Practical 3: Integrative analysis. ChIP-seq & RNA-seq together

Vladimir Teif (vteif@essex.ac.uk)

In this practical session we will continue the p53 binding story started previously, integrating the data that we have obtained with respect to p53 binding (ChIP-seq) and gene expression (RNA-seq).

Summary of the previous practicals. *Our previous practicals werebased on the data reported in the study entitled "Integrative genomic analysis reveals widespread enhancer regulation by p53 in response to DNA damage" (Younger et al. (2015) Nucleic Acids Res. 43 (9): 4447-4462). The full text of this article is available at http://nar.oxfordjournals.org/content/43/9/4447.long. This paper is about chromatin binding of the tumour suppressor protein p53. The authors have determined genome-wide p53 binding profiles in human and mouse cells. Their main finding was that p53 binding occurs predominantly within transcriptional enhancers. You have previously mapped the p53 ChIP-seq data, called peaks to detect p53 binding sites, and checked the overlapping of p53 binding sites with promoters and enhancers. Now we will perform an integrative analysis combing the p53 protein binding data with gene expression changes for the same mouse cells treated with a drugs doxorubicin.*

The data generated by the authors of the article that we use in our practicals are available at the following GEO accession number: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55727.

Plan for this practical:

1 Understand results of differential gene expression analysis based on RNA-seq
2 Determine whether strength of p53 binding at gene promoters correlates with gene expression
3 Perform Gene Ontology (GO) analysis using DAVID
4 Perform Gene Ontology (GO) analysis using GOrilla
5 Perform Gene Ontology (GO) analysis using EnrichR
6 Make a list of TFs co-binding with p53 and construct their interaction network

Task 1. Understand results of differential gene expression analysis based on RNA-seq.

Some people say that 90% of bioinformatics is data conversion from one format to another. Bioinformaticians do not agree with this and cannot tell you what constitutes the remaining 10% ☺

In this case you are lucky, because I have already processed RNA-seq data from this paper for you, and it is already in a human-readable format, very similar to the BED file format in which we have previously obtained p53 binding peaks. Here is how the differential gene expression data look like:

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
Ccng1	10253.9565478971	2.1733415985728	0.0496837029960621	43.7435510542573	0	0
Plau	2868.70628875291	2.29644968758308	0.0591563933877861	38.8199745804182	0	0
Adamts5	2965.14805964652	-3.5324246968983	0.0745015980961711	-47.4140795253604	0	0
Nr4a1	1953.34530631308	3.1725957908854	0.0746751971051276	42.4852683872937	0	0
Ptx3	10991.9420032442	-2.54241242151884	0.0486014725308458	-52.3114278050991	0	0
Icam1	4478.63735905254	2.23008961534929	0.0578959006213177	38.5189554254582	0	0
Notch3	2249.90055725676	2.73860762232716	0.0732767783235185	37.3734719918519	1.05419963864197e-305	1.6483164349909e-302
Epha2	2135.073342786	2.45779451307348	0.0672508448471497	36.5466711780298	2.01401713077518e-292	2.75542718704179e-289
Crip2	1442.08969261518	2.94651026539472	0.0842783818167692	34.9616378705605	8.61827045902016e-268	1.04807744637751e-264
Il6st	12913.6159391834	-1.45668667038069	0.042416124701033	-34.3427571624717	1.80658921536963e-258	1.97731189622206e-255
Mt2	2187.97154447509	-1.97523511200763	0.0620548411881086	-31.8304756597482	2.45248102528859e-222	2.44021862016215e-219
Mki67	8680.16437898843	-1.79983281997386	0.0568558109468329	-31.6560926667095	6.25200113502026e-220	5.70234603523307e-217
Ckap2	5255.93442864738	1.7628614755545	0.0558089037520847	31.5874592947663	5.48925058747985e-219	4.62152674461284e-216

As you can see, the first column gives us the name of the gene, the third column gives expression log2 fold change between two cell conditions, and the fifth column gives the P value. These are perhaps the most interesting columns from the point of view of what changes and how much is the change upon cell treatment.

Now let us look at the file containing p53 bound sites that we have created during the first ChIP-seq analysis practical:

chr8	13548925	13549101	+	1050.7	0.888	307	1160.7	3	386.91	0.00E+00	143.98	0.00E+00	0.58
chr12	111963380	111963556	+	1015.2	0.89	319	1121.5	2	560.76	0.00E+00	216.46	0.00E+00	0.55
chr7	139921178	139921354	+	810.2	0.91	292	895.1	5	179.01	0.00E+00	169.87	0.00E+00	0.6
chr8	12634989	12635165	+	654	0.934	157	722.5	1	722.52	0.00E+00	82.94	0.00E+00	1.11
chr4	128252925	128253101	+	600.8	0.864	186	663.7	4	165.92	0.00E+00	52.7	0.00E+00	0.93
chr1	156903370	156903546	+	561.7	0.908	263	620.6	5	124.11	0.00E+00	101.82	0.00E+00	0.65
chr10	90881469	90881645	+	537.8	0.808	241	594.1	2	297.05	0.00E+00	104.72	0.00E+00	0.71
chr7	87100003	87100179	+	525.3	0.969	150	580.4	2	290.18	0.00E+00	65.18	0.00E+00	1.21
chr17	29227791	29227967	+	500.5	0.877	261	552.9	4	138.23	0.00E+00	38.91	0.00E+00	0.65
chr8	23544523	23544699	+	473	0.866	187	522.5	3	174.18	0.00E+00	20.91	0.00E+00	0.89
chr5	140199090	140199266	+	459.7	0.863	266	507.8	4	126.95	0.00E+00	98.01	0.00E+00	0.62
chr10	117154716	117154892	+	449	0.894	234	496.1	4	124.01	0.00E+00	98.41	0.00E+00	0.71
chr1	54901247	54901423	+	444.6	0.914	246	491.2	4	122.79	0.00E+00	135.75	0.00E+00	0.67
chr8	64780293	64780469	+	437.5	0.923	226	483.3	2	241.66	0.00E+00	211.68	0.00E+00	0.73
chr15	85690303	85690479	+	428.6	0.895	231	473.5	0.5	947.02	0.00E+00	82.45	0.00E+00	0.71
chr9	117068448	117068624	+	423.3	0.659	221	467.6	9	51.96	0.00E+00	24.36	0.00E+00	0.74
chr3	32263187	32263363	+	419.7	0.925	231	463.7	4	115.93	0.00E+00	99.63	0.00E+00	0.7
chr8	23545199	23545375	+	419.7	0.745	247	463.7	5	92.74	0.00E+00	17.78	0.00E+00	0.66
chr10	117147028	117147204	+	418.8	0.832	241	462.7	2	231.36	0.00E+00	76.15	0.00E+00	0.67
chr2	167389561	167389737	+	407.3	0.911	198	450	3	149.99	0.00E+00	81.59	0.00E+00	0.84
chr4	149423131	149423307	+	407.3	0.801	236	450	2	224.99	0.00E+00	49.37	0.00E+00	0.69

In the BED file above, each line corresponds to one p53 peak determined in ChIP-seq. The first column gives the chromosome number, the second column – region start, the third column – region end, the fourth column – strand (all peaks are assumed to be on the plus strand, because the strand information actually disappears after we call a peak), the fourth column is the score of the peak (the higher the peak the bigger its score). These are all the columns that we will need.

It is easy to see that the RNA-seq data and ChIP-seq data are represented in quite different formats. For example, the RNA-seq data only contain the gene name, but do not contain the genomic coordinates of this gene. Since the mouse genome is pretty much annotated, it is possible to get genomic coordinates for each gene, but doing this manually would be too much work. We need to need to make some trick in order to add the genomic coordinates to the genes. But before we do this, let us ask ourselves a question: what is it that we want to learn from the combined analysis of RNA-seq and ChIP-seq? May be we have some hypothesis that we want to check?

For example, say, we have the following hypothesis. We guess that p53 binding at regulatory regions should affect the genes associated with those regulatory regions. What are the regulatory regions? Promoters and enhancers. Let us just take the promoters for simplicity. Promoters are the regulatory regions upstream of the gene. There is no consensus among scientists as to how large the promoters are. A good estimate for a promoter size is about 1-2 kb. We have previously used a BED file with coordinates of all mouse promoters, named "promoters_mm9.bed":

chr4	131977322	131979322	-	GXT_12943606	AK049209	GXL_283229	Phactr4
chr4	42215999	42217999	-	GXT_12943623	AK047126	GXL_778728	Gm10931
chr7	109212607	109214607	-	GXT_12944438	AK078509	GXL_287330	Rnf121
chr14	5944054	5946054	-	GXT_12946537	AK084071	GXL_778563	Gm10021
chr17	95233138	95235138	-	GXT_12947170	AK082664	GXL_461852	Gm1976
chr17	95148281	95150281	-	GXT_12947186	AK080683	GXL_473176	Mettl4
chr19	39536565	39538565	-	GXT_12947662	AK050051	GXL_171813	Cyp2c38
chr7	109207990	109209990	-	GXT_12949553	AK034806	GXL_287330	Rnf121
chr7	109212649	109214649	-	GXT_12949662	AK089714	GXL_287330	Rnf121
chrX	67694797	67696797	-	GXT_12950375	AK089806	GXL_216606	AK089806
chr17	95148211	95150211	-	GXT_12951740	AK043389	GXL_473176	Mettl4
chr17	53092628	53094628	-	GXT_12951756	AK040895	GXL_225725	Kcnh8
chr17	33391090	33393090	-	GXT_12951767	AK038946	GXL_660138	Zfp955a
chr17	6957390	6959390	-	GXT_12951785	AK035271	GXL_155066	Ezr
chr4	25541413	25543413	-	GXT_12953332	AK085009	GXL_282468	Fut9

This file contains almost 200,000 promoters in the mouse genome. Interestingly, the number of annotated genes in the mouse genome is just about 60,000.

How is it possible, that there are more promoters than genes? For example, in the table above we can spot three instances of gene Rnf121, which has three different promoters. Indeed, many genes have several alternative transcripts, alternative transcription start sites, and each of these alternative transcription start sites has its own promoter. But the problem is that the file with the results of the differential gene expression quantifies gene expression per gene, not per gene transcript. There is an easy (and dirty) solution to remove some lines from the file promoters_mm9.bed which contain duplicated gene names. By doing so, we keep only one promoter per gene. It is easy to do this in Excel, so I have done it for you. The file promoters_mm9_52k.bed contains one promoter per gene, in total about 52 thousand genes.

After I have added promoter coordinates to the RNA-seq differential expression file, the resulting file promoters_and_DEseq.bed looks like this:

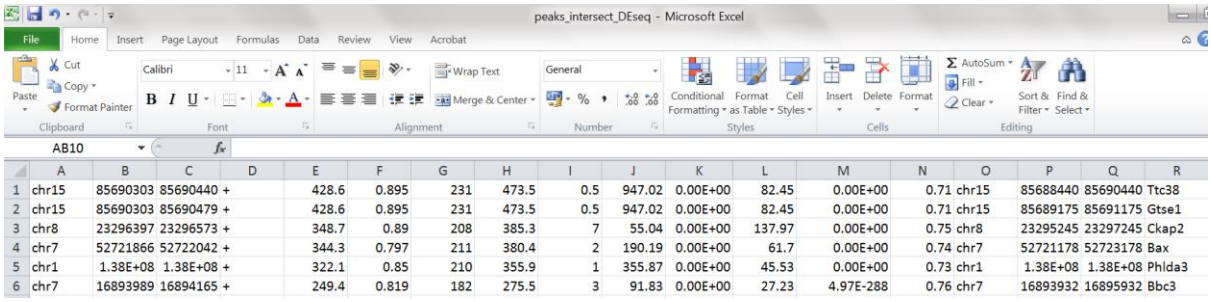
chr4	42215999	42217999	Gm10931	Gm10931	0	NA	NA	NA	NA	NA
chr7	109212607	109214607	Rnf121	Rnf121	0	NA	NA	NA	NA	NA
chr14	5944054	5946054	Gm10021	Gm10021	0	NA	NA	NA	NA	NA
chr17	95148281	95150281	Mettl4	Mettl4	0	NA	NA	NA	NA	NA

Here the first column is the chromosome number, the second column in region start, the third column is region end, then goes the gene name and its differential expression data (in this case of the four genes printed here the expression data is not available, but for most other genes these are available). We can notice that this resembles the BED format which we have seen a lot previously during the ChIP-seq practical. And we know how to find the intersection between two files in BED format. This is what we previously did for the intersection of p53 sites with different genomic features. Now we can intersect p53 sites with the promoters linked to their corresponding gene expression data from RNA-seq. You do not need to do this, because I have already done this for you. The results of this calculation are stored in a new file called peaks_intersect_DEseq.bed. This file is also situated in the shared Galaxy history.

The file peaks_intersect_DEseq.bed finally contains all the information we need to integrate p53 binding ChIP-seq and gene expression RNA-seq data. As I said, I have already prepared this file for you, so that you focus on more interesting steps of the analysis. Now, here is what you have to do:

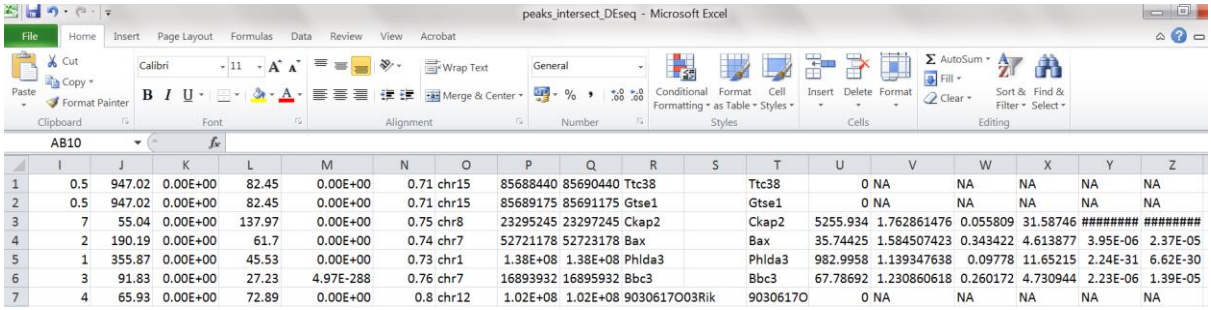
Task 2. Determine whether p53 binding at gene promoters correlates with gene expression

Copy file `peaks_intersect_DEseq.bed` from Galaxy to your computer, and then open it in Excel:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	chr15	85690303	85690440	+	428.6	0.895	231	473.5	0.5	947.02	0.00E+00	82.45	0.00E+00	0.71	chr15	85688440	85690440	Ttc38
2	chr15	85690303	85690479	+	428.6	0.895	231	473.5	0.5	947.02	0.00E+00	82.45	0.00E+00	0.71	chr15	85689175	85691175	Gtse1
3	chr8	23296397	23296573	+	348.7	0.89	208	385.3	7	55.04	0.00E+00	137.97	0.00E+00	0.75	chr8	23295245	23297245	Ckap2
4	chr7	52721866	52722042	+	344.3	0.797	211	380.4	2	190.19	0.00E+00	61.7	0.00E+00	0.74	chr7	52721178	52723178	Bax
5	chr1	1.38E+08	1.38E+08	+	322.1	0.85	210	355.9	1	355.87	0.00E+00	45.53	0.00E+00	0.73	chr1	1.38E+08	1.38E+08	Phlda3
6	chr7	16893989	16894165	+	249.4	0.819	182	275.5	3	91.83	0.00E+00	27.23	4.97E-288	0.76	chr7	16893932	16895932	Bbc3

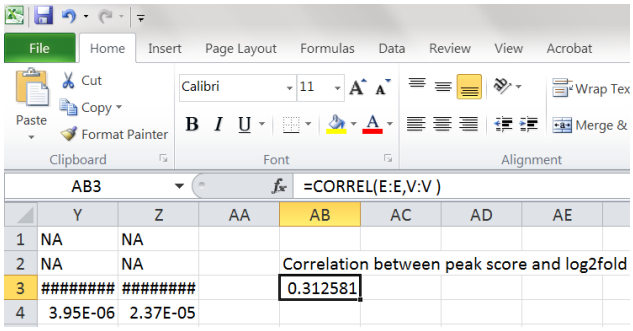
This picture shows only part of the Excel file. Here we can see the information about the peaks. If we scroll more to the right, we will see the second part of the same file:



	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	0.5	947.02	0.00E+00	82.45	0.00E+00	0.71	chr15	85688440	85690440	Ttc38		Ttc38	0	NA	NA	NA	NA	NA
2	0.5	947.02	0.00E+00	82.45	0.00E+00	0.71	chr15	85689175	85691175	Gtse1		Gtse1	0	NA	NA	NA	NA	NA
3	7	55.04	0.00E+00	137.97	0.00E+00	0.75	chr8	23295245	23297245	Ckap2		Ckap2	5255.934	1.762861476	0.055809	31.58746	########	########
4	2	190.19	0.00E+00	61.7	0.00E+00	0.74	chr7	52721178	52723178	Bax		Bax	35.74425	1.584507423	0.343422	4.613877	3.95E-06	2.37E-05
5	1	355.87	0.00E+00	45.53	0.00E+00	0.73	chr1	1.38E+08	1.38E+08	Phlda3		Phlda3	982.9958	1.139347638	0.09778	11.65215	2.24E-31	6.62E-30
6	3	91.83	0.00E+00	27.23	4.97E-288	0.76	chr7	16893932	16895932	Bbc3		Bbc3	67.78692	1.230860618	0.260172	4.730944	2.23E-06	1.39E-05
7	4	65.93	0.00E+00	72.89	0.00E+00	0.8	chr12	1.02E+08	1.02E+08	9030617O03Rik		9030617O	0	NA	NA	NA	NA	NA

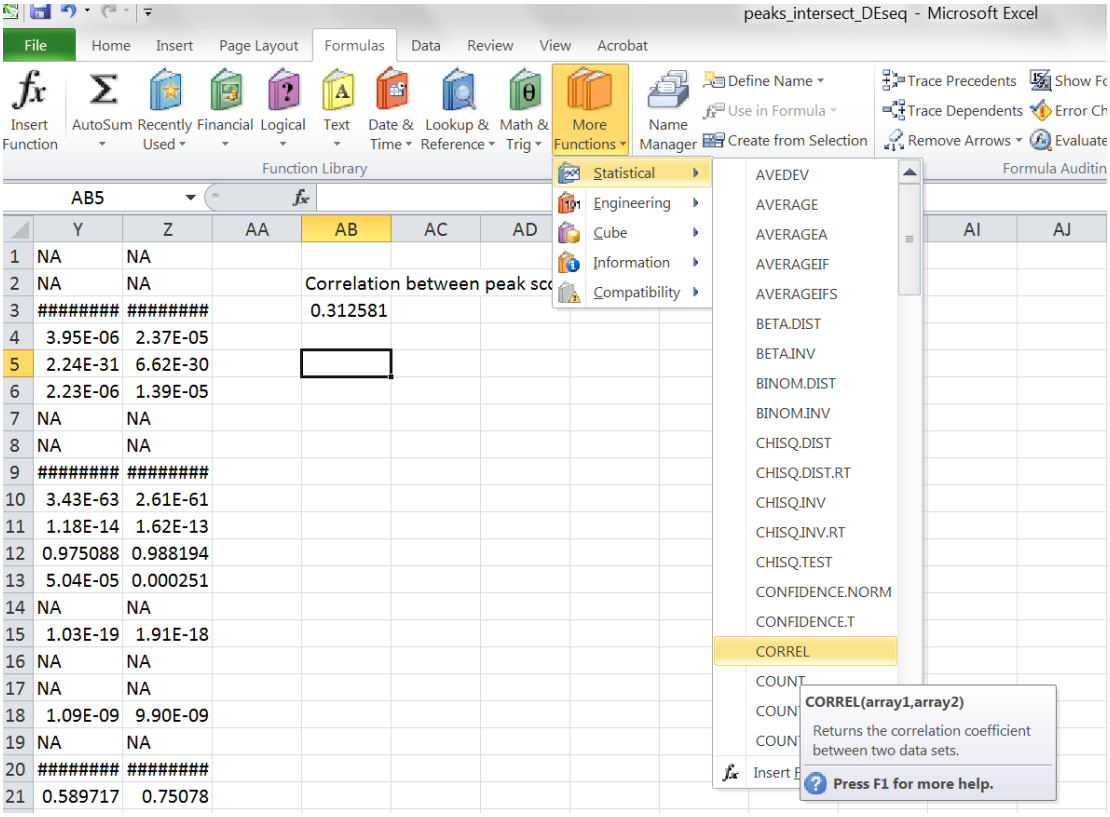
Let us focus on the quantitative characteristics of p53 binding to the promoter and changes of gene expression changes for the corresponding gene. The strength of p53 binding is characterised by the ChIP-seq peak height, which is given by the peak score in column “E”. The change of gene expression is given by the log2 fold change in the column “V”.

The simplest hypothesis that we can text now it this: whether the strength of p53 binding at the promoter is correlated to the change of gene expression? To test this hypothesis we need to calculate the correlation between columns “E” and “V”. This is easy to do in Excel. Just select any empty cell, place there the cursor, and insert there the equation for the correlation between columns “E” and “V”:



	Y	Z	AA	AB	AC	AD	AE
1	NA	NA					
2	NA	NA		Correlation between peak score and log2fold			
3	########	########		0.312581			
4	3.95E-06	2.37E-05					

In case if you are still wondering where to find the CORREL function in Excel, here it is:



	Y	Z	AA	AB	AC	AD	AI	AJ
1	NA	NA						
2	NA	NA		Correlation between peak sc				
3	########	########		0.312581				
4	3.95E-06	2.37E-05						
5	2.24E-31	6.62E-30						
6	2.23E-06	1.39E-05						
7	NA	NA						
8	NA	NA						
9	########	########						
10	3.43E-63	2.61E-61						
11	1.18E-14	1.62E-13						
12	0.975088	0.988194						
13	5.04E-05	0.000251						
14	NA	NA						
15	1.03E-19	1.91E-18						
16	NA	NA						
17	NA	NA						
18	1.09E-09	9.90E-09						
19	NA	NA						
20	########	########						
21	0.589717	0.75078						

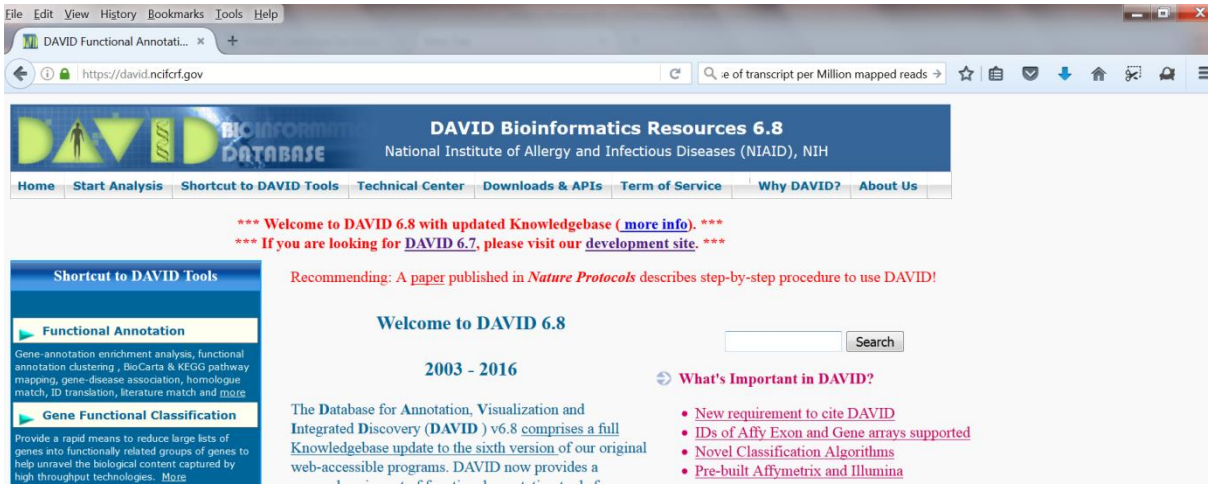
Discuss: Which correlation did you get? What can we say about this correlation? Is it large, small, or moderate? Is it statistically significant? Did you expect it like this at all?

Task 3. Gene Ontology (GO) analysis with DAVID.

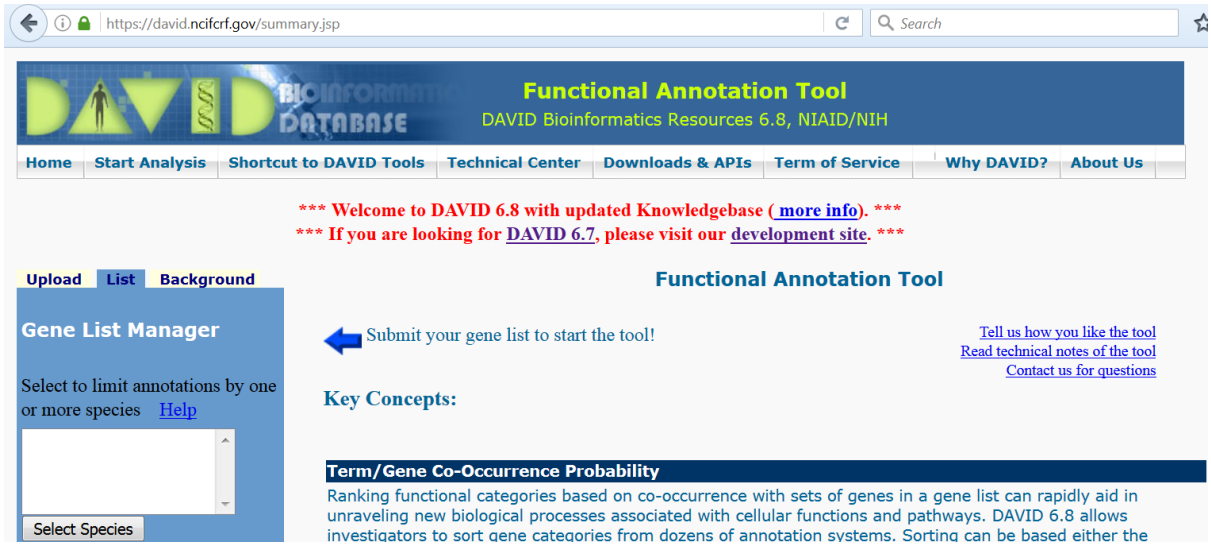
The first type of integrative analysis that I suggest you to try is the easiest to do and also quite a fun thing. Usually wet lab biologists love this type of analysis because it gives them an impression that they understood a lot about the system (in many cases this is an illusion, though). Let's just try it ☺

Let us perform GO analysis for genes which contain bound p53 at their promoters using software DAVID.

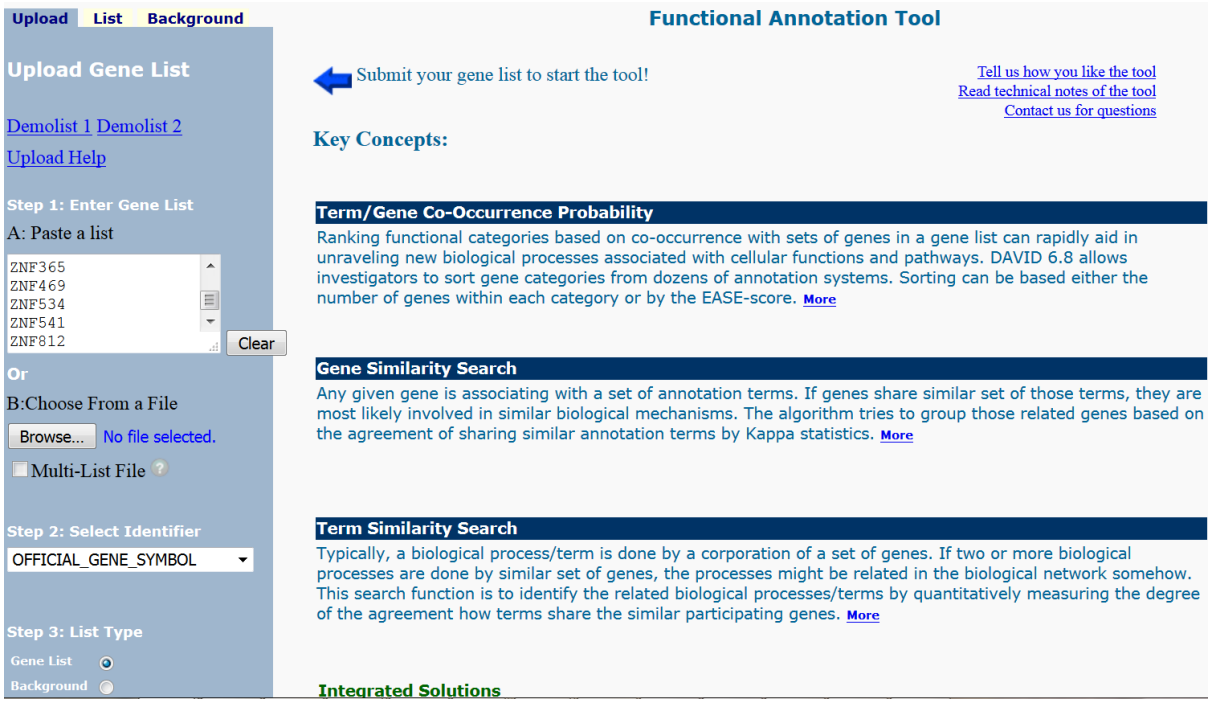
3.1. Please open any Internet browser and go to this web address: https://david.ncifcrf.gov:



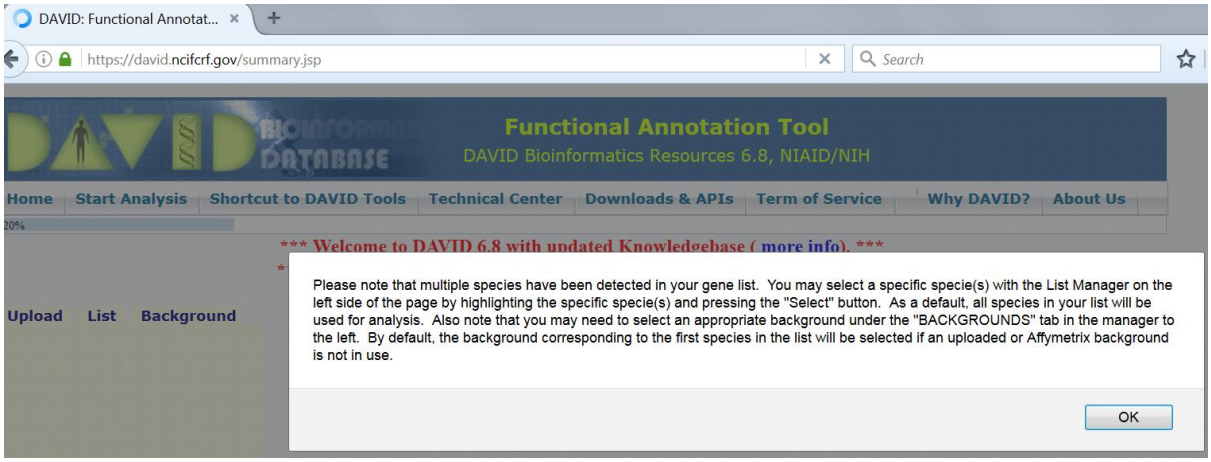
3.2. Select "Functional annotation":



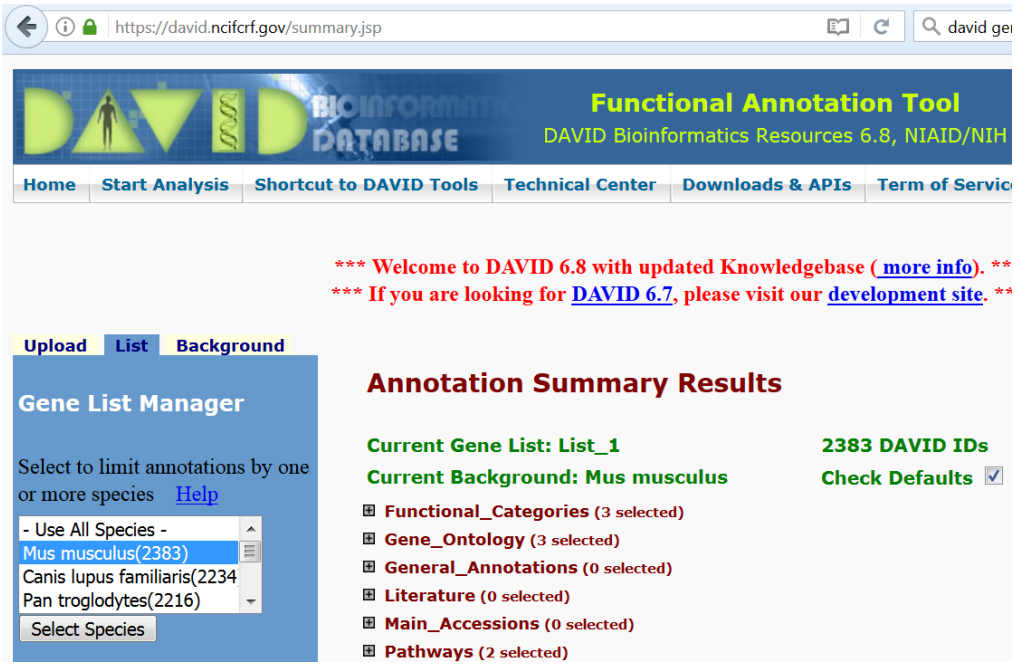
3.3. Select the "upload" link, then under "step 1" paste in the gene list manager your list of genes from the corresponding column in the file peaks_intersect_DEseq.bed opened in Excel. Under "step 2" select "official gene name", and under "step 3" select "gene list":



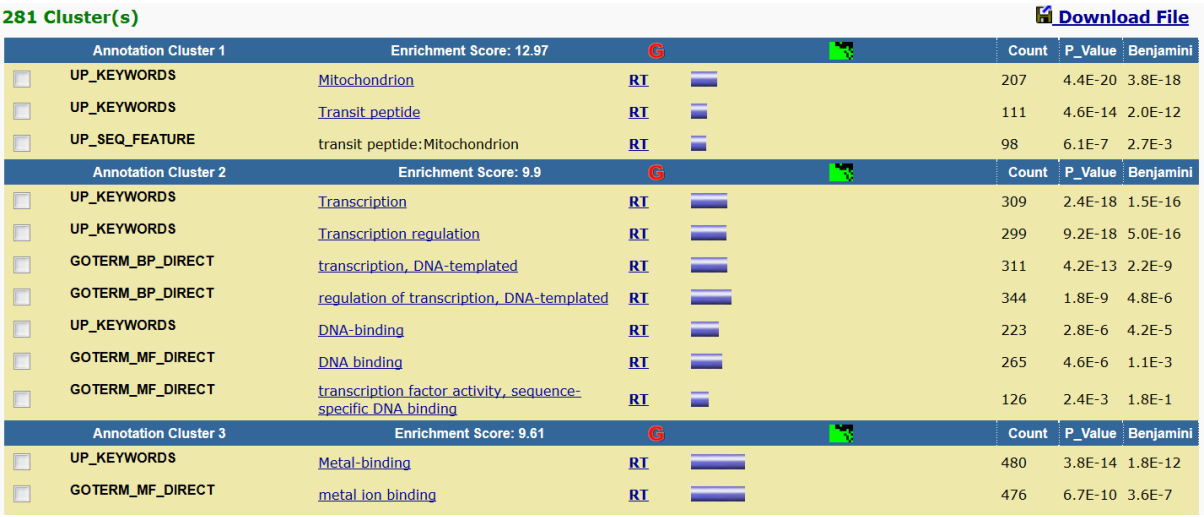
3.4. Under “Step 4” press “submit list”. You will receive the following notification:



3.5. Click “OK”, and then highlight “Mus Musculus” and press button “Select species”:



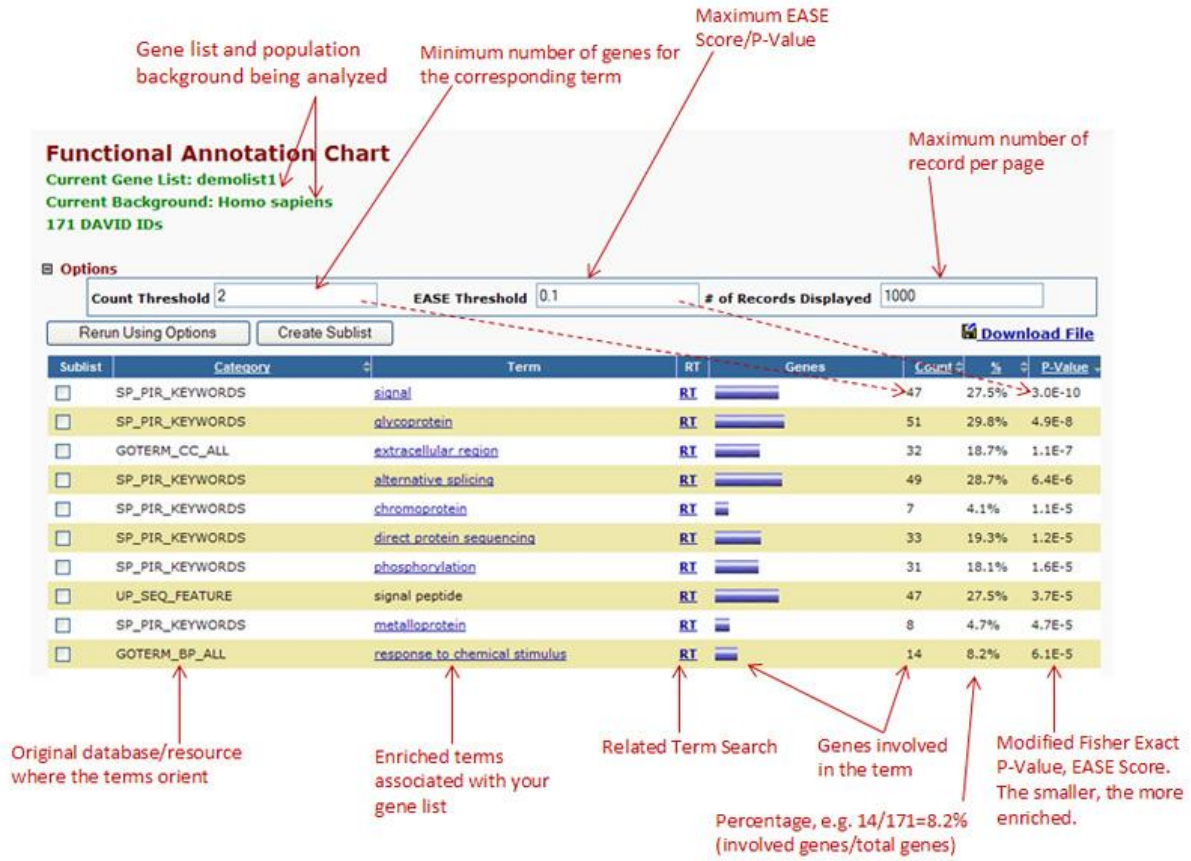
3.6. Then click “Functional annotation clustering”:



281 Cluster(s) — Download File

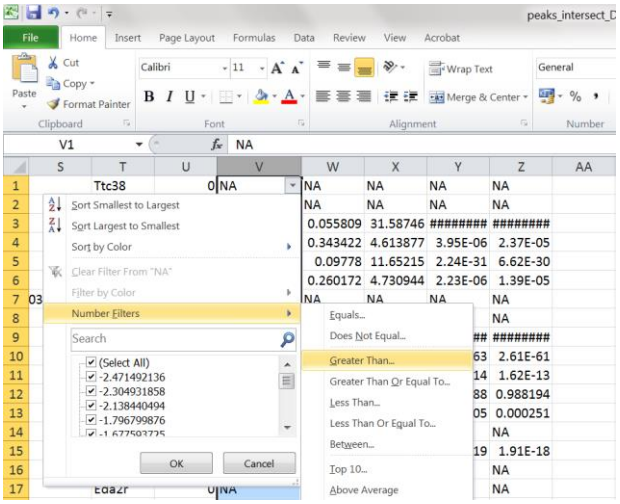
Annotation Cluster 1	Enrichment Score: 12.97				Count	P_Value	Benjamini
UP_KEYWORDS	Mitochondrion	RT			207	4.4E-20	3.8E-18
UP_KEYWORDS	Transit peptide	RT			111	4.6E-14	2.0E-12
UP_SEQ_FEATURE	transit peptide:Mitochondrion	RT			98	6.1E-7	2.7E-3
Annotation Cluster 2	Enrichment Score: 9.9				Count	P_Value	Benjamini
UP_KEYWORDS	Transcription	RT			309	2.4E-18	1.5E-16
UP_KEYWORDS	Transcription regulation	RT			299	9.2E-18	5.0E-16
GOTERM_BP_DIRECT	transcription, DNA-templated	RT			311	4.2E-13	2.2E-9
GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT			344	1.8E-9	4.8E-6
UP_KEYWORDS	DNA-binding	RT			223	2.8E-6	4.2E-5
GOTERM_MF_DIRECT	DNA binding	RT			265	4.6E-6	1.1E-3
GOTERM_MF_DIRECT	transcription factor activity, sequence-specific DNA binding	RT			126	2.4E-3	1.8E-1
Annotation Cluster 3	Enrichment Score: 9.61				Count	P_Value	Benjamini
UP_KEYWORDS	Metal-binding	RT			480	3.8E-14	1.8E-12
GOTERM_MF_DIRECT	metal ion binding	RT			476	6.7E-10	3.6E-7

Understanding DAVID's output:

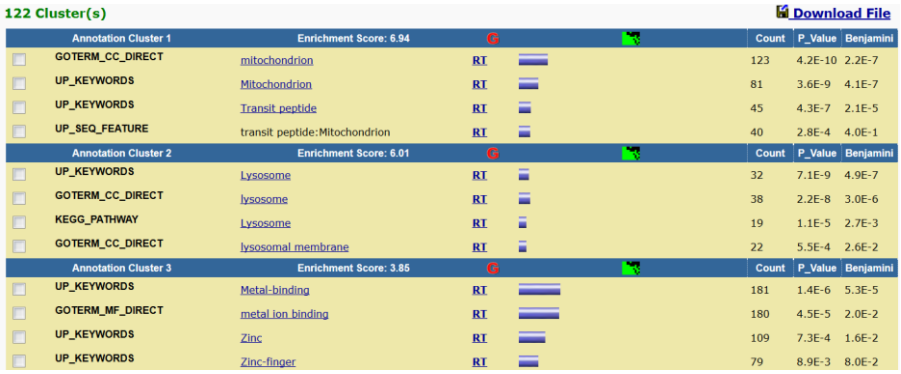


Discuss the results of the DAVID's calculation. What new information did we learn?

3.7. On the previous steps (3.1-3.6) we have analysed all genes that are bound by p53 at their promoters. Now let's narrow down this list. Please go back to the Excel file and select only those genes which have p53 at their promoters and their expression was significantly **up-regulated** upon treatment (log2 fold change >1):



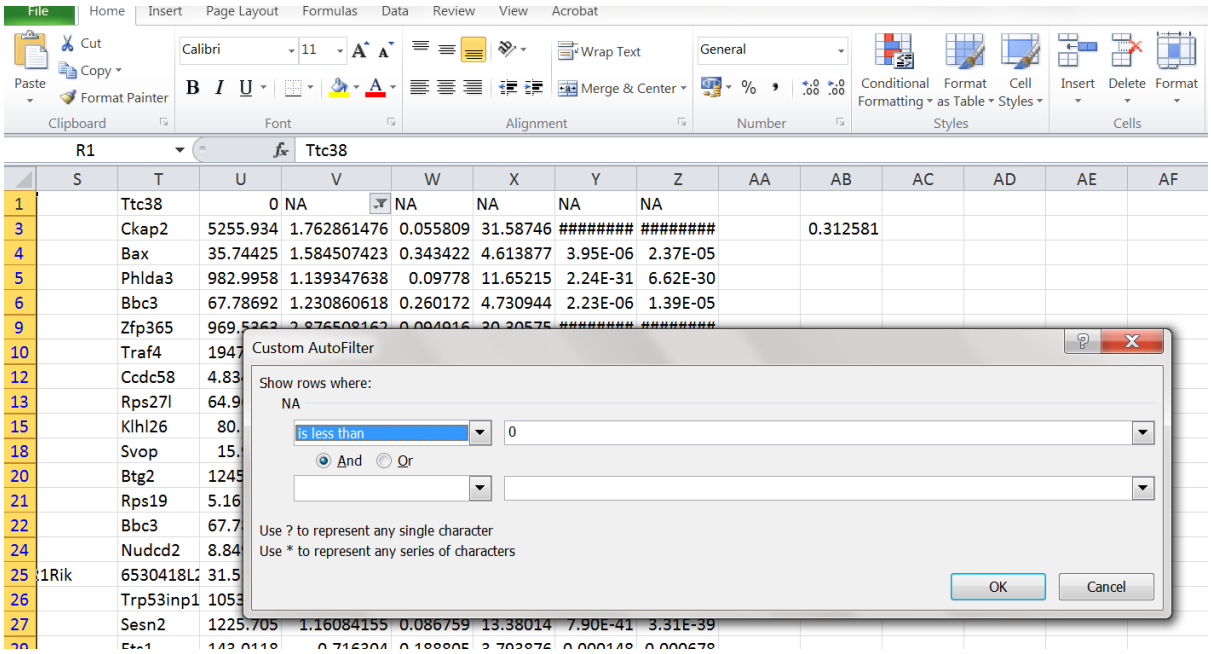
3.8. Now submit them again to DAVID and repeat steps 3.2-3.6 in DAVID as above:



122 Cluster(s) — Download File

Annotation Cluster 1	**Enrichment Score: 6.94**			**Count**	**P_Value**	**Benjamini**
GOTERM_CC_DIRECT	mitochondrion	RT		123	4.2E-10	2.2E-7
UP_KEYWORDS	Mitochondrion	RT		81	3.6E-9	4.1E-7
UP_KEYWORDS	Transit peptide	RT		45	4.3E-7	2.1E-5
UP_SEQ_FEATURE	transit peptide:Mitochondrion	RT		40	2.8E-4	4.0E-1
Annotation Cluster 2	**Enrichment Score: 6.01**			**Count**	**P_Value**	**Benjamini**
UP_KEYWORDS	Lysosome	RT		32	7.1E-9	4.9E-7
GOTERM_CC_DIRECT	lysosome	RT		38	2.2E-8	3.0E-6
KEGG_PATHWAY	Lysosome	RT		19	1.1E-5	2.7E-3
GOTERM_CC_DIRECT	lysosomal membrane	RT		22	5.5E-4	2.6E-2
Annotation Cluster 3	**Enrichment Score: 3.85**			**Count**	**P_Value**	**Benjamini**
UP_KEYWORDS	Metal-binding	RT		181	1.4E-6	5.3E-5
GOTERM_MF_DIRECT	metal ion binding	RT		180	4.5E-5	2.0E-2
UP_KEYWORDS	Zinc	RT		109	7.3E-4	1.6E-2
UP_KEYWORDS	Zinc-finger	RT		79	8.9E-3	8.0E-2

3.9. Now let's do the same type of analysis but only for the genes which contain p53 at their promoters and are **down-regulated** upon treatment (expression log2 fold change <0):



	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1		Ttc38	0	NA	NA	NA	NA	NA						
3		Ckap2	5255.934	1.762861476	0.055809	31.58746	########	########		0.312581				
4		Bax	35.74425	1.584507423	0.343422	4.613877	3.95E-06	2.37E-05						
5		Phlda3	982.9958	1.139347638	0.09778	11.65215	2.24E-31	6.62E-30						
6		Bbc3	67.78692	1.230860618	0.260172	4.730944	2.23E-06	1.39E-05						
9		Zfp365	969.5363	2.876508162	0.094916	30.30575	########	########						
10		Traf4	1947											
12		Ccdc58	4.83											
13		Rps27l	64.9											
15		Klhl26	80.											
18		Svop	15.											
20		Btg2	1245											
21		Rps19	5.16											
22		Bbc3	67.7											
24		Nudcd2	8.84											
25	1Rik	6530418L2	31.5											
26		Trp53inp1	1053											
27		Sesn2	1225.705	1.16084155	0.086759	13.38014	7.90E-41	3.31E-39						

Repeat steps 4.2-4.6 using the set of downregulated genes.

Here is what we get for the downregulated p53-dependent genes:

	Annotation Cluster 1	Enrichment Score: 4.82			Count	P_Value	Benjamini
	UP_KEYWORDS	Cell cycle	RT		43	1.4E-7	7.5E-6
	GOTERM_BP_DIRECT	cell cycle	RT		43	7.0E-7	1.7E-3
	UP_KEYWORDS	Cell division	RT		26	4.8E-5	1.4E-3
	GOTERM_BP_DIRECT	mitotic nuclear division	RT		22	1.1E-4	8.4E-2
	UP_KEYWORDS	Mitosis	RT		20	1.2E-4	3.2E-3
	GOTERM_BP_DIRECT	cell division	RT		26	1.8E-4	7.2E-2
	Annotation Cluster 2	**Enrichment Score: 4.33**			**Count**	**P_Value**	**Benjamini**
	UP_KEYWORDS	Mitochondrion	RT		63	2.4E-8	1.5E-6
	UP_KEYWORDS	Transit peptide	RT		30	2.3E-4	4.3E-3
	UP_SEQ_FEATURE	transit peptide:Mitochondrion	RT		26	1.8E-2	9.7E-1
	Annotation Cluster 3	**Enrichment Score: 3.66**			**Count**	**P_Value**	**Benjamini**
	UP_KEYWORDS	Protein transport	RT		39	9.0E-7	3.1E-5
	GOTERM_BP_DIRECT	protein transport	RT		39	1.0E-5	1.3E-2
	UP_KEYWORDS	Transport	RT		71	9.0E-3	7.5E-2
	GOTERM_BP_DIRECT	transport	RT		70	2.8E-2	7.9E-1
	Annotation Cluster 4	**Enrichment Score: 2.75**			**Count**	**P_Value**	**Benjamini**
	UP_KEYWORDS	Endoplasmic reticulum	RT		49	1.6E-4	3.5E-3
	GOTERM_CC_DIRECT	endoplasmic reticulum	RT		57	4.5E-3	1.1E-1
	GOTERM_CC_DIRECT	endoplasmic reticulum membrane	RT		34	7.8E-3	1.8E-1

We can see that the genes responsible for the cell cycle are downregulated after treatment.

Discuss: What does this mean? Probably, the cells are struggling with doxorubicin-induced DNA damage and cannot enter the cell cycle? Would this be consistent with doxorubicin action leading to cell apoptosis? How is this related to p53 binding?

4) Perform GO analysis with GOrilla

4.1. Go to this web address: http://cbl-gorilla.cs.technion.ac.il

4.2. Select the following options, and then click the button "Search enriched GO terms":

Step 1: Choose organism – Mus musculus

Step 2: Choose running mode ◉ Two unranked lists of genes (target and background lists)

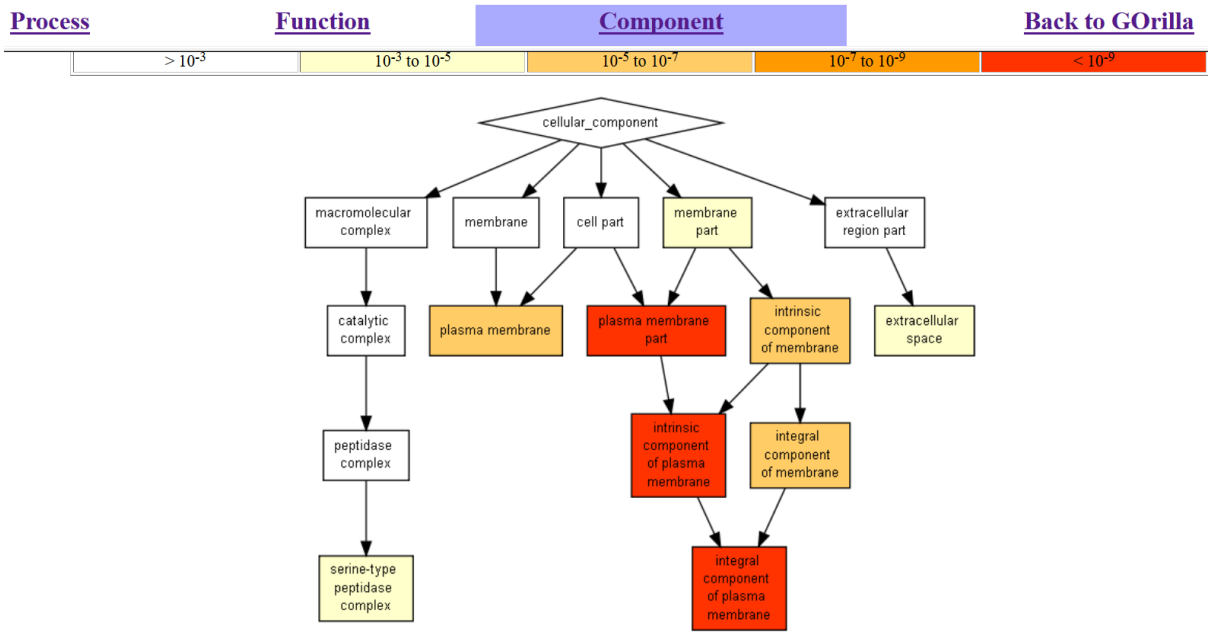
Step 3: Paste a ranked list of gene/protein names

Target set – paste here your list of upregulated genes.

Background set – paste here ALL the gene names of the mouse genome (you can get this list from the following file that needs to be copied from the cluster to your computer: `/storage/projects/BS312/promoters_mm9_52k.bed`)

Step 4: Choose an ontology – ALL

4.3. Study the results calculated by GOrilla. When the figure is larger than the screen, use arrows to see it all. You will obtain figures like this one:



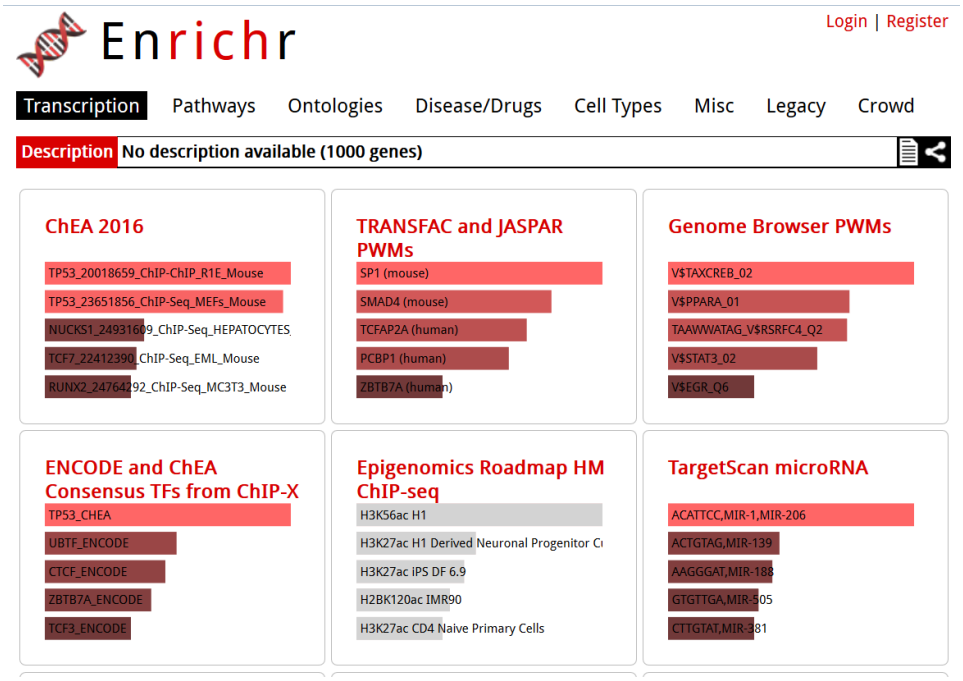
Discuss: what new information did we learn with Gorilla?

5. Perform Gene Ontology enrichment analysis using EnrichR

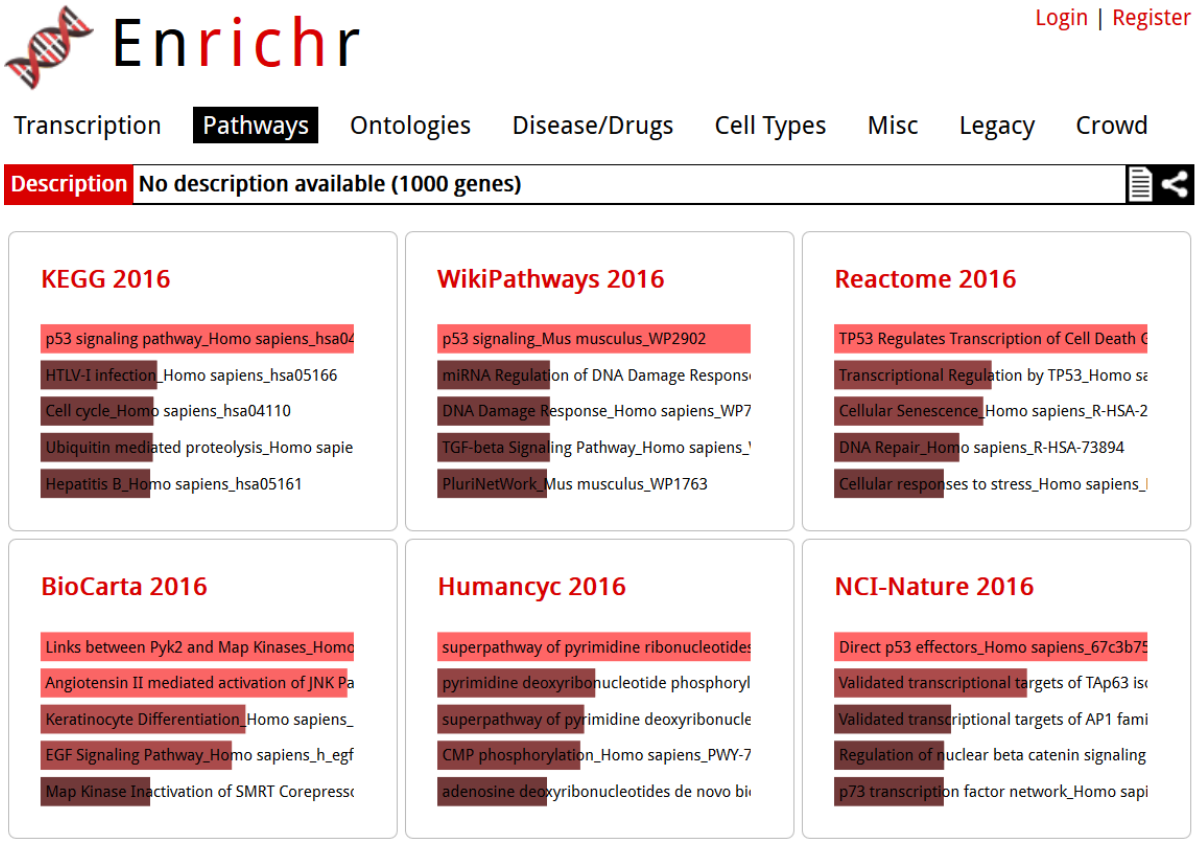
5.1. Open http://amp.pharm.mssm.edu/Enrichr/. Prepare on your computer the BED file with all p53 peaks that you have determined previously (you can download it from Galaxy).

Upload your BED file with all p53 peaks to EnrichR using the "Browse" button; select "mouse mm9", then click "submit":

EnrichR will calculate for you the enrichments of many different genomic features at the regions submitted in your BED file. E.g., this is the "Transcription" panel:



Interestingly, EnrichR finds p53 and p53-related features as top hits. Importantly, EnrichR does not know which experiment we are working on, it only knows the genomic coordinates of the peaks obtained after ChIP-seq. If these peaks look to EnrichR like p53 binding, then this means that our analysis is correct and our peaks indeed represent p53 binding. Convincingly, the "Pathways" panel of EnrichR is almost completely devoted to p53 binding:



Discuss: What new information did you learn with EnrichR?

6. Make a list of TFs co-binding with p53 and construct their interaction network.

Remember our file with DNA sequences of all p53-bound peaks that we submitted for analysis to MEME? Now it's time to check whether your analysis is finished. Here are the results of my MEME analysis: http://meme-suite.org/info/status?service=MEMECHIP&id=appMEMECHIP_5.2.01606168126273-1866210168

6.1. Let's finish the downstream MEME analysis (submit each of the motifs found by MEME to TOMTOM) and collect all TFs whose binding motifs appear to be enriched in p53-bound peaks.

5-minute split in breakout rooms: each group should collect their own list of TFs based on MEME.

Then all groups will meet in the main room and we will see which group collected more TF names.

For example, these are TFs that I have found: TP53, TP73, TP63, EGR2, EGR4, SMAD3, TFAP2A, TFAP2B, TFAP2C, EWSR1, FLI1, GATA4, PBX3, GATA1, E2F4

6.2. Use PathwayCommons visualiser to construct a network of interactions of TFs which you found: https://www.pathwaycommons.org/pcviz

For example, here is the network that I have constructed:

https://www.pathwaycommons.org/pcviz/#pathsbetween/TP53,TP73,TP63,EGR2,EGR4,SMAD3,TFAP2A,TFAP2B,TFAP2C,EWSR1,FLI1,GATA4,PBX3,GATA1,E2F4