Anirudh Rao
BE21B004

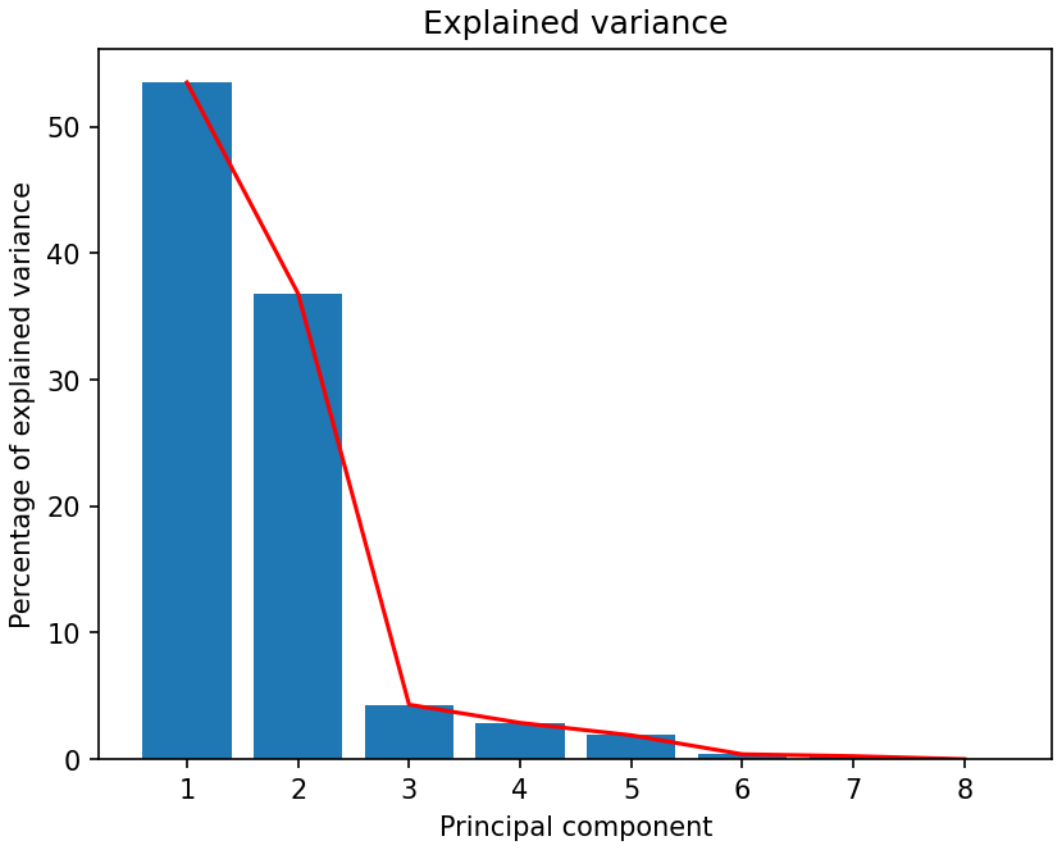
BT4110 – Computational Biology Lab

Transcriptome Data Analysis Assignment

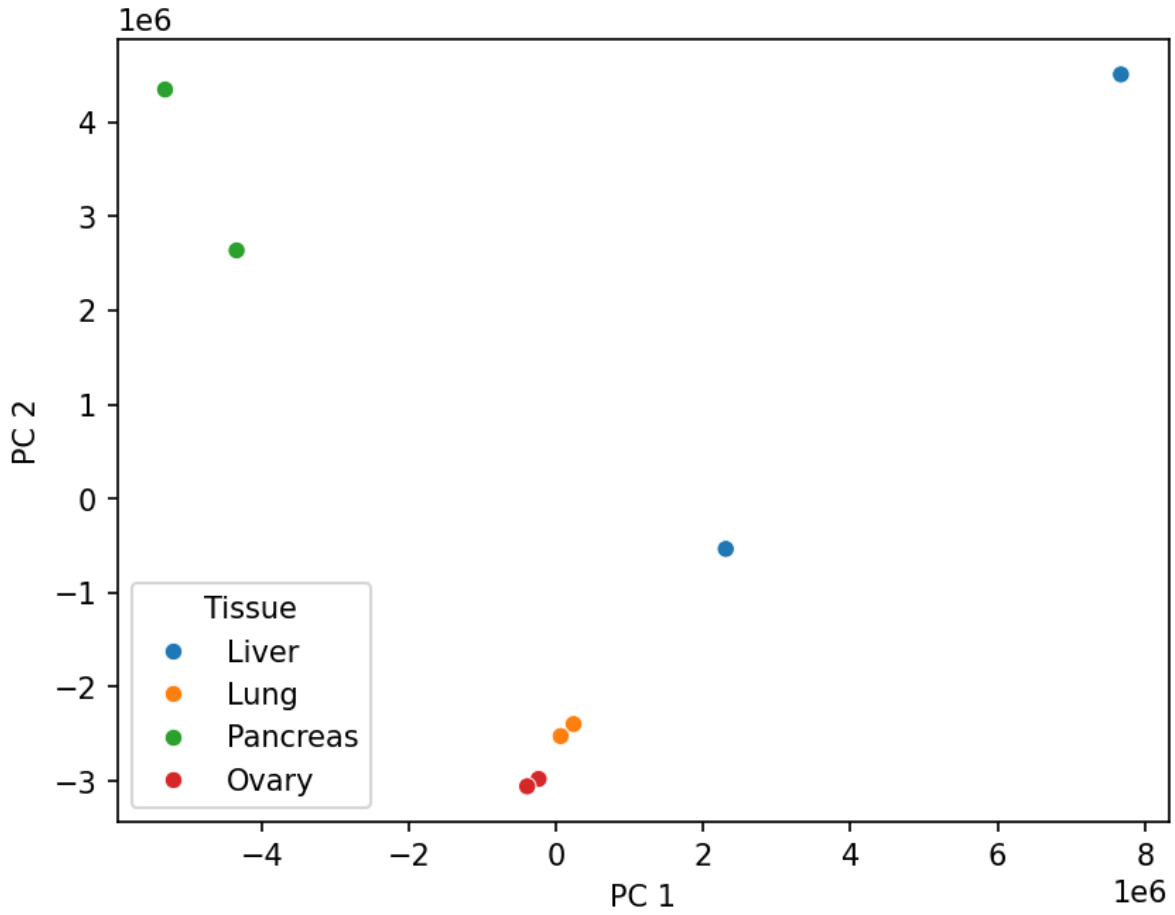
1

After scaling the gene expression data, PCA was performed using sklearn. Based on this, 8 principal components were obtained. The percentage of variance explained by each of the components is shown below:

Principal component	Explained variance (%)
PC 1	5.35E+01
PC 2	3.68E+01
PC 3	4.31E+00
PC 4	2.87E+00
PC 5	1.89E+00
PC 6	3.86E-01
PC 7	2.38E-01
PC 8	4.43E-28



When the first two PCs are plotted against each other, clustering of tissue types can be observed.



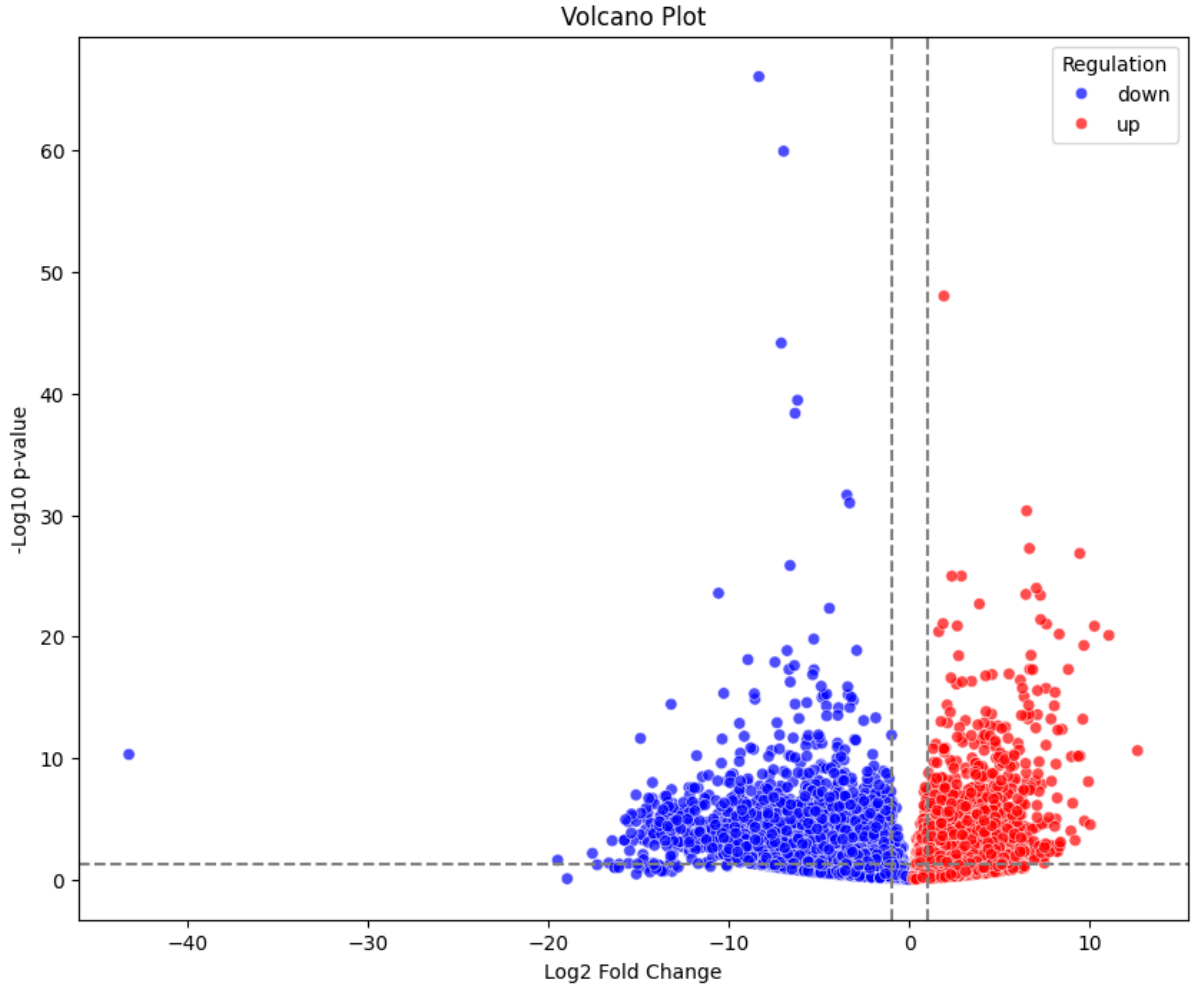
We can see that the ovary samples and lung samples are quite close to each other. The pancreas samples are also relatively close to each other. However, the liver samples are quite distant from each other. Overall, the PCA effectively captures the tissue types from the gene expression data.

2

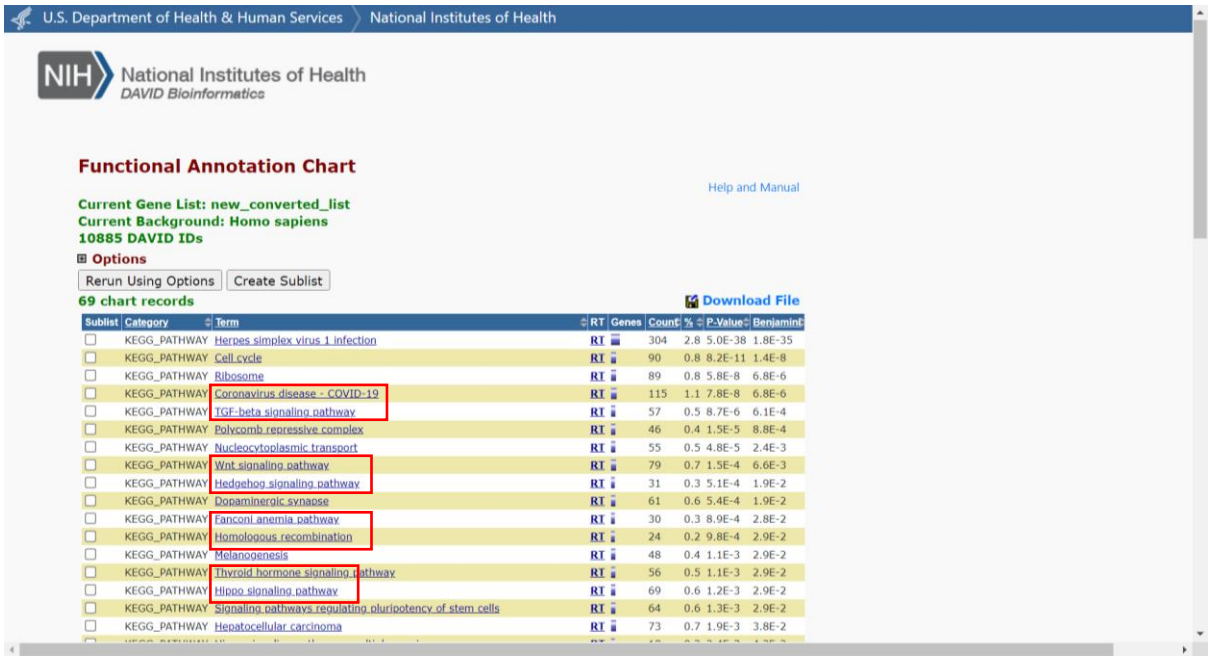
PyDESeq2 was used to identify upregulated and downregulated genes in the ovary samples as compared to those from other tissues. This was done by suitably modifying the given metadata file.

Tissue	group
X5EGGE	Other
X5EQMM	Other
X5GICA	Other
X5N9D6	Other
X5P9GB	Other
X5N9ES	Other
X5P9GS	Ovary
X5N9E1	Ovary

Based on the output of DESeq2, 17355 genes are upregulated and 37237 genes are downregulated in the ovary samples as compared to the other tissues. This can be visualised using a volcano plot.

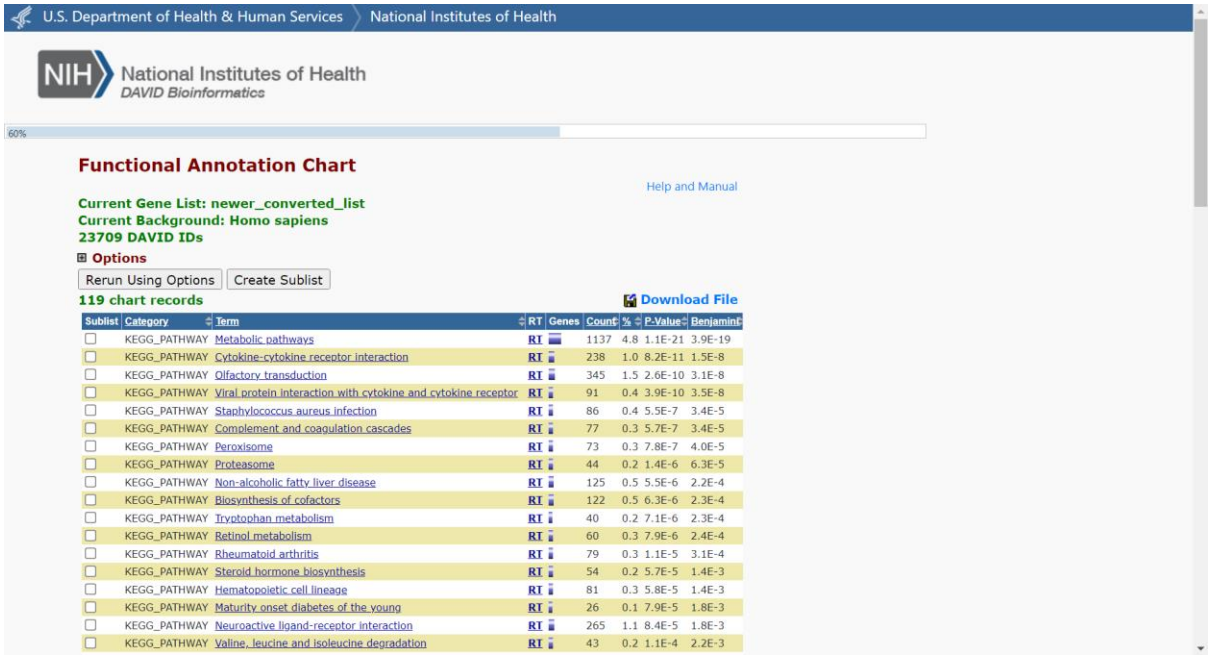


The list of upregulated genes was fed into DAVID for functional enrichment analysis. After mapping the genes to their names in DAVID DB, their pathway functional annotation was performed using the KEGG Pathway database. Of the 17355 upregulated genes, 10885 were available in the KEGG database.



- The herpes simplex virus has been shown to infect the ovary.
- The ovary is the site of gametogenesis via meiosis, which employs the cell cycle pathway.
- Meiosis also requires homologous recombination, which is shown to be enriched.
- The TGF-β signalling pathway, the Wnt pathway, and the hedgehog pathway, have all been shown to play a role in ovarian follicle development.
- The Polycomb repressive complex has been shown to play an important role in the epigenetic regulation of ovary function.
- The proteins associated with the Fanconi anaemia pathway are also known to be involved in follicle development in the ovaries.

The list of downregulated genes was then fed into DAVID. After mapping the genes to their names in DAVID DB, their pathway functional annotation was performed using the KEGG Pathway database. Of the 37237 upregulated genes, 23709 were available in the KEGG database.



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	KEGG_PATHWAY	Metabolic pathways	RT		1137	4.8	1.1E-21	3.9E-19
☐	KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT		238	1.0	8.2E-11	1.5E-8
☐	KEGG_PATHWAY	Olfactory transduction	RT		345	1.5	2.6E-10	3.1E-8
☐	KEGG_PATHWAY	Viral protein interaction with cytokine and cytokine receptor	RT		91	0.4	3.9E-10	3.5E-8
☐	KEGG_PATHWAY	Staphylococcus aureus infection	RT		86	0.4	5.5E-7	3.4E-5
☐	KEGG_PATHWAY	Complement and coagulation cascades	RT		77	0.3	5.7E-7	3.4E-5
☐	KEGG_PATHWAY	Peroxisome	RT		73	0.3	7.8E-7	4.0E-5
☐	KEGG_PATHWAY	Proteasome	RT		44	0.2	1.4E-6	6.3E-5
☐	KEGG_PATHWAY	Non-alcoholic fatty liver disease	RT		125	0.5	5.5E-6	2.2E-4
☐	KEGG_PATHWAY	Biosynthesis of cofactors	RT		122	0.5	6.3E-6	2.3E-4
☐	KEGG_PATHWAY	Tryptophan metabolism	RT		40	0.2	7.1E-6	2.3E-4
☐	KEGG_PATHWAY	Retinol metabolism	RT		60	0.3	7.9E-6	2.4E-4
☐	KEGG_PATHWAY	Rheumatoid arthritis	RT		79	0.3	1.1E-5	3.1E-4
☐	KEGG_PATHWAY	Steroid hormone biosynthesis	RT		54	0.2	5.7E-5	1.4E-3
☐	KEGG_PATHWAY	Hematopoietic cell lineage	RT		81	0.3	5.8E-5	1.4E-3
☐	KEGG_PATHWAY	Maturity onset diabetes of the young	RT		26	0.1	7.9E-5	1.8E-3
☐	KEGG_PATHWAY	Neuroactive ligand-receptor interaction	RT		265	1.1	8.4E-5	1.8E-3
☐	KEGG_PATHWAY	Valine, leucine and isoleucine degradation	RT		43	0.2	1.1E-4	2.2E-3

None of the pathways identified seem to be directly and specifically associated with the ovaries. Thus, DESeq2 has accurately identified biologically relevant upregulated and downregulated genes in the ovary compared to other tissue types.