Progress on Female Mouse Liver RNAseq Data Analysis

2021/09/01

Workflow

- **Samples** → From Vanessa: raw read files for RNA extracted from female cohort mice livers, three treatment groups: BC16, BC12, CTRL (only analyzed BC16 vs. CTRL first so I can more easily learn the process with simpler experimental design), 4-5 biological replicates per group with 2 technical replicates (?) per biological replicate

- **Sequence alignment** → HISAT2

- **Format Conversion** → SAMtools

- **Count Genomic Features** → featureCounts

- **Differential Expression Analysis** → DESeq2

Workflow — *Sequence Alignment*

bash

```bash
# run HISAT2 to align seqence reads to reference genome
hisat2
    -q  # (query files are in .fq format)
    -p 16  # (use 16 threads for alignment)
    --pen-noncansplice 1000000  # (penalty for a non-canonical splice site)
    -x path/to/index  # (Index filename prefix (minus trailing .X.ht2))
    -1 input.fq  # (Files with #1 mates, paired with files in -2)
    -2 input.fq  # (Files with #2 mates, paired with files in -1)
    -S output.sam  # (File for SAM output)

# log the alignment results, importantly the alignment rate (should be >70%)

# after this point, .fq files can be gzipped again to save space
```

Kim, D., Paggi, J.M., Park, C. *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).

alignment rate	
BC16_fliv_N_1	86.59%
BC16_fliv_N_2	86.54%
BC16_fliv_O_1	90.82%
BC16_fliv_O_2	90.68%
BC16_fliv_P_1	93.44%
BC16_fliv_P_2	93.21%
BC16_fliv_Q_1	86.32%
BC16_fliv_Q_2	86.23%
BC16_fliv_R_1	84.52%
BC16_fliv_R_2	84.66%
CTRL_fliv_B_1	85.05%
CTRL_fliv_B_2	84.90%
CTRL_fliv_D_1	92.73%
CTRL_fliv_D_2	92.54%
CTRL_fliv_E_1	89.17%
CTRL_fliv_E_2	89.32%
CTRL_fliv_F_1	92.49%
CTRL_fliv_F_2	92.49%

Workflow — *Format Conversion*

bash

```bash
# convert .sam files to more compact binary .bam files
samtools view
    -b --threads 16  # (output in binary .bam format, use 16 threads for compression)
    input.sam > output.bam  # (input .sam file and output .bam file)

# after this point, .sam files can be deleted to save space

# sort the .bam files using samtools sort
samtools sort
    --threads 16  # (use 16 threads for sorting)
    input.bam -o output.sort.bam  # (input .bam and output sorted .bam)

# after this step we can delete the unsorted .bam files to save space
```

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H,
Twelve years of SAMtools and BCFtools, *GigaScience* (2021) 10(2) giab008

Workflow — *Count Reads*

bash

```bash
# count reads to genomic features
featureCounts
    -p  # (fragments or templates will be counted instead of reads)
    -t exon  # (specify feature type in GTF annotation)
    -a gencode.VM25.annotation.gtf  # (name of annotation file, GTF format)
    -g gene_name  # (specify attribute type in GTF annotation)
    -T 16  # (number of threads)
    -o output.txt  # (name of output file including read counts)
    f1.sort.bam f2.sort.bam ... fN.sort.bam  # (list all sorted .bam files to use)
```

- Total of 55,292 genes in output reads file
- ~25,000 genes with zero-sum rows removed → used for downstream analyses

Yang Liao, Gordon K. Smyth, Wei Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features
Bioinformatics, Volume 30, Issue 7, 1 April 2014, Pages 923–930, https://doi.org/10.1093/bioinformatics/btt656

Workflow — *Differential Expression Analysis*

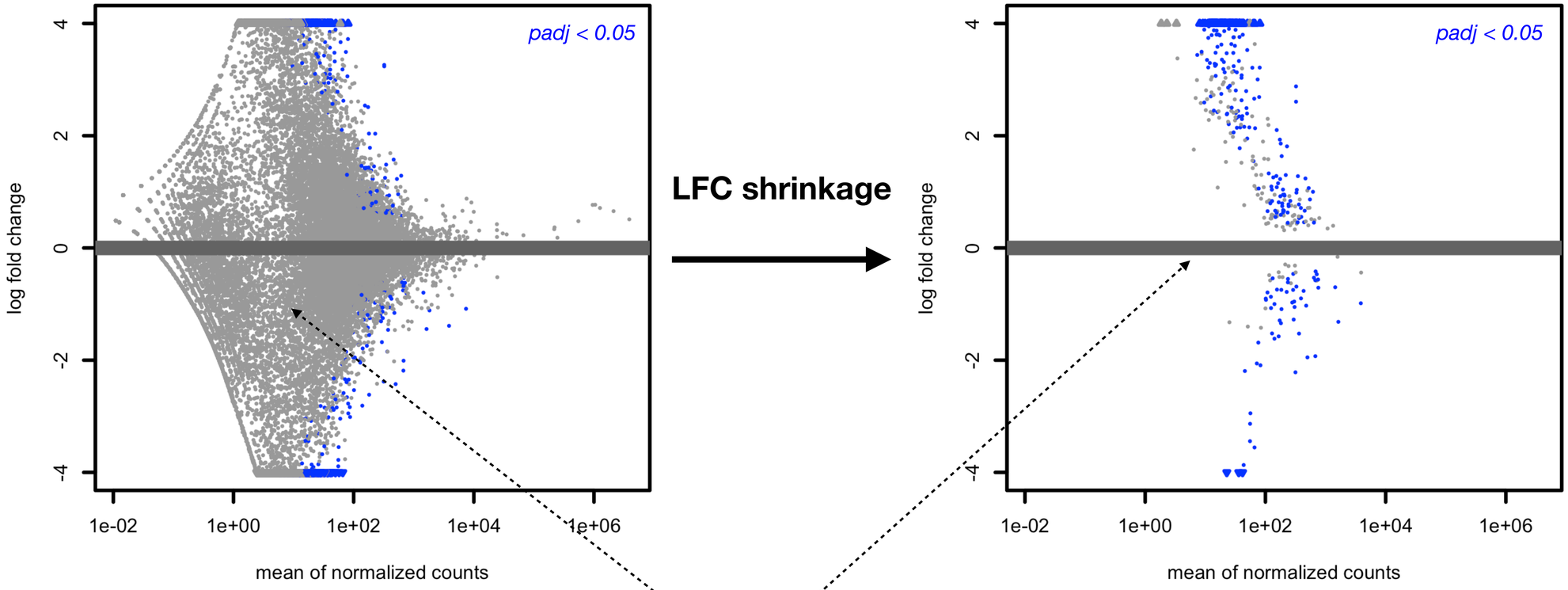
R

```r
# load raw cound data from csv and initialize DESeq data instance
deseq_data <- init_deseq_data(counts_df_from_csv())

# perform differential expression analysis
deseq_data <- DESeq(deseq_data)
print(head(deseq_data))

# gather results
deseq_res <- results(deseq_data)
summary(deseq_res)
print(mcols(deseq_res)$description)

# plot mean count vs. LFC (no shrinkage)
plotMA_(deseq_res, alpha = 0.05, ylim = c(-4, 4), png = "MA_plot_no_shrinkage.png")

# export results to .csv (no shrinkage)
write.csv(as.data.frame(deseq_res),
          file = "fliv_deseq2_results.csv")
write.csv(subset(as.data.frame(deseq_res), padj < 0.05),
          file = "fliv_deseq2_results_padj<0.05.csv")

# apply LFC shrinkage (Zhu, Ibrahim, and Love 2018)
deseq_res_norm <- lfcShrink(deseq_data, coef = 2, type = "apeglm")
summary(deseq_res_norm)
print(mcols(deseq_res_norm)$description)
```

- computes statistical information about differentially expressed genes between the experimental treatments (BC16 and CTRL)
- performs variance-stabilizing transformations on raw count data, corrects Log2FCs
- computes and corrects p-values for group comparison of gene expression
- ~5000 genes with corrected p-value (BC16 vs. CTRL) < 0.05

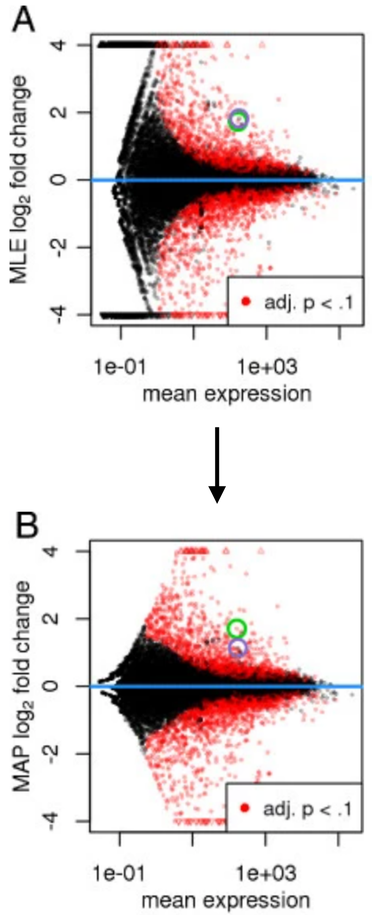
Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 (2014). https://doi.org/10.1186/s13059-014-0550-8

MA plots show LFC shrinkage reduces excess variance in LFCs from low-abundance genes



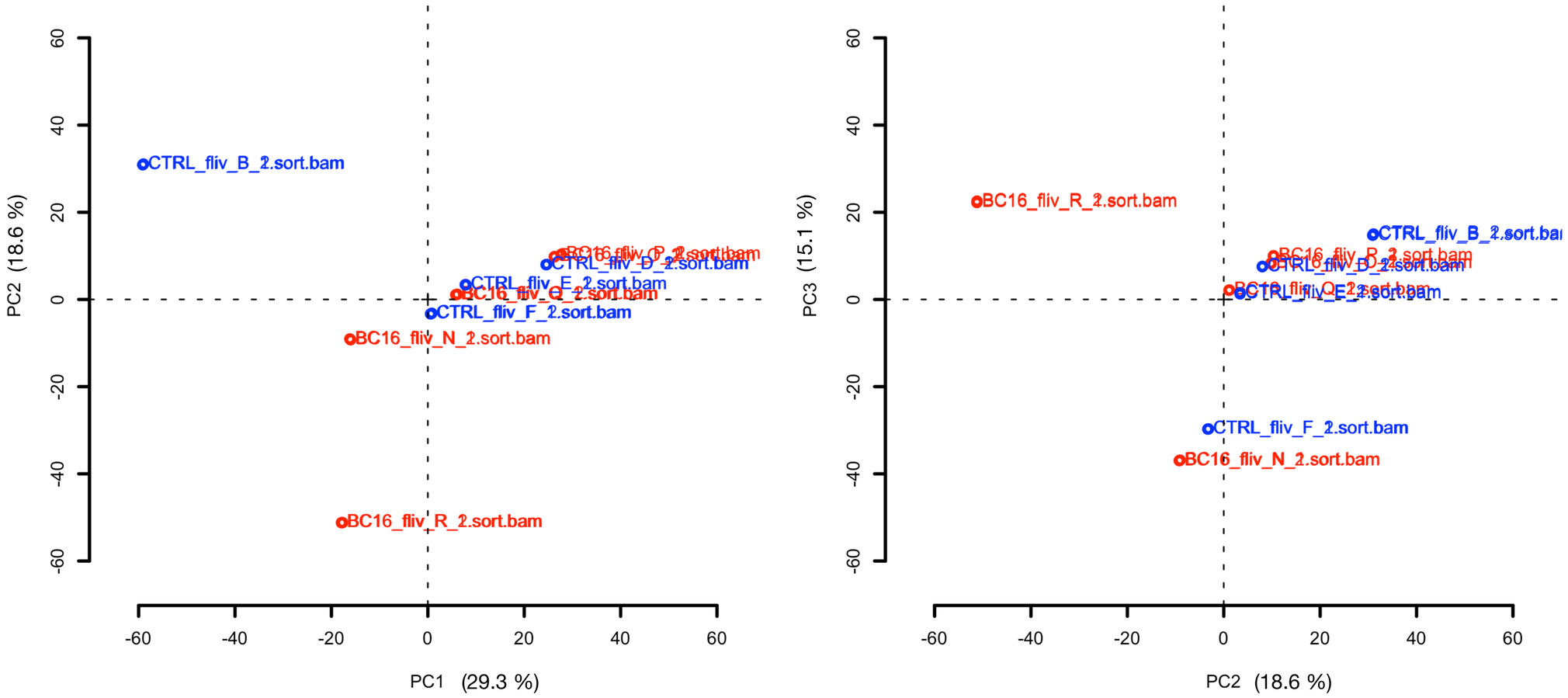
LFC shrinkage

low abundance, high-variance LFCs with non-significant p-values go away? get shrunken to 0?

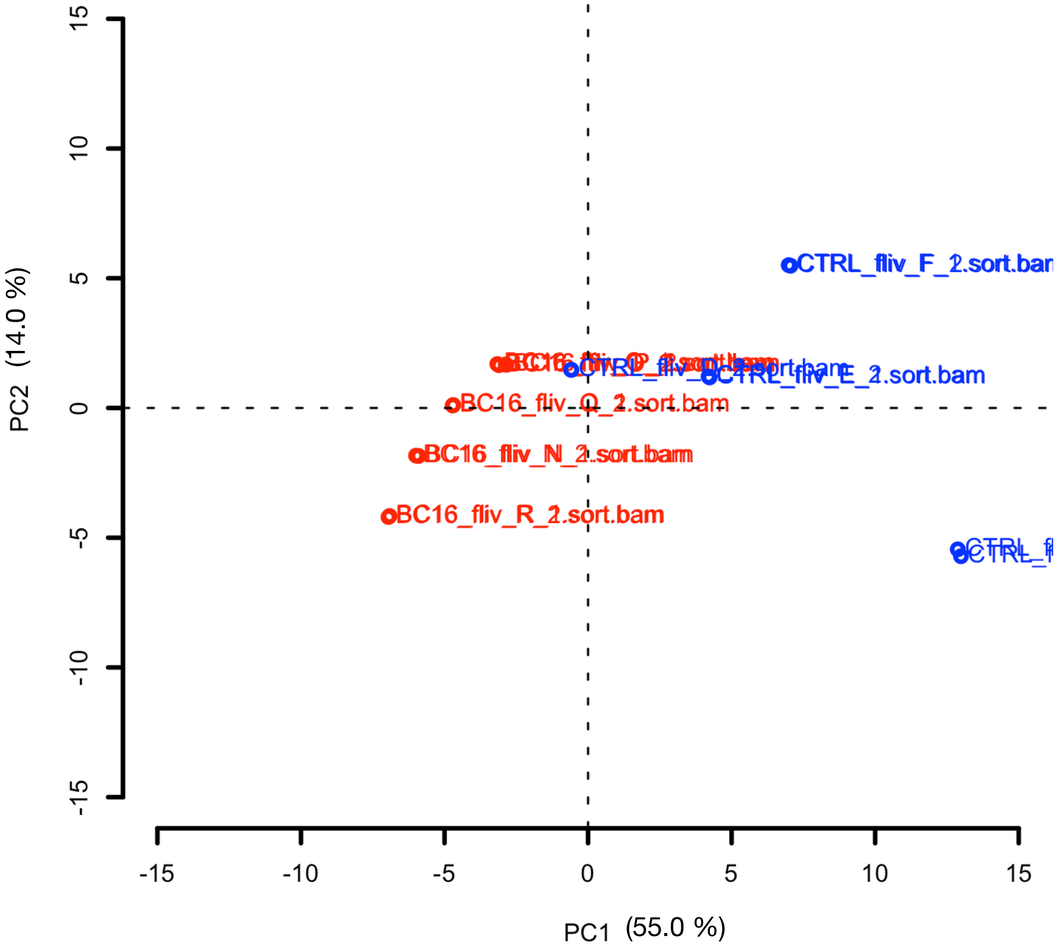


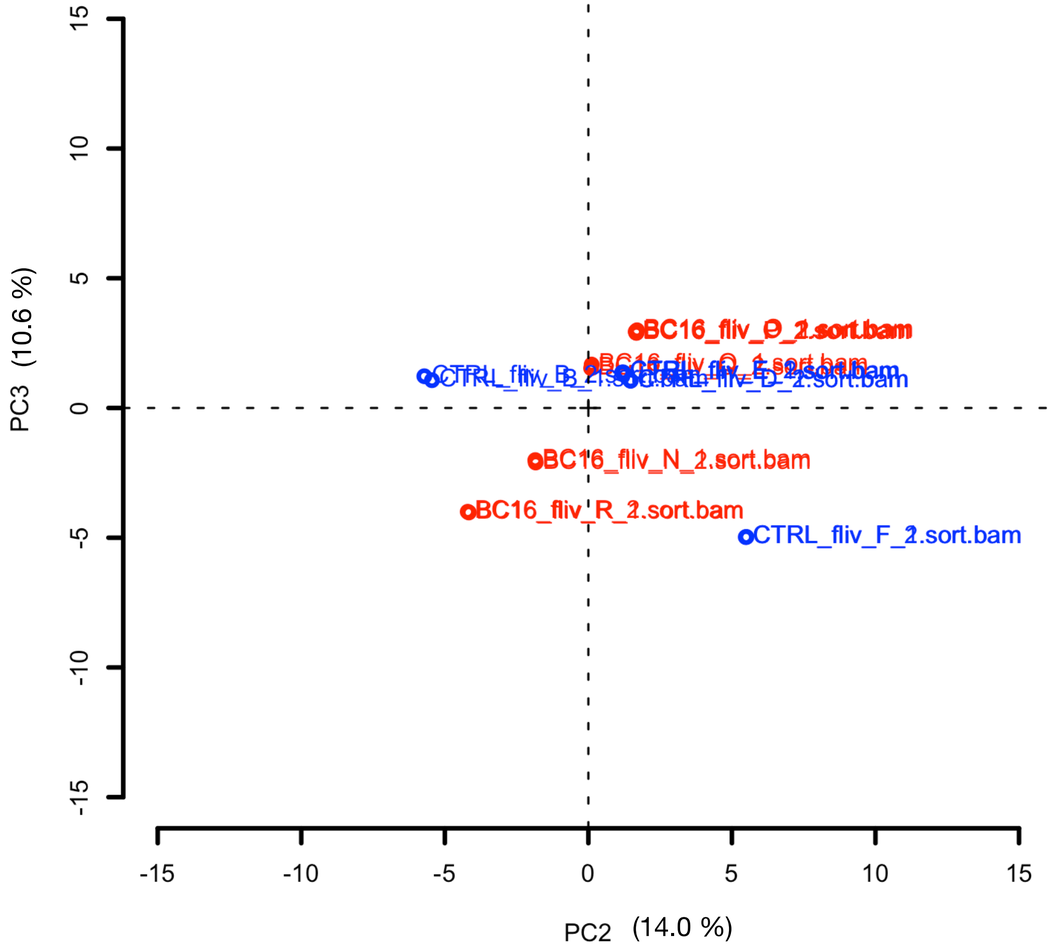
LFC shrinkage example from DESeq2 paper

PCA on all features shows no distinct difference between BC16 and CTRL samples



PCA on all genes with adj. p-value < 0.05 shows group separation between BC16 and CTRL samples





conditions (BC16 and CTRL) separate along PC1, as expected

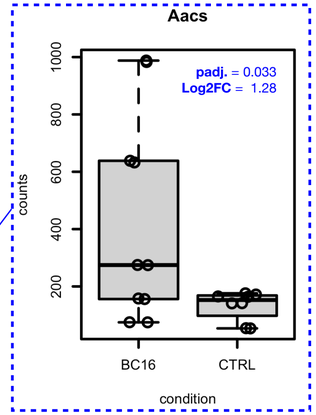
Some previously identified genes (neonatal mouse brain) also had significantly altered expression in this data

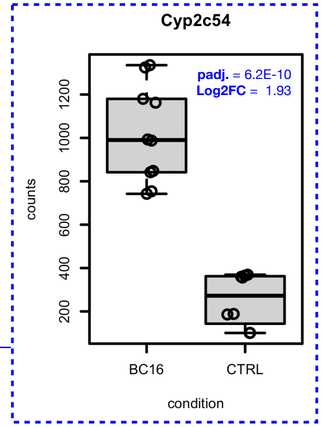
Table 2 (Herron, *et al. Tox. Sci.* 2019)

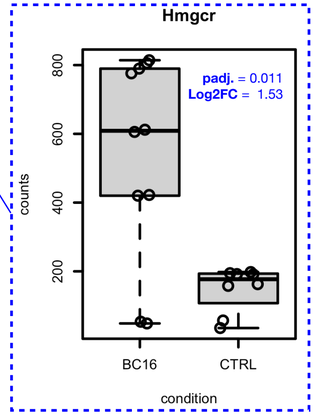
Gene ID	Description	BAC C16	
		Log2(FC)	Adjusted *p* Value
AACS	Acetoacetyl-CoA synthetase	0.35	3.84E–02
ACSL4	Acyl-CoA synthetase long-chain family member 4		
ALB	Albumin	–3.69	9.98E–03
CYP51	Cytochrome P450, family 51		
DLK1	Delta like non–canonical Notch ligand 1		
ELOVL6	ELOVL family member 6, elongation of long-chain fatty acids		
HMGCR	3-hydroxy-3-methylglutaryl-CoenzymeA reductase		
HMGCS2	3-hydroxy-3-methylglutaryl-CoenzymeA synthase 2	0.61	9.98E–03
IDI1	Isopentenyl-diphosphate delta isomerase		
INSIG1	Insulin induced gene 1		
LDLR	Low-density lipoprotein receptor	0.73	9.98E–03
MSMO1	Methylsterol monooxygenase 1	0.48	9.98E–03
MT2	Metallothionein 2		
MTHFD2	Methylenetetrahydrofolate cyclohydrolase	–0.53	9.98E–03
PCSK9	Proprotein convertase subtilisin/kexin type 9		
SCD	Stearoyl-Coenzyme A desaturase		
SQLE	Squalene epoxidase	0.45	9.98E–03

$n = 4$ biological replicates per condition; adjusted $p < .05$.

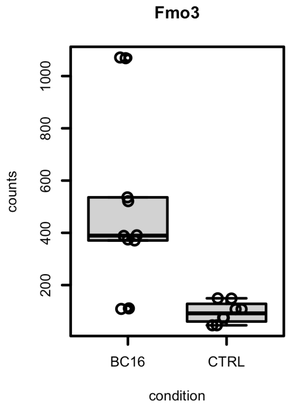
significantly up-regulated in BAC C12 in previously reported data (omitted)







another notable individual gene



FMO3 is the major FMO isoform in the liver and is known to metabolize trimethylamine … is it possible for trimethylamine or other similar amines to be metabolites of BACs?

Future Directions

- Hierarchical clustering analysis on significant DEGs from DESeq results (clustvis, and/or other standalone package), PCA loadings from significant DEG data can give some similar insight

- Ingenuity pathway analysis to look for significantly altered pathways

- work up BAC-C12 data, repeat analysis steps for BC12 vs. CTRL groups, compare/contrast DEGs for treatments with BC16 and BC12

- develop some scripts for automating portions of the data prep/analysis, most of the processing pipeline is pretty straightforward and some validation between steps can be automated as well (e.g. checking alignment rate)