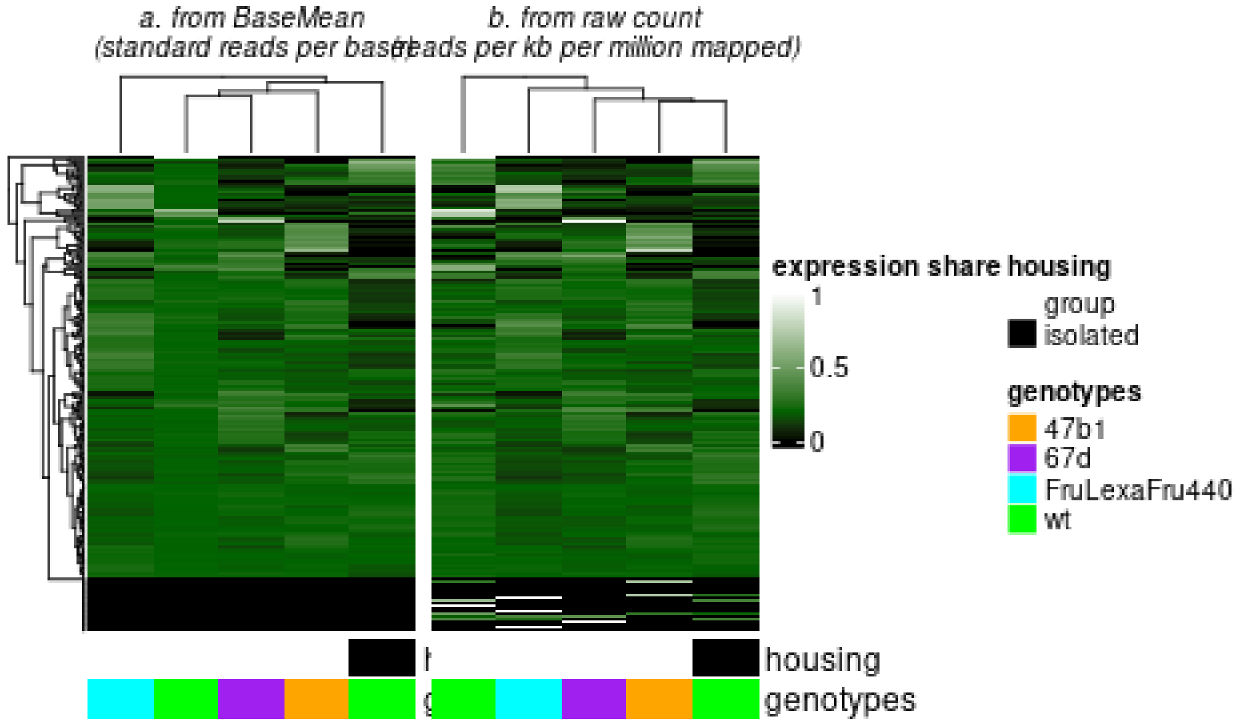
Figure 61 . Absolute Expression Heatmap
for Ion Channel Activity Genes
(simultaneous housing/genotype model, multi only)



```
## png
##   2
```

To try to better display relative differences between samples, a relative expression share was calculated for each gene:

Figure 62 Relative Expression Heatmap for Ion Channel Activity Genes (simultaneous housing/genotype model)



```
## png
##   2
```

3.4.8 Fruitless-less

Out of data quality concerns, the contrast was rerun with the FruLexa/Fru440 samples excluded.

Every gene that can be analyzed using the Without counts can be analyzed using the With counts; there are 0 that cannot. On the other hand, there are 339 genes which can be analyzed using the With counts but not the Without:

Table 82. Genes Lost When FruLexa/Fru440 Counts are Excluded
genes which no longer pass minimum count threshold

aligner	count
multi	318
rando	315
uniq	315

In 0 cases were these genes significant (padj < 0.01) in the With tests.

A gene with significance values in both tests may gain significance when FruLexa/Fru440 samples are dropped, lose significance, maintain significance while switching direction, or remain unchanged. No switches were seen, but moderate numbers (up to ~5%) gained significance.

Table 83. Changes in Differential Expression Significance
when FruLexa/Fru440 samples are dropped

	change		
	gain	loss	none
47b1			
multi	666	18	13243
rando	653	18	13100
uniq	658	15	13020
67d			
multi	1007	15	12905
rando	1000	16	12755
uniq	982	17	12694
isolated			
multi	132	3	13792
rando	133	3	13635
uniq	129	3	13561

In some cases, the significance increase was very large:

Table 84. Top 10 Biggest Significance Changes
when FruLexa/Fru440 samples are dropped

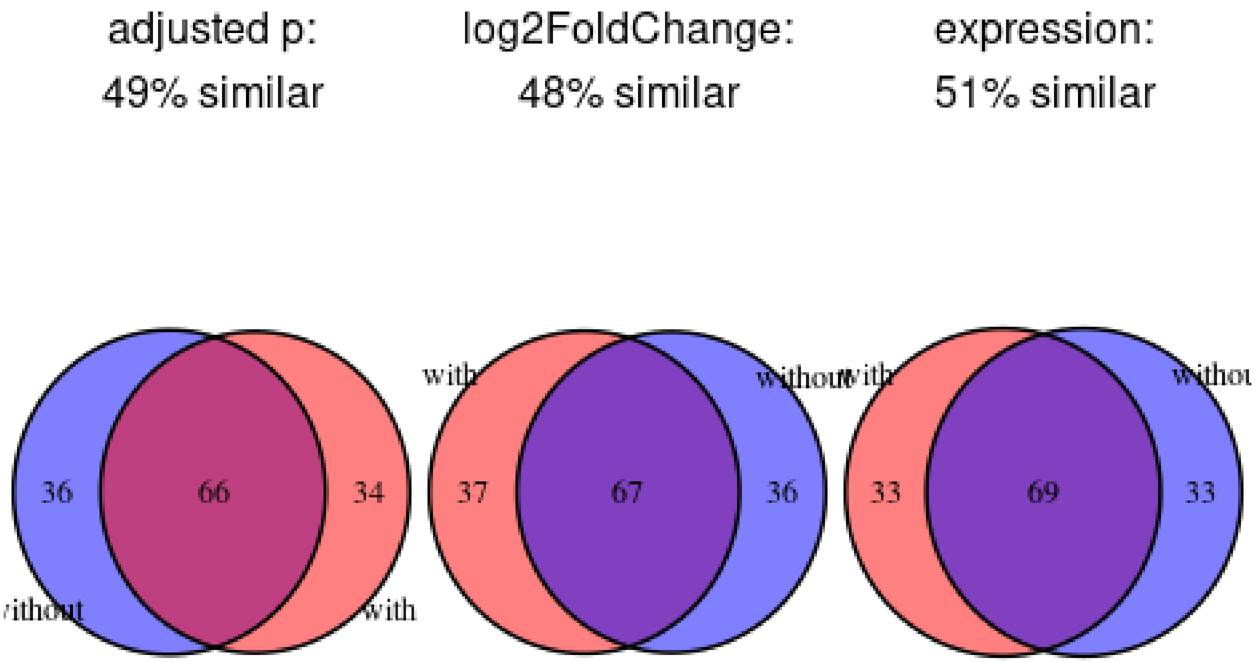
	effect size (l2fc)		adjusted p	
	with	without	with	without
47b1 - multi				
csw	0.56	0.67	0.032	1.26×10^{-16}
kek1	0.78	0.91	0.014	2.49×10^{-16}
CAH1	0.44	0.49	0.041	1.77×10^{-13}
Urod	0.57	0.67	0.036	1.62×10^{-12}
Jheh3	−0.52	−0.59	0.034	4.41×10^{-12}
CG13251	−0.29	−0.30	0.041	2.10×10^{-15}
Nrx-1	0.46	0.58	0.125	7.66×10^{-14}
RIC-3	−0.45	−0.52	0.064	4.85×10^{-12}
pyd	0.55	0.61	0.011	4.02×10^{-23}
Fhos	−0.34	−0.60	0.396	1.02×10^{-16}
67d - multi				
csw	0.55	0.65	0.030	7.77×10^{-16}
Spn	−0.38	−0.41	0.021	1.18×10^{-14}
CG16935	−0.46	−0.57	0.113	1.20×10^{-13}
CG13251	−0.26	−0.27	0.057	4.06×10^{-13}
Sp7	0.34	0.38	0.110	1.20×10^{-12}
CG12814	0.61	0.70	0.016	3.24×10^{-15}
RIC-3	−0.55	−0.61	0.016	1.96×10^{-16}
Ir8a	−0.47	−0.55	0.077	6.41×10^{-14}
pyd	0.41	0.46	0.065	1.75×10^{-13}
Fhos	−0.29	−0.51	0.421	2.27×10^{-12}

isolated - multi				
Trp1	0.06	0.28	0.307	6.69×10^{-6}
CG15202	0.03	0.41	0.311	3.58×10^{-7}
CG9498	0.62	0.74	0.032	4.33×10^{-7}
Ugt301D1	0.32	0.42	0.066	2.50×10^{-7}
Loxl2	0.04	0.45	0.185	2.07×10^{-6}
ELOVL	0.46	0.53	0.017	1.91×10^{-7}
CG9717	0.02	0.52	0.324	2.07×10^{-6}
vir-1	0.15	0.23	0.182	4.64×10^{-6}
CG31459	0.00	0.00	0.228	1.04×10^{-6}
Cda5	−0.03	−0.52	0.191	2.76×10^{-6}

3.4.8.1 Perturbation to Housing Contrast

To see how much exclusion of the FruLexa/Fru440 alters the big picture results in the group vs. isolated contrast, we can look at how well the top-100 lists agree (similarity is calculated as size of the intersection divided by size of the union; lists are pooled across all aligners and thus may have more than 100 unique elements)

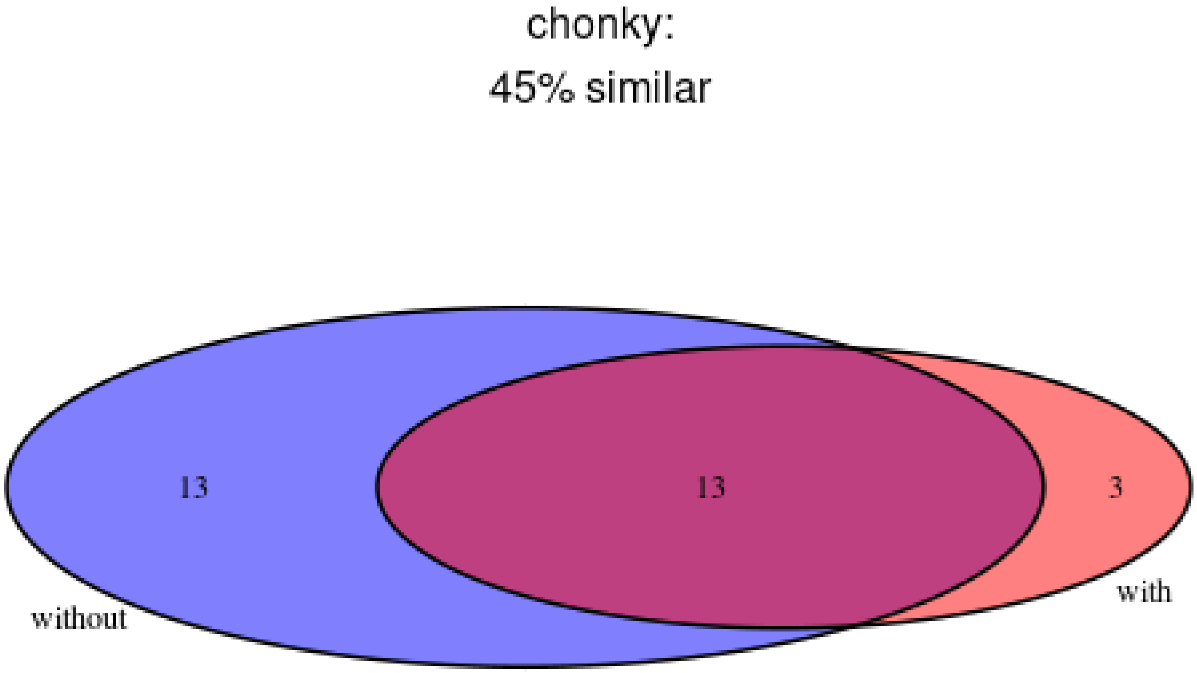
Figure 64 . Similarity of Housing Contrast Top 100 Lists,
with/without FruLexaFru440 samples (pooled alignment strategies)



```
## png
##   2
```

Comparing the “chonky” gene lists (padj < 0.01, abs(l2fc)>1):

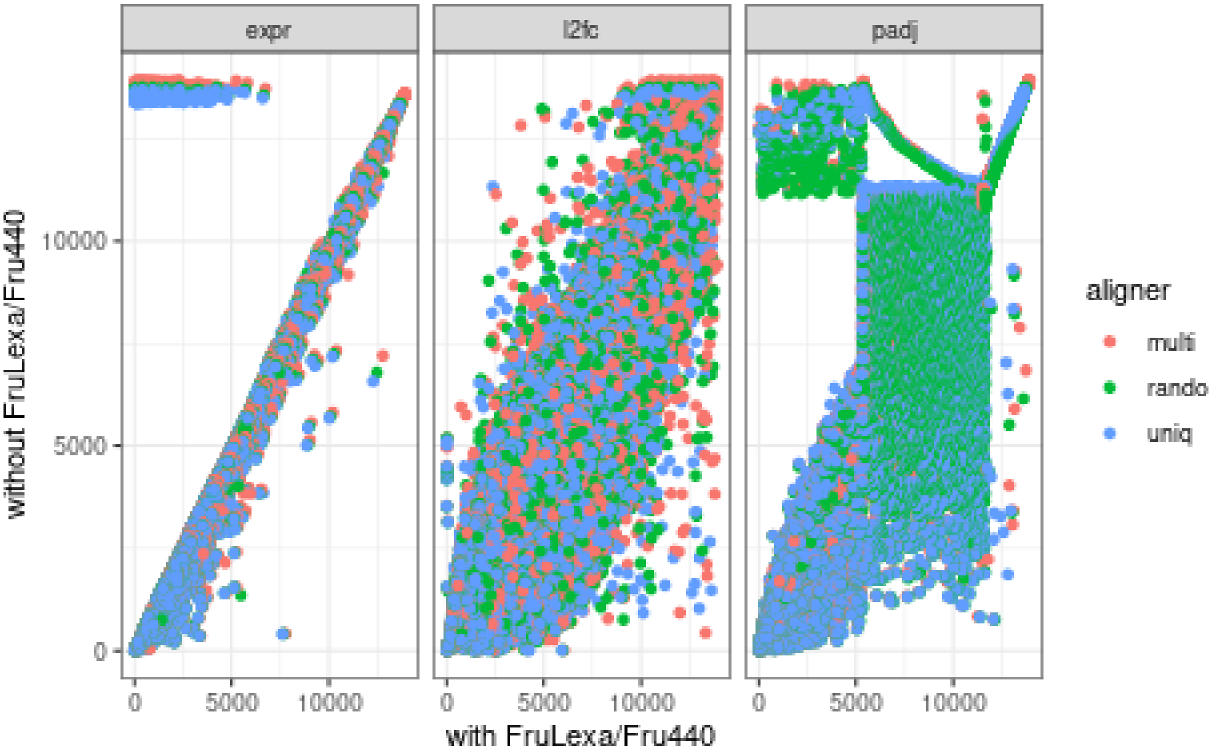
Figure 65 . Similarity of Housing Contrast Chonky Lists Lists,
with/without FruLexaFru440 samples (pooled alignment strategies)



```
## png
##   2
```

We can also look at the rank correlations:

Figure 66. Rank correlations of expression, effect size, and significance (housing contrasts, with/without FruLexa)

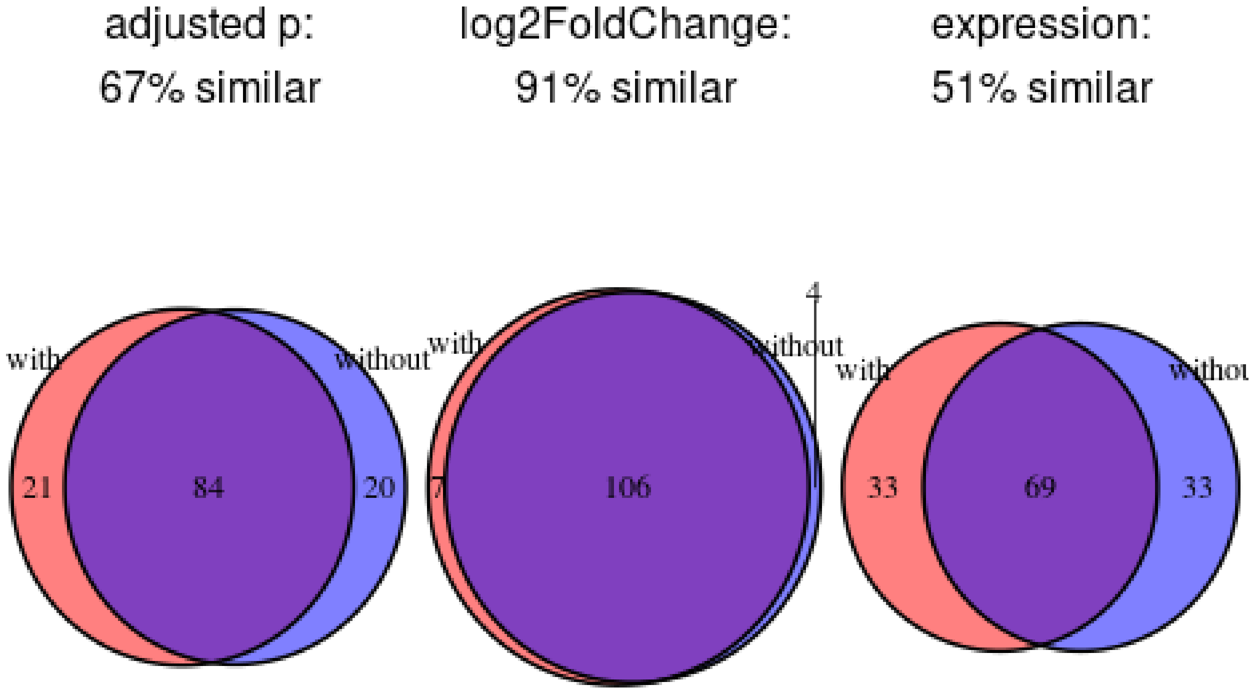


```
## png
##   2
```

3.4.8.2 Perturbation to 47b1 Contrast

To see how much exclusion of the FruLexa/Fru440 alters the big picture results in the 47b1 vs. wt contrast, we can look at how well the top-100 lists agree (similarity is calculated as size of the intersection divided by size of the union; lists are pooled across all aligners and thus may have more than 100 unique elements)

Figure 67 . Similarity of 47b1 Contrast Top 10 Lists,
with/without FruLexaFru440 samples (pooled alignment strategies)



adjusted p: 67% similar

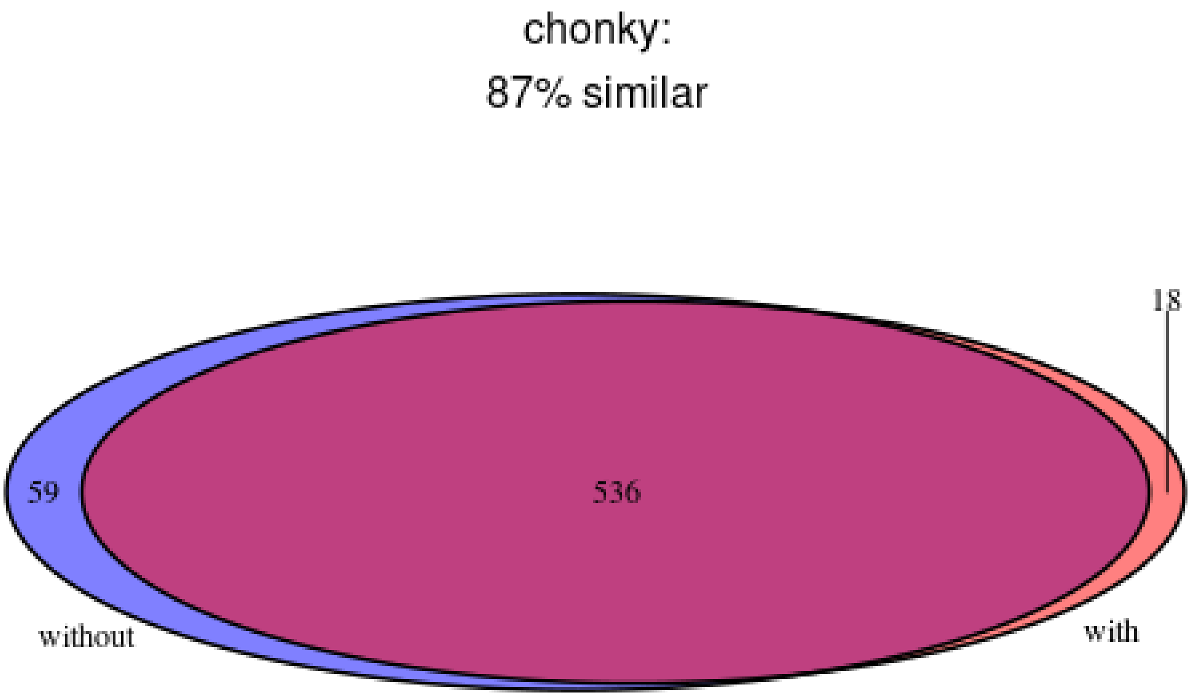
log2FoldChange: 91% similar

expression: 51% similar

```
## png
##   2
```

Comparing the “chonky” gene lists (padj < 0.01, abs(l2fc)>1):

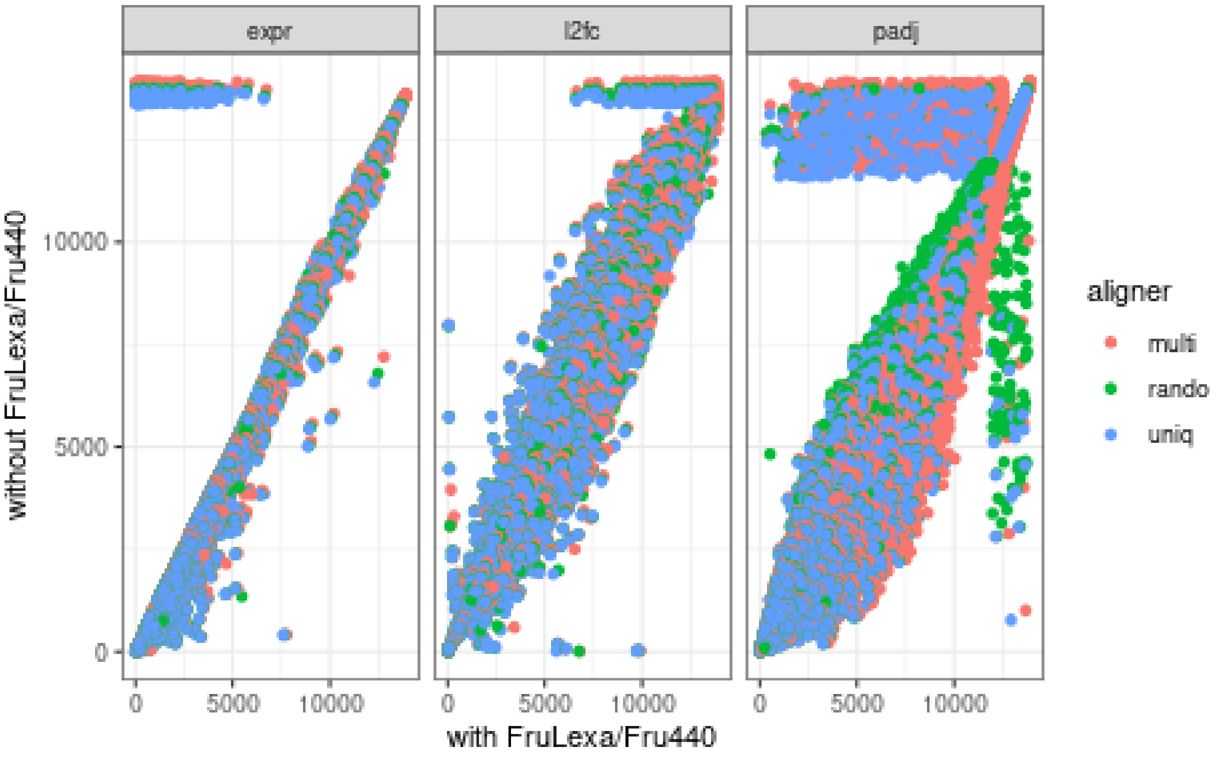
Figure 68 . Similarity of 47b Contrast Chonky Lists Lists,
with/without FruLexaFru440 samples (pooled alignment strategies)



```
## png
##   2
```

We can also look at the rank correlations:

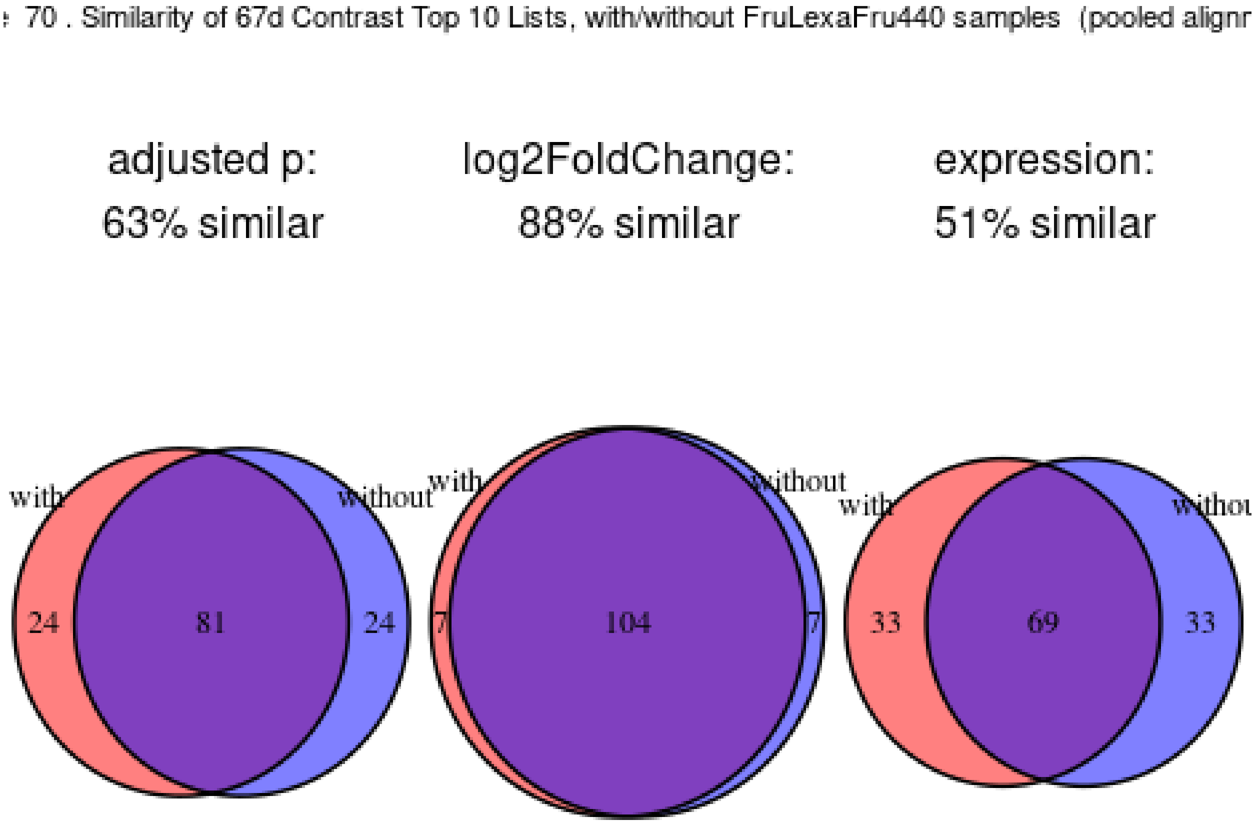
Figure 69. Rank correlations of expression, effect size, and significance (47b1 contrasts, with/without FruLexa)



```
## png
##   2
```

3.4.8.3 Perturbation to 67d Contrast

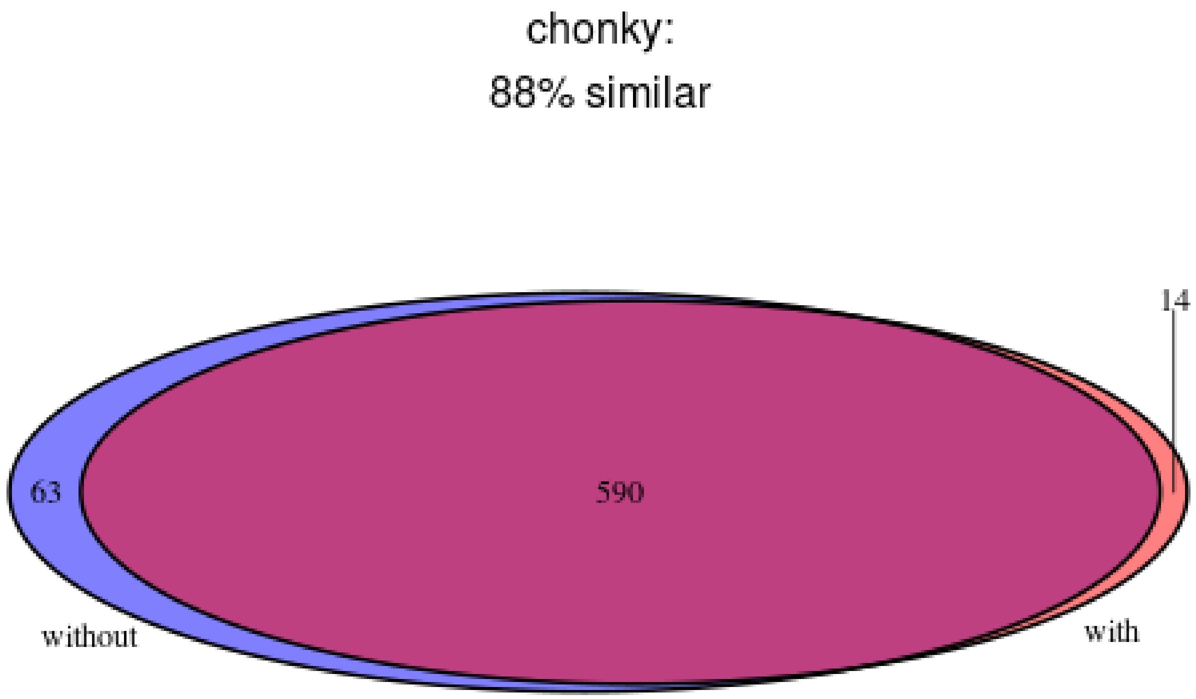
To see how much exclusion of the FruLexa/Fru440 alters the big picture results in the 67d vs. wt contrast, we can look at how well the top-100 lists agree (similarity is calculated as size of the intersection divided by size of the union; lists are pooled across all aligners and thus may have more than 100 unique elements)



```
## png
##   2
```

Comparing the “chonky” gene lists (padj < 0.01, abs(l2fc)>1):

Figure 71 . Similarity of 67d Contrast Chonky Lists Lists,
with/without FruLexaFru440 samples (pooled alignment strategies)

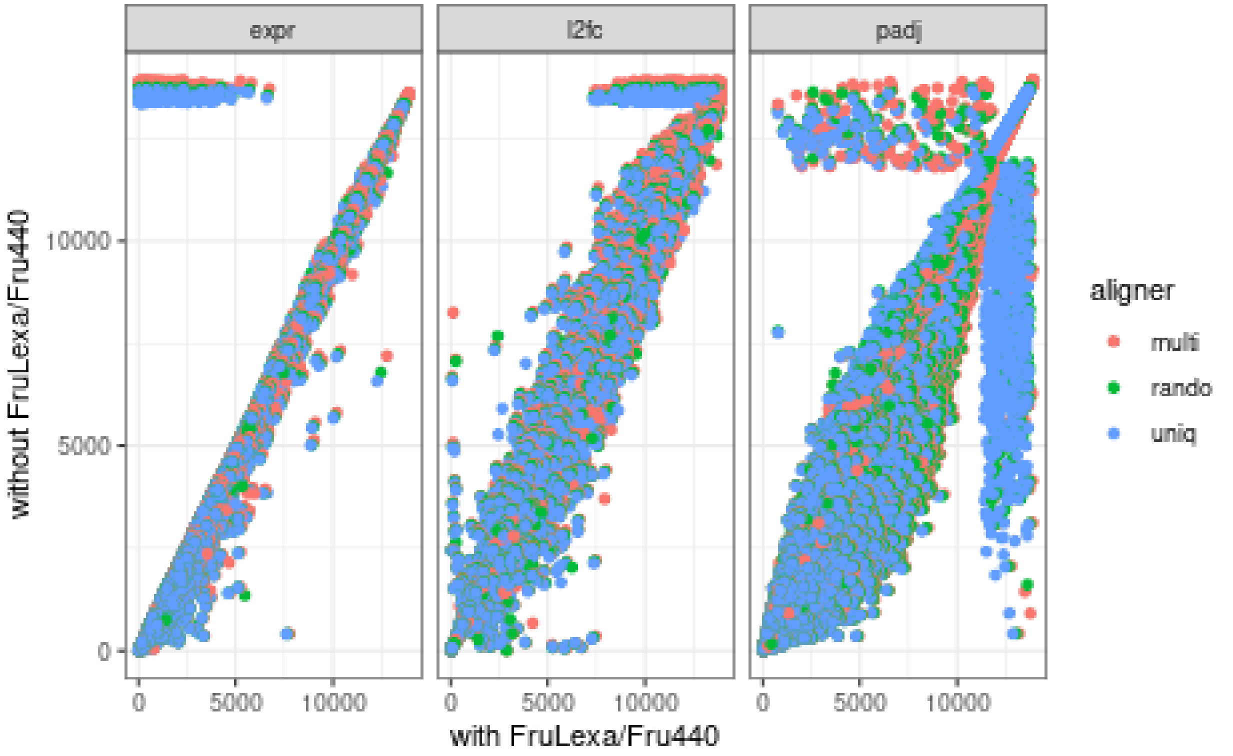


chonky:
88% similar

```
## png
##   2
```

We can also look at the rank correlations:

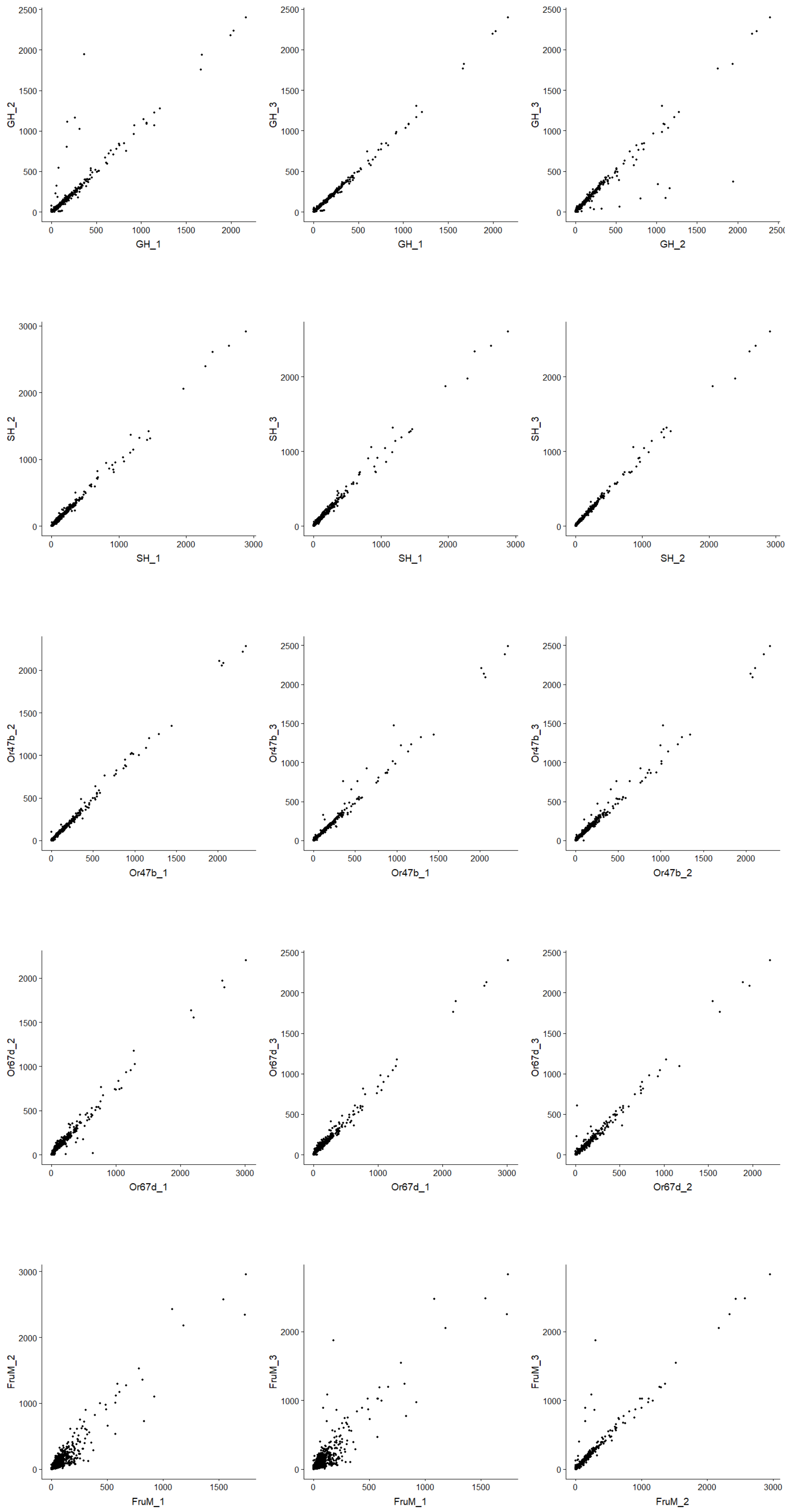
Figure 72. Rank correlations of expression, effect size, and significance (67d contrasts, with/without FruLexa)



```
## png
##   2
```

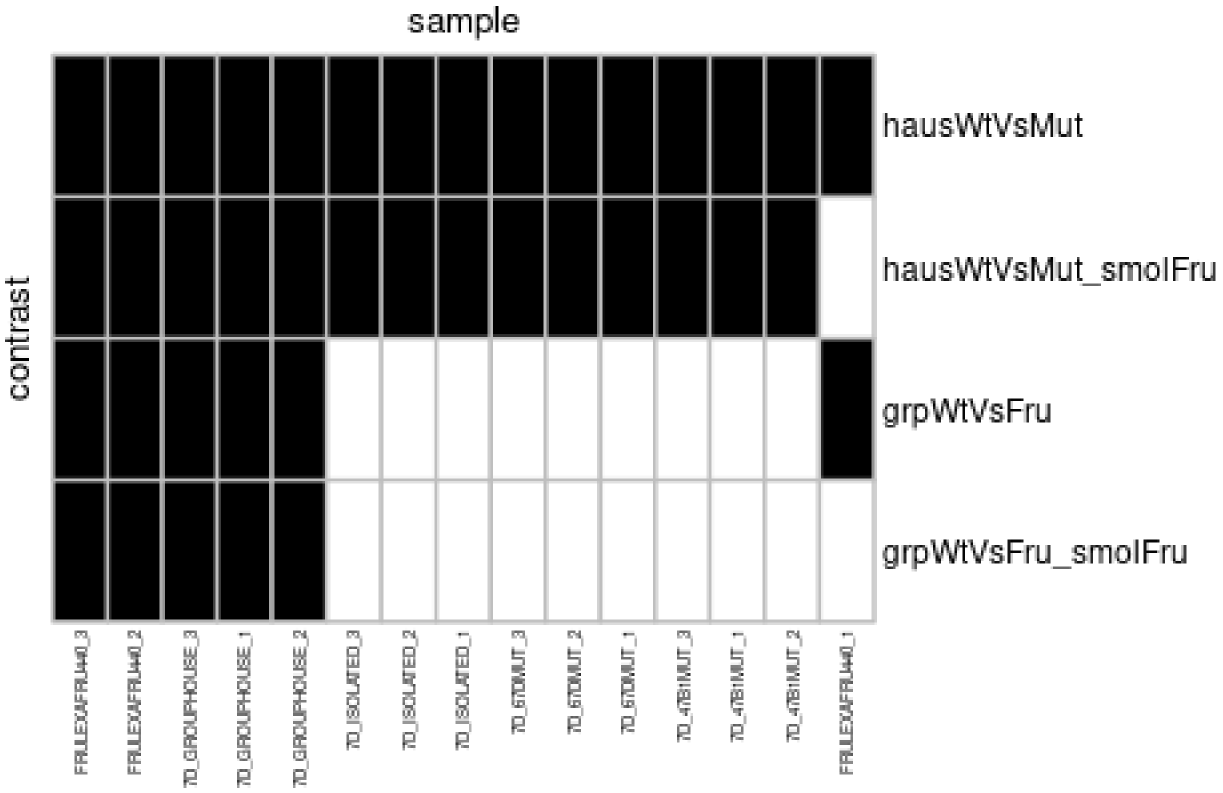
3.4.9 Reduced FruLexaFru440 Samples

FruLexaFru440 replicate 1 has been flagged as specifically problematic, possiby b/c of sex contamination.



source: qichen duan , 21 Dec

Figure 73. RNASeq Samples Used in Reduced Fruitless Comparisons



```
## png
##   2
```

The same pairwise comparisons as above are repeated on a reduced model, in which replicate 1 is excluded. The data are found here: "results/tables/supp/hausWtVsMut_smolFru.allAligners.DESeq2.MpBC.reformatted.tsv"

Every gene that can be analyzed using the Reduced counts can be analyzed using the With counts; there are 0 that cannot. On the other hand, there are 60 genes which can be analyzed using the Full counts but not the Reduced:

Table 85. Genes Lost When FruLexa/Fru440 replicate 1 Counts are Excluded
genes which no longer pass minimum count threshold

aligner	count
multi	58
rando	56
uniq	57

In 0 cases were these genes significant (padj < 0.01) in the With tests.

A gene with significance values in both tests may gain significance when FruLexa/Fru440 samples are dropped, lose significance, maintain significance while switching direction, or remain unchanged. No switches were seen, but moderate numbers (up to ~5%) gained significance.

Table 86. Changes in Differential Expression Significance

when FruLexa/Fru440 replicate 1 is dropped

	change		
	gain	loss	none
47b1			
multi	558	7	13362
rando	558	7	13206
uniq	552	6	13135
67d			
multi	887	3	13037
rando	887	3	12881
uniq	872	3	12818
FruLexaFru440			
multi	849	279	12799
rando	848	286	12637
uniq	843	271	12579
isolated			
multi	116	5	13806
rando	119	5	13647
uniq	120	5	13568

In some cases, the significance increase was very large:

Table 87. Top 10 Biggest Significance Changes

when FruLexa/Fru440 replicate 1 is dropped

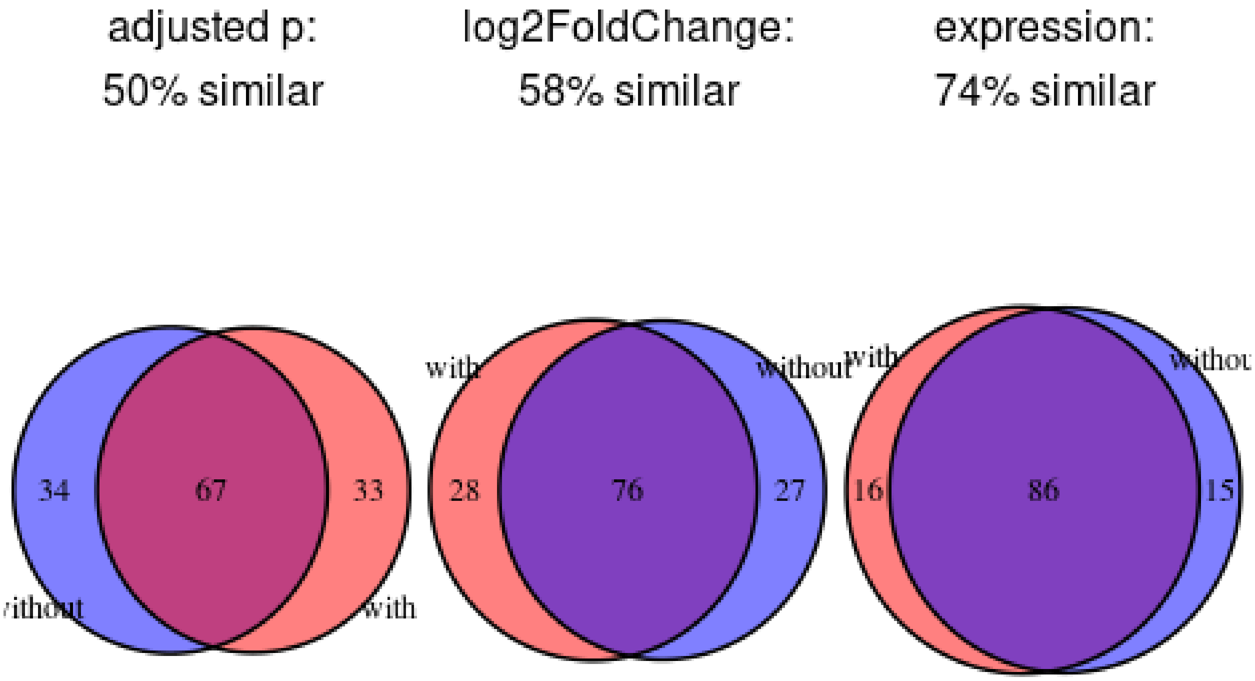
	effect size (l2fc)		adjusted p	
	full	reduced	full	reduced
47b1 - multi				
csw	0.56	0.67	3.23×10^{-2}	1.59×10^{-17}
kek1	0.78	0.91	1.37×10^{-2}	3.68×10^{-17}
CAH1	0.44	0.49	4.07×10^{-2}	6.74×10^{-14}
Urod	0.57	0.66	3.64×10^{-2}	3.11×10^{-12}
CG13251	-0.29	-0.30	4.11×10^{-2}	3.71×10^{-17}
Nrx-1	0.46	0.58	1.25×10^{-1}	1.89×10^{-13}
CG30026	-0.54	-0.62	3.56×10^{-2}	1.11×10^{-11}
RIC-3	-0.45	-0.52	6.37×10^{-2}	4.40×10^{-13}
CG42541	0.76	0.87	1.21×10^{-2}	8.68×10^{-13}
pyd	0.55	0.61	1.05×10^{-2}	2.31×10^{-22}
67d - multi				
csw	0.55	0.65	2.99×10^{-2}	7.65×10^{-17}
Rab10	0.25	0.28	1.74×10^{-1}	2.25×10^{-12}
lqf	0.34	0.38	7.73×10^{-2}	6.32×10^{-12}
CG16935	-0.46	-0.57	1.13×10^{-1}	4.23×10^{-12}
CG13251	-0.26	-0.27	5.72×10^{-2}	2.78×10^{-14}
CG13252	0.43	0.56	1.51×10^{-1}	2.82×10^{-11}

Sp7	0.34	0.39	1.10×10^{-1}	2.91×10^{-11}
CG12814	0.61	0.70	1.64×10^{-2}	4.05×10^{-16}
RIC-3	−0.55	−0.61	1.62×10^{-2}	1.06×10^{-17}
pyd	0.41	0.46	6.47×10^{-2}	3.36×10^{-13}
FruLexaFru440 - multi				
X11L	−0.13	−0.52	5.95×10^{-1}	5.30×10^{-17}
CG15270	−0.36	−0.71	6.86×10^{-2}	1.46×10^{-23}
Pdcd4	−0.19	−0.48	3.26×10^{-1}	1.61×10^{-16}
CG17572	0.47	0.85	5.38×10^{-2}	1.71×10^{-16}
SP2353	−0.28	−0.86	2.73×10^{-1}	4.79×10^{-16}
CG16711	−0.34	−0.58	2.43×10^{-2}	1.22×10^{-23}
CG13251	−0.22	−0.40	9.81×10^{-2}	1.18×10^{-23}
CG34417	−0.25	−0.54	1.56×10^{-1}	6.10×10^{-19}
nwk	−0.22	−0.51	2.33×10^{-1}	1.43×10^{-16}
Der-1	0.39	0.67	5.65×10^{-2}	1.01×10^{-16}
isolated - multi				
CG15270	−0.02	−0.28	5.73×10^{-1}	1.18×10^{-4}
CG15202	0.03	0.41	3.11×10^{-1}	5.10×10^{-7}
CG9498	0.62	0.74	3.20×10^{-2}	8.01×10^{-7}
Loxl2	0.04	0.45	1.85×10^{-1}	3.88×10^{-6}
Cln3	0.02	0.39	4.88×10^{-1}	1.26×10^{-4}
ELOVL	0.46	0.52	1.72×10^{-2}	9.26×10^{-7}
CG10550	0.37	0.43	2.31×10^{-2}	4.89×10^{-6}
CG9717	0.02	0.51	3.24×10^{-1}	8.50×10^{-5}
vir-1	0.15	0.23	1.82×10^{-1}	1.50×10^{-5}
Cda5	−0.03	−0.52	1.91×10^{-1}	8.22×10^{-6}

3.4.9.1 Perturbation to Housing Contrast

To see how much exclusion of the FruLexa/Fru440 replicate 1 alters the big picture results in the group vs. isolated contrast, we can look at how well the top-100 lists agree (similarity is calculated as size of the intersection divided by size of the union; lists are pooled across all aligners and thus may have more than 100 unique elements)

Figure 74 . Similarity of Housing Contrast Top 100 Lists,
with/without FruLexaFru440 replicate 1 (pooled alignment strategies)

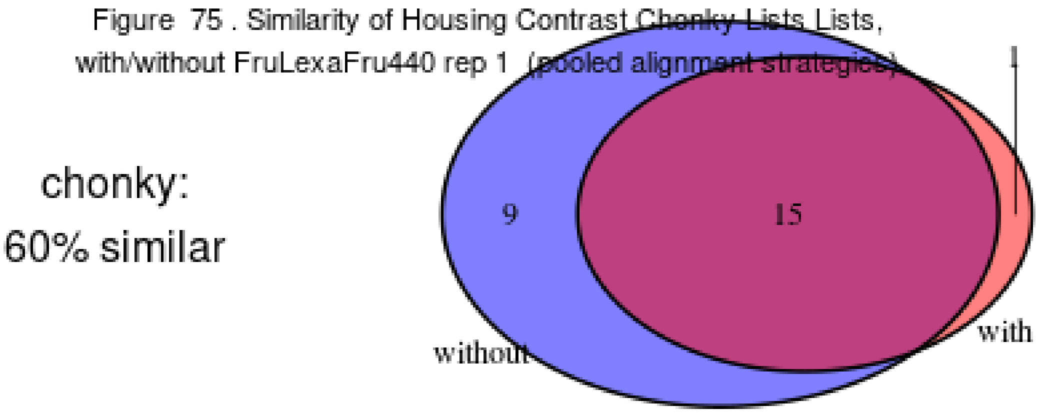

adjusted p:
50% similar

log2FoldChange:
58% similar

expression:
74% similar

```
## png
##   2
```

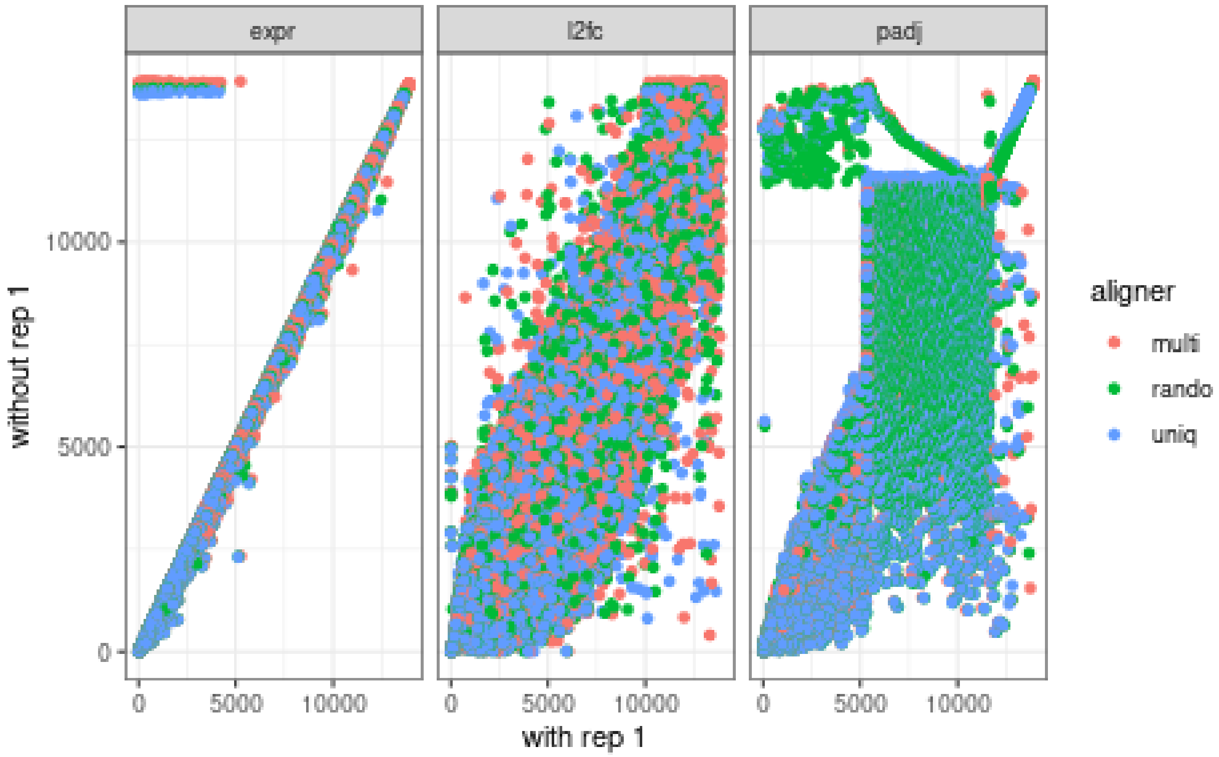
Comparing the “chonky” gene lists (padj < 0.01, abs(l2fc)>1):



```
## png
##   2
```

We can also look at the rank correlations:

Figure 76. Rank correlations of expression, effect size, and significance (housing contrasts, with/without problematic FruLexaFru440 replicate)

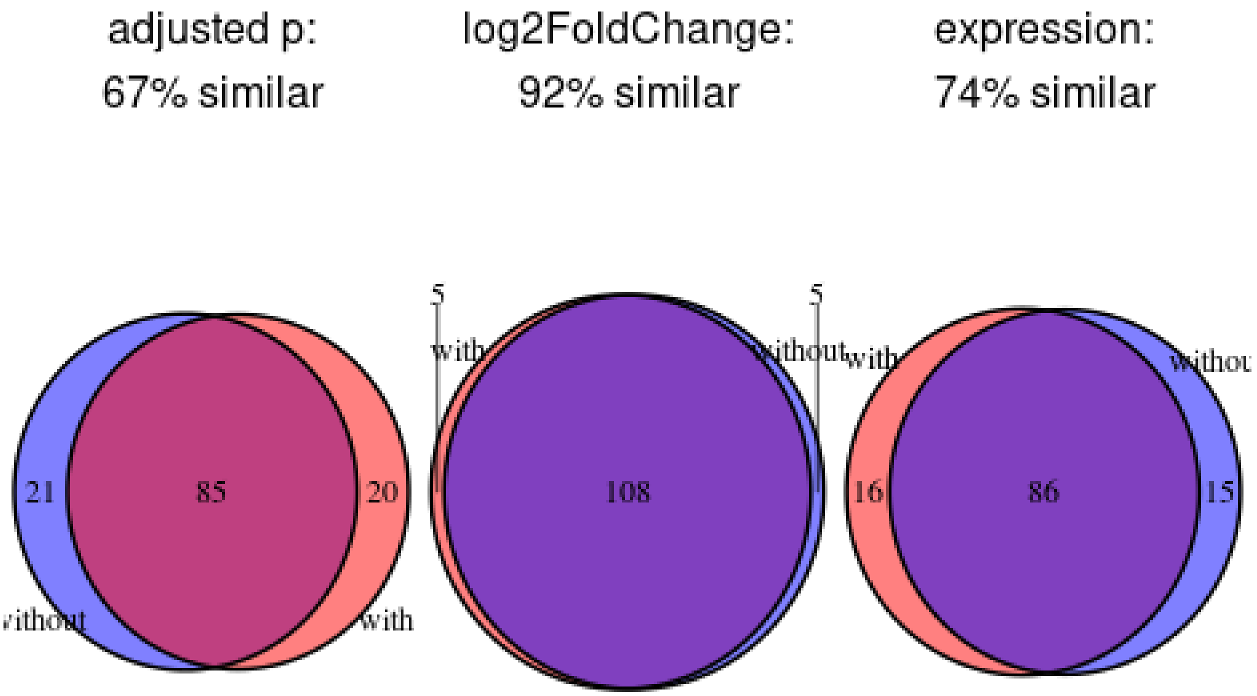


```
## png
##   2
```

3.4.9.2 Perturbation to 47b1 Contrast

To see how much exclusion of the FruLexa/Fru440 replicate 1 alters the big picture results in the group vs. isolated contrast, we can look at how well the top-100 lists agree (similarity is calculated as size of the intersection divided by size of the union; lists are pooled across all aligners and thus may have more than 100 unique elements)

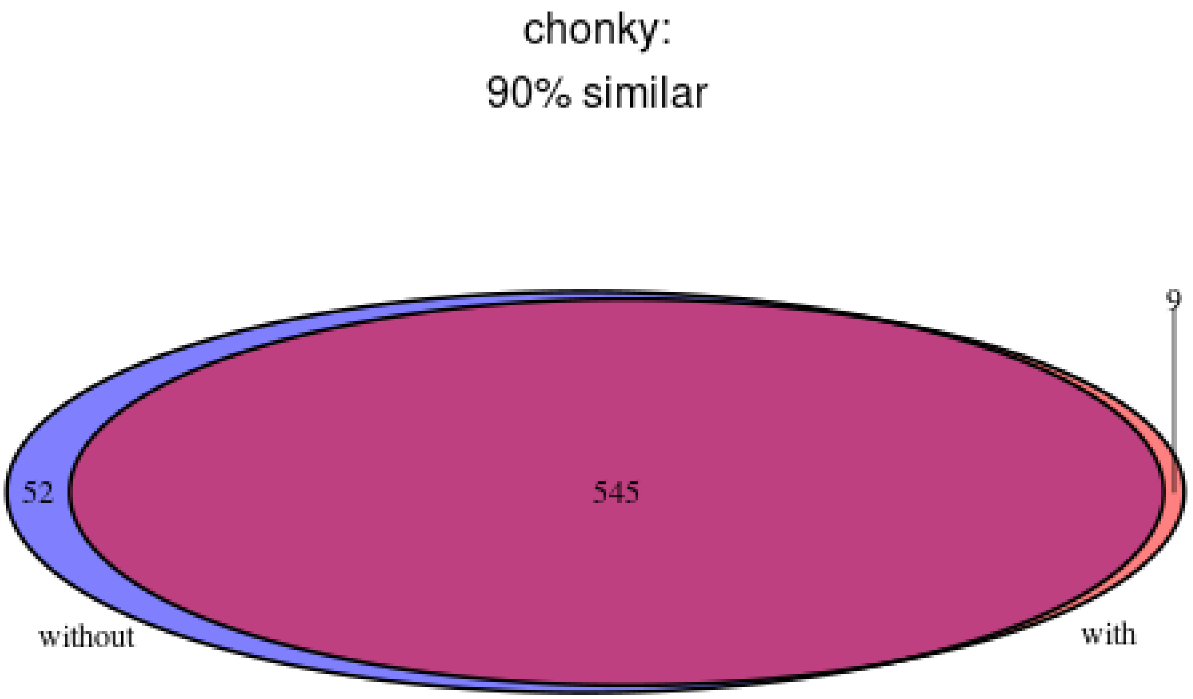
Figure 77 . Similarity of 47b1 Contrast Top 100 Lists,
with/without FruLexaFru440 replicate 1 (pooled alignment strategies)



```
## png
##   2
```

Comparing the “chonky” gene lists (padj < 0.01, abs(l2fc)>1):

Figure 78 . Similarity of 67d Contrast Chonky Lists Lists,
with/without FruLexaFru440 replicate 1 (pooled alignment strategies)

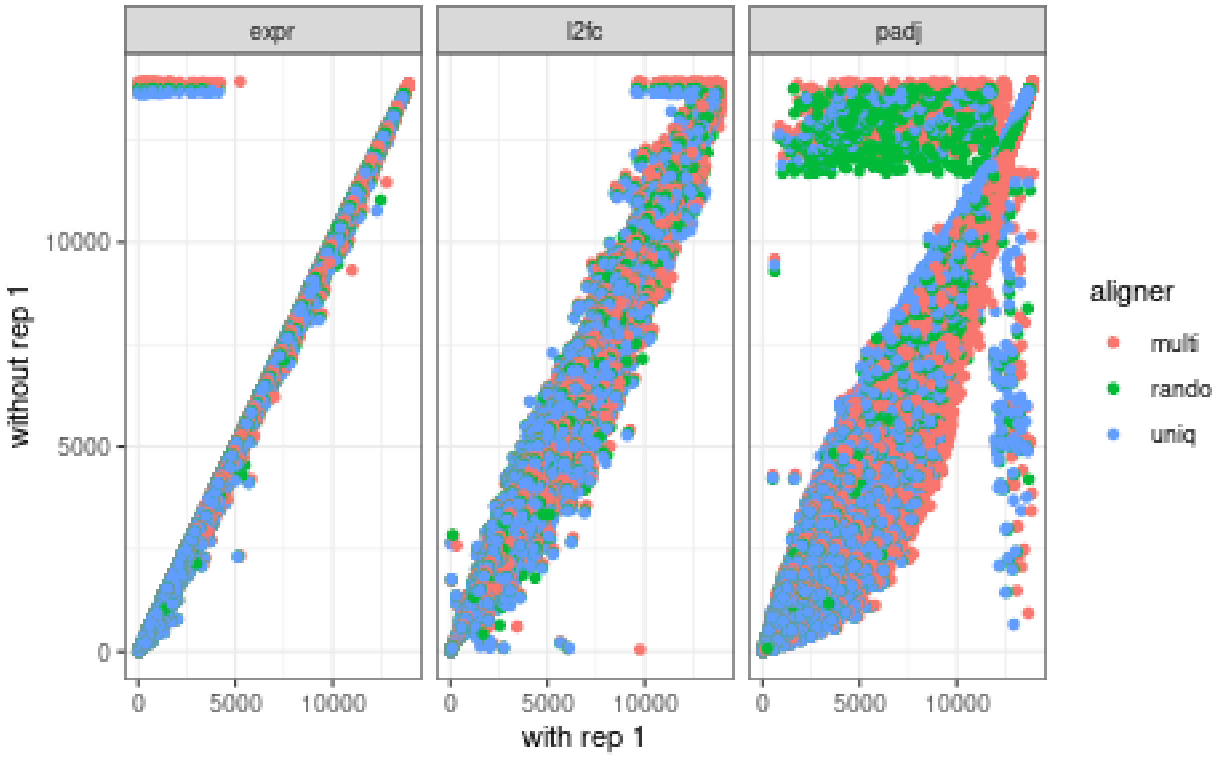


chonky:
90% similar

```
## png
##   2
```

We can also look at the rank correlations:

Figure 79. Rank correlations of expression, effect size, and significance (47b1 contrasts, with/without problematic FruLexaFru440 replicate)

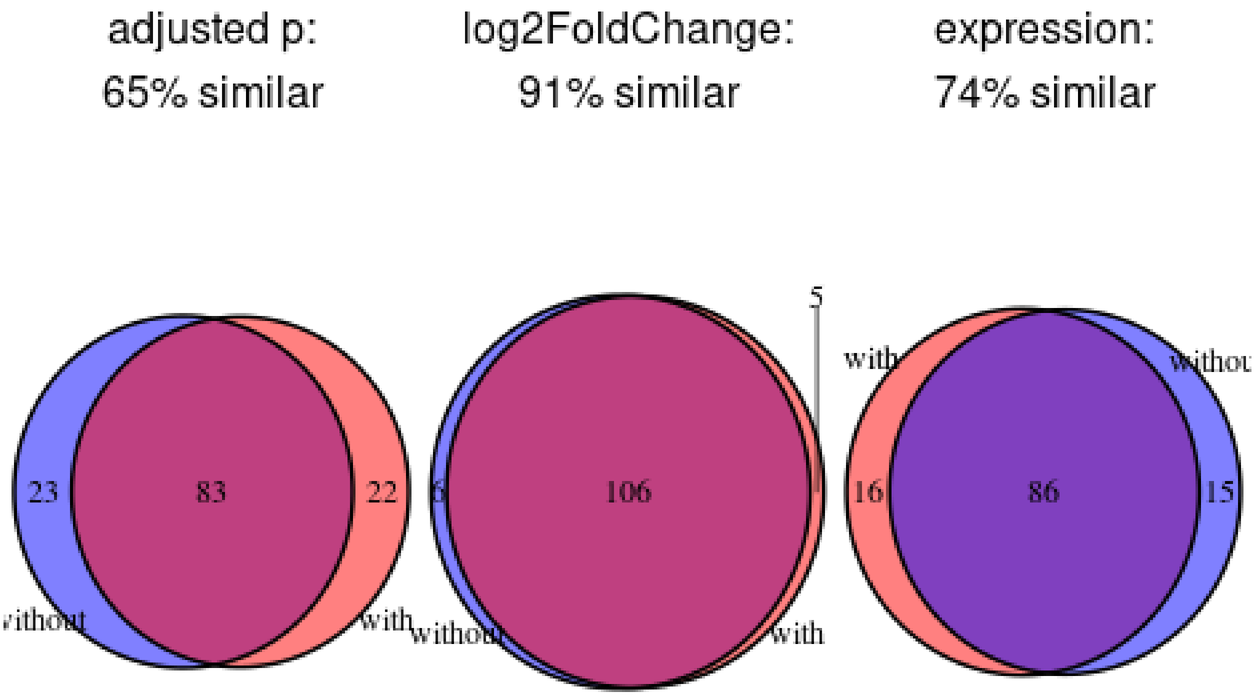


```
## png
##   2
```

3.4.9.3 Perturbation to 67d Contrast

To see how much exclusion of the FruLexa/Fru440 replicate 1 alters the big picture results in the 67d vs. wt contrast, we can look at how well the top-100 lists agree (similarity is calculated as size of the intersection divided by size of the union; lists are pooled across all aligners and thus may have more than 100 unique elements)

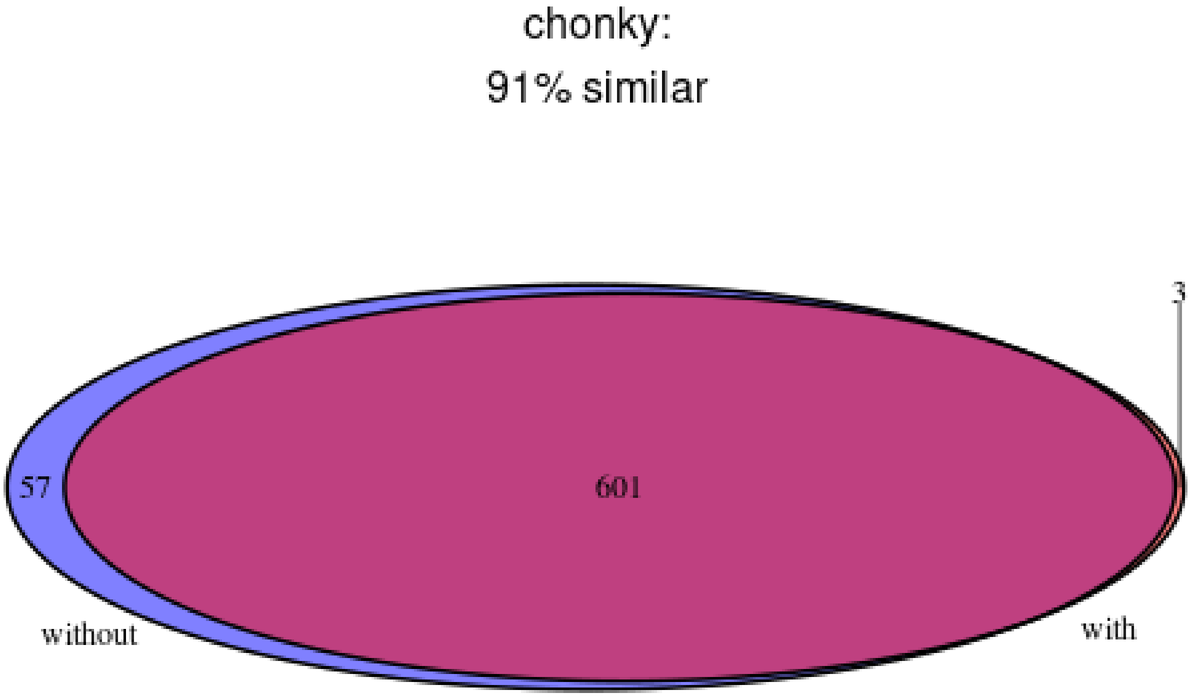
Figure 80 . Similarity of 67d Contrast Top 100 Lists,
with/without FruLexaFru440 replicate 1 (pooled alignment strategies)



```
## png
##   2
```

Comparing the “chonky” gene lists (padj < 0.01, abs(l2fc)>1):

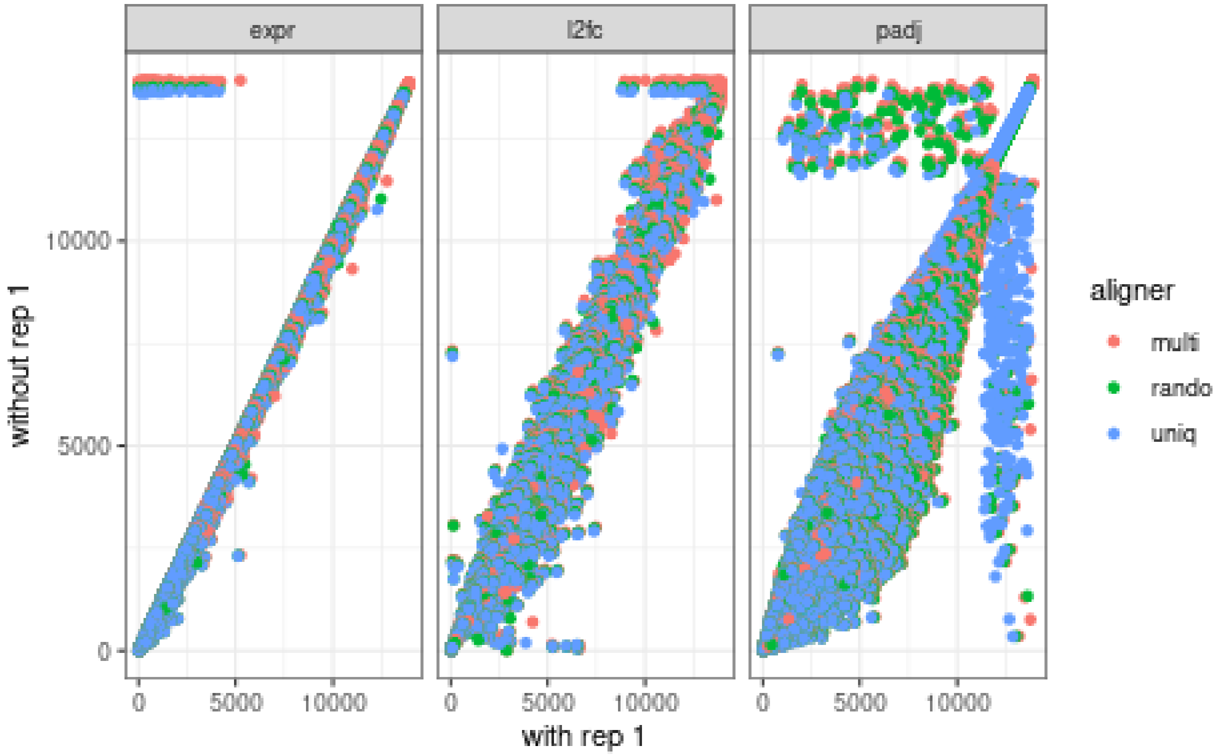
Figure 81 . Similarity of 67d Contrast Chonky Lists Lists,
with/without FruLexaFru440 replicate 1 (pooled alignment strategies)



```
## png
##   2
```

We can also look at the rank correlations:

Figure 82. Rank correlations of expression, effect size, and significance (67d contrasts, with/without problematic FruLexaFru440 replicate)

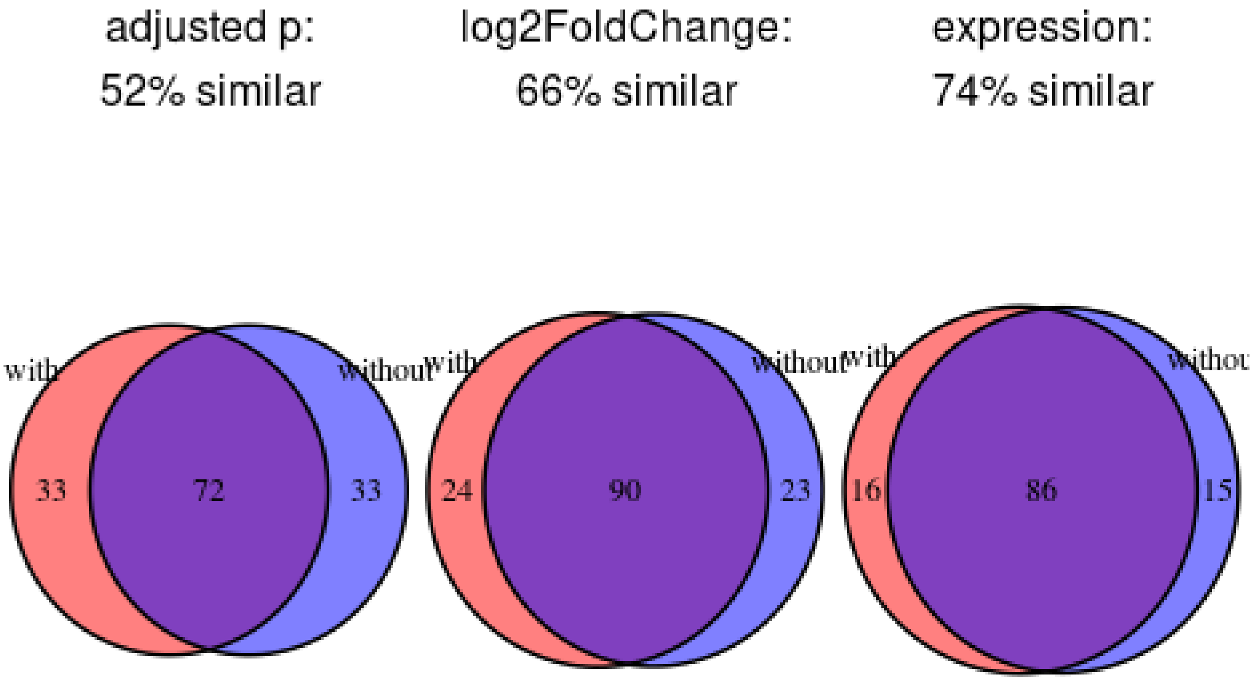


```
## png
##   2
```

3.4.9.4 Perturbation to FruLexa/Fru440 Contrast

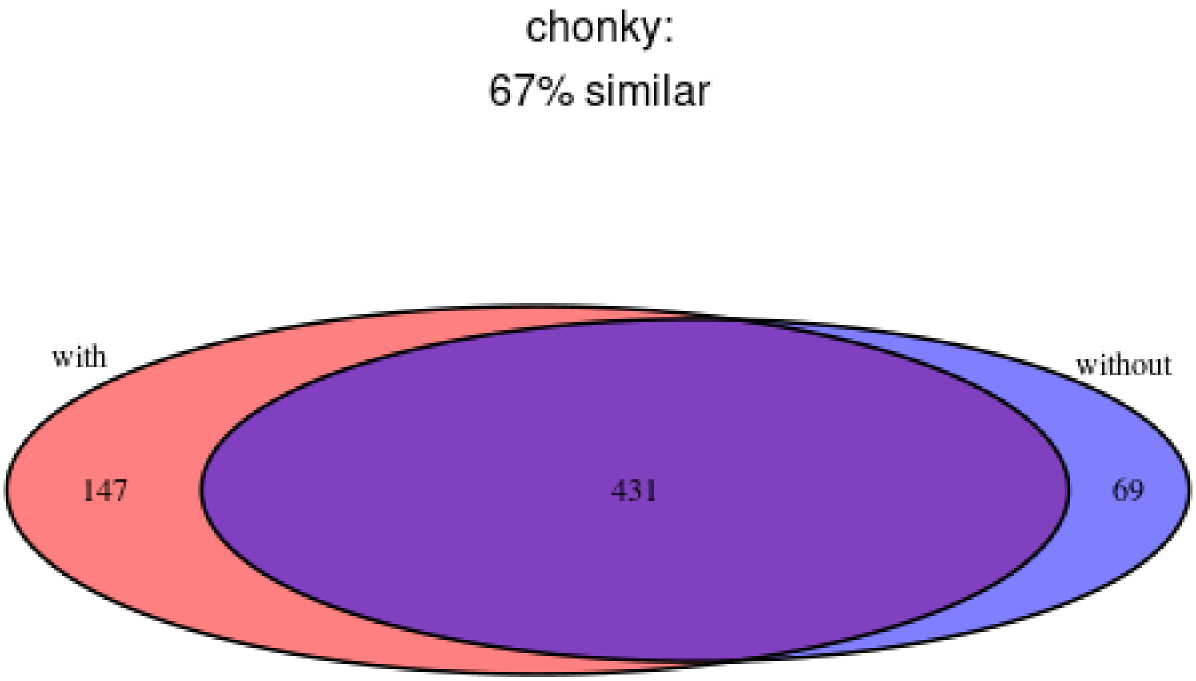
To see how much exclusion of the FruLexa/Fru440 replicate 1 alters the big picture results in the FruLexaFru440 vs. wt contrast, we can look at how well the top-100 lists agree (similarity is calculated as size of the intersection divided by size of the union; lists are pooled across all aligners and thus may have more than 100 unique elements)

Figure 83 . Similarity of FruLexa440 Contrast Top 100 Lists,
with/without FruLexaFru440 replicate 1 (pooled alignment strategies)



```
## png
##   2
```

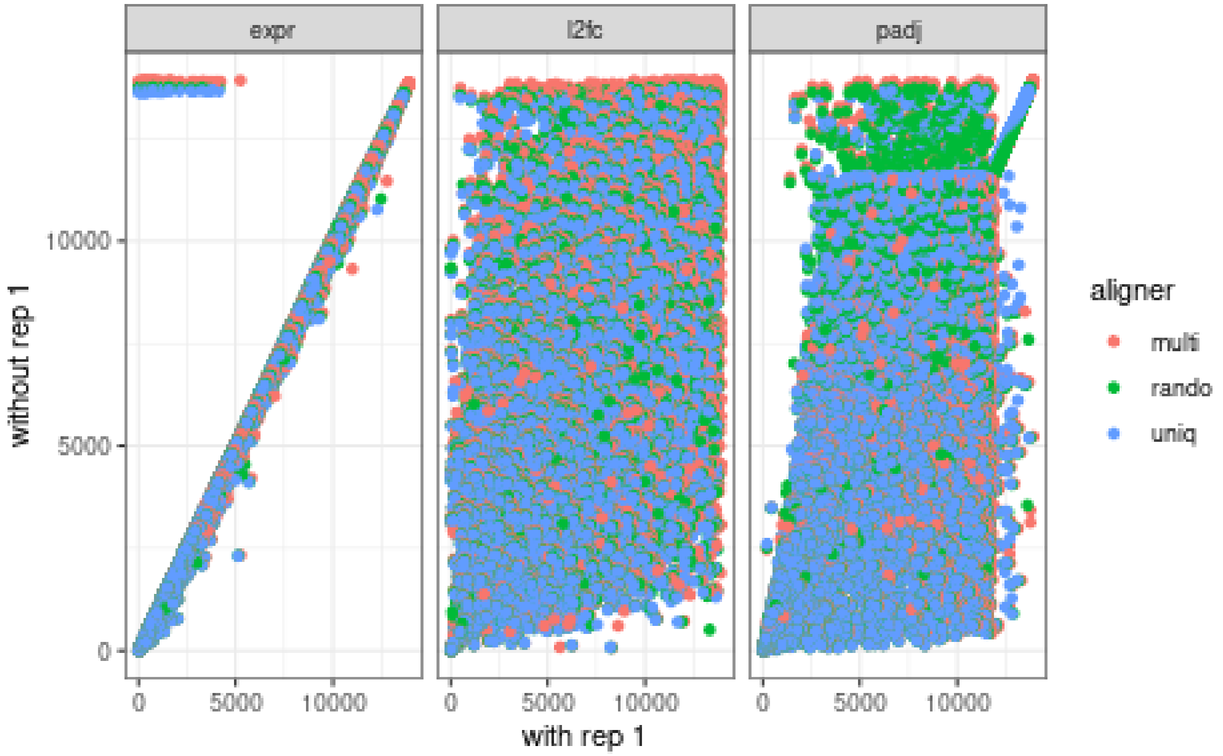
Figure 84 . Similarity of Fru Contrast Chonky Lists Lists,
with/without FruLexaFru440 replicate 1 (pooled alignment strategies)



```
## png
##   2
```

We can also look at the rank correlations:

Figure 85. Rank correlations of expression, effect size, and significance (67d contrasts, with/without problematic FruLexaFru440 replicate)



```
## png
##   2
```

3.4.9.5 Perturbation to FruLexa/Fru440 Contrast (single-factor)

People have objections to using the 2-factor model so if we consider single-factor models, only the FruLexaFru440 contrast actually matters (since no Fru mutant replicates are included in any other contrast, ever). Intuitively I would expect that more (presumably high-quality) samples included in the model would better buffer it to the effects of dropping replicates. Dropping replicate 1 from hausWtVsMut changes the sample size from to 15 to 14, whereas in the single-factor grpWtVsFru changes it from 6 to 5.

The data are found here: “results/tables/supp/grpWtVsFru_smolFru.allAligners.DESeq2.MpBC.reformatted.tsv”

Every gene that can be analyzed using the Reduced counts can be analyzed using the With counts; there are 0 that cannot. On the other hand, there are 0 genes which can be analyzed using the Full counts but not the Reduced

In 0 cases were these genes significant (padj < 0.01) in the With tests.

A gene with significance values in both tests may gain significance when FruLexa/Fru440 samples are dropped, lose significance, maintain significance while switching direction, or remain unchanged. No switches were seen, but moderate numbers (up to ~5%) gained significance.

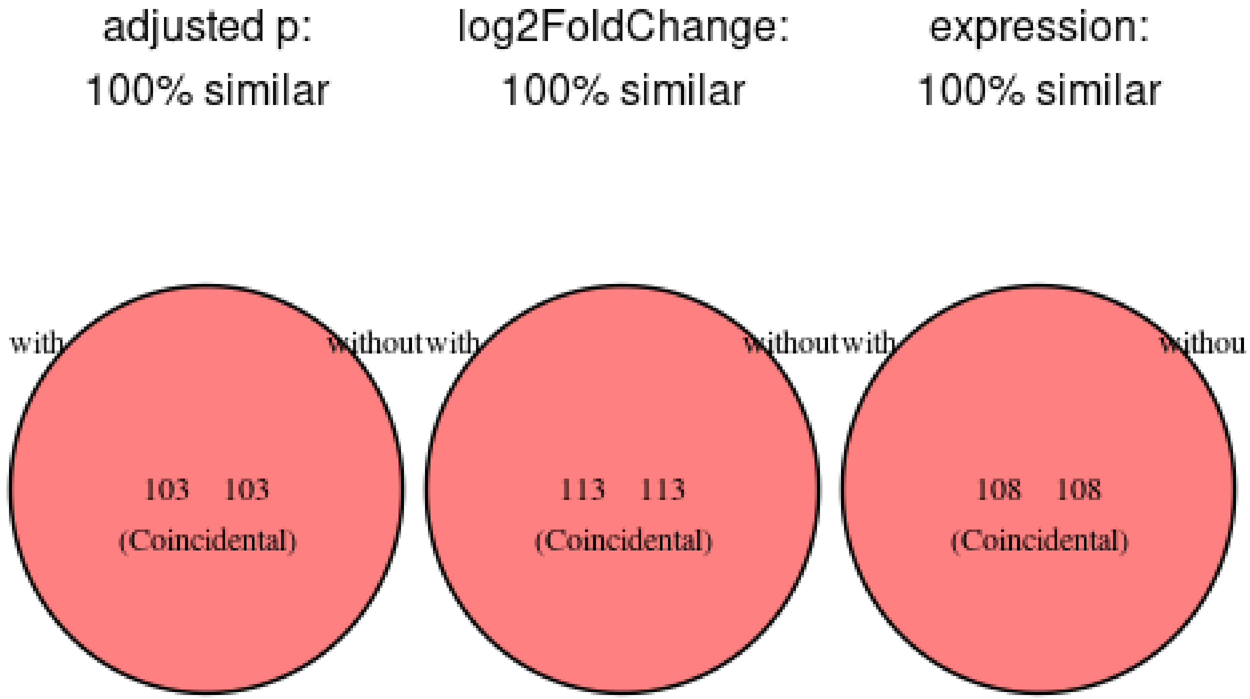
Table 88. Changes in Differential Expression Significance
when FruLexa/Fru440 replicate 1 is dropped; single-factor

none

	FruLexaFru440
multi	12949
rando	12812
uniq	12734

To see how much exclusion of the FruLexa/Fru440 replicate 1 alters the big picture results in the FruLexaFru440 vs. wt contrast, we can look at how well the top-100 lists agree (similarity is calculated as size of the intersection divided by size of the union; lists are pooled across all aligners and thus may have more than 100 unique elements)

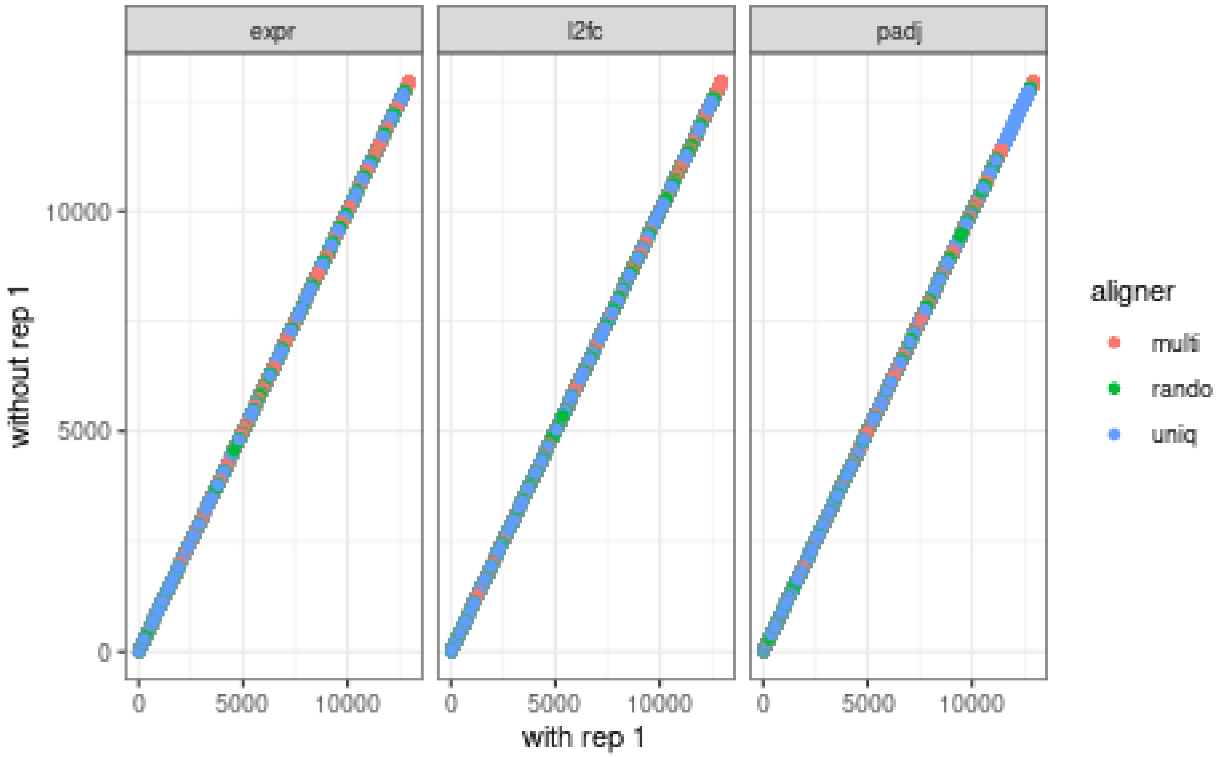
Figure 86 . Similarity of FruLexa440 Contrast Top 100 Lists,
with/without FruLexaFru440 replicate 1 (single-factor model)



```
## png
##   2
```

We can also look at the rank correlations:

Figure 87. Rank correlations of expression, effect size, and significance (single-factor FruLexaFru440 contrast, with/without problematic replicate



```
## png
##   2
```

3.5 Comparing Expression Changes from Housing with Expression Changes from Genotype

We want to see if the difference in life history creates similar changes in expression as various mutations. This was done using the differential expression data from the genotype & housing simultaneous model. This circumvents the joining step in earlier versions. Earlier versions comparing results from two distinct models readjustd the p-values with a Bonferroni correction using n=2; in the current iteration in which both p-values are coming from the same model, this step is skipped. Candidate genes of interest are then collected by filtering this joint comparison for genes which show a significant change in both contrasts. These candidates are further classified as to whether the expression changes are in the same direction (ie, both enriched or both depleted) or not (ie, one enriched and the other depleted).

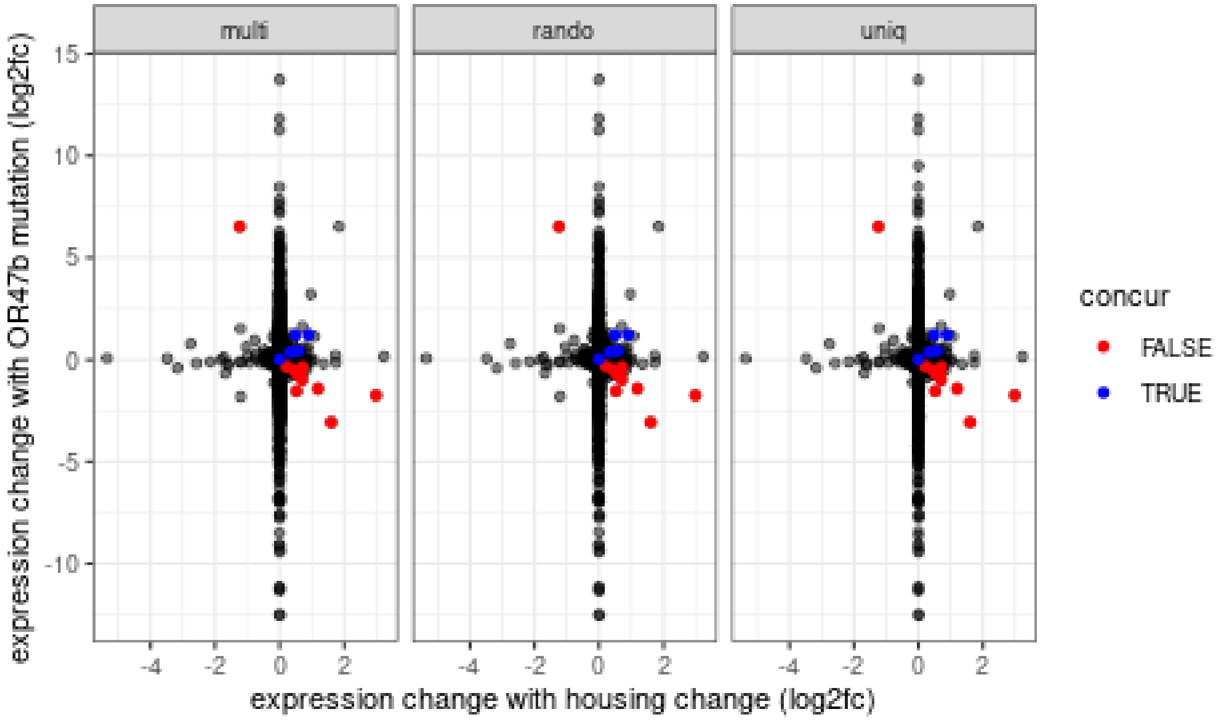
Average significance for gene is currently computed as exp((ln(p1)+ln(p2))/2). (Better to apply stouffer's?)

look at NAs in fulljoin (gene dropout may be interesting…)

3.5.1 Housing & OR47b

Here is a scatterplot of the log2 fold change of the 47b & wt contast vs the housing contrast (wt group & wt isolated). The upper right quadrant contains genes which are enriched in both cases; the lower left contains genes which are depleted in both cases. The other two quadrants contain mismatches between expression patterns. Significant changes are highlighted accordingly.

Figure 88. Scatterplot of Expression Changes in OR47b mutants vs Expression Changes in Housing
(Significant Similarities and Differences Highlighted)



```
## png
##   2
```

Of the mututally significant genes, fewer have the same direction of change than not:

Table 89. Number of Genes with Significant Changes in Both Contrasts, by Shared Direction of Change
change in housing vs OR47b

	multi	rando	uniq
Agree	5	5	5
Disagree	11	11	11

Of those mutually significant genes with the same direction of change, the top 10 most significant agree well across alignment strategy:

Table 90. Top Ten Most Significant Gene
in difference expression between housing and OR47b

	multi						
rank	name	mean expression	mean readusted p	housing l2fc	mutation l2fc	name	mean expression
1	jv	0.16	5.67×10^{-23}	0.488	1.216	jv	0.16
2	CG12986	0.20	1.24×10^{-6}	0.896	1.220	CG12986	0.20

3	CG43147	0.03	5.11×10^{-4}	0.000	0.031	CG43147	0.03
4	CG13659	0.60	6.49×10^{-4}	0.576	0.447	CG13659	0.60
5	PGRP-LB	0.76	1.52×10^{-3}	0.340	0.378	PGRP-LB	0.76

When mutually significant genes with the same direction of change are ranked by the magnitude of their mean log2FoldChange, the top 10 agree well across alignment strategy:

Table 91. Top Ten Largest Magnitude Changes In Significant Ge
in difference expression between housing and OR47b contrants

	multi				rando			
rank	name	mean l2fc	mean expression	mean readusted p	name	mean l2fc	mean expression	mean
1	CG12986	1.058	0.20	1.24×10^{-6}	CG12986	1.059	0.20	
2	jv	0.852	0.16	5.67×10^{-23}	jv	0.852	0.16	3
3	CG13659	0.512	0.60	6.49×10^{-4}	CG13659	0.512	0.60	
4	PGRP-LB	0.359	0.76	1.52×10^{-3}	PGRP-LB	0.360	0.76	
5	CG43147	0.016	0.03	5.11×10^{-4}	CG43147	0.015	0.03	

Of those mutually significant genes with different directions of change, the top 10 most significant agree well across alignment strategy.

Table 92. Top Ten Most Significant Genes of
in difference expression between housing and OR47b contr

	multi							
rank	name	mean expression	mean readusted p	housing l2fc	OR47b l2fc	name	mean expression	mea
1	CG6912	0.73	0.00	−1.238	6.522	CG6912	0.73	
2	Obp84a	0.75	9.42×10^{-29}	0.524	−1.535	Obp84a	0.75	
3	CG11852	0.14	1.40×10^{-9}	1.600	−3.066	CG11852	0.14	
4	amd	2.00	1.30×10^{-7}	1.190	−1.418	amd	2.00	
5	CG10050	0.47	2.60×10^{-6}	0.701	−0.986	CG10050	0.47	
6	Fer2LCH	8.34	3.17×10^{-6}	0.202	−0.320	magu	0.40	
7	magu	0.40	3.34×10^{-6}	0.713	−0.353	Fer2LCH	8.34	
8	Or92a	3.09	4.51×10^{-5}	0.422	−0.487	Or92a	3.09	
9	CG13332	0.56	5.49×10^{-5}	0.548	−0.723	CG13332	0.56	
10	Dh44-R2	0.05	1.67×10^{-4}	0.681	−0.785	Dh44-R2	0.05	

When mutually significant genes with different directions of change are ranked by the magnitude of their difference in log2FoldChange, the top 10 genes agree well across alignment strategy:

Table 93. Top Ten Most Serious Significant Differences betw
in difference expression between housing and OR47b contrants

	multi				rando		
rank	name	l2fc difference	mean expression	mean readusted p	name	l2fc difference	mean expression
1	CG6912	−7.761	0.73	0.00	CG6912	−7.762	0.73
2	Jhe	4.735	0.50	3.04×10^{-4}	Jhe	4.731	0.50
3	CG11852	4.666	0.14	1.40×10^{-9}	CG11852	4.667	0.14
4	amd	2.608	2.00	1.30×10^{-7}	amd	2.604	2.00

5	Obp84a	2.059	0.75	9.42×10^{-29}	Obp84a	2.058	0.75
6	CG10050	1.687	0.47	2.60×10^{-6}	CG10050	1.691	0.47
7	Dh44-R2	1.466	0.05	1.67×10^{-4}	Dh44-R2	1.465	0.05
8	CG13332	1.271	0.56	5.49×10^{-5}	CG13332	1.269	0.56
9	magu	1.066	0.40	3.34×10^{-6}	magu	1.064	0.40
10	Or92a	0.909	3.09	4.51×10^{-5}	Or92a	0.907	3.09

The full joined comparisons can be found in the tables folder: $results/tables/supp/housingContrast_and_47bContrast.multi.ts$

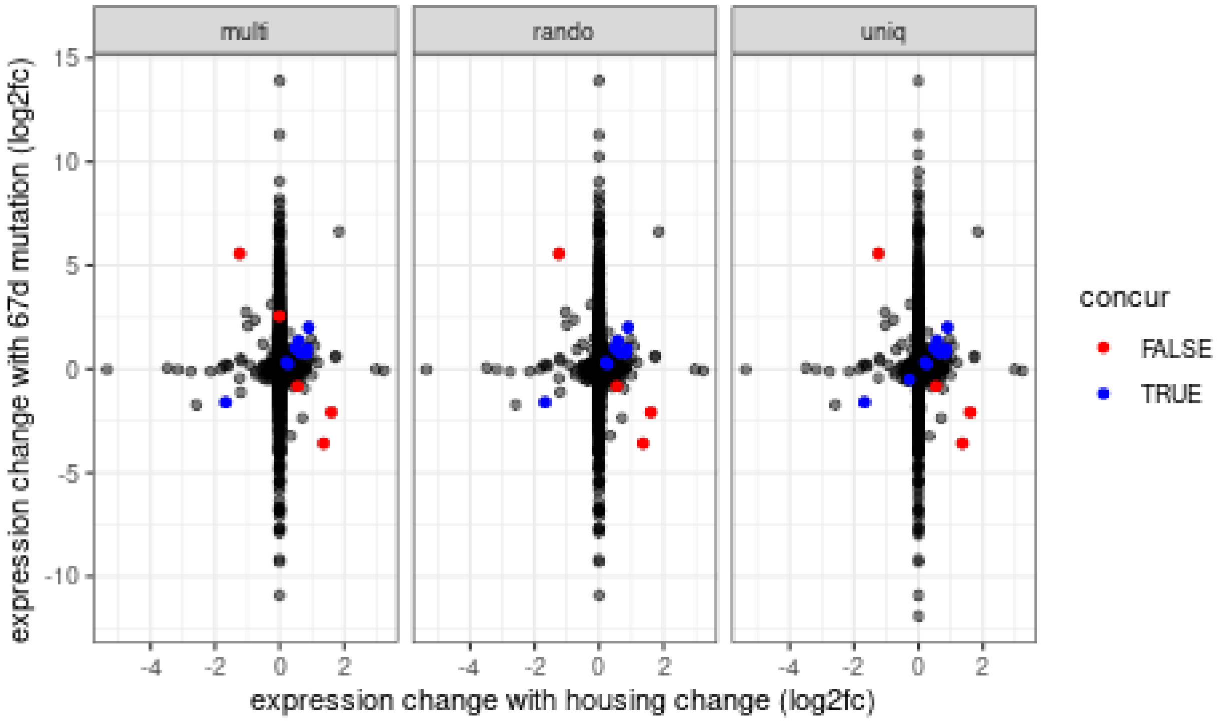
$results/tables/supp/housingContrast_and_47bContrast.rando.tsv$

$results/tables/supp/housingContrast_and_47bContrast.uniq.tsv$

3.5.2 Housing & 67d

Here is a scatterplot of the log2 fold change of the 67d & wt contast vs the housing contrast (wt group & wt isolated). The upper right quadrant contains genes which are enriched in both cases; the lower left contains genes which are depleted in both cases. The other two quadrants contain mismatches between expression patterns. Significant changes are highlighted accordingly.

Figure 89. Scatterplot of Expression Changes in 67d mutants vs Expression Changes in Housing (Significant Similarities and Differences Highlighted)



```
## png
##   2
```

Of the mutually significant genes, slightly more have the same direction of change than not:

Table 94. Number of Genes with Significant Changes in Both Contrasts, by Shared Direction of Change

change in housing vs 67d

	multi	rando	uniq
Agree	9	9	10
Disagree	5	4	4

Of those mutually significant genes with the same direction of change, the top 10 most significant agree well across alignment strategy:

Table 95. Top Ten Most Significant Genes of Ag

in difference expression between housing and 67d contrants

	multi					ran		
rank	name	mean expression	mean readusted p	housing l2fc	67d l2fc	name	mean expression	mean
1	jv	0.16	3.23×10^{-16}	0.488	1.006	jv	0.16	2
2	CG13659	0.60	1.86×10^{-15}	0.576	1.364	CG13659	0.60	1
3	CG12986	0.20	4.02×10^{-14}	0.896	2.012	CG12986	0.20	3
4	CG31288	2.72	3.57×10^{-10}	0.852	0.980	CG31288	2.72	3
5	Fer2LCH	8.34	3.28×10^{-6}	0.202	0.330	Fer2LCH	8.34	
6	CG31272	0.19	1.22×10^{-5}	0.826	0.775	CG31272	0.19	
7	hgo	0.07	3.45×10^{-4}	-1.674	-1.602	hgo	0.07	
8	Oatp33Ea	0.04	7.82×10^{-4}	0.636	0.782	Oatp33Ea	0.04	
9	CG32276	4.00	7.27×10^{-3}	0.264	0.235	CG32276	4.00	
10	NA	NA	NA	NA	NA	NA	NA	

When mutually significant genes with the same direction of change are ranked by the magnitude of their mean log2FoldChange, the top 10 agree relatively well across alignment strategy:

Table 96. Top Ten Largest Magnitude Changes In Significant Ge

in difference expression between housing and 67d contrants

	multi				rando			
rank	name	mean l2fc	mean expression	mean readusted p	name	mean l2fc	mean expression	mean r
1	hgo	-1.638	0.07	3.45×10^{-4}	hgo	-1.638	0.07	
2	CG12986	1.454	0.20	4.02×10^{-14}	CG12986	1.454	0.20	3
3	CG13659	0.970	0.60	1.86×10^{-15}	CG13659	0.971	0.60	1
4	CG31288	0.916	2.72	3.57×10^{-10}	CG31288	0.913	2.72	3
5	CG31272	0.800	0.19	1.22×10^{-5}	CG31272	0.800	0.19	
6	jv	0.747	0.16	3.23×10^{-16}	jv	0.746	0.16	2
7	Oatp33Ea	0.709	0.04	7.82×10^{-4}	Oatp33Ea	0.709	0.04	
8	Fer2LCH	0.266	8.34	3.28×10^{-6}	Fer2LCH	0.266	8.34	
9	CG32276	0.249	4.00	7.27×10^{-3}	CG32276	0.248	4.00	
10	NA	NA	NA	NA	NA	NA	NA	

Of those mutually significant genes with different directions of change, the top 10 most significant agree well across alignment strategy.

Table 97. Top Ten Most Significant Genes of Disa
in difference expression between housing and 67d contrasts

	multi					rand		
rank	name	mean expression	mean readusted p	housing l2fc	67d l2fc	name	mean expression	mean re
1	CG6912	0.73	1.49×10^{-112}	−1.238	5.569	CG6912	0.73	1.10
2	MtnB	0.76	4.73×10^{-47}	1.361	−3.586	MtnB	0.76	3.4
3	CG11852	0.14	3.29×10^{-6}	1.600	−2.087	CG11852	0.14	3
4	CG13332	0.56	1.56×10^{-5}	0.548	−0.825	CG13332	0.56	1
5	Amy-d	0.08	1.92×10^{-3}	−0.008	2.558	NA	NA	

When mutually significant genes with different directions of change are ranked by the magnitude of their difference in log2FoldChange, the top 10 genes agree well across alignment strategy:

Table 98. Top Ten Most Serious Significant Differences betw
in difference expression between housing and 67d contrants

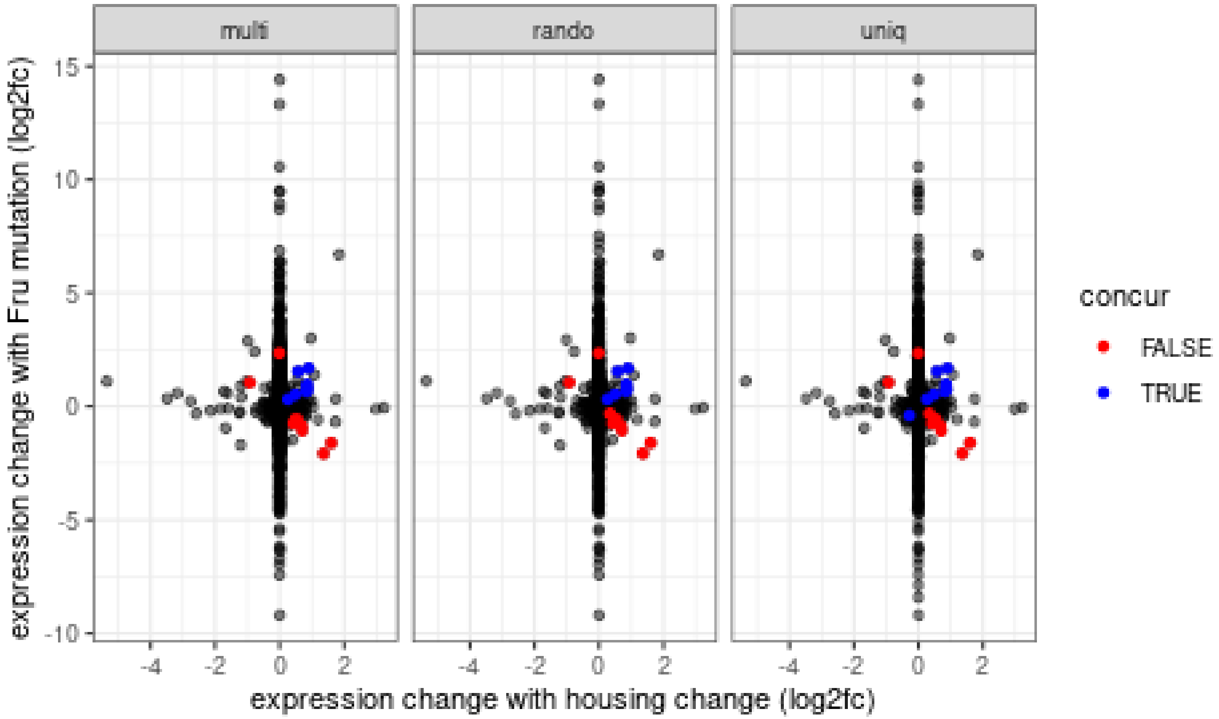
	multi				rando		
rank	name	l2fc difference	mean expression	mean readusted p	name	l2fc difference	mean expression
1	CG6912	−6.807	0.73	1.49×10^{-112}	CG6912	−6.805	0.73
2	MtnB	4.948	0.76	4.73×10^{-47}	MtnB	4.946	0.76
3	CG11852	3.686	0.14	3.29×10^{-6}	CG11852	3.688	0.14
4	Amy-d	−2.566	0.08	1.92×10^{-3}	CG13332	1.372	0.56
5	CG13332	1.373	0.56	1.56×10^{-5}	NA	NA	NA

The full joined comparisons can be found in the tables folder: $results/tables/supp/housingContrast_and_67dContrast.multi.ts$ $results/tables/supp/housingContrast_and_67dContrast.rando.tsv$ $results/tables/supp/housingContrast_and_67dContrast.un$

3.5.3 Housing & FruLexFru440

Here is a scatterplot of the log2 fold change of the Fru & wt contast vs the housing contrast (wt group & wt isolated). The upper right quadrant contains genes which are enriched in both cases; the lower left contains genes which are depleted in both cases. The other two quadrants contain mismatches between expression patterns. Significant changes are highlighted accordingly.

Figure 90. Scatterplot of Expression Changes in Fru mutants vs Expression Changes in Housing
(Significant Similarities and Differences Highlighted)



```
## png
##   2
```

Of the mutually significant genes, about the same number have the same direction of change as not:

Table 99. Number of Genes with Significant Changes in Both Contrasts, by Shared Direction of Change
change in housing vs Fru

	multi	rando	uniq
Agree	7	7	8
Disagree	9	10	10

Of those mutually significant genes with the same direction of change, the top 10 most significant agree well across alignment strategy:

Table 100. Top Ten Most Significant Genes of Ag
in difference expression between housing and Fru contrants

	multi					rand		
rank	name	mean expression	mean readusted p	housing l2fc	Fru l2fc	name	mean expression	mean re
1	CG13659	0.60	5.13×10^{-19}	0.576	1.544	CG13659	0.60	3.3
2	CG12986	0.20	4.03×10^{-10}	0.896	1.687	CG12986	0.20	3.5

3	CG31288	2.72	4.48×10^{-10}	0.852	0.979	CG31288	2.72
4	jv	0.16	2.69×10^{-6}	0.488	0.521	jv	0.16
5	CG31272	0.19	6.19×10^{-5}	0.826	0.661	CG31272	0.19
6	CG42806	1.07	8.62×10^{-5}	0.869	0.751	CG42806	1.07
7	CG32276	4.00	6.44×10^{-4}	0.264	0.317	CG32276	4.00
8	NA	NA	NA	NA	NA	NA	NA

The last visible column of this table is cut off; what shows of it, rows 3 to 7: 3.8, 2., 6., 8., 6.

When mutually significant genes with the same direction of change are ranked by the magnitude of their mean log2FoldChange, the top 10 agree well across alignment strategy.

Table 101. Top Ten Largest Magnitude Changes In Significant Ge
in difference expression between housing and Fru contrants

	multi				rando			
rank	name	mean l2fc	mean expression	mean readusted p	name	mean l2fc	mean expression	mean rea
1	CG12986	1.292	0.20	4.03×10^{-10}	CG12986	1.292	0.20	3.5
2	CG13659	1.060	0.60	5.13×10^{-19}	CG13659	1.062	0.60	3.3
3	CG31288	0.915	2.72	4.48×10^{-10}	CG31288	0.913	2.72	3.8
4	CG42806	0.810	1.07	8.62×10^{-5}	CG42806	0.810	1.07	8.
5	CG31272	0.743	0.19	6.19×10^{-5}	CG31272	0.744	0.19	6.
6	jv	0.504	0.16	2.69×10^{-6}	jv	0.504	0.16	2.
7	CG32276	0.290	4.00	6.44×10^{-4}	CG32276	0.291	4.00	6.
8	NA	NA	NA	NA	NA	NA	NA	

Of those mutually significant genes with different directions of change, the top 10 most significant agree well across alignment strategy.

Table 102. Top Ten Most Significant Genes of D
in difference expression between housing and Fru contrants

	multi					ranc		
rank	name	mean expression	mean readusted p	housing l2fc	Fru l2fc	name	mean expression	mean r
1	MtnB	0.76	4.28×10^{-28}	1.361	−2.073	MtnB	0.76	3
2	Or92a	3.09	2.23×10^{-8}	0.422	−0.750	Or92a	3.09	2
3	CG10050	0.47	1.65×10^{-6}	0.701	−1.056	CG10050	0.47	
4	CG11852	0.14	6.94×10^{-5}	1.600	−1.612	CG11852	0.14	6
5	TotC	0.29	1.32×10^{-4}	−0.009	2.348	TotC	0.29	
6	Dh44-R2	0.05	1.57×10^{-4}	0.681	−0.857	Dh44-R2	0.05	
7	T48	0.39	2.52×10^{-4}	0.484	−0.575	T48	0.39	2
8	Gbs-70E	0.18	9.84×10^{-4}	−0.923	1.057	Gbs-70E	0.18	9
9	CG13332	0.56	1.07×10^{-3}	0.548	−0.574	CG13332	0.56	
10	NA	NA	NA	NA	NA	PGRP-LB	0.76	8

When mutually significant genes with different directions of change are ranked by the magnitude of their difference in log2FoldChange, the top 10 genes agree well across alignment strategy.

Table 103. Top Ten Most Serious Significant Differences be
in difference expression between housing and Fru contrasts

multi	rando

rank	name	l2fc difference	mean expression	mean readusted p	name	l2fc difference	mean expression
1	MtnB	3.434	0.76	4.28×10^{-28}	MtnB	3.431	0.76
2	CG11852	3.212	0.14	6.94×10^{-5}	CG11852	3.210	0.14
3	TotC	−2.357	0.29	1.32×10^{-4}	TotC	−2.356	0.29
4	Gbs-70E	−1.981	0.18	9.84×10^{-4}	Gbs-70E	−1.983	0.18
5	CG10050	1.757	0.47	1.65×10^{-6}	CG10050	1.762	0.47
6	Dh44-R2	1.539	0.05	1.57×10^{-4}	Dh44-R2	1.541	0.05
7	Or92a	1.172	3.09	2.23×10^{-8}	Or92a	1.169	3.09
8	CG13332	1.122	0.56	1.07×10^{-3}	CG13332	1.120	0.56
9	T48	1.059	0.39	2.52×10^{-4}	T48	1.056	0.39
10	NA	NA	NA	NA	PGRP-LB	0.659	0.76

Full data are in the tables folder:

$results/tables/supp/housingContrast_and_FruContrast.multi.tsv$ $results/tables/supp/housingContrast_and_FruContrast.rc$
$results/tables/supp/housingContrast_and_FruContrast.uniq.tsv$

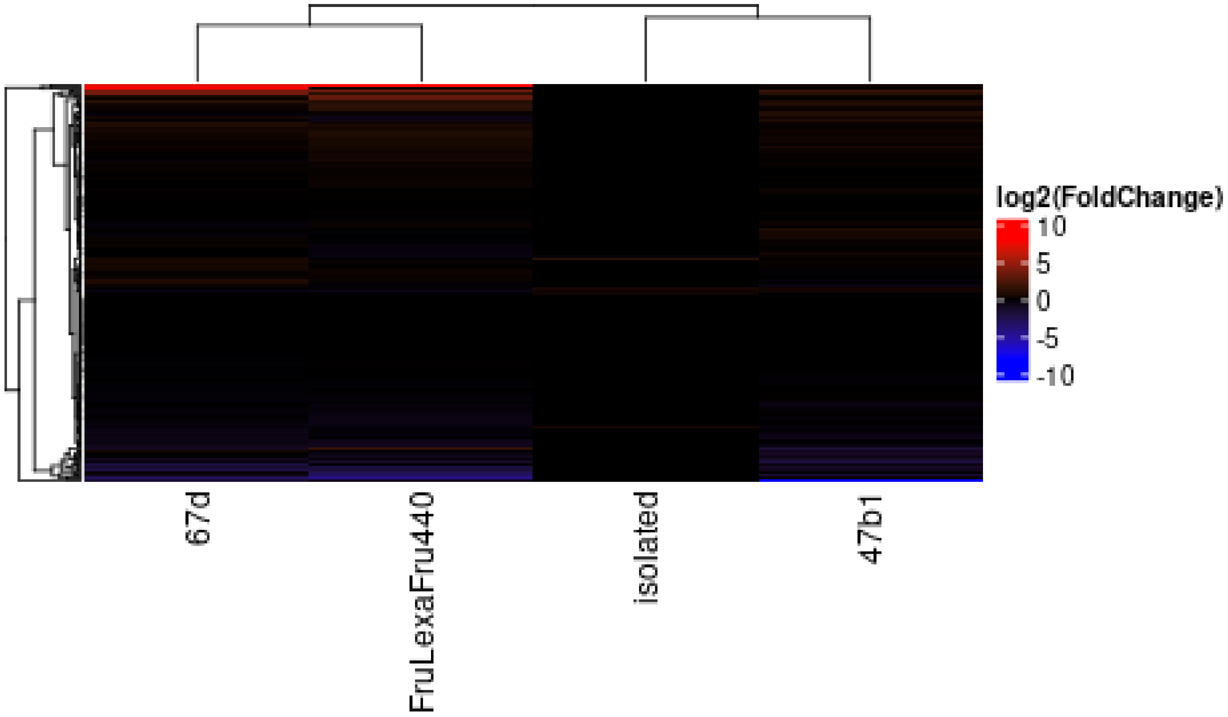
3.5.4 Overview (Heatmaps)

We can also display changes in gene expression as a heatmap. Increases in expression are show in red, and decreases in blue. Significance of change is not currently indicated.

3.5.4.1 Ion Channel Activity

Here is a heat map specific to the Ion Channel Activity genes

Figure 91 . Heatmap Displaying Difference in Expression in Different Experimental Contrasts (Ion Channel Activity Genes)(multi alignment)



```
## png
##    2
```

3.6 Comparing Expression Changes Between Mutants

do this

3.6.1 Fru & 67d

do this

3.6.2 Fru & 47b

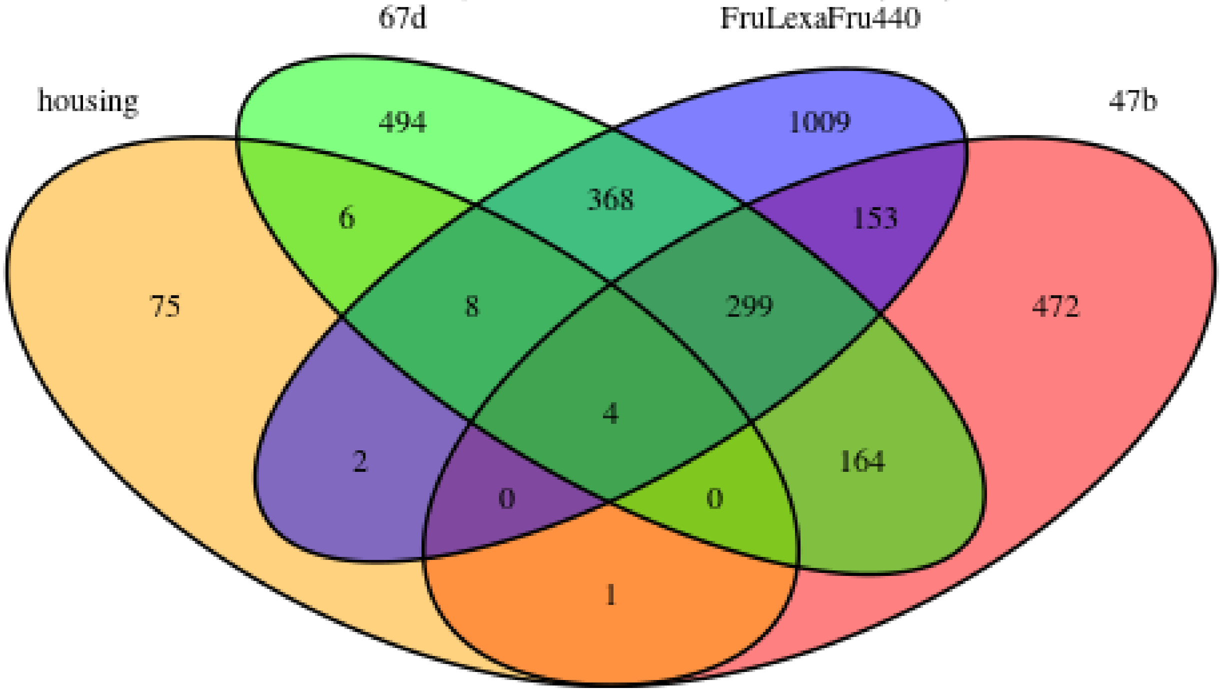
do this

3.6.3 47b & 67d

do this

3.7 Mutually Significant Differential Expression Overview

Figure 92 . Venn Diagram: # genes with shared significant change, by experimental contrast intersection (multi)



```
## null device
##           1
```

```
## null device
##           1
```

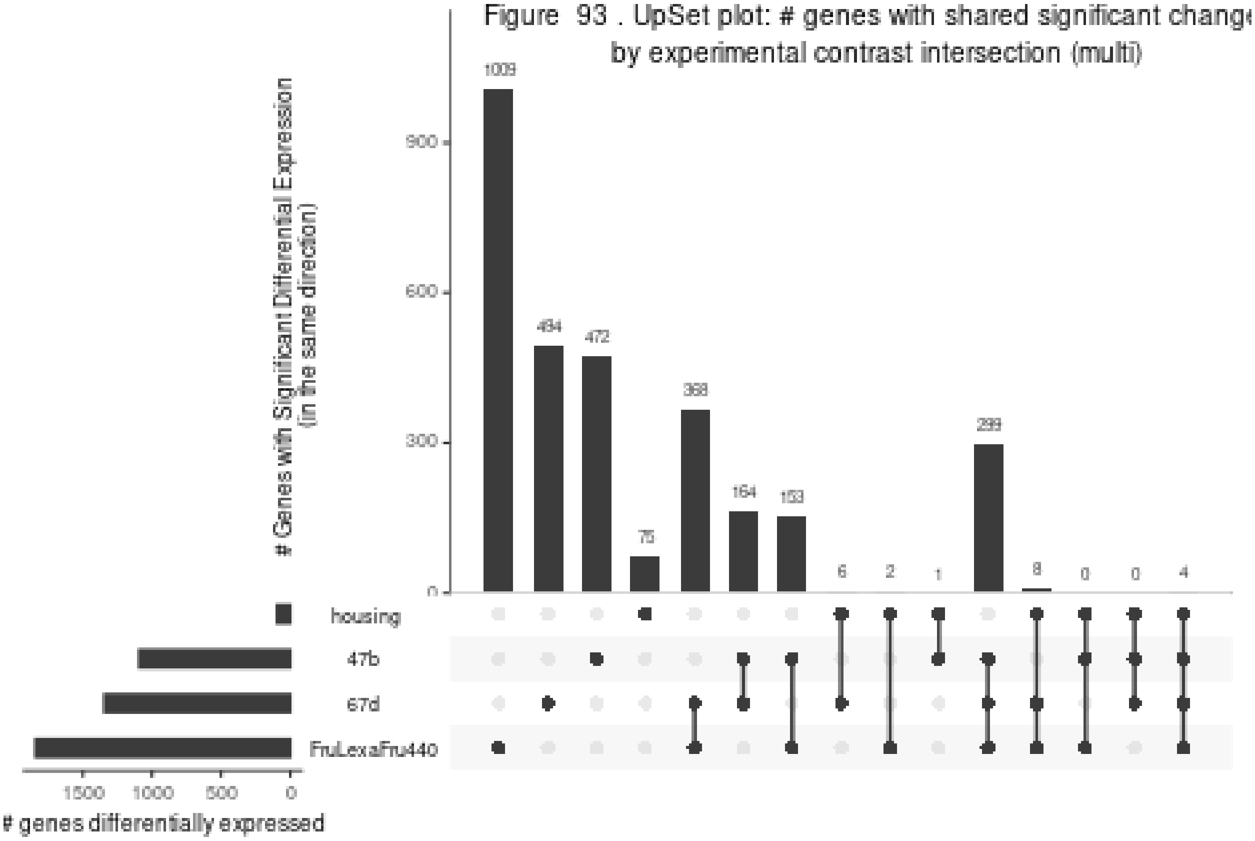


Figure 93 . UpSet plot: # genes with shared significant change by experimental contrast intersection (multi)

```
## null device
##           1
```

```
## null device
##           1
```

The two genes with the same behavior across all experimental contrasts are javelin and CG13659. Both are enriched in all cases:

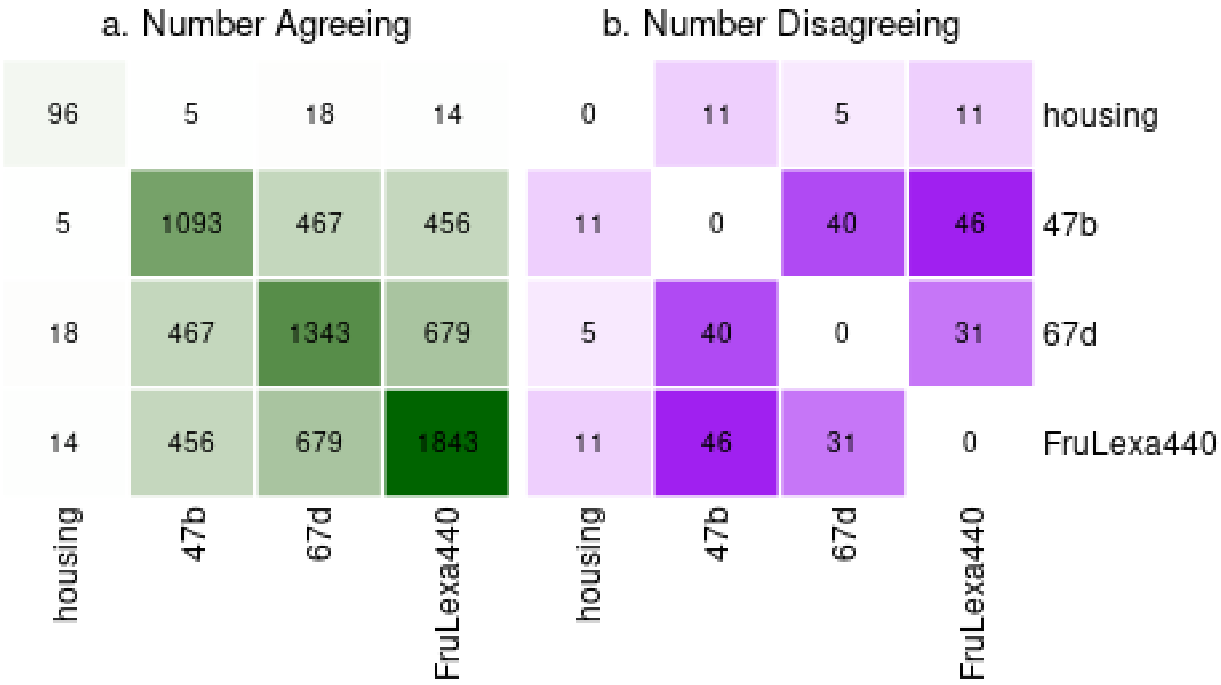
Table 104. Genes sharing significant differential expression in all four contrasts

	47b vs wt		67d vs wt		FruLexaFru440 vs wt		isc
	log2FoldChange	padj	log2FoldChange	padj	log2FoldChange	padj	log2FoldC
multi							
CG12986	1.220	5.45×10^{-9}	2.012	1.44×10^{-24}	1.687	1.44×10^{-16}	
CG13659	0.447	2.03×10^{-3}	1.364	4.15×10^{-27}	1.544	3.17×10^{-34}	

CG43147	0.031	4.09×10^{-5}	0.028	3.71×10^{-4}	0.038	1.24×10^{-6}
jv	1.216	3.47×10^{-40}	1.006	2.81×10^{-27}	0.521	1.95×10^{-7}
rando						
CG12986	1.221	4.51×10^{-9}	2.012	1.05×10^{-24}	1.689	1.10×10^{-16}
CG13659	0.446	1.83×10^{-3}	1.364	2.59×10^{-27}	1.546	1.50×10^{-34}
CG43147	0.030	4.12×10^{-5}	0.028	3.85×10^{-4}	0.038	1.31×10^{-6}
jv	1.219	1.71×10^{-40}	1.006	1.99×10^{-27}	0.521	1.66×10^{-7}
uniq						
CG12986	1.221	4.56×10^{-9}	2.013	9.44×10^{-25}	1.689	1.04×10^{-16}
CG13659	0.447	1.77×10^{-3}	1.365	1.75×10^{-27}	1.547	9.31×10^{-35}
CG43147	0.029	4.48×10^{-5}	0.028	3.98×10^{-4}	0.037	1.46×10^{-6}
jv	1.217	2.48×10^{-40}	1.006	2.39×10^{-27}	0.523	1.73×10^{-7}

results shown are for multi only; very similar across aligner strategies

Figure 94 Heatmap of Pairwise Comparisons between Contrasts:
# significant genes with the same (left)
or different (right) directions of change
(2-factor models)



a. Number Agreeing

	housing	47b	67d	FruLexa440
housing	96	5	18	14
47b	5	1093	467	456
67d	18	467	1343	679
FruLexa440	14	456	679	1843

b. Number Disagreeing

	housing	47b	67d	FruLexa440
housing	0	11	5	11
47b	11	0	40	46
67d	5	40	0	31
FruLexa440	11	46	31	0

```
## png
##   2
```

3.8 Focus on Fruitless

Table 105a. Differential Expression of Fruitless
(single factor)

	significance (p)	effect size (l2fc)
housing		
multi	0.47	0.06
rando	0.47	0.06
uniq	0.47	0.06
47b		
multi	0.83	0.04
rando	0.82	0.04
uniq	0.83	0.04
67d		
multi	0.55	0.11
rando	0.55	0.11
uniq	0.55	0.11
Fru		
multi	0.51	0.20
rando	0.51	0.20
uniq	0.51	0.20

Table 105b. Differential Expression of Fruitless
(multifactor)

	effect size (l2fc)	significance (p)
47b1		
multi	0.02	0.98
rando	0.02	0.98
uniq	0.02	0.98
67d		
multi	0.07	0.87
rando	0.07	0.87
uniq	0.07	0.87
FruLexaFru440		
multi	0.33	0.26
rando	0.33	0.26
uniq	0.33	0.26
isolated		
multi	0.00	1.00
rando	0.00	1.00
uniq	0.00	1.00

Changes in splicing of Fruitless are of special interest, and feature counting/differential expression testing was performed on an annotation which considers all available exons separately. In this way, changes in exon use by treatment might be detected.

3.8.1 By Exon

3.8.1.1 Ambiguous Read Assignment: None

The default featureCounts settings ignore ambiguously assigned reads. Because some exons overlap and because junction-spanning reads will be considered ambuiguous in this context, some relevant reads might be being ignored and deflating the power in these tests. Several exons were filtered out entirely based on low read count number. Here are the results from this assignment strategy.

Table 106. Number of Fruitless Exons Available For Analysis
('none' counting, by aligner)

aligner	count	frac	total
multi	14	63.6%	22
rando	14	63.6%	22
uniq	14	63.6%	22

The only exons with even marginally significant differential expression in any contrast are 18 20, and 22, in the FruLexa/Fru400 contrast:

Table 107. Differential Use of Fruitless Exons, by Contrast
('none'counting, multi only)

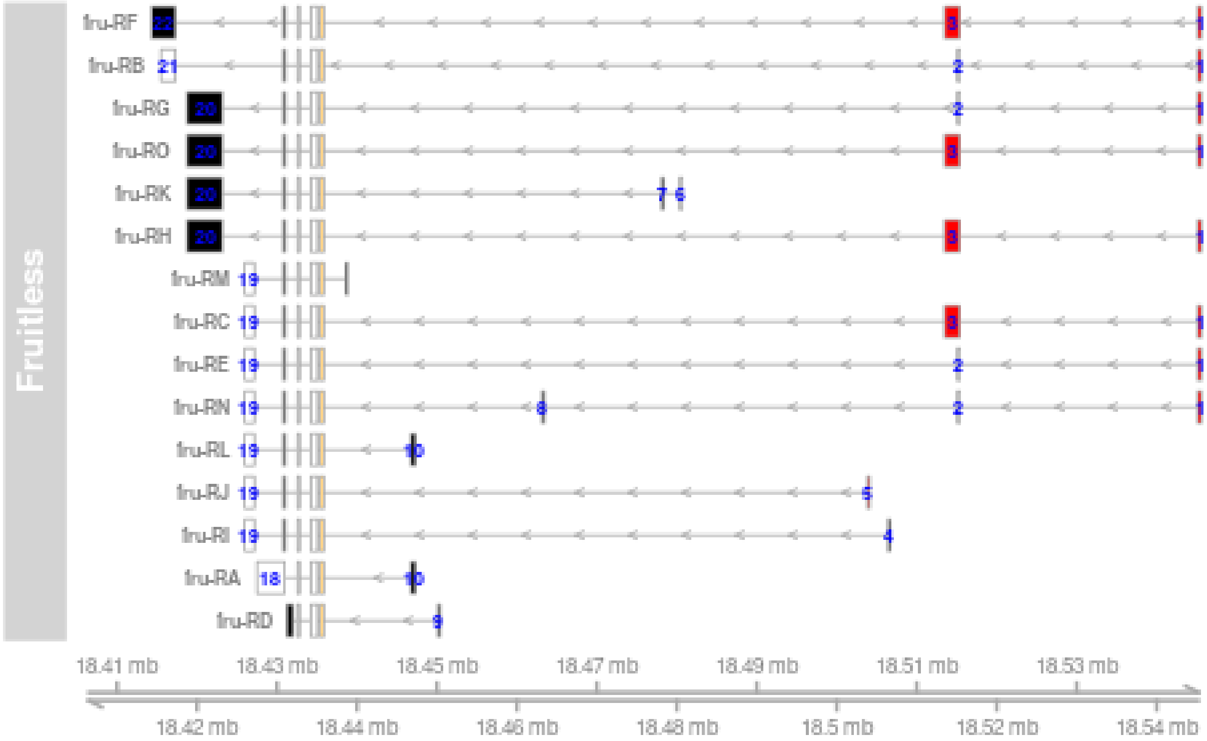
	47b		67d		Fru		wt
	log2FoldChange	adjusted p	log2FoldChange	adjusted p	log2FoldChange	adjusted p	log2FoldCha
exon_1	0.0000	9.61×10^{-1}	0.0000	7.82×10^{-1}	1.4928	3.04×10^{-2}	0.0
exon_10	0.0000	3.79×10^{-1}	0.0000	7.82×10^{-1}	−0.1628	8.73×10^{-1}	−0.0
exon_11	0.0000	9.61×10^{-1}	0.0000	8.90×10^{-1}	0.0452	9.02×10^{-1}	0.0
exon_13	0.0000	9.61×10^{-1}	0.0000	7.82×10^{-1}	0.7224	7.01×10^{-2}	0.0
exon_16	0.0000	3.77×10^{-1}	0.0000	4.04×10^{-1}	0.2955	1.95×10^{-1}	0.0
exon_17	0.0000	3.77×10^{-1}	0.0000	7.82×10^{-1}	1.3600	5.20×10^{-2}	−0.0
exon_20	0.0000	9.61×10^{-1}	0.0000	9.68×10^{-1}	0.1060	9.02×10^{-1}	0.0
exon_22	0.0000	3.77×10^{-1}	0.0000	7.82×10^{-1}	−0.0539	9.02×10^{-1}	0.0
exon_3	0.0000	9.61×10^{-1}	0.0000	7.82×10^{-1}	1.0176	3.04×10^{-2}	−0.0
exon_4	0.0000	9.61×10^{-1}	0.0000	7.82×10^{-1}	−0.0476	9.02×10^{-1}	0.0
exon_5	0.0000	3.79×10^{-1}	0.0000	4.04×10^{-1}	2.1427	2.13×10^{-3}	−0.0
exon_7	0.0000	9.61×10^{-1}	0.0000	7.82×10^{-1}	−0.0335	9.02×10^{-1}	−0.0
exon_8	0.0000	9.61×10^{-1}	0.0000	7.82×10^{-1}	−0.1860	8.73×10^{-1}	0.0
exon_9	0.0000	9.61×10^{-1}	0.0000	8.90×10^{-1}	0.0911	9.02×10^{-1}	−0.0

Table 108. Fru exons with significantly (padj<0.05) differential use
('none' counting)

	Fru	
aligner	log2FoldChange	adjusted p
exon_1		
multi	1.49	0.03
rando	1.49	0.03
uniq	1.49	0.03

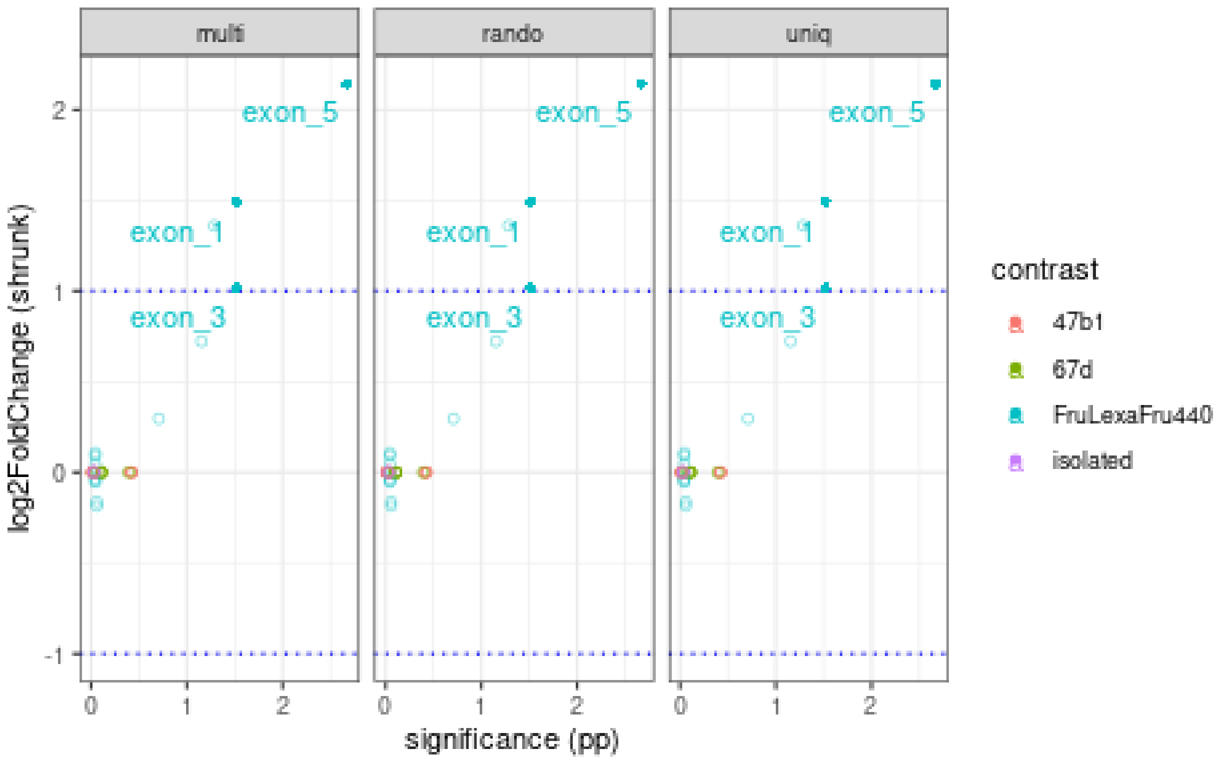
exon_3		
multi	1.02	0.03
rando	1.02	0.03
uniq	1.02	0.03
exon_5		
multi	2.14	0.002
rando	2.14	0.002
uniq	2.14	0.002

Figure 95. Fruitless gene model: exons with any significant change detected highlighted (any contrast, any aligner, ambiguous assigned to none)



```
## png
##   2
```

Figure 96. Volcano Plot: Fold Change vs. Significance
(fruitless exons, 'none' counting strategy)



```
## png
##   2
```

3.8.1.2 Ambiguous Read Assignment: All

Here, ambiguous reads have been assigned to every feature they overlap, rather than none.

Table 109. Number of Fru Exons Available For Analysis
(by aligner)

aligner	count	frac	total
multi	18	81.8%	22
rando	18	81.8%	22
uniq	18	81.8%	22

only the FruLexa/Fru440 contrast had significantly different exon use:

Table 110. Differential Use of Fru Exons, by Contrast
('all' counting, multi only)

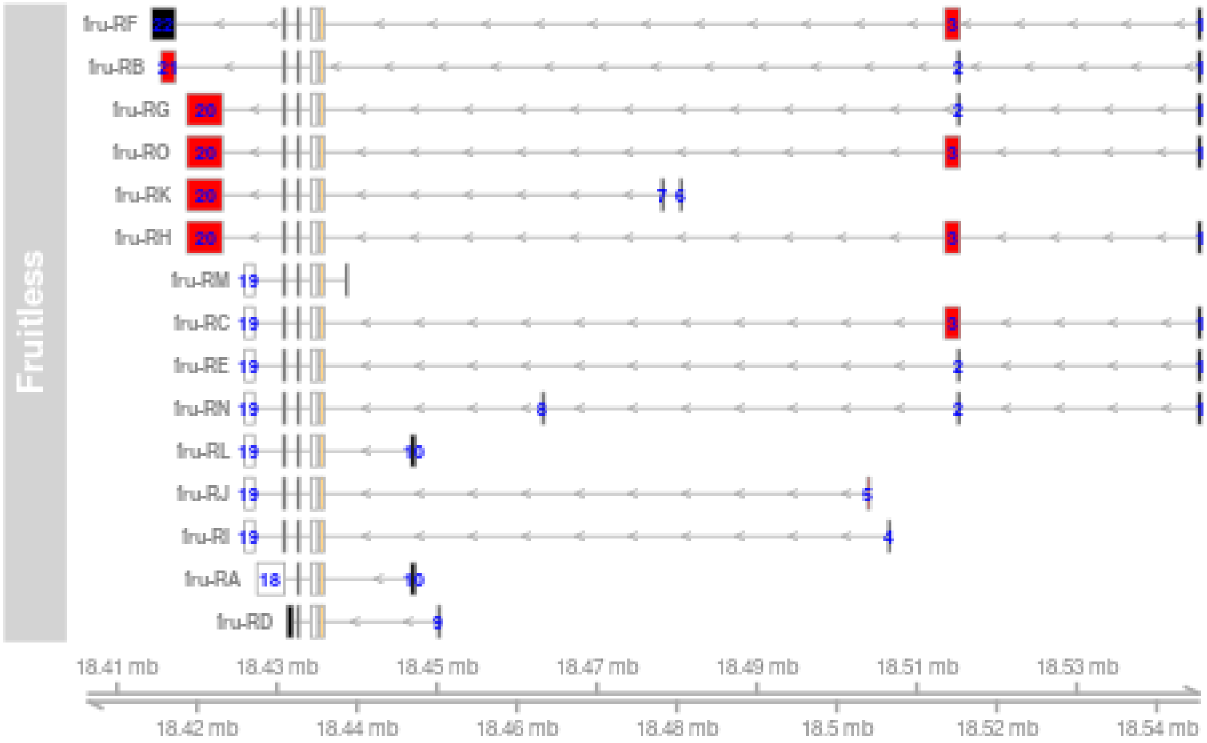
	47b		67d		Fru		wt
	log2FoldChange	adjusted p	log2FoldChange	adjusted p	log2FoldChange	adjusted p	log2FoldCha
exon_1	0.00	9.51×10^{-1}	0.00	9.36×10^{-1}	0.82	7.33×10^{-2}	0

exon_10	0.03	3.35×10^{-1}	-0.01	8.25×10^{-1}	-0.18	4.41×10^{-1}
exon_11	-0.02	7.03×10^{-1}	0.00	9.36×10^{-1}	-0.01	9.61×10^{-1}
exon_13	-0.01	7.03×10^{-1}	0.00	8.27×10^{-1}	0.66	5.74×10^{-2}
exon_15	0.00	7.03×10^{-1}	0.00	8.27×10^{-1}	0.73	1.40×10^{-1}
exon_16	0.00	3.35×10^{-1}	0.00	5.70×10^{-1}	0.52	1.91×10^{-1}
exon_17	0.00	3.35×10^{-1}	0.00	9.36×10^{-1}	1.33	7.99×10^{-2}
exon_2	0.00	8.58×10^{-1}	0.00	9.36×10^{-1}	0.46	1.65×10^{-1}
exon_20	-0.01	7.03×10^{-1}	0.00	9.36×10^{-1}	-1.92	4.23×10^{-6}
exon_21	-0.01	7.03×10^{-1}	0.00	9.36×10^{-1}	-3.52	1.38×10^{-6}
exon_22	-0.04	3.35×10^{-1}	-0.01	8.25×10^{-1}	-0.33	1.91×10^{-1}
exon_3	0.00	8.50×10^{-1}	0.00	8.27×10^{-1}	1.03	2.58×10^{-2}
exon_4	-0.01	8.50×10^{-1}	-0.01	5.70×10^{-1}	-0.08	6.86×10^{-1}
exon_5	0.01	3.35×10^{-1}	0.00	5.70×10^{-1}	1.61	1.70×10^{-3}
exon_6	0.02	3.35×10^{-1}	0.00	8.27×10^{-1}	-0.10	8.44×10^{-1}
exon_7	0.00	9.56×10^{-1}	0.00	9.36×10^{-1}	0.02	9.61×10^{-1}
exon_8	0.02	7.03×10^{-1}	-0.01	8.27×10^{-1}	-0.18	4.84×10^{-1}
exon_9	-0.01	8.58×10^{-1}	-0.01	8.27×10^{-1}	-0.05	8.44×10^{-1}

Table 111. Fru exons with significantly (padj<0.05) different use
('all' counting, by aligner)

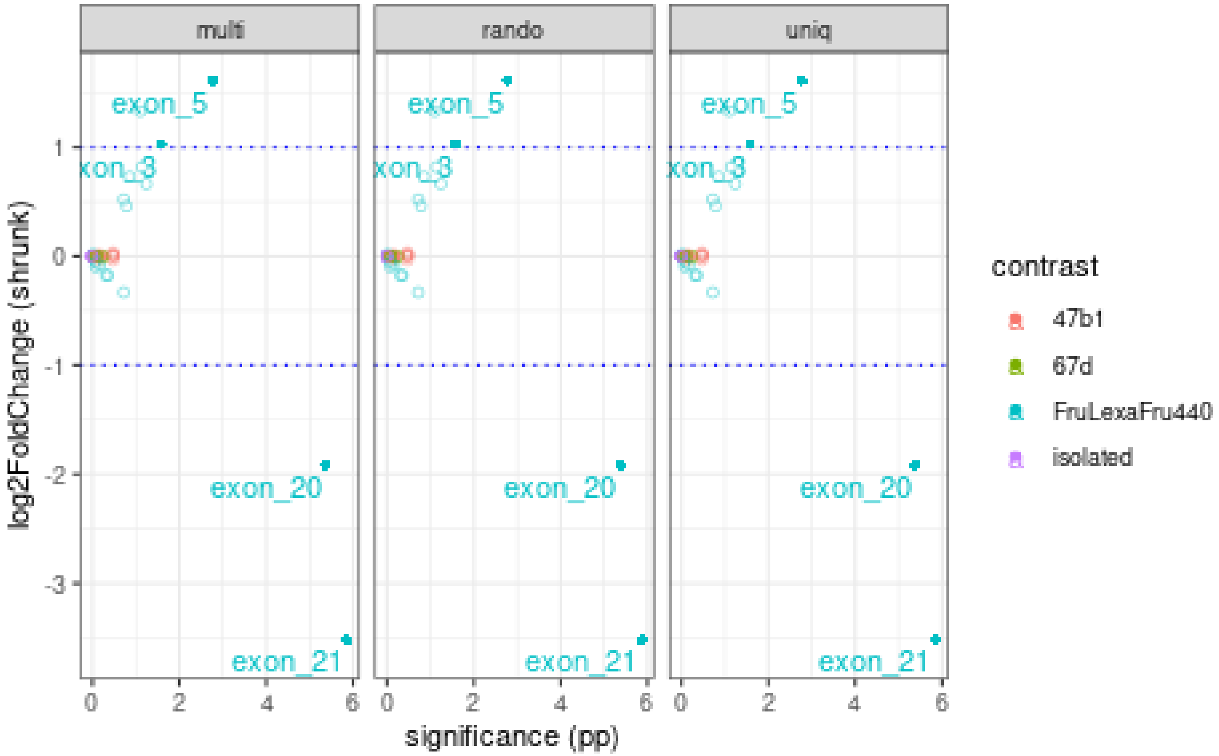
	Fru	
aligner	log2FoldChange	adjusted p
exon_20		
multi	-1.92	4.23×10^{-6}
rando	-1.92	4.23×10^{-6}
uniq	-1.92	4.23×10^{-6}
exon_21		
multi	-3.52	1.38×10^{-6}
rando	-3.52	1.38×10^{-6}
uniq	-3.52	1.38×10^{-6}
exon_3		
multi	1.03	2.58×10^{-2}
rando	1.03	2.58×10^{-2}
uniq	1.03	2.58×10^{-2}
exon_5		
multi	1.61	1.70×10^{-3}
rando	1.61	1.70×10^{-3}
uniq	1.61	1.70×10^{-3}

Figure 97. Fruitless gene model: exons with any significant change detected highlighted (any contrast, any aligner, ambiguous assigned to all)



```
## png
##   2
```

Figure 98. Volcano Plot: Fold Change vs. Significance
(fruitless exons, 'all' counting strategy, padj < 0.05)



```
## png
##   2
```

3.8.2 By Exon Junction

When the *_SplicedOnly alignments were counted ("all" strategy) against the fru_junct annotation:

Table 112. Number of Fru Exons Available For Analysis
(spliced reads counted by splice junction)

	count	fraction
multi	15	68.2%
rando	15	68.2%
uniq	15	68.2%

Table 113. Differential Exon Use in Fruitless, by Contrast
Junction-based, 'all' counting (Multi only)

	47b		67d		Fru		wt
	log2FoldChange	adjusted p	log2FoldChange	adjusted p	log2FoldChange	adjusted p	log2FoldCha
exon_1	−0.01	8.29×10^{-1}	0.00	8.78×10^{-1}	0.00	1.80×10^{-2}	
exon_10	−0.15	3.92×10^{-1}	0.00	8.39×10^{-1}	0.00	7.17×10^{-1}	

exon_12	−0.15	3.92×10^{-1}	0.00	8.39×10^{-1}	0.00	7.17×10^{-1}
exon_13	0.00	9.76×10^{-1}	0.00	8.78×10^{-1}	0.00	9.99×10^{-1}
exon_14	−0.09	4.77×10^{-1}	0.00	8.39×10^{-1}	0.00	4.35×10^{-1}
exon_15	0.07	5.32×10^{-1}	0.00	8.78×10^{-1}	0.00	9.90×10^{-1}
exon_16	0.01	4.13×10^{-1}	0.00	5.91×10^{-1}	0.00	4.35×10^{-1}
exon_17	0.10	4.13×10^{-1}	0.00	8.99×10^{-1}	0.00	9.09×10^{-1}
exon_18	0.69	1.14×10^{-1}	0.00	7.59×10^{-1}	0.00	5.86×10^{-1}
exon_19	−0.06	4.13×10^{-1}	0.00	8.39×10^{-1}	0.00	5.95×10^{-1}
exon_2	−0.09	4.13×10^{-1}	0.00	8.78×10^{-1}	0.00	5.99×10^{-4}
exon_20	0.02	6.19×10^{-1}	0.00	8.39×10^{-1}	0.00	9.09×10^{-1}
exon_21	−0.03	6.77×10^{-1}	0.00	3.86×10^{-1}	0.00	5.86×10^{-1}
exon_22	−0.03	6.77×10^{-1}	0.00	3.86×10^{-1}	0.00	5.86×10^{-1}
exon_3	−0.01	8.46×10^{-1}	0.00	8.78×10^{-1}	0.00	1.80×10^{-2}

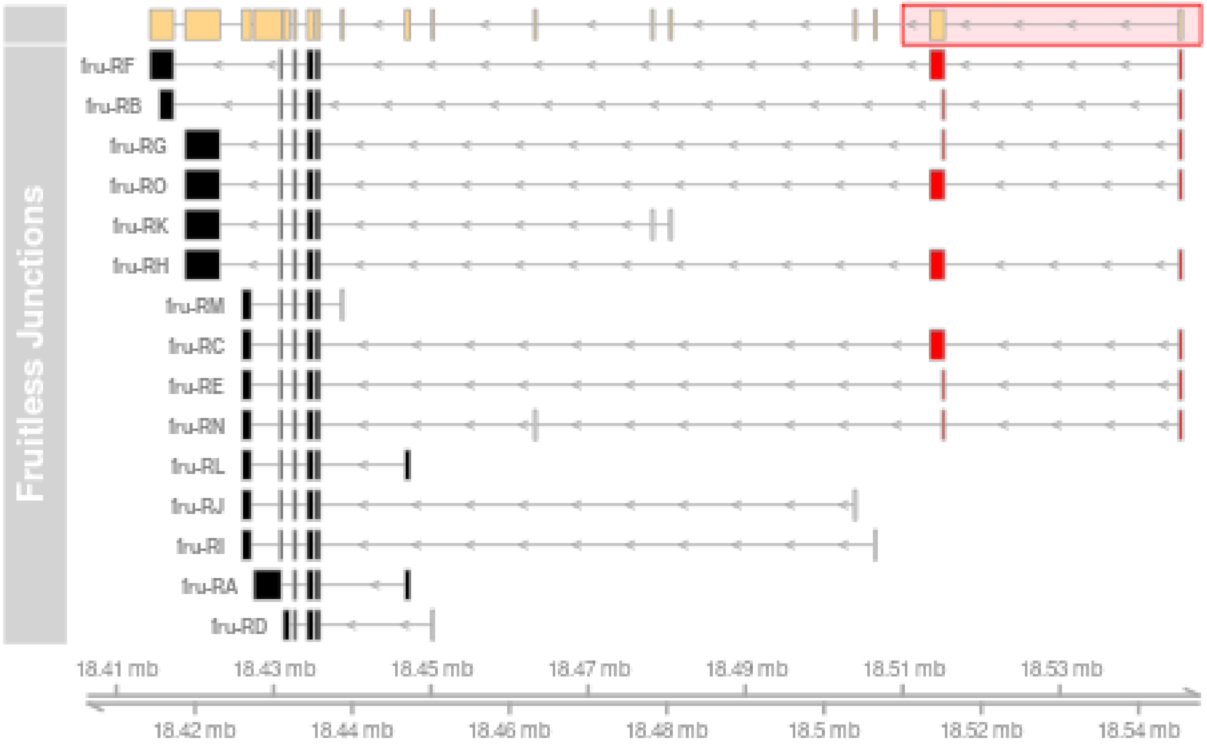
Exons 1,2, and 3, the most 5' of exons, are less used in the FruLexa/Fru440 contrast; however, exons 1 and 3 bizarrely low effect sizes given their significance:

Table 114. Fru exons with significantly (padj<0.05) different use

Junction-based, 'all' counting

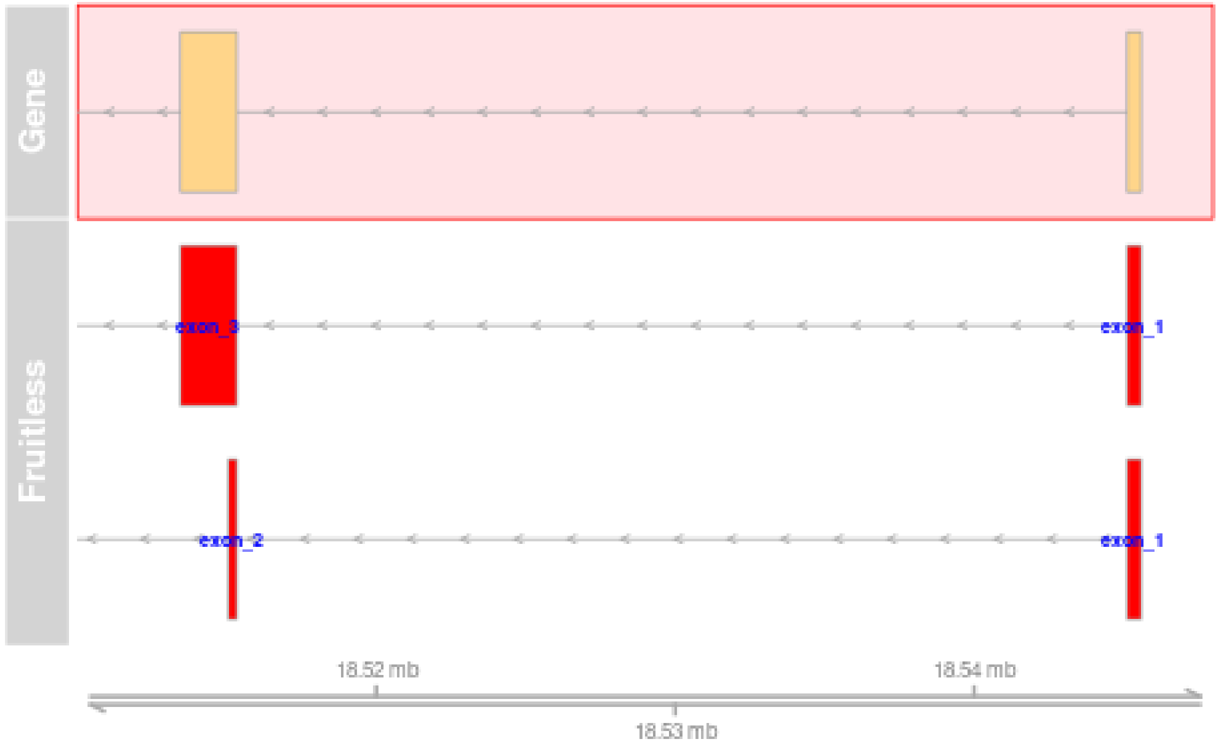
	Fru	
aligner	log2FoldChange	adjusted p
exon_1		
multi	0.00	0.018
rando	0.00	0.018
uniq	0.00	0.018
exon_2		
multi	0.00	0.001
rando	0.00	0.001
uniq	0.00	0.001
exon_3		
multi	0.00	0.018
rando	0.00	0.018
uniq	0.00	0.018

Figure 99 Fruitless exons with significant change in use (measured by junction)



```
## png
##   2
```

igure 99 a. Fruitless exons with significant change in use (measured by junction) (detail



```
## png
##   2
```

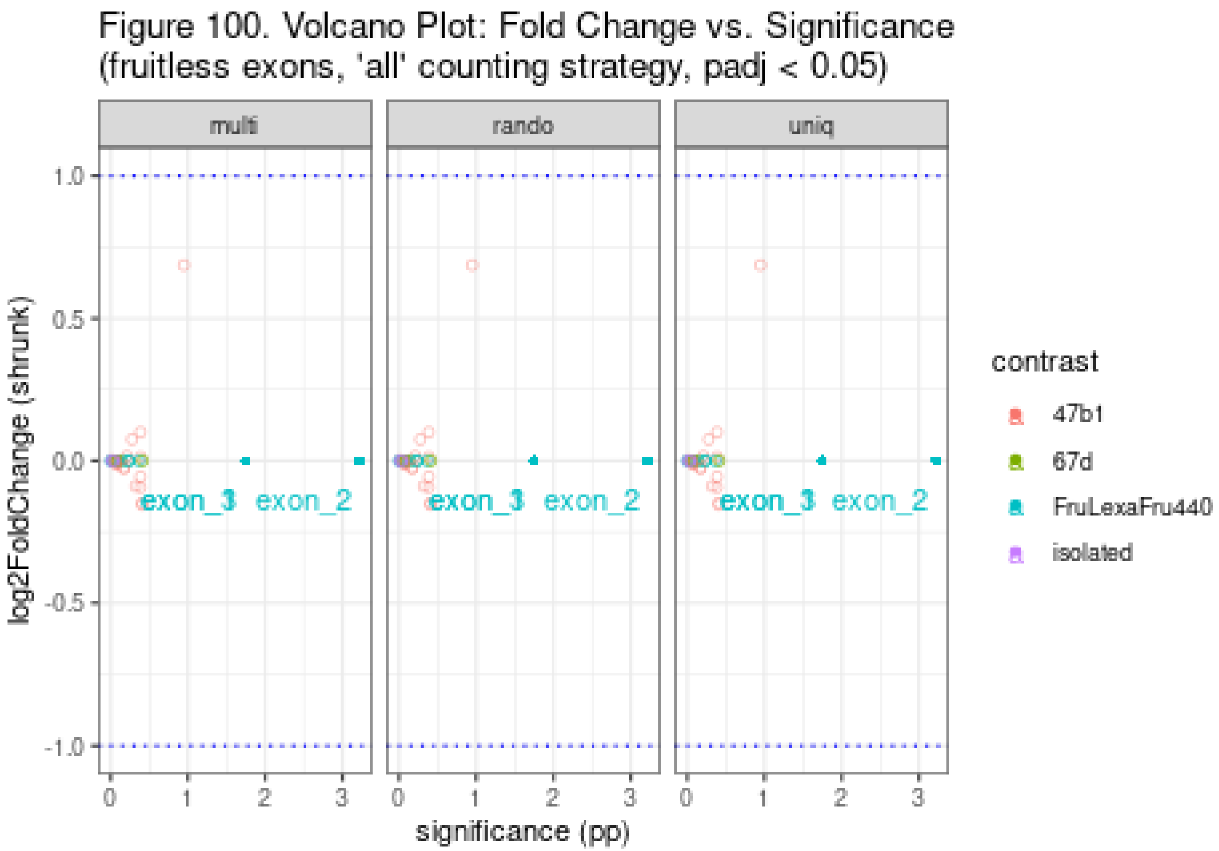


Figure 100. Volcano Plot: Fold Change vs. Significance
(fruitless exons, 'all' counting strategy, padj < 0.05)

```
## png
##   2
```

3.8.3 By Intron

When the *_SplicedOnly alignments were counted(“all” strategy) against the fru_introns annotation:

Table 115. Number of Fru Introns Available For Analysis
(spliced reads counted by intron boundaries)

	count	fraction
multi	19	86.4%
rando	19	86.4%
uniq	19	86.4%

Table 116. Differential Intron Use in Fruitless, by Contrast
(Multi only)

	47b		67d		Fru		w
	log2FoldChange	adjusted p	log2FoldChange	adjusted p	log2FoldChange	adjusted p	log2FoldCh
intron_1	0.37	5.56×10^{-1}	0.00	1.00×10^{0}	−2.12	1.73×10^{-2}	
intron_10	−0.02	9.32×10^{-1}	0.00	1.00×10^{0}	0.00	9.83×10^{-1}	

intron_11	-0.02	9.32×10^{-1}	0.00	1.00×10^{0}	0.00	9.83×10^{-1}
intron_12	-0.02	9.32×10^{-1}	0.00	1.00×10^{0}	0.00	9.83×10^{-1}
intron_13	0.65	3.43×10^{-2}	0.00	1.00×10^{0}	0.00	8.10×10^{-1}
intron_14	0.37	5.56×10^{-1}	0.00	1.00×10^{0}	0.00	8.10×10^{-1}
intron_15	0.30	5.56×10^{-1}	0.00	8.38×10^{-1}	0.00	9.24×10^{-1}
intron_16	2.10	1.75×10^{-6}	0.00	7.56×10^{-1}	0.00	2.79×10^{-1}
intron_17	2.10	1.75×10^{-6}	0.00	7.56×10^{-1}	0.00	2.79×10^{-1}
intron_18	0.23	6.48×10^{-1}	0.00	8.38×10^{-1}	0.00	8.10×10^{-1}
intron_19	0.23	6.48×10^{-1}	0.00	8.38×10^{-1}	0.00	8.10×10^{-1}
intron_20	0.25	6.48×10^{-1}	0.00	8.38×10^{-1}	0.00	8.10×10^{-1}
intron_3	-0.29	6.48×10^{-1}	0.00	1.00×10^{0}	0.00	2.09×10^{-3}
intron_4	-0.03	9.32×10^{-1}	0.00	1.00×10^{0}	0.00	9.83×10^{-1}
intron_5	-0.02	9.32×10^{-1}	0.00	1.00×10^{0}	0.00	9.83×10^{-1}
intron_6	-0.02	9.32×10^{-1}	0.00	1.00×10^{0}	0.00	9.83×10^{-1}
intron_7	-0.02	9.32×10^{-1}	0.00	1.00×10^{0}	0.00	9.83×10^{-1}
intron_8	-0.02	9.32×10^{-1}	0.00	1.00×10^{0}	0.00	9.83×10^{-1}
intron_9	-0.02	9.32×10^{-1}	0.00	1.00×10^{0}	0.00	9.83×10^{-1}

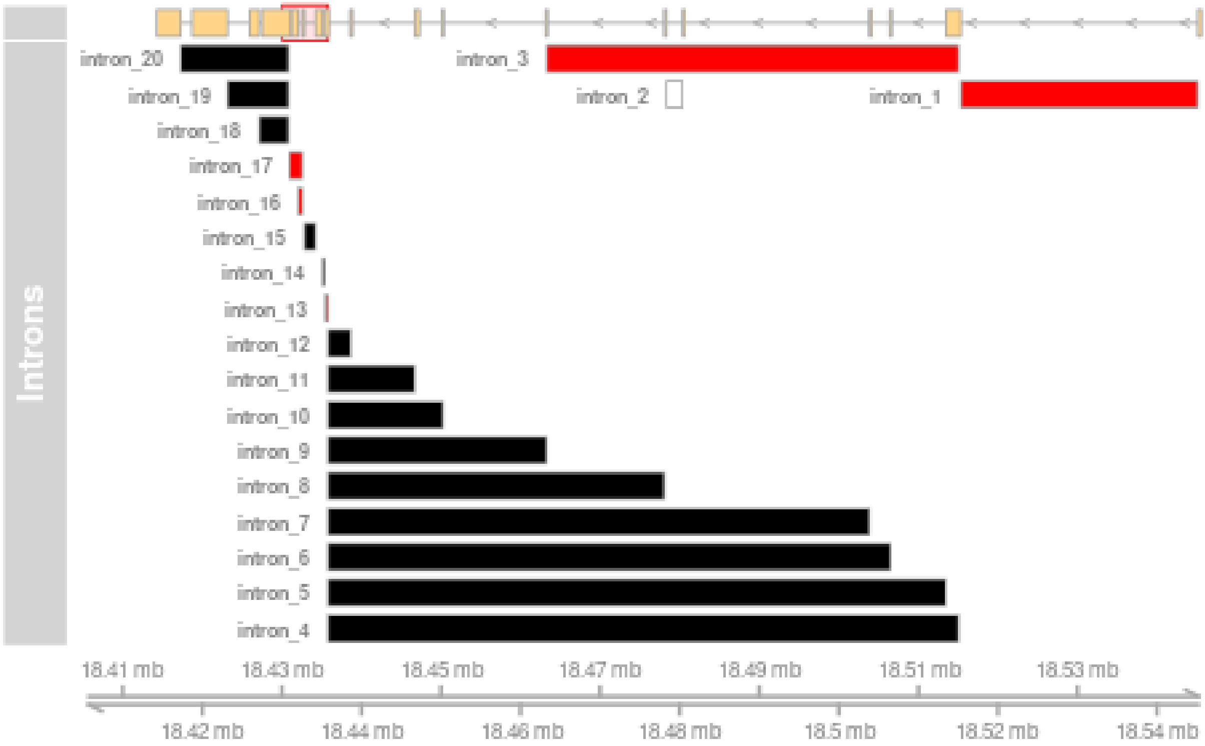
Introns 1 and 3 come up significant in the FruLexa/Fru440 contrast, though they have bizarrely small effect sizes. Introns 16 and 17 come up significant in the 47b contrast.

Table 117. Fru introns with significantly (padj<0.05) different use

(by aligner)

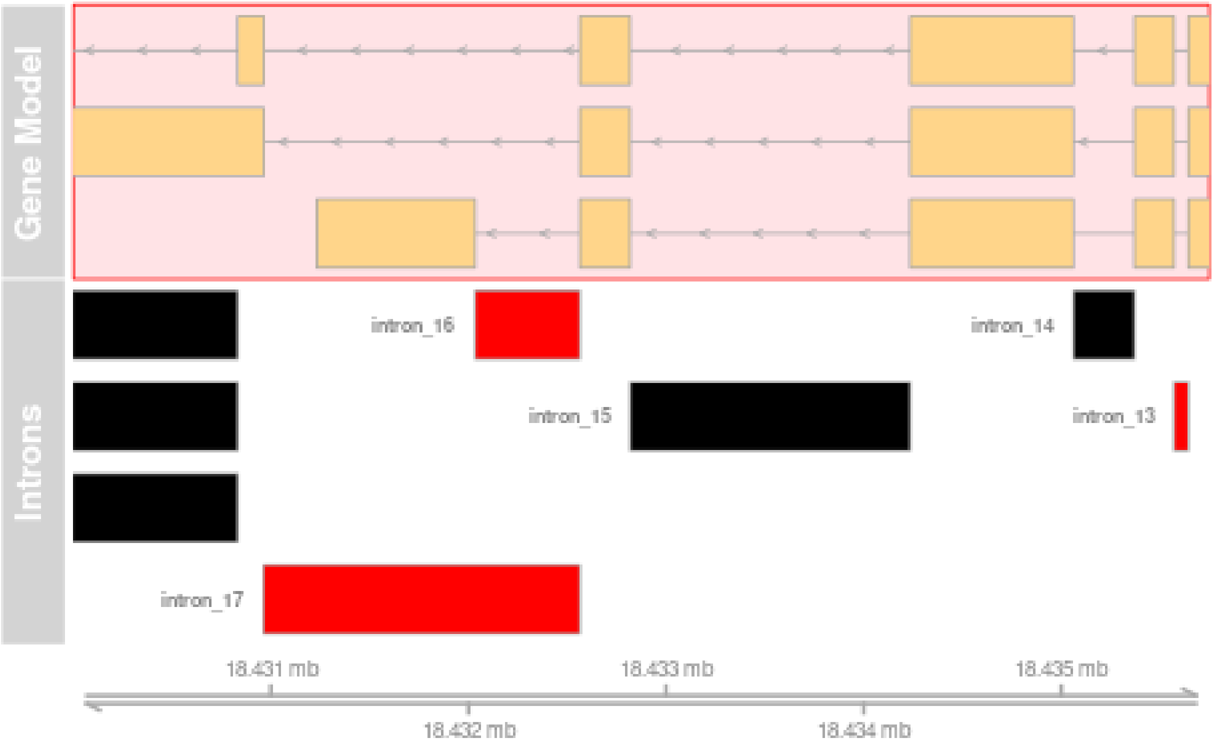
	log2FoldChange	adjusted p
intron_1 - FruLexaFru440		
multi	-2.12	1.73×10^{-2}
rando	-2.12	1.73×10^{-2}
uniq	-2.12	1.73×10^{-2}
intron_13 - 47b1		
multi	0.65	3.43×10^{-2}
rando	0.65	3.43×10^{-2}
uniq	0.65	3.43×10^{-2}
intron_16 - 47b1		
multi	2.10	1.75×10^{-6}
rando	2.10	1.75×10^{-6}
uniq	2.10	1.75×10^{-6}
intron_17 - 47b1		
multi	2.10	1.75×10^{-6}
rando	2.10	1.75×10^{-6}
uniq	2.10	1.75×10^{-6}
intron_3 - FruLexaFru440		
multi	0.00	2.09×10^{-3}
rando	0.00	2.09×10^{-3}
uniq	0.00	2.09×10^{-3}

Figure 101 Fruitless introns with significant change



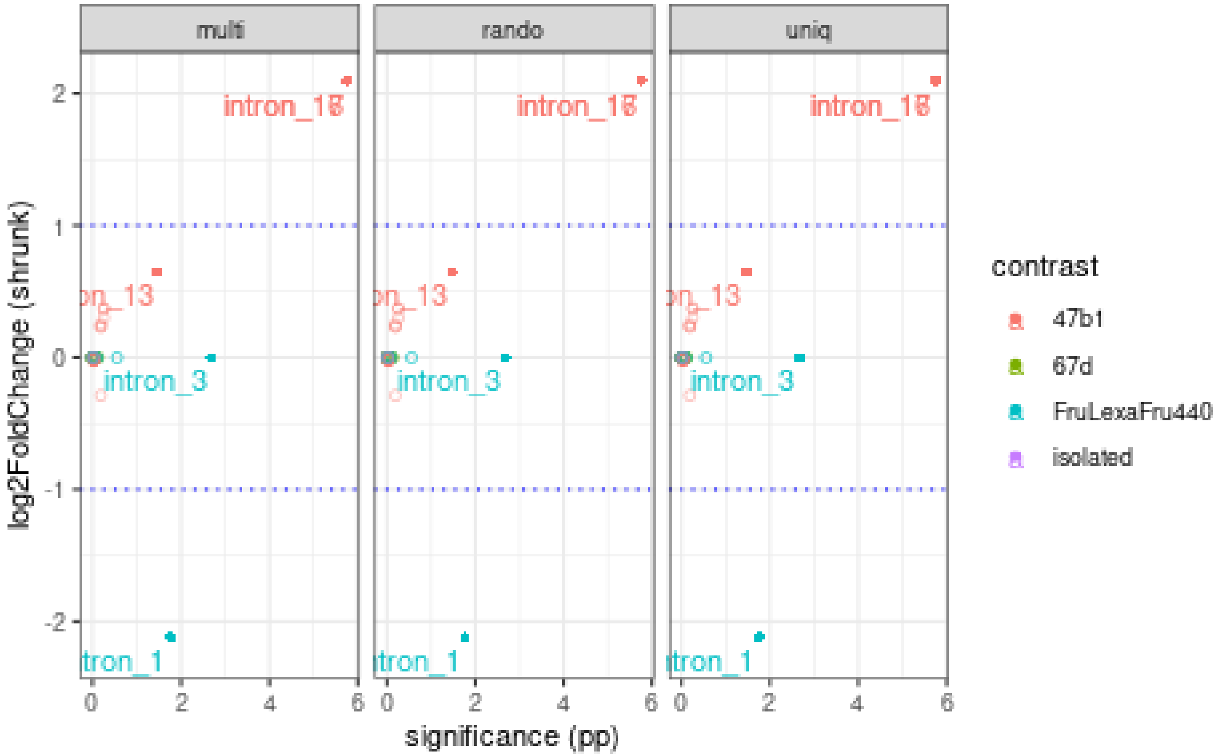
```
## png
##   2
```

Figure 101 a. Fruitless introns with significant change (detail)



```
## png
##   2
```

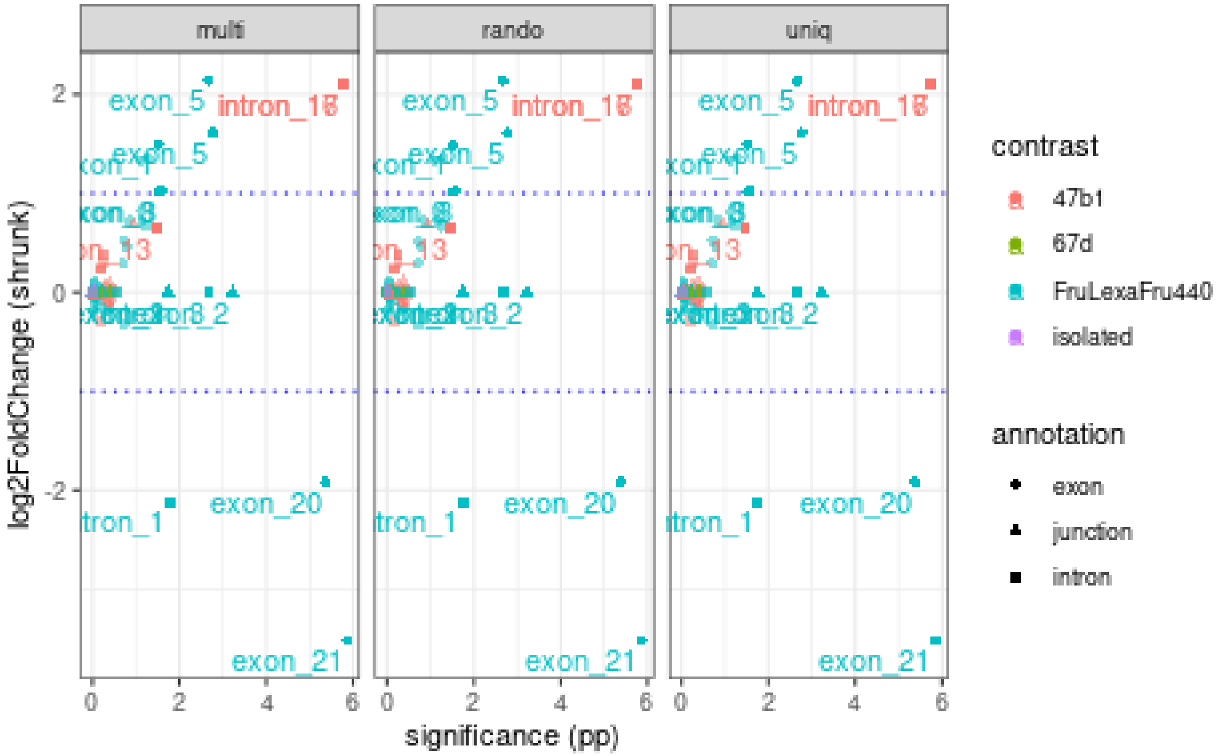
Figure 102. Volcano Plot: Fold Change vs. Significance
(fruitless introns, 'all' counting strategy, padj < 0.05)



```
## png
##   2
```

3.8.4 Overall

Figure 103. Volcano Plot: Fold Change vs. Significance
(all fruitless breakdowns, 'all' counting strategy, padj < 0.05)

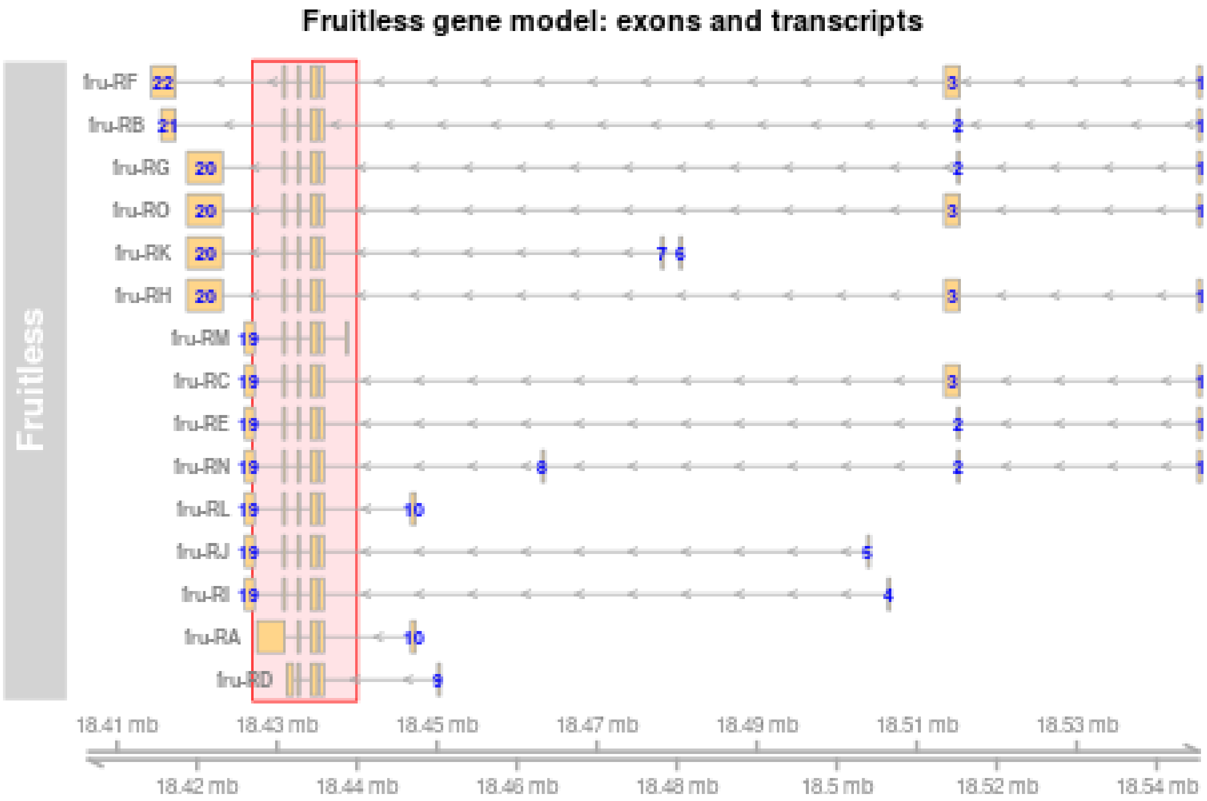


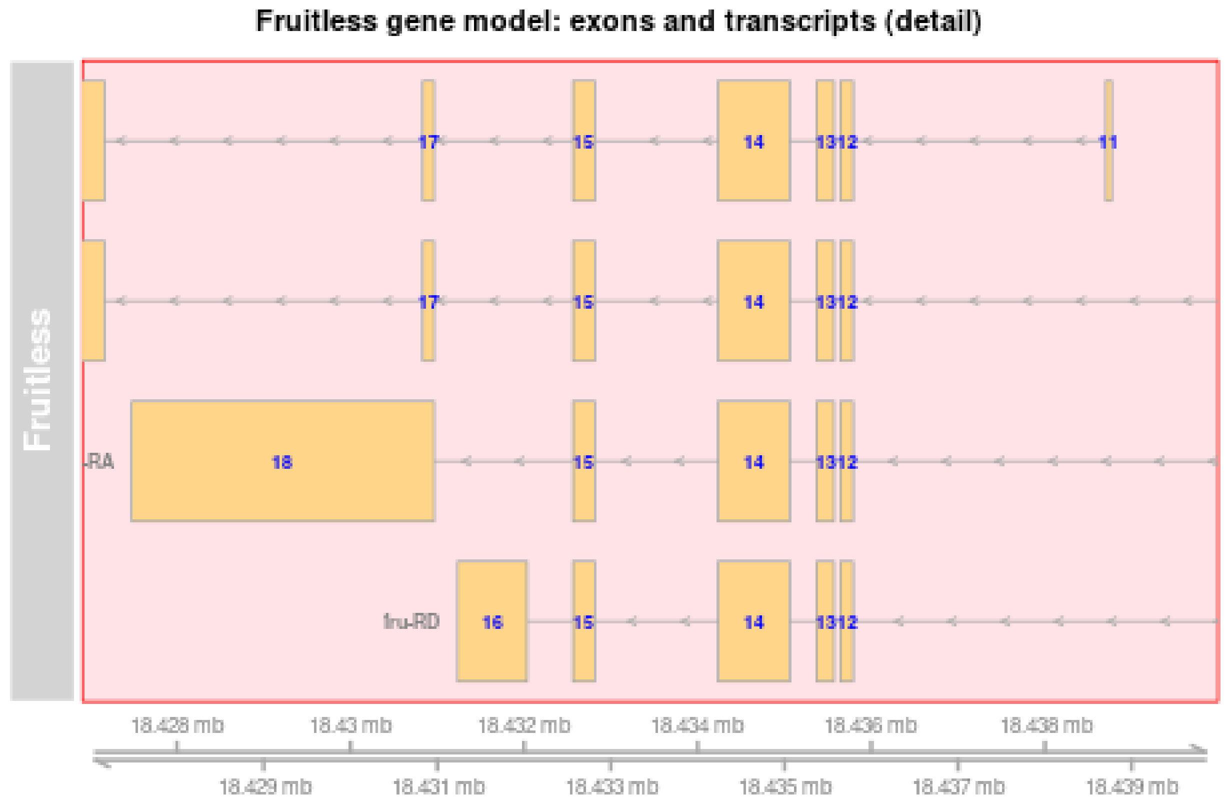
```
## png
##   2
```

Reexamining the underlying gene models, we can try to interpret these results:

The results from both strategies for handling ambiguous assignments indicated increases in the use of 3' exons 18, 20, and 21/22, in the FruLexa/Fru440 contrast, which would mean an increase in transcripts RA, RG/RO/RK/RH, and maybe RF/RB. Measured by junction, a decrease in the use of exon 2 (and maybe exon 3) in the FruLexa/Fru440 contrast was detected. On the one hand, this is hard to reconcile with the previous observation, since all but RA and RK include either exon 2 or 3, and the RK-specific exons 6 and 7 don't show any compensating increase in use. Also, exon 19 or 16, which are the 3' exons which would be used instead of 18/20/21/22, never show a compensating decrease.

The results from the intron-based analysis technically supports the decrease in the use of the 5' exons 1 and 2 but the effect size is bizarrely low. the 47b1 contrast results are more sensible, indicating an enhanced use of the most 5' exon 1. It appears to specifically differentiate the use of exons 2 and 3, specifically finding an increase in the intron between exons 2 and 8, ie, an increased use of transcript RN.

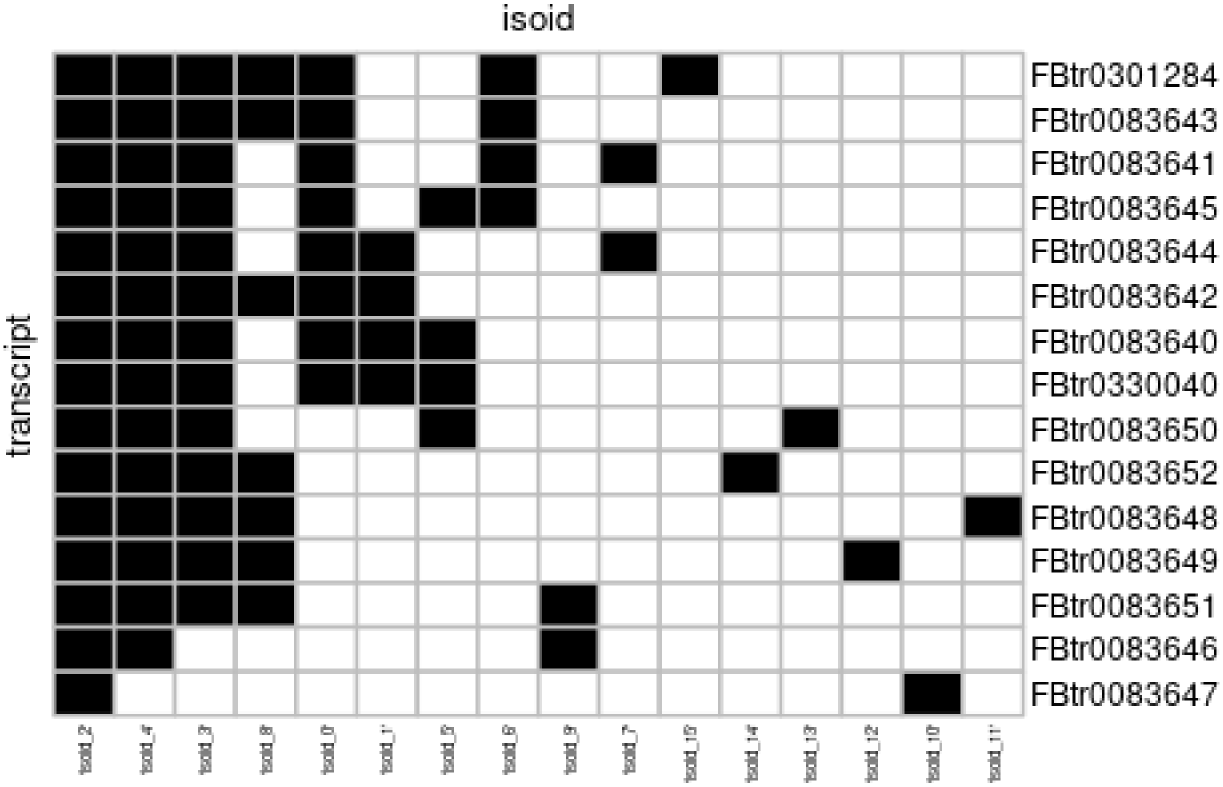

Fruitless gene model: exons and transcripts
Fruitless
fru-RF 22 3 1
fru-RB 21 2 1
fru-RG 20 2 1
fru-RO 20 3 1
fru-RK 20 7 6
fru-RH 20 3 1
fru-RM 19
fru-RC 19 3 1
fru-RE 19 2 1
fru-RN 19 8 2 1
fru-RL 19 10
fru-RJ 19 5
fru-RI 19 4
fru-RA 10
fru-RD 9
18.41 mb
18.43 mb
18.45 mb
18.47 mb
18.49 mb
18.51 mb
18.53 mb
18.42 mb
18.44 mb
18.46 mb
18.48 mb
18.5 mb
18.52 mb
18.54 mb



3.8.5 edgeHog

An experimental approach here is similar to the junction/intron assignment, except the splice sites have been grouped according to which subset of transcripts contain them. For example, the constitutive exons would be assigned to a group representing all transcripts, whereas transcript-specific exon junction would contribute to a group representing only that transcript. The gene model representing a subset of transcripts is called an "isoid".

Figure 105 . Transcript Subsets and their Isoid Representations (Fruitless)



```
## png
##   2
```

Am I handling 2-sidedness correctly? double check this

The "SplicedOnly" reads were counted against the isoids, with assignment to all annotations overlapped. These counts were used with DESeq2 and the hausWtVsMut contrast. Since ONLY the Fru counts are used, this normalizes any difference in overall expression of Fru between treatments

Table 118. Significant Changes in Fru Transcript Use
by Stouffer's Test on DESeq2 + Isoids

	housing		FruLexaFru440		67d		47b	
transcript	Z	p	Z	p	Z	p	Z	p
FBtr0083640	NA	NA	NA	NA	NA	NA	NA	NA
FBtr0083641	0.25	8.0×10^{-1}	1.50	1.3×10^{-1}	1.51	1.3×10^{-1}	0.42	6.7×10^{-1}
FBtr0083642	NA	NA	NA	NA	NA	NA	NA	NA
FBtr0083643	0.13	9.0×10^{-1}	1.37	1.7×10^{-1}	0.68	5.0×10^{-1}	0.57	5.7×10^{-1}
FBtr0083644	NA	NA	NA	NA	NA	NA	NA	NA
FBtr0083645	-0.14	8.9×10^{-1}	1.52	1.3×10^{-1}	0.51	6.1×10^{-1}	0.40	6.9×10^{-1}
FBtr0083646	0.27	7.9×10^{-1}	0.13	9.0×10^{-1}	0.17	8.7×10^{-1}	0.42	6.8×10^{-1}
FBtr0083647	-0.79	4.3×10^{-1}	-1.29	2.0×10^{-1}	-1.70	8.9×10^{-2}	-1.55	1.2×10^{-1}
FBtr0083648	NA	NA	NA	NA	NA	NA	NA	NA
FBtr0083649	NA	NA	NA	NA	NA	NA	NA	NA
FBtr0083650	NA	NA	NA	NA	NA	NA	NA	NA

FBtr0083651	0.30	7.7×10^{-1}	0.17	8.7×10^{-1}	0.23	8.2×10^{-1}	0.55	5.8×10^{-1}
FBtr0083652	NA	NA	NA	NA	NA	NA	NA	NA
FBtr0301284	NA	NA	NA	NA	NA	NA	NA	NA
FBtr0330040	NA	NA	NA	NA	NA	NA	NA	NA

The only vaguely significant change is in the FruLexaFru440 treatment, in which FBtr0083647 is depleted approximately 4 fold.

3.8.6 DEXSeq

All approaches to differential exon use detection were based around a standard featureCounts -> DESeq2 subpipeline, with modifications made to the input reads & annotations and/or downstream analysis. To compare these results to an established tool, DEXSeq (Anders, Reyes, and Huber 2012) was used. Another difference is that while the other methods have analyzed the Fruitless locus in isolation, this tool was run on the entire annotation and Fruitless results extracted later.

DEXSeq divides the exons in the annotation into non-overlapping intervals by exon start/end points:

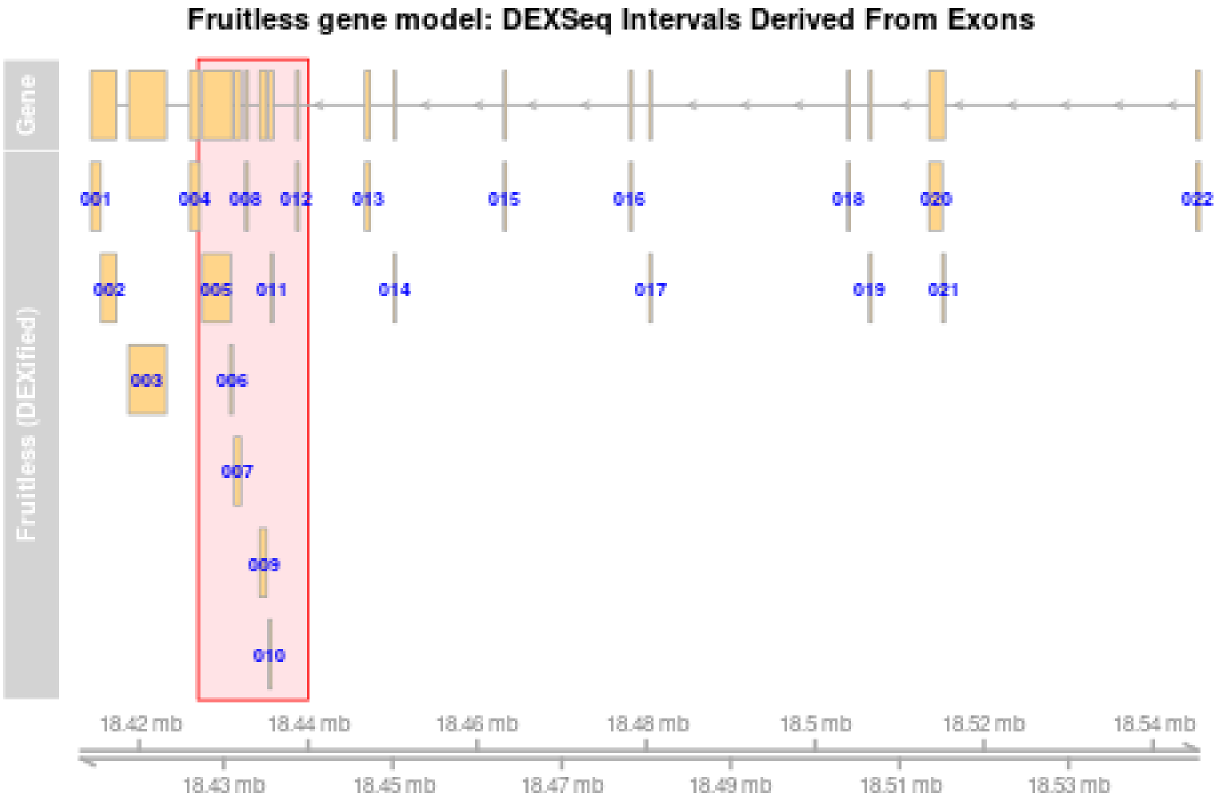


Figure 105 a. Fruitless gene model: DEXSeq Intervals Derived From Exons (detail)

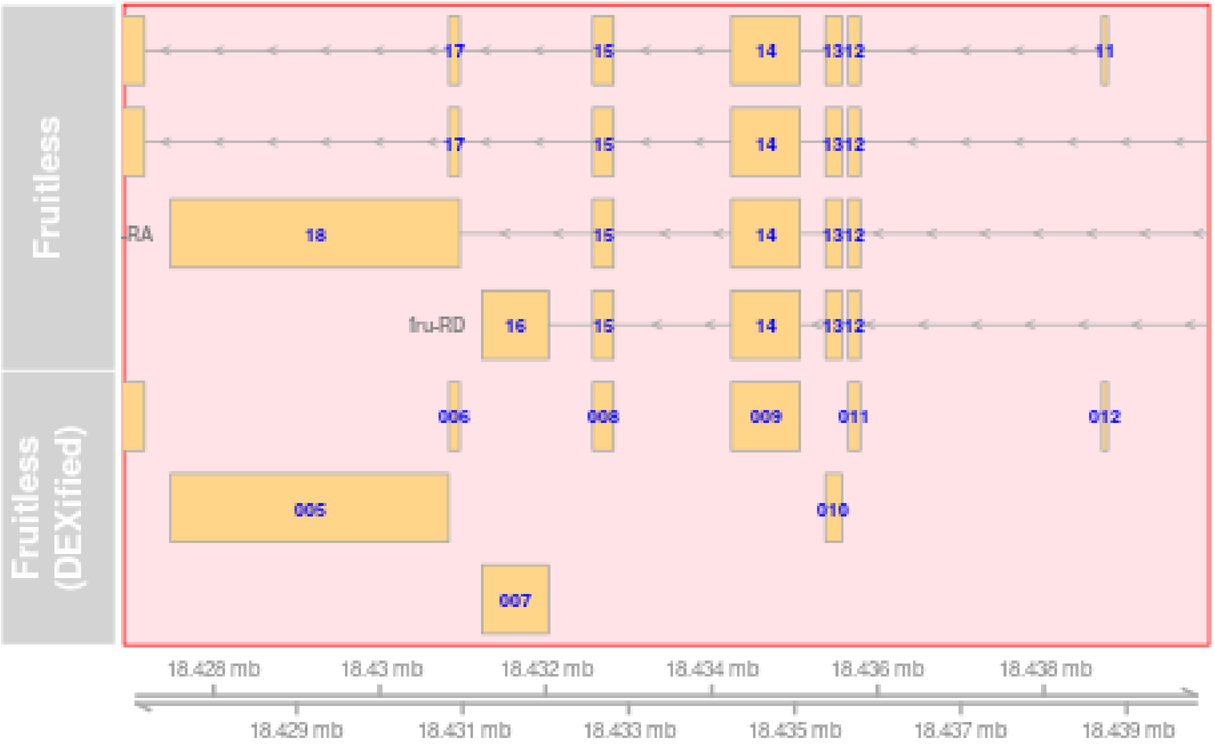


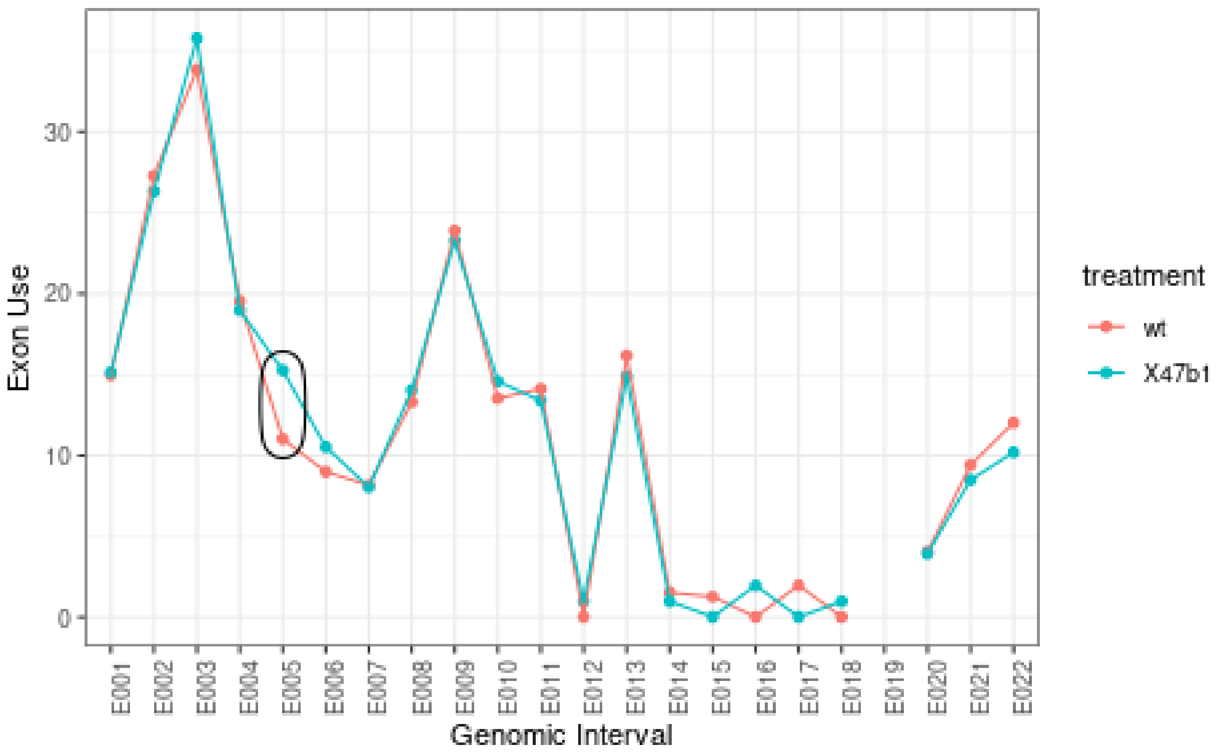
Table 119. DEXSeq Test for Differential Exon Use
Fruitless (FBgn0004652) Exons

internal name	genomic locus	log2FoldChange	padj
grpWtVs47b			
E005	chr3R:18427480-18430831	0.96	1.37×10^{-3}
grpWtVsFru			
E021	chr3R:18515052-18515343	−4.28	5.57×10^{-12}
grpWtVs67d			
E005	chr3R:18427480-18430831	1.25	2.47×10^{-8}

The interval E005 corresponds to the 3' end unique to exon_18, and a significant increase in its use is detected in the 47b and 67d treatments. E021 corresponds to exon_2/the shared 5' end of exon_3, and a significant decrease in its use is detected in the FruLexaFru440 treatment.

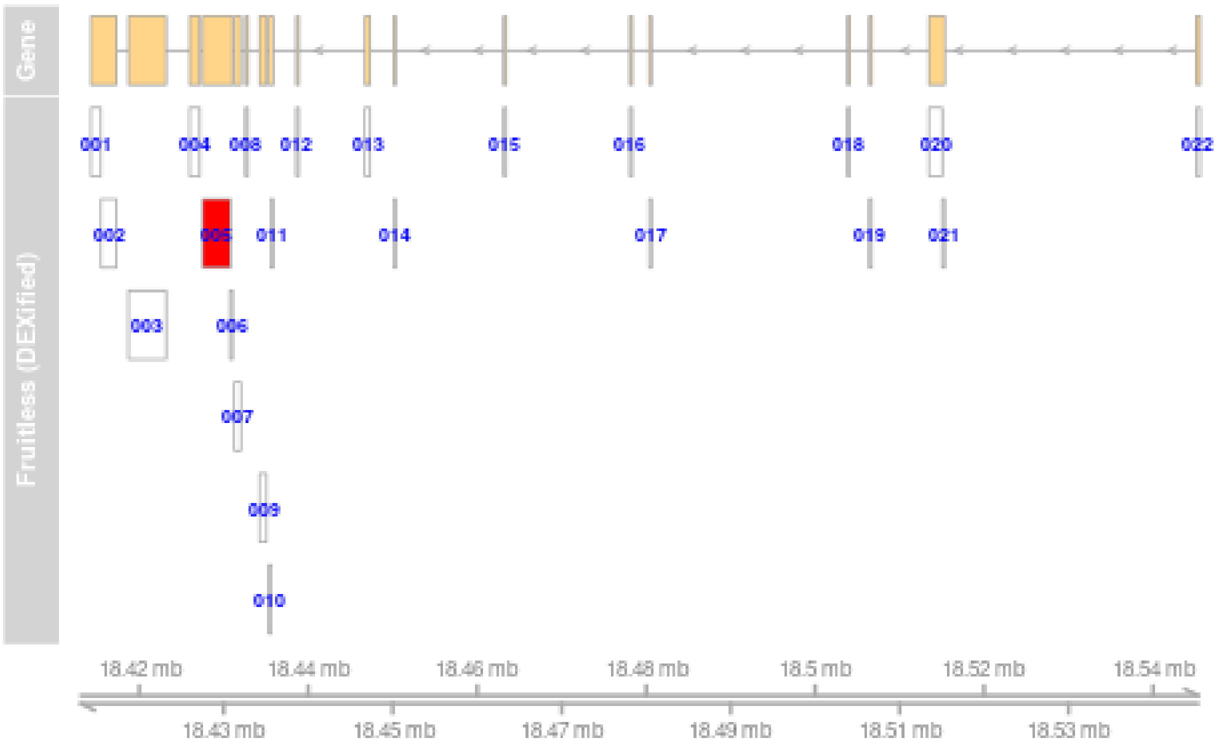
Figure 106. DEXSeq Estimate of Exon Use

47b1 contrast; significant (adjusted p<0.01) Differences Circled



```
## png
##   2
```

Fruitless gene model: DEXSeq Intervals Derived From Exons



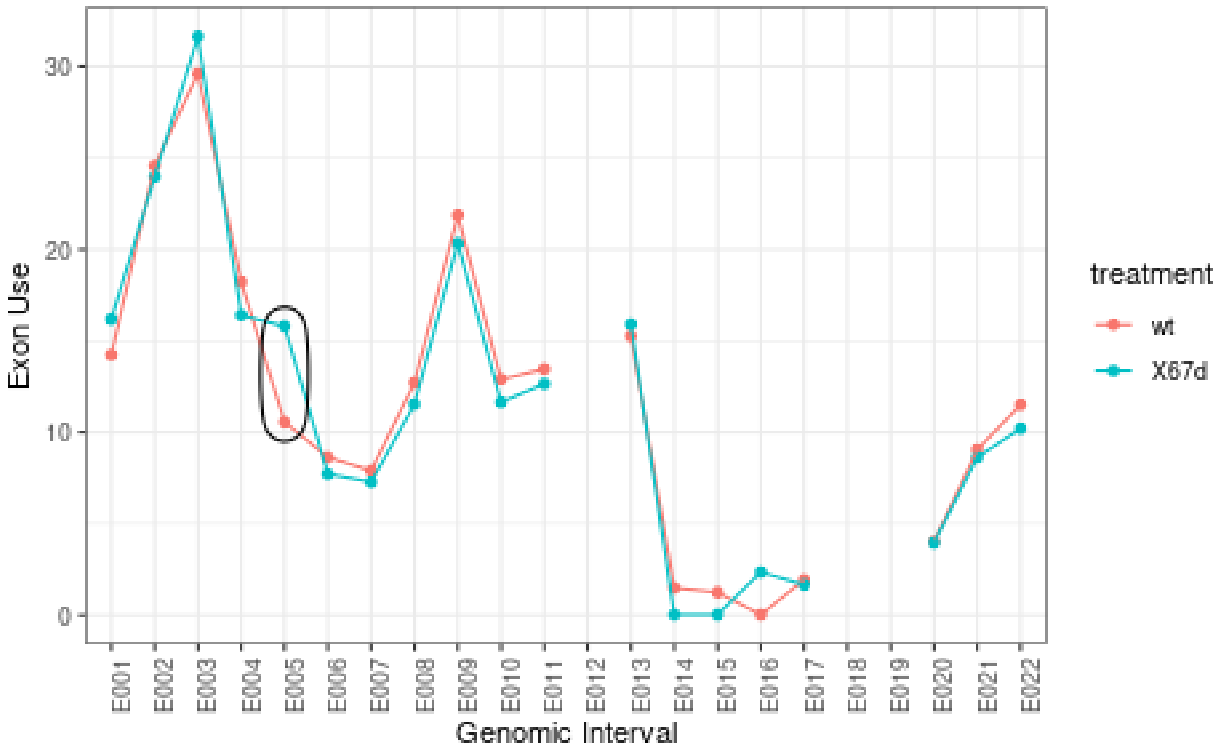


Figure 107. DEXSeq Estimate of Exon Use

67d contrast; significant (adjusted p<0.01) Differences Circled

```
## png
##   2
```

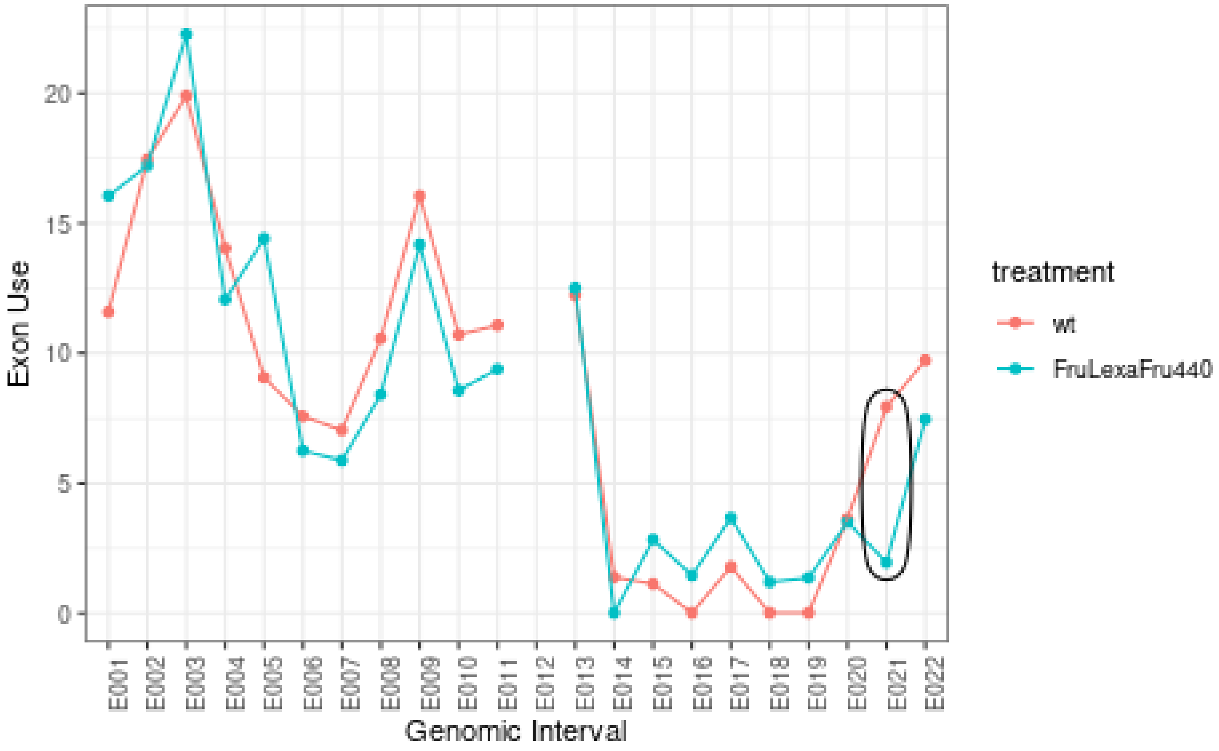


Figure 108. DEXSeq Estimate of Exon Use

FruLexaFru440 contrast; significant (adjusted p<0.01) Differences Circled

```
## png
##   2
```

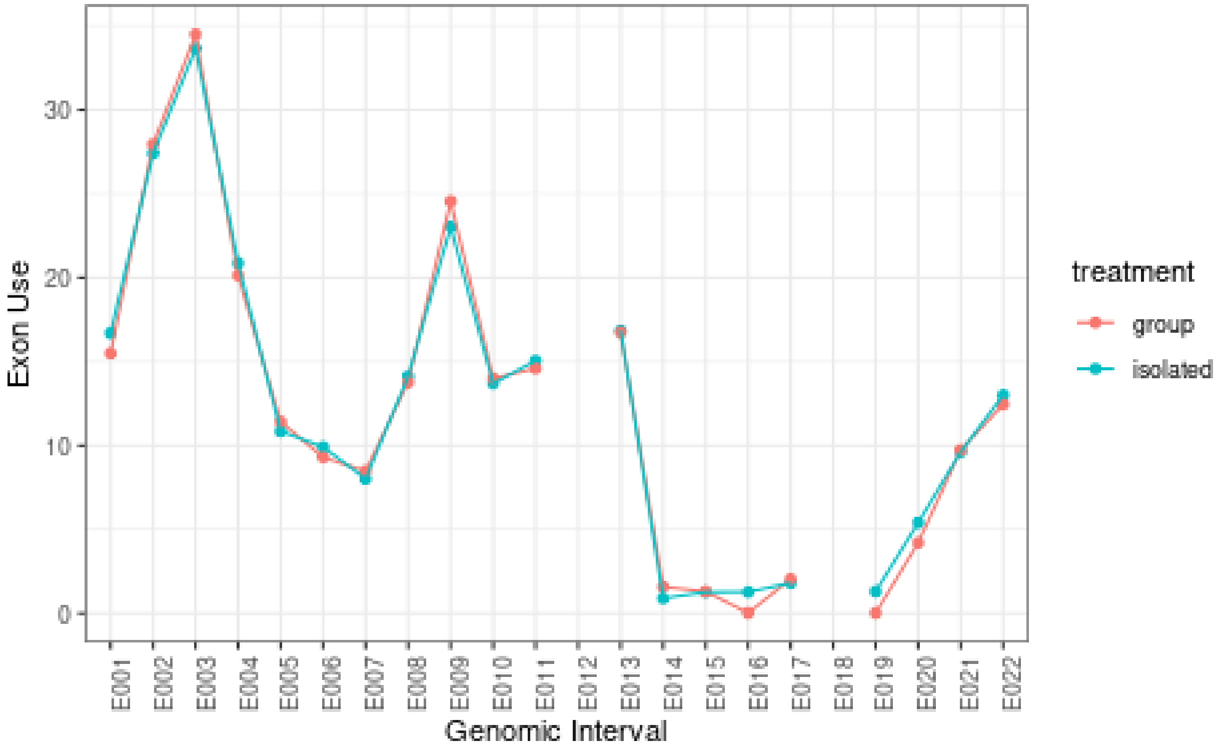


Figure 109. DEXSeq Estimate of Exon Use

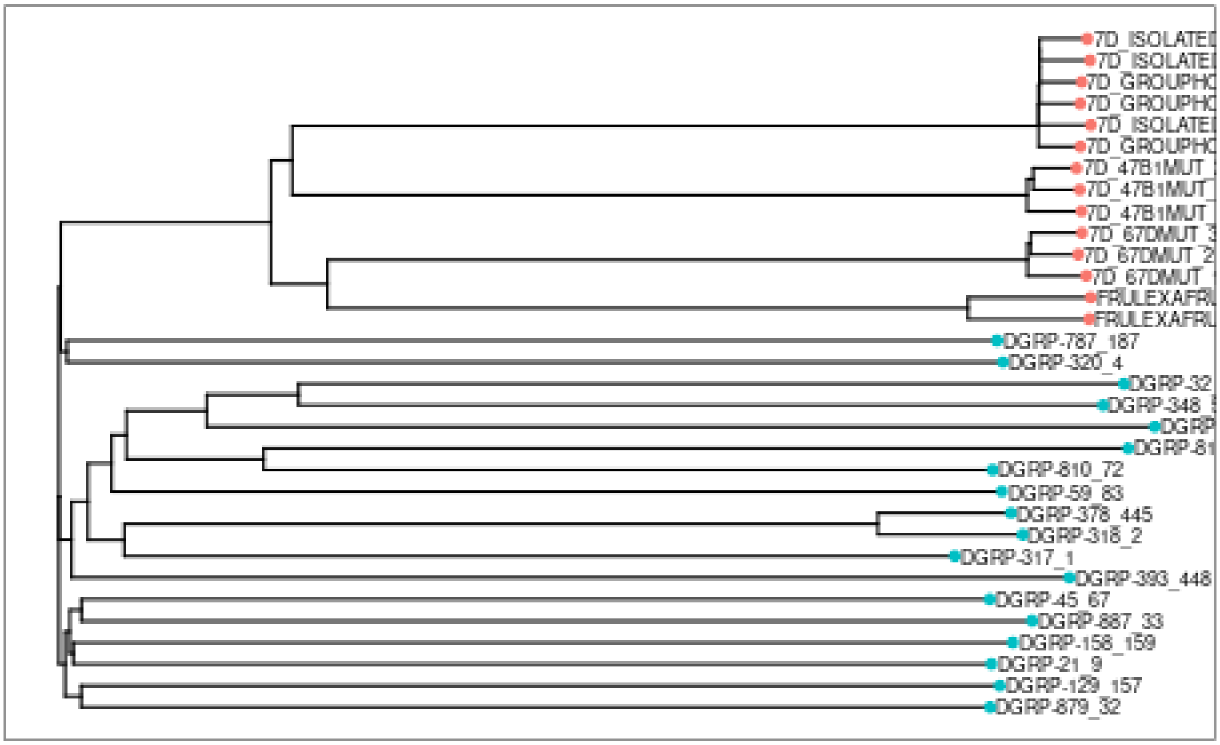
Housing contrast; significant (adjusted p<0.01) Differences Circled

```
## png
##   2
```

3.9 Genetic Distance: between vs within

Figure 109. Genetic Distance: this study vs DGRP

problematic Fru excluded; universally transcribed regions on autosomes; tree built with clustalw

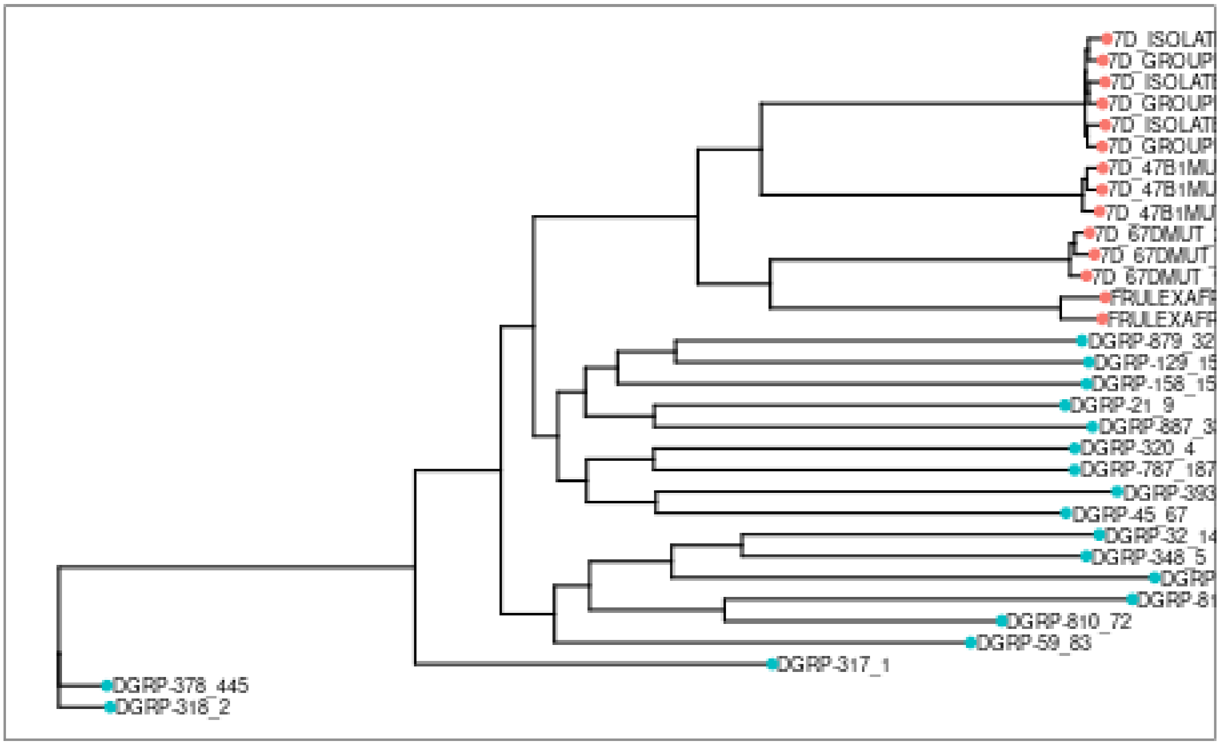


```
## png
##   2
```

```
## png
##   2
```

Figure 109. Genetic Distance: this study vs DGRP

problematic Fru excluded; universally transcribed regions on autosomes; tree built with phyML



```
## png
##   2
```

4 Bibliography

```
##
## To cite ggplot2 in publications, please use
##
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2016.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2016},
##     isbn = {978-3-319-24277-4},
##     url = {https://ggplot2.tidyverse.org},
##   }

##
##   Zhu, A., Ibrahim, J.G., Love, M.I. Heavy-tailed prior distributions
##   for sequence count data: removing the noise and preserving large
##   differences Bioinformatics (2018)
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Heavy-tailed prior distributions for sequence count data: removing the noise and preserv
##     author = {Anqi Zhu and Joseph G. Ibrahim and Michael I. Love},
##     year = {2018},
##     journal = {Bioinformatics},
##     doi = {10.1093/bioinformatics/bty895},
##   }
```

```
##
##   Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change
##   and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550
##   (2014)
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2},
##     author = {Michael I. Love and Wolfgang Huber and Simon Anders},
##     year = {2014},
##     journal = {Genome Biology},
##     doi = {10.1186/s13059-014-0550-8},
##     volume = {15},
##     issue = {12},
##     pages = {550},
##   }
```

```
##
## To cite the biomaRt package in publications use:
##
##   Mapping identifiers for the integration of genomic datasets with the
##   R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman,
##   Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).
##
##   BioMart and Bioconductor: a powerful link between biological
##   databases and microarray data analysis. Steffen Durinck, Yves Moreau,
##   Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang
##   Huber, Bioinformatics 21, 3439-3440 (2005).
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.
```

```
##
## To cite package 'topGO' in publications use:
##
##   Adrian Alexa and Jorg Rahnenfuhrer (2021). topGO: Enrichment Analysis
##   for Gene Ontology. R package version 2.46.0.
##
```

```
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {topGO: Enrichment Analysis for Gene Ontology},
##     author = {Adrian Alexa and Jorg Rahnenfuhrer},
##     year = {2021},
##     note = {R package version 2.46.0},
##   }
##
## ATTENTION: This citation information has been auto-generated from the
## package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

```
##
## The methods within the code package can be cited as:
##
##   Gu, Z. (2014) circlize implements and enhances circular visualization
##   in R. Bioinformatics.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {circlize implements and enhances circular visualization in R},
##     author = {Zuguang Gu and Lei Gu and Roland Eils and Matthias Schlesner and Benedikt Brors},
##     journal = {Bioinformatics},
##     volume = {30},
##     issue = {19},
##     pages = {2811-2812},
##     year = {2014},
##   }
##
## This free open-source software implements academic research by the
## authors and co-workers. If you use it, please support the project by
## citing the appropriate journal articles.
```

```
##
## The methods within the code package can be cited as:
##
##   Gu, Z. (2016) Complex heatmaps reveal patterns and correlations in
##   multidimensional genomic data. Bioinformatics.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Complex heatmaps reveal patterns and correlations in multidimensional genomic data},
##     author = {Zuguang Gu and Roland Eils and Matthias Schlesner},
##     journal = {Bioinformatics},
##     year = {2016},
##   }
##
## This free open-source software implements academic research by the
## authors and co-workers. If you use it, please support the project by
## citing the appropriate journal articles.
```

Anders, Simon, Alejandro Reyes, and Wolfgang Huber. 2012. "Detecting differential usage of exons from RNA-seq data." *Genome Research* 22 (10): 2008–17. https://doi.org/10.1101/gr.133744.111.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. "Fastp: An ultra-fast all-in-one FASTQ preprocessor." *Bioinformatics* 34 (17): i884–i890. https://doi.org/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. https://doi.org/10.1093/bioinformatics/btr330.

Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of Samtools and Bcftools." *GigaScience* 10 (2). Oxford University Press: 1–4. https://doi.org/10.1093/gigascience/giab008.

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

Huang, Wen, Andreas Massouras, Yutaka Inoue, Jason Peiffer, Miquel Ràmia, Aaron M. Tarone, Lavanya Turlapati, et al. 2014. "Natural Variation in Genome Architecture Among 205 Drosophila Melanogaster Genetic Reference Panel Lines." *Genome Research* 24 (7): 1193–1208. https://doi.org/10.1101/gr.171546.113.

Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, et al. 2007. "Clustal W and Clustal X Version 2.0." *Bioinformatics* 23 (21): 2947–8. https://doi.org/10.1093/bioinformatics/btm404.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features." *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology* 15 (12): 1–21. https://doi.org/10.1186/s13059-014-0550-8.

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

Shiao, Meng Shin, Jia Ming Chang, Wen Lang Fan, Mei Yeh Jade Lu, Cedric Notredame, Shu Fang, Rumi Kondo, and Wen Hsiung Li. 2015. "Expression divergence of chemosensory genes between Drosophila sechellia and its sibling species and its implications for host shift." *Genome Biology and Evolution* 7 (10): 2843–58. https://doi.org/10.1093/gbe/evv183.

Wang, Kai, Darshan Singh, Zheng Zeng, Stephen J Coleman, Yan Huang, Gleb L Savich, Xiaping He, et al. 2010. "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery." *Nucleic Acids Research* 38 (18): e178. https://doi.org/10.1093/nar/gkq622.

Wang, Li-Gen, Tommy Tsan-Yuk Lam, Shuangbin Xu, Zehan Dai, Lang Zhou, Tingze Feng, Pingfan Guo, et al. 2020. "Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data." *Molecular Biology and Evolution* 37 (2): 599–603. https://doi.org/10.1093/molbev/msz240.

Yu, Guangchuang. 2020. "Using Ggtree to Visualize Data on Tree?Like Structures." *Current Protocols in Bioinformatics* 69 (1). https://doi.org/10.1002/cpbi.96.