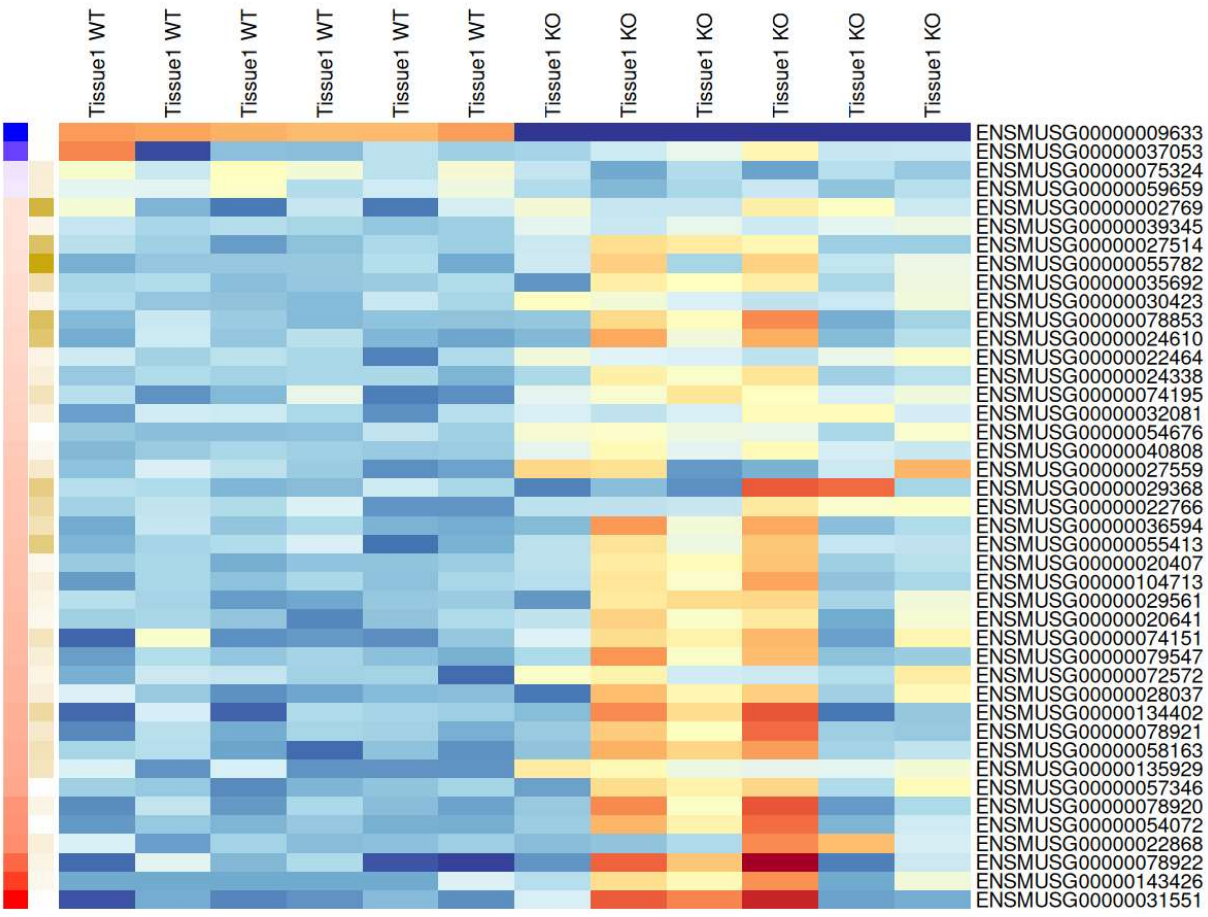
RNA processing and differential expression

The FASTQ data was aligned to the genome and processed using the nf-core framework.

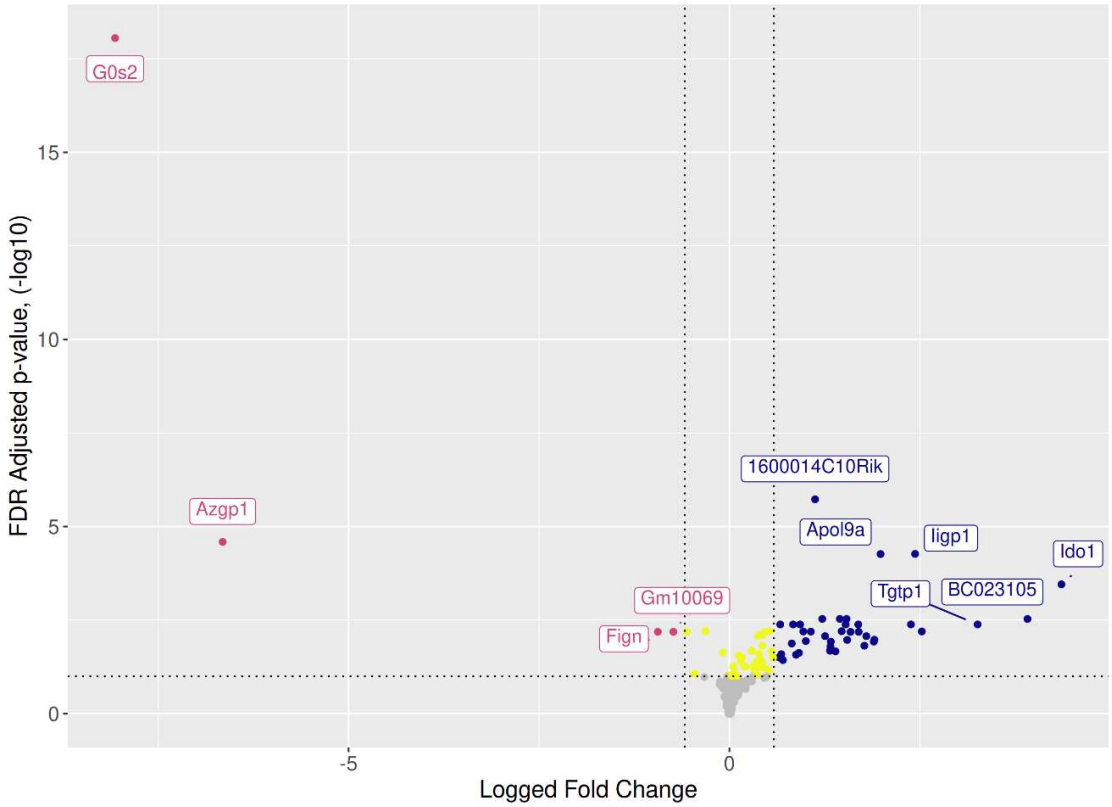
The data comprised two tissue types - Tissue1 and Tissue2, with WT and Gene1 KO for each tissue (4 conditions in total). The samples underwent QC and after removing failing samples, we compared Gene1 KO vs WT for each tissue.

This resulted in the following numbers of genes differentially expressed (comparing each tissue to the WT sample of the same tissue)

contrast	Down	Up	Total
KO Tissue1	7	83	90
KO Tissue2	194	100	294



This heatmap shows some of the differentially expressed genes in Tissue1. The blue/red column shows the magnitude of changes in expression (down and up, respectively). The light brown column shows p-value (lighter is more significant).



This volcano plot shows the differentially expressed genes in Tissue1.

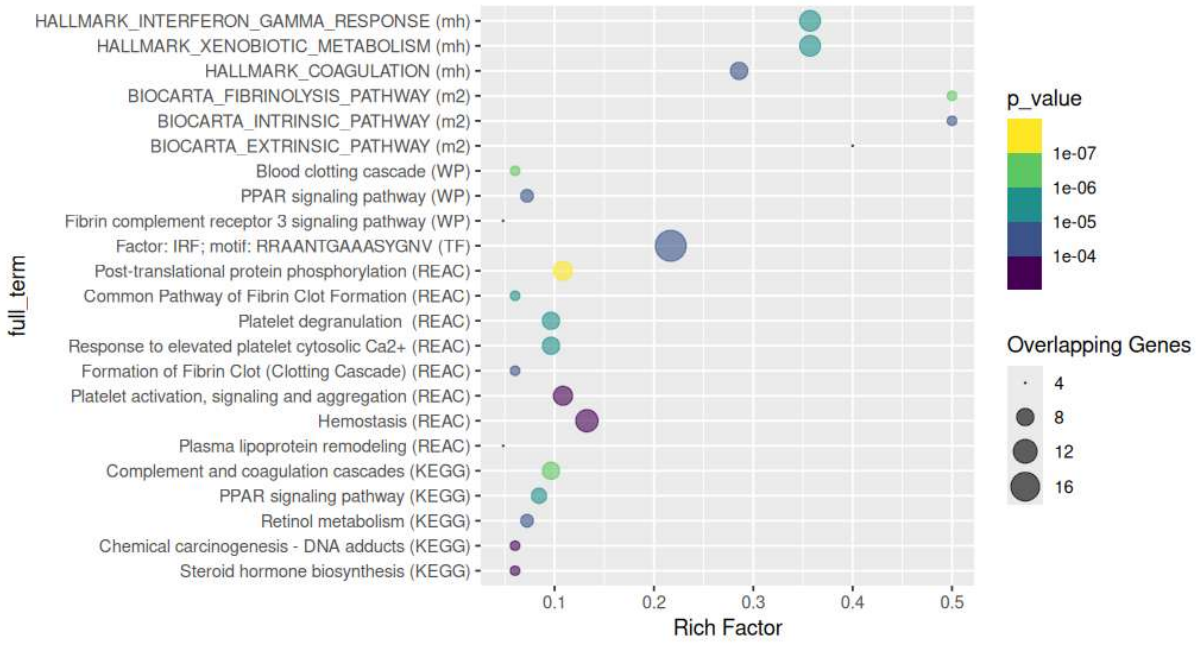
Blue and pink represent upregulated and downregulated genes with a significant p-value and fold change above or below a threshold, respectively. Yellow genes have statistically significant expression change, with a low fold change.

An interactive HTML version of this figure, as well as the textual table is available in full report.

TISSUE2 deliberately not shown to save space.

Functional Enrichment

Pathway analysis was performed separately on the up- and down-regulated differentially expressed genes using the gprofiler R package. A figure showing a subset of the enriched gene sets is shown.



The gene set name appears on the left, where each row is a gene set. The Rich Factor is the ratio of the number of differentially expressed genes in this pathway term to the total number of all genes annotated. The number of annotated genes is represented by circle size, while the colour represents the adjusted p-value.

Biological relevance: It looks like after KO, we have upregulated of blood clotting and steroid/fat metabolism.

No particular functional category of genes was detected as being enriched in the downregulated genes in Tissue1.

Full tables and figures are available in full report.

Methods

Creation of sequencing libraries

Messenger RNA was purified from total RNA using poly-T oligo-attached magnetic beads. After fragmentation, the first strand cDNA was synthesized using random hexamer primers, followed by the second strand cDNA synthesis using either dTTP for non strand specific library or dUTP for strand specific library.

Data processing and alignment

The sequencing data files were processed using the nf-core/rnaseq pipeline (doi: 10.5281/zenodo.1400710). In brief, this pipeline takes a samplesheet and FASTQ files as input performs QC using FastQC, trimming and (pseudo-)alignment, and removes possible contamination (genome reads, ribosomal reads, etc). It then produces a gene expression matrix using STAR aligner and salmon and an extensive QC report. A JSON file detailing all parameters is provided in the Supplementary Information.

The quantified gene expression was processed using the nf-core/diffabundance pipeline (doi: 10.5281/zenodo.7568000), comparing the gene expression of KO Tissue2 against WT Tissue2, and KO Tissue1 against WT Tissue1 using DESeq2. Contrasts file provided in Supplementary Information.

When using the nf-core/diffabuncance pipeline, the following parameters were different from the default values:

	Current	Default
filtering_min_proportion	0.25	
deseq2_min_replicates_for_replace	5	7
exploratory_n_features	1000	500
gprofiler2_evcodes	TRUE	FALSE
gsea_run	FALSE	TRUE
differential_max_pval	0.05	1
differential_min_fold_change	1.5	2
gprofiler2_min_diff	3	1
differential_max_qval	0.1	0.05
filtering_min_abundance	5	10

All parameters not listed here were kept as default values for this pipeline on the Seqara platform. A JSON file detailing all parameters is provided in Supplementary Information.

Functional enrichment was performed using goprofiler2, running it manually after the initial results from the nf-core/diffabundance pipeline running gprofiler2.

References

- Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316-319. doi: 10.1038/nbt.3820
- Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology, 38(3), 276-278. doi: 10.1038/s41587-020-0439-x
- Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J., & Bioconda Team. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods, 15(7), 475-476. doi: 10.1038/s41592-018-0046-7
- da Veiga Leprevost, F., Grüning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J., Bai, M., Jimenez, R. C., Sachsenberg, T., Pfeuffer, J., Vera Alvarez, R., Griss, J., Nesvizhskii, A. I., & Perez-Riverol, Y. (2017). BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics (Oxford, England), 33(16), 2580-2582. doi: 10.1093/bioinformatics/btx192
- Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12):550. PubMed PMID: 25516281; PubMed Central PMCID: PMC4302049.
- Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H (2020). “gprofiler2- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler.” F1000Research, 9 (ELIXIR)(709). R package version 0.2.2.
- *Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23;1(6):417-425*.
- BioCarta: http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways (defunct), see the original publication: https://doi.org/10.1089/152791601750294344. Note also special terms for these gene sets in the MSigDB license.