### Methods

RNA-seq data was processed using the fastp toolkit1 to trim low-quality bases and sequencing adapters from the 3’ end of reads, then mapped to GRCh38 (downloaded from Ensembl, version 106)2 using the STAR RNA-seq alignment tool3,and reads aligning to a single genomic location were summarized across genes. For genes having an overlap of at least 10 reads, gene counts were normalized and differential expression was carried out using the DESeq24 Bioconductor5 package implemented for the R programming environment6. Consistent with the recommendation of the DESeq authors, independent filtering7 was utilized prior to calculating adjusted p-values8 and moderated log2 fold-changes were derived using the ashr package9. Gene set enrichment analysis10 was performed to identify gene ontology terms and pathways associated with altered gene expression for each of the comparisons performed.

### Materials

*DifferentialExpression.xlsx*

Differential expression results for each of the comparisons.

*PCA.pdf*

A plot of the the first two components from a principal components analysis.

*Heatmap-<comparison>.pdf*

These images show a heatmap of the top 100 differentially expressed genes for each comparison (by p-value). Gene expression has been z-score normalized and the samples and genes are clustered by correlation distance with complete linkage.

*GSEA-<gene-set>.xlsx*

For each comparison, a table of gene-sets for either the C2 (called *GeneOntologies*) or C4 (called *Pathways*) collection of the MSigDB project.

More information about the MSigDB collections can be found here:

https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp

More information about specific gene-sets can be found here:

https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp

*GSEA-<gene-set>-<comparison>.pdf*

This figure shows the normalized enrichment score for the 10 most significant (by p-value) gene-sets with a positive or negative normalized expression score (NES).

### References

1. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i890.
2. *Kersey PJ, Staines DM, Lawson D, et al. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res. 2012;40(Database issue):D91-D97.*
3. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
4. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
5. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015;12(2):115-121.
6. www.r-project.org
7. Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods*. 2016;13(7):577-580.
8. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57(1):289-300.
9. Stephens M. False discovery rates: a new deal. *Biostat*. Published online October 17, 2016:kxw041.
10. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102(43):15545-15550.