Functional enrichment analysis, clustering and classification of gene expression profiles

Cluster analysis is used to identify genes with similar expression patterns over experiment conditions or possibly related functions. We will first perform k-means clustering and check the functions of the genes in the same clusters.

Number of Homo sapien annotated genes related to molecular function is 54878			
GO Term	p-value	counts	definition
GO:0016905	0.00057	3 / 15	Catalysis of the reaction: ATP + myosin-heavy-chain = ADP + myosin-heavy-chain phosphate." [EC:2.7....
GO:0004686	0.00083	3 / 17	Catalysis of the phosphorylation of eukaryotic elongation factor-2." [GOC:jl, PMID:11904175]...
GO:0004689	0.00116	3 / 19	Catalysis of the reaction: 4 ATP + 2 phosphorylase b = 4 ADP + phosphorylase a." [EC:2.7.11.19]...
GO:0005251	0.00155	5 / 77	Catalysis of the transmembrane transfer of a potassium ion by a delayed rectifying voltage-gated ch...
GO:0005094	0.00184	2 / 6	Prevents the dissociation of GDP from the small GTPase Rho, thereby preventing GTP from binding." [...
GO:0004879	0.00489	6 / 1251	A ligand-dependent receptor found in the nucleus of the cell." [GOC:ai]...
GO:0005252	0.00575	4 / 66	Catalysis of the transmembrane transfer of a potassium ion by an open rectifier voltage-gated chann...
GO:0005249	0.00601	5 / 107	Catalysis of the transmembrane transfer of a potassium ion by a voltage-gated channel." [GOC:mtg_tr...
GO:0001653	0.00609	6 / 1223	Combining with an extracellular or

			intracellular peptide to initiate a change in cell activity." [G...
GO:0015026	0.00609	6 / 1223	Combining with an extracellular or intracellular messenger, and in cooperation with a nearby primar...

The top 10 enriched functions and find evidences how these functions are related with ovarian serous cystadenocarcinoma (OV)

0016905:
This term is a catalyzer of the reaction: ATP + myosin-heavy-chain = ADP + myosin-heavy-chain phosphate
The relation to OV is that this set contains gene TNFRSF4, a Tumor necrosis factor receptor superfamily member 4. This is related to OV as this gene is a cell signaling protein involved in acute phase reaction to inflammation (precursor to tumor)

0004686
This term is defined to aid in catalysis of the phosphorylation of eukaryotic elongation factor-2. Obviously protein synthesis regulation would be an important factor is the development of a tumor. This is a possible tie to ovarian serous cystadenocarcinoma

0004689
The geneontology website has this to say about this annotation “Note that, in addition to forming the root of the molecular function ontology, this term is recommended for use for the annotation of gene products whose molecular function is unknown. When this term is used for annotation, it indicates that no information was available about the molecular function of the gene product annotated as of the date the annotation was made”, so not much can be gleaned from this function

0005251
This term has various annotations including among ATPase activator activity, the nucleoplasm, cytosol. It has no high level behaviors tied to it that would obviously tie to an ovarian tumour, but that is not to discount that there may not be some tie

0005094
This term is tied to the dissociation of GDP from the small GTPase Rho, thereby preventing GTP from binding. As GTP can act as a substrate for both the synthesis of RNA during the transcription process and of DNA during DNA replication, it’s prevention can have links to DNA mistranscription. A tie back to how DNA mutations can result in tumours.

0004879

In gene Ddit3 which is within this function, we have an agent of stress response as a result of a stress acting at the endoplasmic reticulum. Stress of the endoplasmic reticulum has a connection to cisplatin resistance in human **ovarian cancer** cells

0005252
This term has various annotations including among cytosol, the nucleoplasm, and the extracellular regions. It has no high level behaviors tied to it that would obviously tie to an ovarian tumour, but that is not to discount that there may not be some tie.

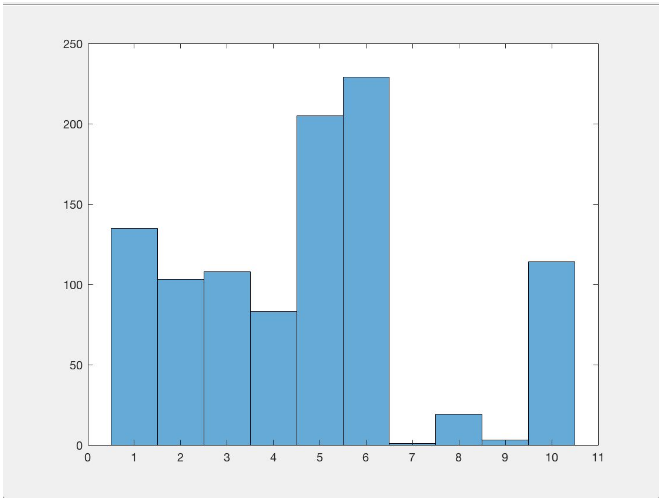
0005249
This term has to do with the catalysis of the transmembrane transfer of a potassium ion by a voltage-gated channel. The cells and functions it is associated with have no clear ties to ovarian cancer.

0001653
This term has ties to the positive and negative regulation of dna transcription, as well as protein binding. If there was a connection to an occurrence of ovarian cancer it would be in the generation of new dna transcripts.

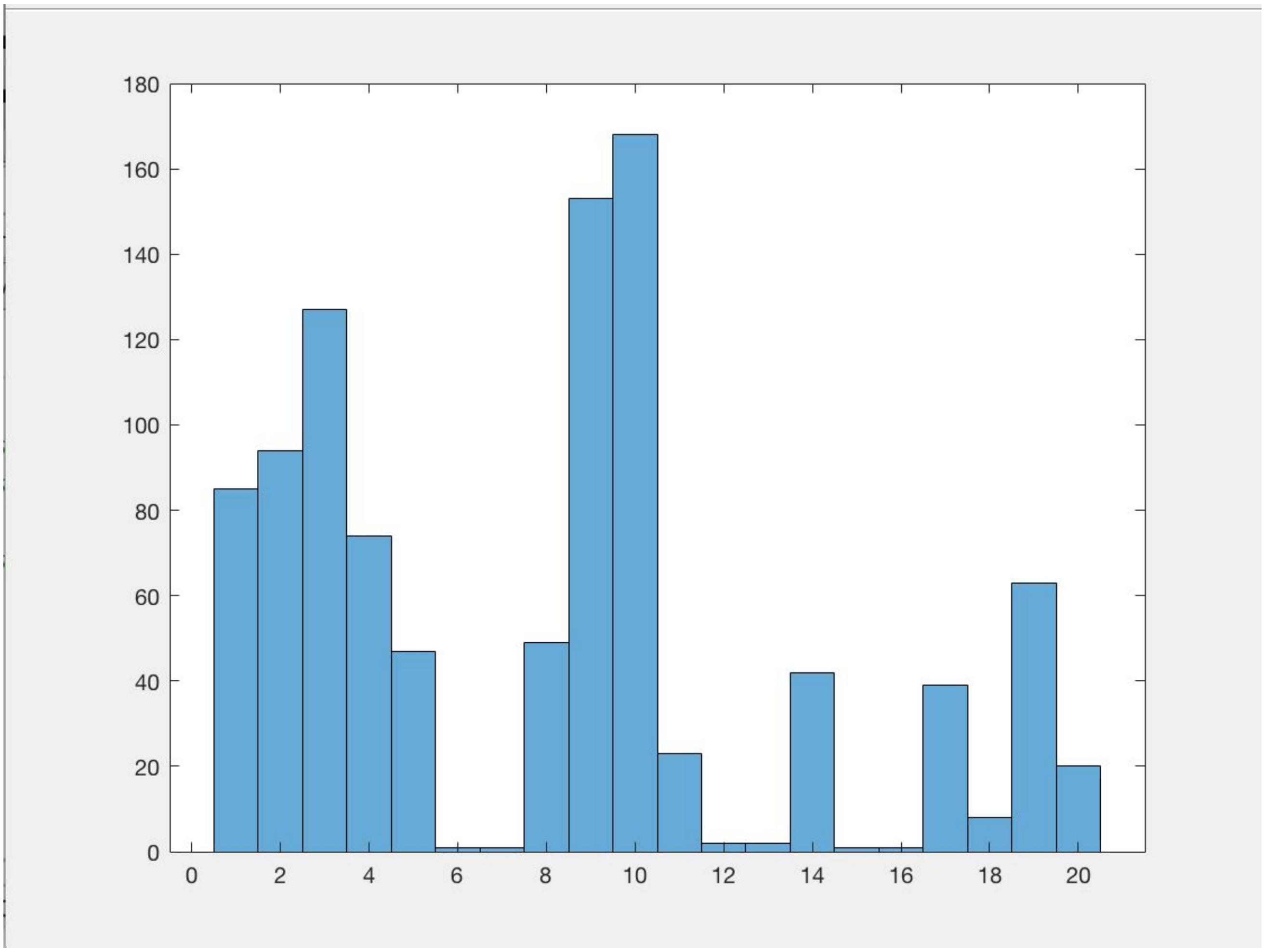
0015026
This term is tied to the behavior of Combining with an extracellular or intracellular messenger, and in cooperation with a nearby primary receptor, initiating a change in cell activity. Becoming a tumor is a major change in cell activity, and the genes tied to this term may be involved in the occurrence of a tumour.

Clustering Accuracies

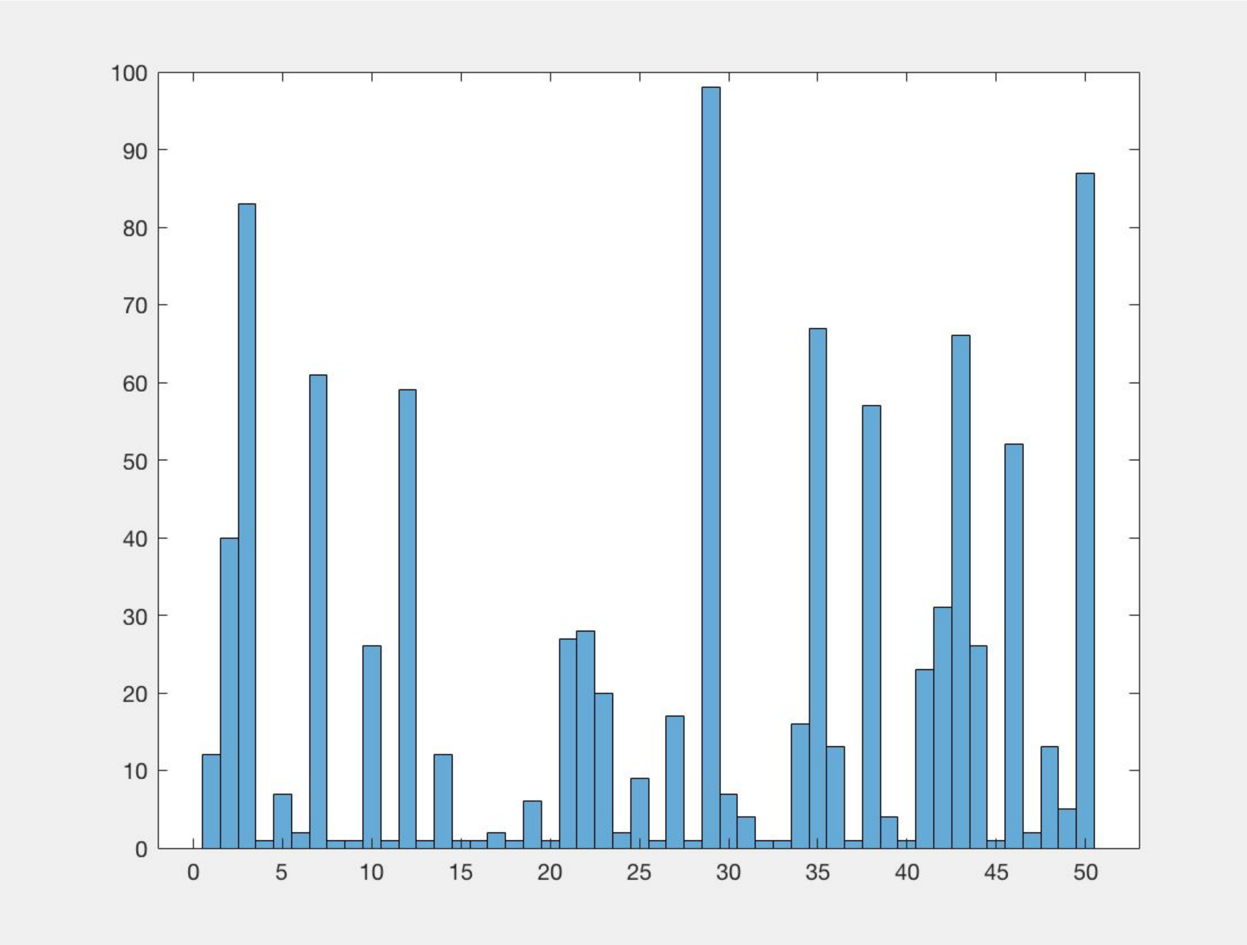
K = 10



K = 20



K = 50



Selected one cluster from the k-means clustering with k = 20. Performed functional enrichment analysis of the genes in the cluster. Described the enriched functions of the genes in the cluster. Report the list of genes in the cluster and the enriched functions.

Selected cluster 3,

List of genes

TEX261'	TJP1'	CASC4'	PCYOX1'	RBP1'	C11orf2'	TMEM33'
ADIPOR2'	AGFG1'	TCEB3'	ARSD'	SERINC3'	UNC5B'	LSM14A'
ITCH'	ARID5B'	EIF1AX'	ALCAM'	SEC24C'	C11orf31'	VPS24'
ARFGEF2'	GLUD1'	GPS1'	ANKFY1'	ATP2C1'	FAM89B'	H2AFX'

UPK3B'	GSK3B'	EIF4G3'	KIAA0196'	UBE2G1'	MCM3'	UFC1'
LMO4'	FH'	PSMC4'	TBCD'	KEAP1'	PTPRU'	FUT8'
ZNF148'	PAFAH1B1'	KIF1C'	KDM5A'	ATP6V1C1'	TBL2'	MRPL12'
KIF5B'	PPP1R14B'	SNX1'	XAB2'	RARRES2'	MED29'	ECH1'
PYGB'	PSMA1'	AP3M1'	ARIH1'	SLC3A2'	AFTPH'	MLL4'
ADAM10'	LAMB3'	KPNA1'	DDX21'	HGS'	PPP2CB'	ERC1'
ZBTB4'	TAB2'	ANXA4'	TMEM106C'	SIPA1L3'	C9orf123'	C11orf68'
DNAJC4'	ELMO2'	ZNF146'	MANF'	FDPS'	WDTC1'	SSH1'
THOC4'	TMEM205'	EPHB2'	STRAP'	SH3PXD2A'	IPO8'	CLPTM1L'
RALGAPB'	SEMA3F'	PHGDH'	DNAJC5'	HADH'	GNA12'	
C20orf3'	DUS1L'	FYTTD1'	FAM195B'	PYCR2'	TMED9'	
NDRG3'	BMI1'	PPFIA1'	ZNF358'	ZC3H13'	ARHGAP21'	
BEX4'	MAPKAP1'	CYB561'	ITGB8'	MTOR'	CASP2'	
SNRPA'	SERINC1'	AAK1'	VKORC1L1'	GRLF1'	GPR108'	
ELAVL1'	HBP1'	FAM127B'	ANAPC11'	CNOT8'	STRN4'	

Enriched Functions

Number of Homo sapien annotated genes related to molecular function is 54878
GO Term p-value counts definition

GO:0005488	0	69 / 7899	The selective, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule.
GO:0004930	0.0003	0 / 860	A receptor that binds an extracellular ligand and transmits the signal to a heterotrimeric c G-protein complex.G-protein complex.
GO:0046872	0.00077	20 / 3700	Interacting selectively with any metal ion.
GO:0005515	0.00128	70 / 7185	Interacting selectively with any protein or protein complex
GO:0004300	0.00156	2 / 7	Catalysis of the reaction: (3S)-3-hydroxyacyl-CoA = trans-2(or 3)-enoyl-CoA + H2O.
GO:0004984	0.0017	0 / 685	Combining with airborne compounds to initiate a change in cell activity.
GO:0017124	0.00301	6 / 169	Interacting selectively with a SH3 domain (Src homology 3) of a protein
GO:0030275	0.00344	4 / 73	Interacting selectively with a LRR domain (leucine rich repeats) of a protein.
GO:0005509	0.00465	19 / 3275	Interacting selectively with calcium ions (Ca2+)." [GOC:ai]...

GO:0016836	0.00585	3 / 43	Catalysis of the cleavage of a carbon-oxygen bond by elimination of water." [EC:4.2.1]...

Gene Ontology object with 16 Terms.
Biograph object with 16 nodes and 15 edges.

Figure 3.1)

K	Accuracy (%)
1	62.16
3	64.86
5	62.16

Figure 2)

K	Accuracy (%)
1	67.57
3	70.27
5	70.27

One explanation for why a first-1000 feature trained model is better than a model trained on all the features is the concept of overfitting your data. The concept of overfitting is when a model "corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably"-(src: https://en.wikipedia.org/wiki/Overfitting). The 1000 feature trained model must generalize better to the test data and it is not as closely fit to the training data as the exhaustively trained model is.

Figure 3. 72.97% is the classification accuracy