LabBook 25_03_16

Claire Green

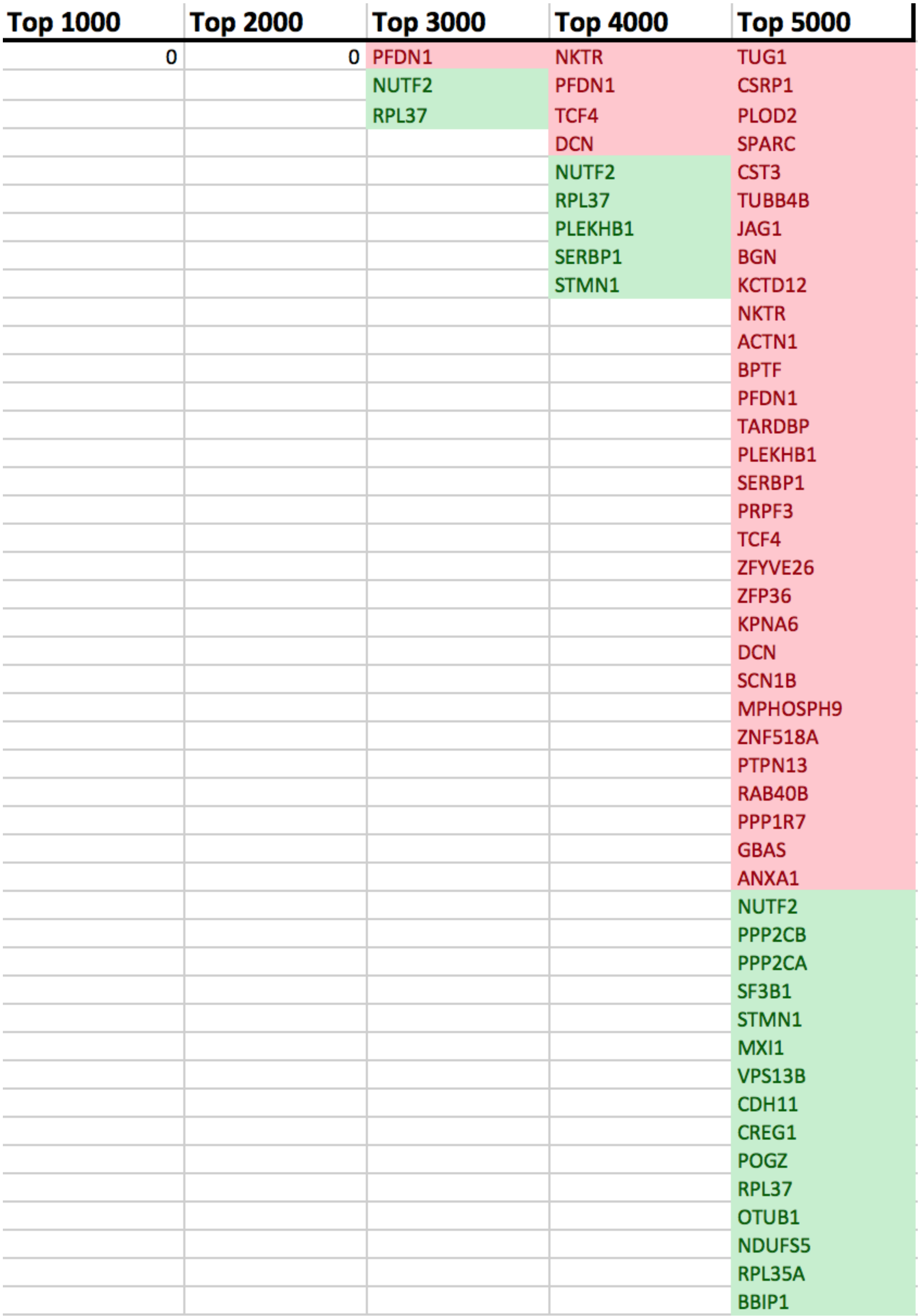
Monday

I completed the intersect analysis using the ensembl IDs and gained some more genes:

Top 1000	Top 2000	Top 3000		Top 4000		Top 5000	
0	0	ENSG00000102898	NUTF2	ENSG00000114857	NKTR	ENSG00000011465	DCN
		ENSG00000145592	RPL37	ENSG00000102898	NUTF2	ENSG00000021300	PLEKHB1
				ENSG00000145592	RPL37	ENSG00000025800	KPNA6
				ENSG00000021300	PLEKHB1	ENSG00000051825	MPHOSPH9
				ENSG00000142864	SERBP1	ENSG00000072110	ACTN1
				ENSG00000117632	STMN1	ENSG00000072121	ZFYVE26
				ENSG00000196628	TCF4	ENSG00000101384	JAG1
				ENSG00000011465	DCN	ENSG00000101439	CST3
						ENSG00000102898	NUTF2
						ENSG00000104695	PPP2CB
						ENSG00000105711	SCN1B
						ENSG00000113140	SPARC
						ENSG00000113575	PPP2CA
						ENSG00000114857	NKTR
						ENSG00000115524	SF3B1
						ENSG00000115685	PPP1R7
						ENSG00000117360	PRPF3
						ENSG00000117632	STMN1
						ENSG00000119950	MXI1
						ENSG00000120948	TARDBP

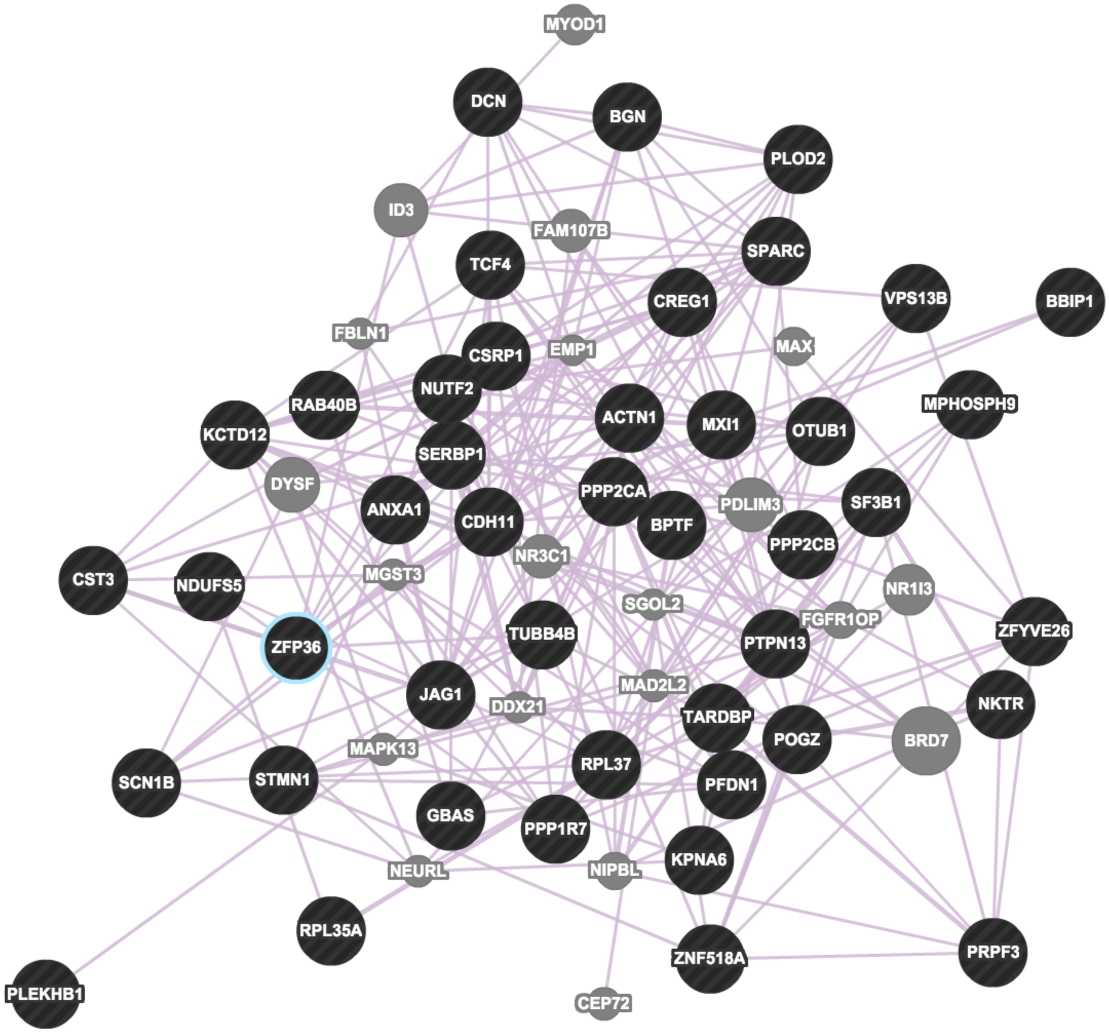
Top 1000	Top 2000	Top 3000	Top 4000	Top 5000	
				ENSG00000128016	ZFP36
				ENSG00000132549	VPS13B
				ENSG00000135046	ANXA1
				ENSG00000140937	CDH11
				ENSG00000141542	RAB40B
				ENSG00000142864	SERBP1
				ENSG00000143162	CREG1
				ENSG00000143442	POGZ
				ENSG00000145592	RPL37
				ENSG00000146729	GBAS
				ENSG00000152952	PLOD2
				ENSG00000159176	CSRP1
				ENSG00000163629	PTPN13
				ENSG00000167770	OTUB1
				ENSG00000168653	NDUFS5
				ENSG00000171634	BPTF
				ENSG00000177853	ZNF518A
				ENSG00000178695	KCTD12
				ENSG00000182492	BGN
				ENSG00000182899	RPL35A
				ENSG00000188229	TUBB4B
				ENSG00000196628	TCF4
				ENSG00000214413	BBIP1
				ENSG00000253352	TUG1

When combined with the HGNC generated genes, I now have an intersect list of 45 genes at 5000 threshold (red is HGNC, green is new ensemblID). Using random permutation testing, the number of genes selected at 3000, 4000 and 5000 top genes is significantly more than would be expected by chance.



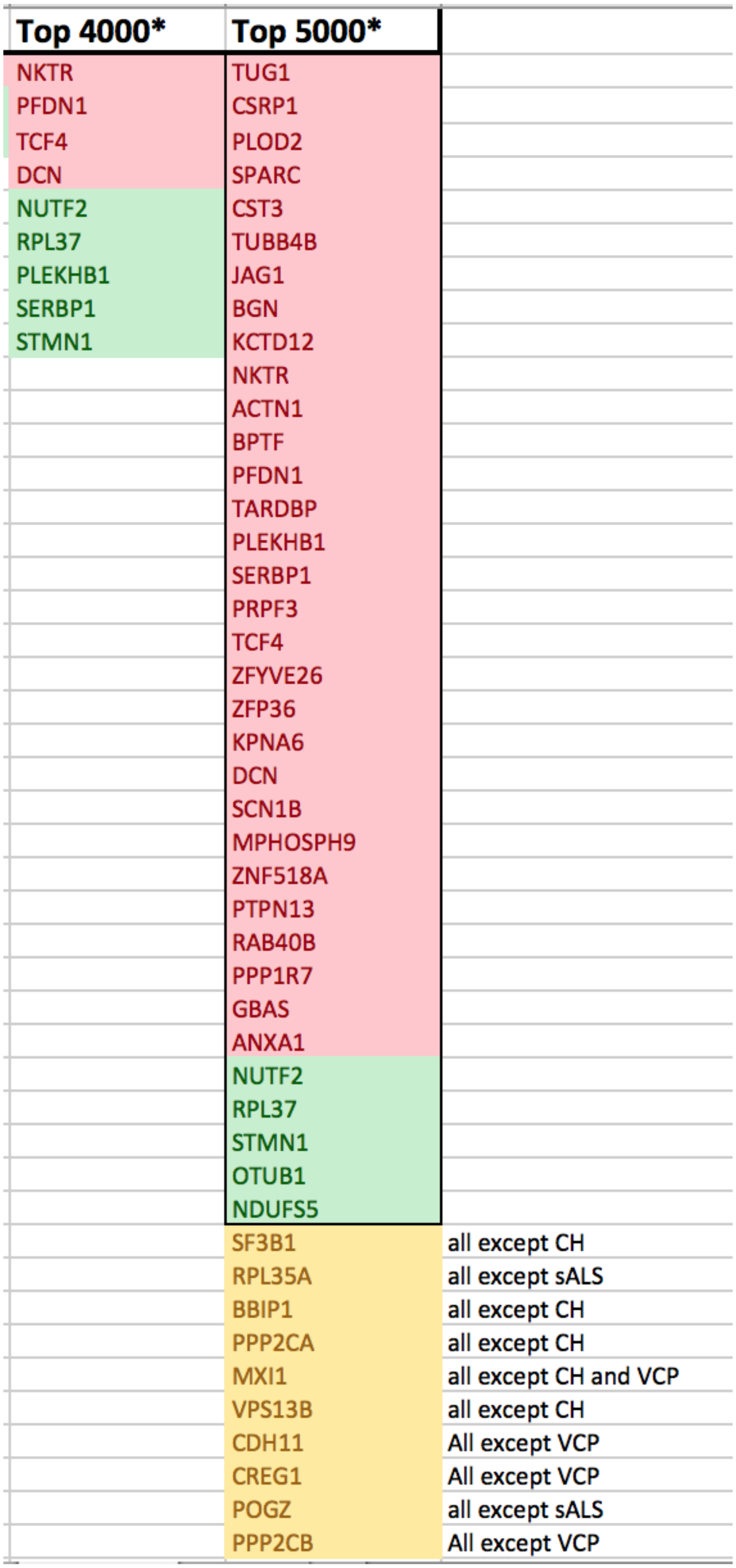
Top 1000	Top 2000	Top 3000	Top 4000	Top 5000
0	0	PFDN1	NKTR	TUG1
		NUTF2	PFDN1	CSRP1
		RPL37	TCF4	PLOD2
			DCN	SPARC
			NUTF2	CST3
			RPL37	TUBB4B
			PLEKHB1	JAG1
			SERBP1	BGN
			STMN1	KCTD12
				NKTR
				ACTN1
				BPTF
				PFDN1
				TARDBP
				PLEKHB1
				SERBP1
				PRPF3
				TCF4
				ZFYVE26
				ZFP36
				KPNA6
				DCN
				SCN1B
				MPHOSPH9
				ZNF518A
				PTPN13
				RAB40B
				PPP1R7
				GBAS
				ANXA1
				NUTF2
				PPP2CB
				PPP2CA
				SF3B1
				STMN1
				MXI1
				VPS13B
				CDH11
				CREG1
				POGZ
				RPL37
				OTUB1
				NDUFS5
				RPL35A
				BBIP1

This is the geneMANIA output. It is clear that there is a high level of coexpression according to current literature. It is important to note that TUG1 was not recognised potentially because it is a non-protein coding RNA.



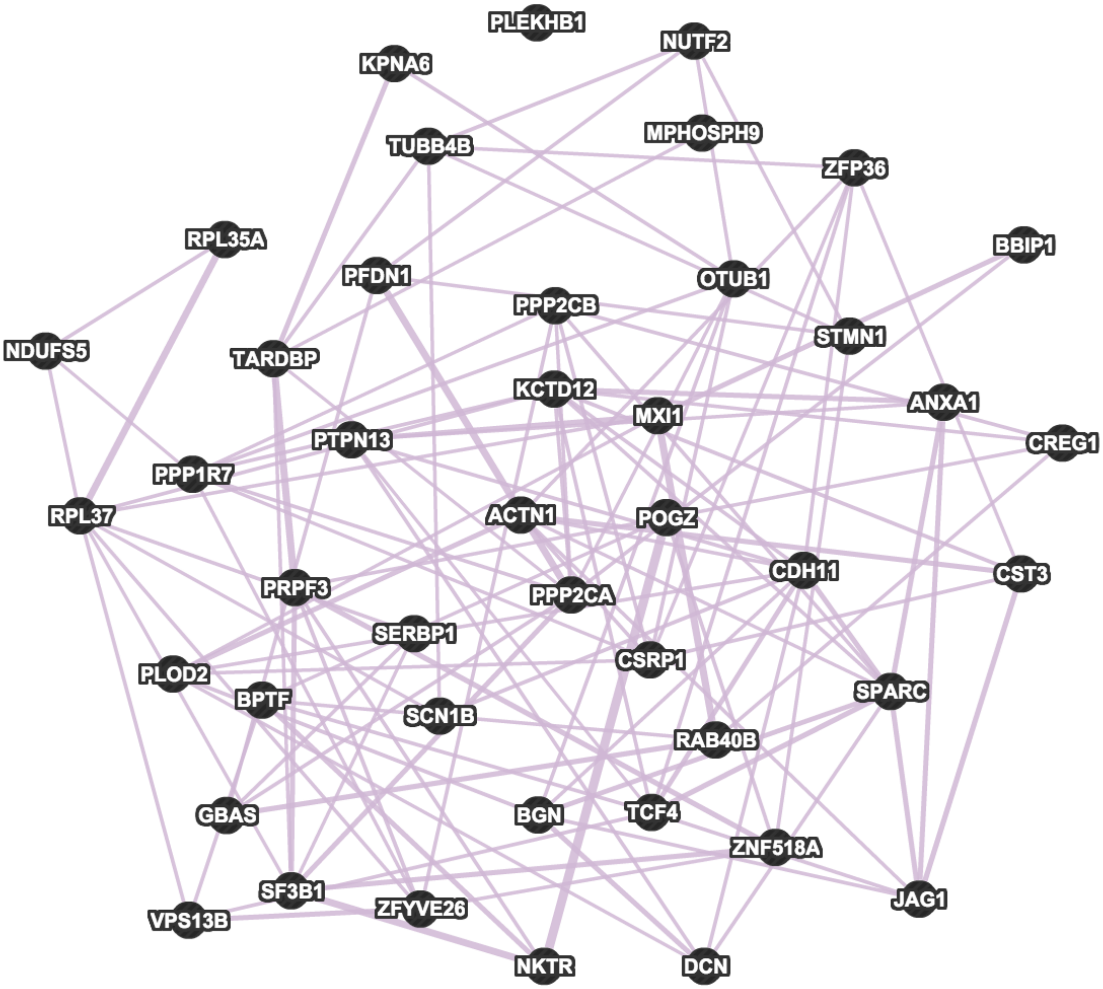
Tuesday

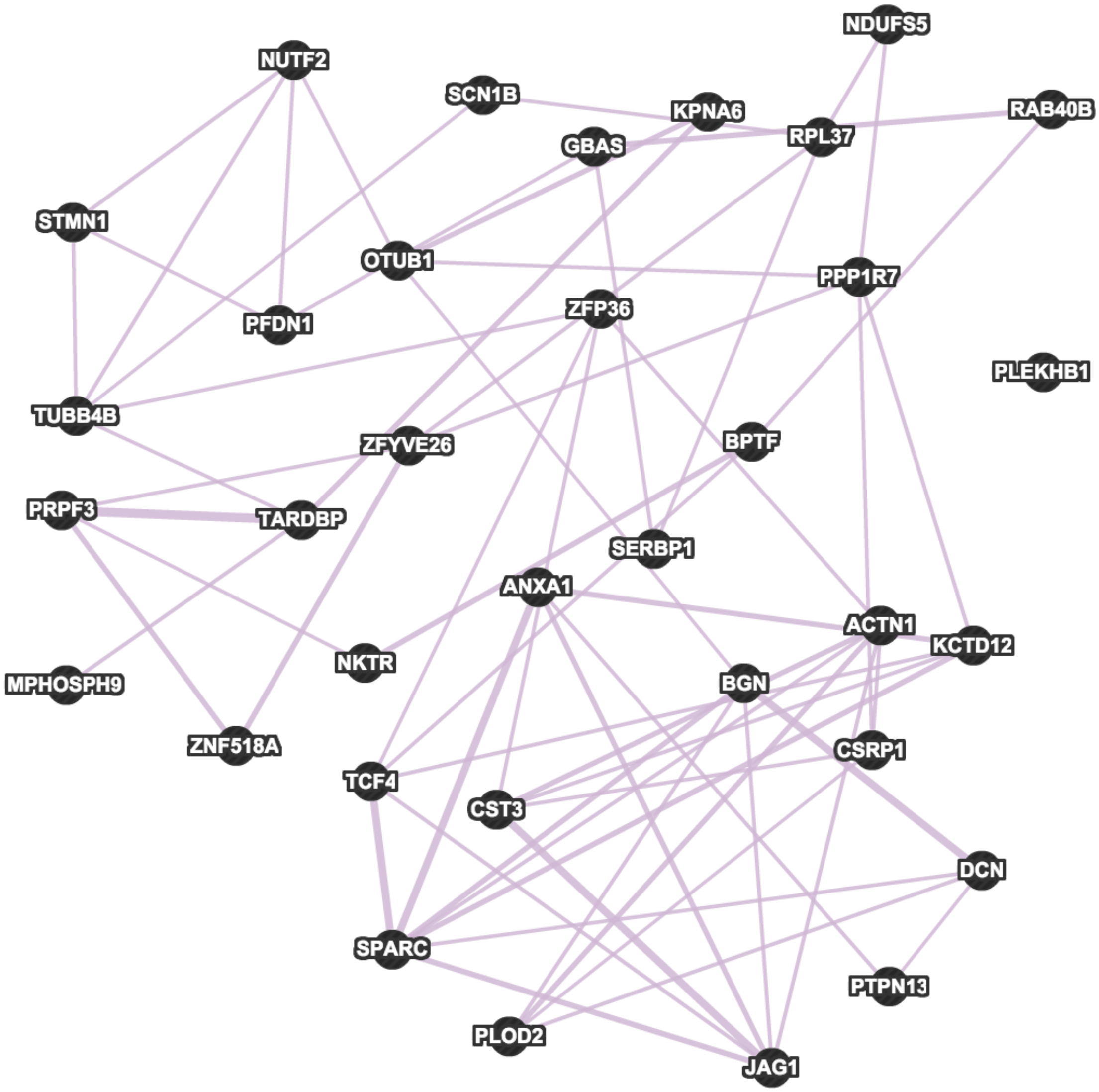
Since doing the intersect experiment left me with genes generated from using Ensembl IDs rather than HGNC symbols, I wanted to go through the files I had used to generate the consensus from HGNC symbols to check the extra genes were present but were just suffering from naming issues. As it turned out, 10 of the genes that were present in all of the ensembl ID lists were not present in all of the HGNC lists. It appears that the CHMP2B data set seems to be the most common culprit, potentially because there are only 3 patients.



Top 4000*	Top 5000*	
NKTR	TUG1	
PFDN1	CSRP1	
TCF4	PLOD2	
DCN	SPARC	
NUTF2	CST3	
RPL37	TUBB4B	
PLEKHB1	JAG1	
SERBP1	BGN	
STMN1	KCTD12	
	NKTR	
	ACTN1	
	BPTF	
	PFDN1	
	TARDBP	
	PLEKHB1	
	SERBP1	
	PRPF3	
	TCF4	
	ZFYVE26	
	ZFP36	
	KPNA6	
	DCN	
	SCN1B	
	MPHOSPH9	
	ZNF518A	
	PTPN13	
	RAB40B	
	PPP1R7	
	GBAS	
	ANXA1	
	NUTF2	
	RPL37	
	STMN1	
	OTUB1	
	NDUFS5	
	SF3B1	all except CH
	RPL35A	all except sALS
	BBIP1	all except CH
	PPP2CA	all except CH
	MXI1	all except CH and VCP
	VPS13B	all except CH
	CDH11	All except VCP
	CREG1	All except VCP
	POGZ	all except sALS
	PPP2CB	All except VCP

To understand if any of these genes were important, I put the remaining 35 geens into geneMANIA, and asked it to add the 100 most co-expressed genes. Of the added genes, none of which corresponded to the 10 genes identified. However, when I compared the geneMANIA results with and without those 10 genes, the network score was 10% higher with those genes than without. So I may consider keeping them depending on what the others say.





The genes clearly coexpress with multiple other genes in the list, and technically they were just as validly generated as the rest, I just don't know quite how this happened. I'm really starting to learn how delicate data is. . .

Functions

DEGS alone:

Feature	FDR	Genes in network	Genes in genome
extracellular matrix organization	3.03E-01	6	290
regulation of RNA stability	3.03E-01	3	34
extracellular structure organization	3.03E-01	6	291
regulation of mRNA stability	3.03E-01	3	33
3'-UTR-mediated mRNA stabilization	8.71E-01	2	10
dermatan sulfate metabolic process	9.40E-01	2	13
chondroitin sulfate catabolic process	9.40E-01	2	13
dermatan sulfate biosynthetic process	9.40E-01	2	12

DEGS with 10 coexpressed genes

Feature	FDR	Genes in network	Genes in genome
extracellular matrix organization	4.94E-02	7	290
'de novo' posttranslational protein folding	4.94E-02	4	38
'de novo' protein folding	4.94E-02	4	43
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	4.94E-02	5	116
extracellular structure organization	4.94E-02	7	291
mRNA catabolic process	4.94E-02	6	190
nuclear-transcribed mRNA catabolic process	4.94E-02	6	179
RNA catabolic process	7.07E-02	6	215
cellular protein complex disassembly	7.97E-02	5	135
mRNA binding	1.22E-01	4	77

Today I also worked on my abstract for the presentation I have to do on the 13th April.

Identification of a Molecular Signature for TDP-43 Pathology

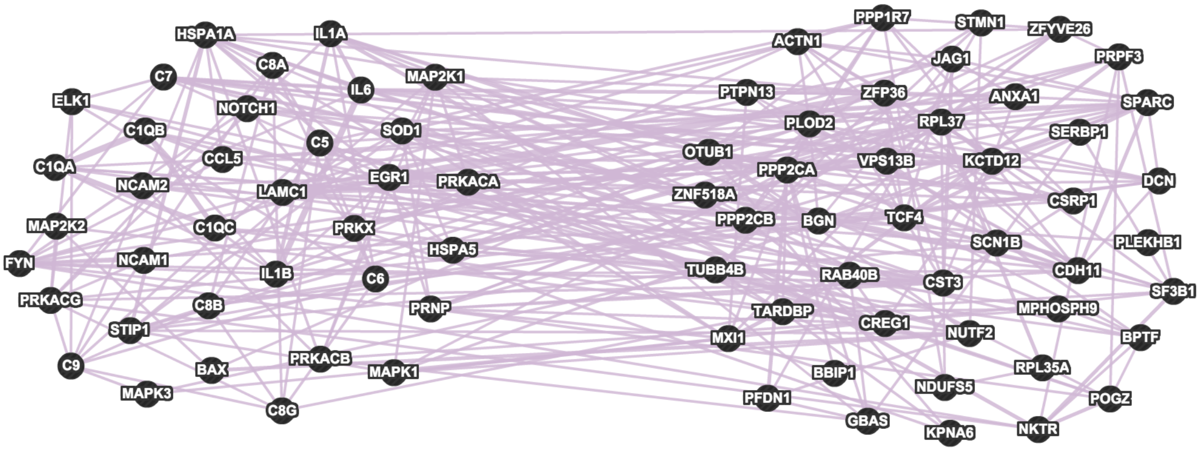
Green, C., Cooper-Knock, J. & Hide, W.

TDP-43 pathology is a histological hallmark for many neurodegenerative conditions, including amyotrophic lateral sclerosis and Alzheimer's disease. In affected individuals, TDP-43 protein is exported from the nucleus and aggregated into toxic cytoplasmic inclusions. These inclusions are believed to contribute to the neurodegenerative process, but the nature of this contribution is currently unclear, largely due to the variability in disease phenotypes. Consequently, this project has two aims; firstly, to identify cellular functions that are consistently dysregulated across all instances of TDP-43 pathology (thus generating a molecular signature), and secondly to identify specific components of that signature which drive the disease process, and are thus potential therapeutic targets. To generate the molecular signature, differential gene expression analysis was performed using the R limma package on 7 independent data sets: 5 from microarray experiments and 2 from RNA-Seq. The most differentially expressed genes from each data set were intersected, leaving a consensus of 45. Initial analysis of global co-expression indicates these genes are tightly co-expressed, and enrich for functions involved in cellular structure and RNA stability. High levels of co-expression suggest a functional relationship between the genes, and this function appears to relate to processes known to be associated with TDP-43 protein and neurodegeneration. Future investigations will involve validation of this signature in a separate cohort to ensure its robustness, as well as generation of disease-specific co-expression networks - including the incorporation of associated disease loci – as a data-driven approach to identifying both functional dysregulation and key drivers of disease.

Wednesday & Thursday

As I was writing my abstract, I realised that I was misisng something. Although I had run the pathprint experiment and idenitifed the DEGs, I had to tie them back to oneanother somehow. The obvious way to do that would be to show that the genes which are DE are enriched in the list of genes associated with the pathways idenitified by pathprint. Unfortunately, only two genes were shown to be present in one of the pathways, and I'm not sure of the statistics but I don't believe that will be of any significance. Alternatively, it would have been nice if the pathways from pathprint matched the pathways enriched in the gene set, but that's not quite true. So, what I did was this - I took the list of genes from each pathway, and I inputted them into geneMANIA. This allowed for idenitfication of any significantly enriched funcitons within each pathway gene set. From this I found extracellular matrix organization enriched in two pathways ({F2,46} (Static Module), and Complement and coagulation cascades (KEGG)). There was no mention of any RNA-related functions.

My next idea was to look at the coexpression of my gene set with the pathway gene sets. If the genes in my gene set do not appear in the pathways, perhaps they at least coexpress. Below is an example, and the full collection of images can be found in the powerpoint "PPvsDEGs" in the reuslts folder.


PPP1R7
STMN1
ZFYVE26
HSPA1A
IL1A
ACTN1
PRPF3
JAG1
C8A
C7
MAP2K1
IL6
PTPN13
ZFP36
ANXA1
SPARC
ELK1
NOTCH1
RPL37
PLOD2
C1QB
SOD1
SERBP1
C5
OTUB1
CCL5
KCTD12
VPS13B
C1QA
PPP2CA
DCN
EGR1
PRKACA
NCAM2
LAMC1
ZNF518A
CSRP1
TCF4
PPP2CB
BGN
MAP2K2
C1QC
PRKX
SCN1B
PLEKHB1
HSPA5
FYN
NCAM1
TUBB4B
CDH11
IL1B
C6
RAB40B
CST3
SF3B1
PRKACG
C8B
PRNP
TARDBP
MPHOSPH9
STIP1
CREG1
MXI1
NUTF2
BPTF
C9
BAX
PRKACB
MAPK1
RPL35A
BBIP1
NDUFS5
MAPK3
PFDN1
POGZ
C8G
GBAS
KPNA6
NKTR