

Vladimir Teif ([yteif@essex.ac.uk](mailto:yteif@essex.ac.uk))

An updated version of this document will be available at <http://generegulation.info/index.php/teaching>

## Summary

In this practical we will learn how to perform integrative analysis of ChIP-seq and RNA-seq data. We will continue working 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 the following link: <http://nar.oxfordjournals.org/content/43/9/4447.long>. In the previous sessions we have determined the locations of bound p53 and quantified changes of gene expression following activation of DNA repair by doxorubicin. Here we will use the file with p53 peaks from the first practical and the file with RNA-seq quantification by DEseq2 from the second practical and will learn how to combine these.

## Converting DEseq output to a BED-compatible format

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% ☺

Today we have been doing format conversions already; now let us focus on this task a bit more. First of all, let us look at the output of DEseq2 software, summarised in the file `DEseq2_results.txt` that we have downloaded from Galaxy and copied to the directory `/storage/projects/proficio/ChIPseq`. It contains differential gene expression data:

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, 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 `peaks_formatted.bed` 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	0.85
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	0.72
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	0.65

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?

I have one hypothesis. I 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	Mett14
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	Mett14
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 DEseq 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. File [promoters\\_mm9\\_52k.bed](#) contains one promoter per gene, in total about 52 thousand genes.

## Adding promoter coordinates to the differential gene expression data

Let us now take the file with differential gene expression (`DEseq2_results.txt`) and the file with promoters (`promoters_mm9_52k.bed`) and combine them in such a way so that for each gene we would have both the coordinate of its promoter (taken from the file `promoters_mm9_52k.bed`) and the values of its expression fold change (taken from file `DEseq2_results.txt`). The genes in these two files are sorted differently, therefore the script that is doing this has to read each of these large files and pair the corresponding lines from these two files which contain the same gene name. We have recently developed a software package called NucTools (<https://homeveg.github.io/nuctools>), which includes a Perl script `merge2tabs.pl` that is doing exactly this. I have copied the script `merge2tabs.pl` to the directory `/storage/projects/proficio/ChIPseq`. The command calling script `merge2tabs.pl` in order to merge files `promoters_mm9_52k.bed` and `DEseq2_results.txt` is written in your bash file `Task_6_DEseq_to_BED_format.sh` as follows:

```
perl /storage/projects/proficio/ChIPseq/merge2tabs.pl --
table1=/storage/projects/proficio/ChIPseq/promoters_mm9_52k.bed --
table2=/storage/projects/proficio/ChIPseq/DEseq2_results.txt --
output=promoters_and_DEseq.bed --colID_tab1=3 --colID_tab2=0
```

All you need for this step is just to execute the bash file `Task_6_DEseq_to_BED_format.sh`:

```
qsub Task_6_DEseq_to_BED_format.sh
```

After the execution of this bash file you will get a new file named `promoters_and_DEseq.bed`.

The execution of `Task_6_DEseq_to_BED_format.sh` takes about 20 minutes, please wait.

## Intersecting p53 peaks with promoter-based RNA-seq data

After we 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	Mett14	Mett14	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 let us intersect p53 sites with the promoters linked to their corresponding gene expression data from RNA-seq. All we need to do is to intersect regions in the file `peaks_formatted.bed` and `promoters_and_DEseq.bed`. This is realised for you in the next bash file `Task_7_intersect_DEseq.sh`:

```
intersectBed -a
/storage/projects/proficio/ChIPseq/peaks_formatted.bed -b
/storage/projects/proficio/ChIPseq/promoters_and_DEseq.bed -wb >
peaks_intersect_DEseq.bed
```

All we need to do now is just to execute bash file `Task_7_intersect_DEseq.sh`.

```
qsub Task_6_DEseq_to_BED_format.sh
```

The results of this calculation are stored in a new file called `peaks_intersect_DEseq.bed`.

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. Let us copy this file to our local computer using WinSCP, and then open it in Excel:

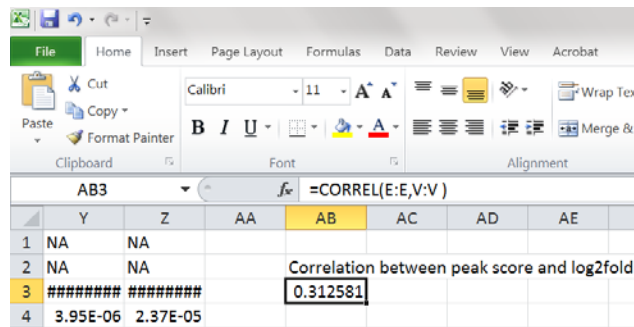
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	chr15	85690303	85690440	+	428.6	0.895	231	473.5	0.5	947.02	0.00E+00	82.45	0.00E+00	0.71	chr15	85688440	85690440	Ttc38
2	chr15	85690303	85690479	+	428.6	0.895	231	473.5	0.5	947.02	0.00E+00	82.45	0.00E+00	0.71	chr15	85689175	85691175	Gtse1
3	chr8	23296397	23296573	+	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:

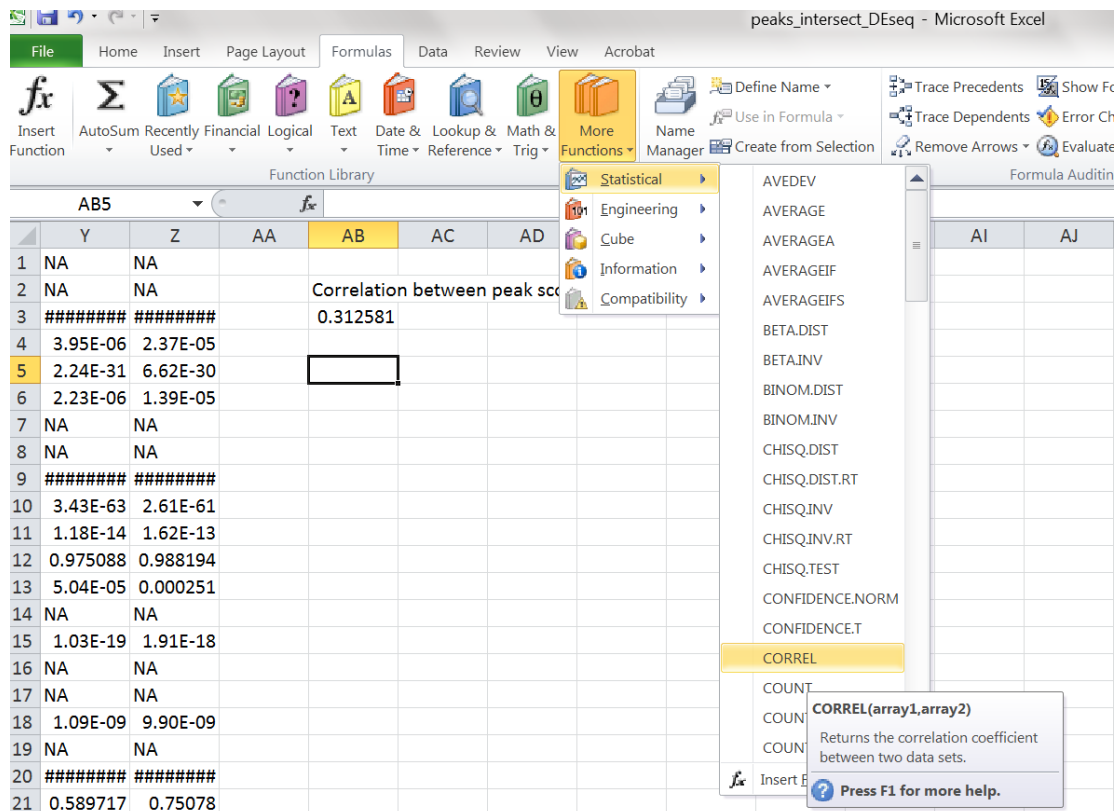
	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	0.5	947.02	0.00E+00	82.45	0.00E+00	0.71	chr15	85688440	85690440	Ttc38		Ttc38	0	NA	NA	NA	NA	NA
2	0.5	947.02	0.00E+00	82.45	0.00E+00	0.71	chr15	85689175	85691175	Gtse1		Gtse1	0	NA	NA	NA	NA	NA
3	7	55.04	0.00E+00	137.97	0.00E+00	0.75	chr8	23295245	23297245	Ckap2		Ckap2	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	90306170	90306170	90306170	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 test now is 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 some empty cell and ask Excel to calculate in this cell the correlation between columns “E” and “V”:



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



So, in my case the correlation is 0.31. Is it the same for you? ☺

What can we say about this correlation? It is a moderate but statistically significant correlation. It tells us that those genes which contain the strongest p53 binding are characterised by the largest changes of gene expression when p53 pathways are induced due to antibiotic treatment. Did you expect to find it? Well, at least this is something non-trivial, and this is a publishable scientific result that we were able to derive from the integrative analysis of ChIP-seq and RNA-seq data.

Did the authors of the paper also notice this result? Yes they did. Interestingly, they came to a similar conclusion using another type of analysis of the same data. Now let us do some other analysis.



## Calculating average aggregate occupancy profiles

If you remember the first overview lecture, there were a lot of plots with average profiles of protein binding around some genomic features. We have calculated in the first ChIP-seq practical p53 occupancy profiles chromosome-wide using HOMER (remember the HOMER tag directories?) Now we can use these to calculate average aggregate occupancy profiles of p53 around some genomic features. Say, we already know that p53 is enriched at promoters and enhancers, let us calculate p53 aggregate occupancy profiles around promoters and enhancers. This is realized for you in bash file `Task_8_average_profiles.sh`:

```
#!/bin/bash
#$ -cwd
#$ -q all.q
#$ -S /bin/bash

cd ~

#Calculate average profile of p53 binding around transcription start
sites (TSS):
annotatePeaks.pl tss mm9 -size 2000 -hist 10 -d HOMER_p53 >
profile_p53_around_TSS.txt

#Calculate average profile of p53 binding around transcriptional
enhancers:
annotatePeaks.pl /storage/projects/proficio/ChIPseq/enhancers_mm9.bed
mm9 -size 2000 -hist 10 -d HOMER_p53 >
profile_p53_around_enhancers.txt

#Calculate average A/T/C/G frequencies around bound p53
annotatePeaks.pl
/storage/projects/proficio/ChIPseq/peaks_formatted.bed mm9 -size 2000
-hist 10 -CpG > profile_CpG_around_p53.txt
```

Here the first command calculates p53 occupancy around transcription start sites (notice the parameter “tss” in the HOMER command line below:

```
annotatePeaks.pl tss mm9 -size 2000 -hist 10 -d HOMER_p53 >
profile_p53_around_TSS.txt
```

Other parameters in this command tell HOMER that the size of the region around TSS should be 2000 base pairs, and that the calculation should be performed with the step 10 base pairs, and the occupancy data should be taken from the tag directory named “HOMER\_p53”, while the results should be placed in the output file named `profile_p53_around_TSS.txt`.

The second command in a similar way tells HOMER to calculate p53 occupancy profiles around enhancers:

```
annotatePeaks.pl /storage/projects/proficio/ChIPseq/enhancers_mm9.bed
mm9 -size 2000 -hist 10 -d HOMER_p53 >
profile_p53_around_enhancers.txt
```

The only difference here is that we need to tell HOMER where to take the enhancer coordinates (/storage/projects/proficio/ChIPseq/enhancers\_mm9.bed).

The third command tells HOMER that we want to calculate the frequencies of A/T/C/G nucleotides around p53 binding sites. This is a useful function in the case if we are testing a hypothesis that our protein of interest binds e.g. primarily inside CpG islands:

```
annotatePeaks.pl
/storage/projects/proficio/ChIPseq/peaks_formatted.bed mm9 -size 2000
-hist 10 -CpG > profile_CpG_around_p53.txt
```

Now we just need to execute the bash file:

```
qsub Task_8_average_profiles.sh
```

This calculation takes just about 5 minutes.

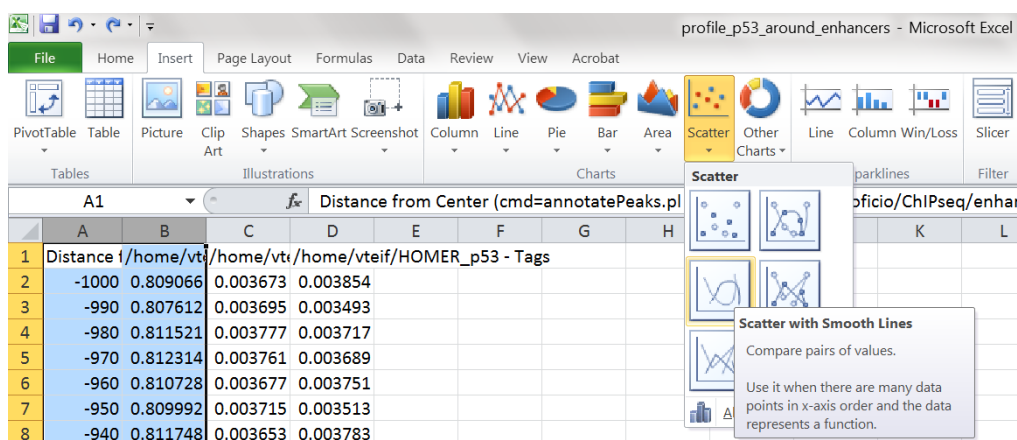
After the calculation is finished, we can locate in our home directory the files named

```
profile_CpG_around_p53.txt
profile_p53_around_enhancers.txt
profile_p53_around_TSS.txt
```

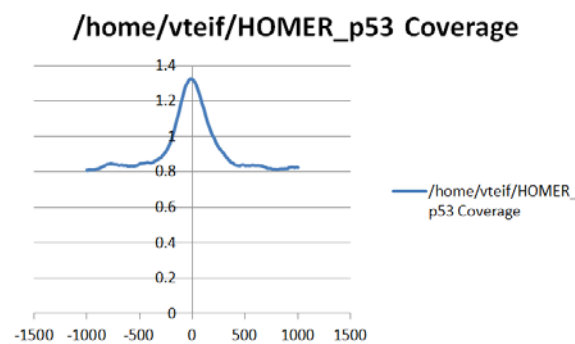
and copy them using WinSCP to our local computer. Then we can open them in Excel:

	A	B	C	D	E	F
1	Distance t/home/vt/home/vt/home/vteif/HOMER_p53 - Tags					
2	-1000	0.809066	0.003673	0.003854		
3	-990	0.807612	0.003695	0.003493		
4	-980	0.811521	0.003777	0.003717		
5	-970	0.812314	0.003761	0.003689		
6	-960	0.810728	0.003677	0.003751		
7	-950	0.809992	0.003715	0.003513		
8	-940	0.811748	0.003653	0.003783		

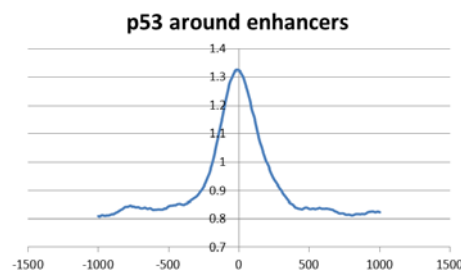
The first column shows the distance from the centre of the feature (e.g. the centre of enhancer). The second column shows the average p53 occupancy at this distance. Then we can make the plot based on these two columns:



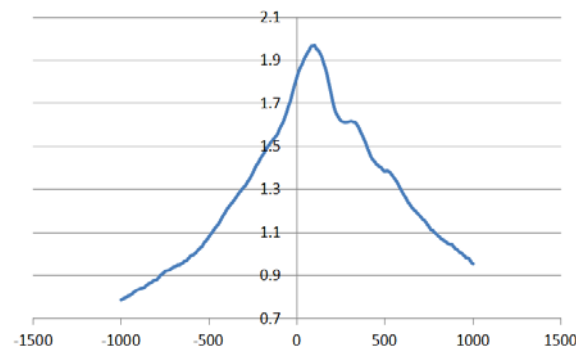
For example, this is how my plot for the average p52 occupancy around enhancers looks:



Of course it is possible to adjust in Excel the scale and correct captions to make it a nicer figure:



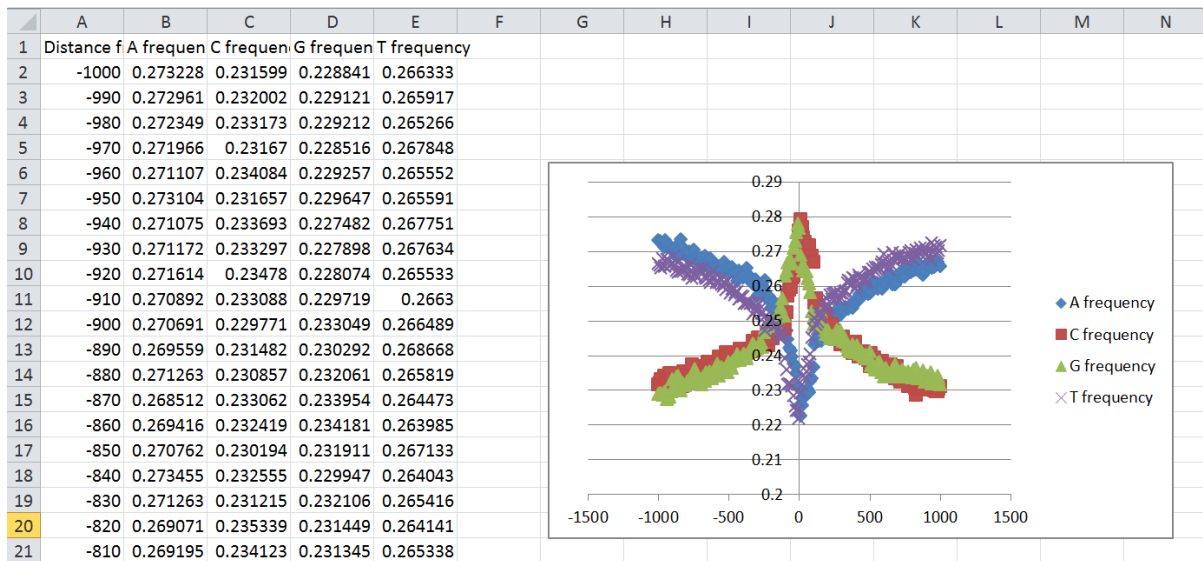
Similarly, we can plot the average aggregate profile of p53 occupancy around transcription start sites:



Notice, that unlike p53 around enhancers, the profile of p53 around enhancers is asymmetric. Should it be expected to be asymmetric? Yes, because genes have directionality, and so do their transcription start sites. But did you expect to see the p53 peak mostly shifted downstream of TSS? Usually it is assumed that the promoter is more upstream than downstream of TSS. In this case it appears that p53 binds more at the part of the promoter downstream of TSS. Is it a scientific discovery? Hmm... may be. Or maybe this is something already well known. Oh, wait, we actually have to read a lot of paper before making scientific claims 😊



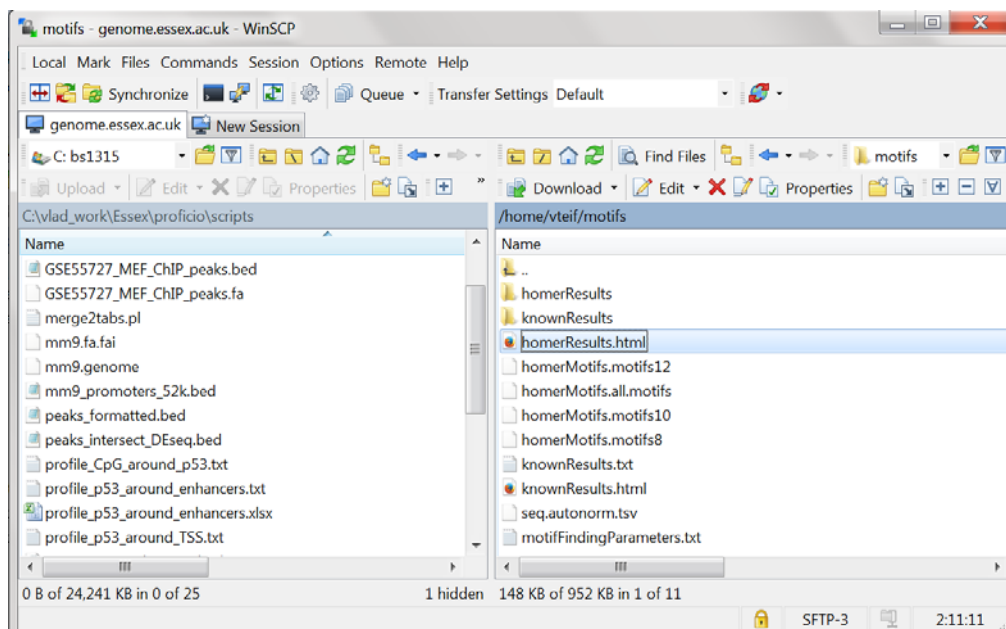
Similarly, we can plot in Excel the nucleotide frequencies around p53 sites:



But what about the fine sequence patterns of p53 binding sites? We will learn them on the next step.

## DNA sequence motif analysis

Remember when we did peak calling in the first ChIP-seq practical we also asked HOMER to calculate for us the DNA sequence motifs corresponding to the bound p53 peaks? Let us now look at these data. Open in WinSCP the directory /motifs inside your home directory:



Now right-click on the file homerResults.html and select “open”. The HTML file will open in your browser:

## Homer *de novo* Motif Results (/home/vteif/motifs/)

[Known Motif Enrichment Results](#)

[Gene Ontology Enrichment Results](#)





If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into [STAMP](#)

More information on motif finding results: [HOMER](#) | [Description of Results](#) | [Tips](#)

Total target sequences = 15377

Total background sequences = 33092

\* - possible false positive

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-3525	-8.118e+03	42.90%	6.51%	42.5bp (68.8bp)	p53(p53)/mES-cMyc-ChIP-Seq(GSE11431)/Homer(0.725) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
2		1e-1059	-2.439e+03	13.08%	1.73%	42.6bp (62.5bp)	ZNF416(Zf)/HEK293-ZNF416.GFP-ChIP-Seq(GSE58341)/Homer(0.870) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
3		1e-411	-9.474e+02	14.75%	5.30%	51.1bp (63.8bp)	Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer(0.991) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>
4		1e-220	-5.072e+02	0.83%	0.01%	47.9bp (23.9bp)	ZFX(Zf)/mES-Zfx-ChIP-Seq(GSE11431)/Homer(0.755) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>	<a href="#">motif file (matrix)</a>

HOMER has found a number of motifs, ranked them based on the P values, and associated found motifs with known transcription factors from its database. The top hit is p53. Surprise, surprise ☺

Remember, HOMER actually did not know what type of experiment was performed. It did not know that it was ChIP-seq with antibody against p53. The only information it had was the DNA sequence motifs most frequently found in the peaks determined based on this experiment. Based on these DNA sequence motifs, HOMER decided that the best matching transcription factor is p53. Bingo!

As the links in this HTML file suggest, you can click on them and get more information. We will let you to play around this file for several minutes independently...

## Gene Ontology (GO) analysis

The last type of integrative analysis that we will learn is the easiest to do and also quite a fun thing. This type of analysis is based on the classification of molecular processes, pathways, and types of molecules into a number of scientific terms such as “apoptosis”, “differentiation”, “cell cycle”, etc. 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). OK, let’s just do it. As the introductory lecture suggested, there are many different online tools to perform GO analysis.

1.9. Let us perform GO analysis for genes which contain bound p53 at their promoters using software DAVID. Go to the internet and open this address: <https://david.ncifcrf.gov>:



File Edit View History Bookmarks Tools Help

DAVID Functional Annotati... x

https://david.ncifcrf.gov

DAVID Bioinformatics Resources 6.8  
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

\*\*\* Welcome to DAVID 6.8 with updated Knowledgebase ([more info](#)). \*\*\*  
\*\*\* If you are looking for DAVID 6.7, please visit our [development site](#). \*\*\*

Recommend: A [paper](#) published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Shortcut to DAVID Tools

Functional Annotation  
Gene annotation enrichment analysis, functional annotation clustering, BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification  
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. [More](#)

Welcome to DAVID 6.8

2003 - 2016

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a

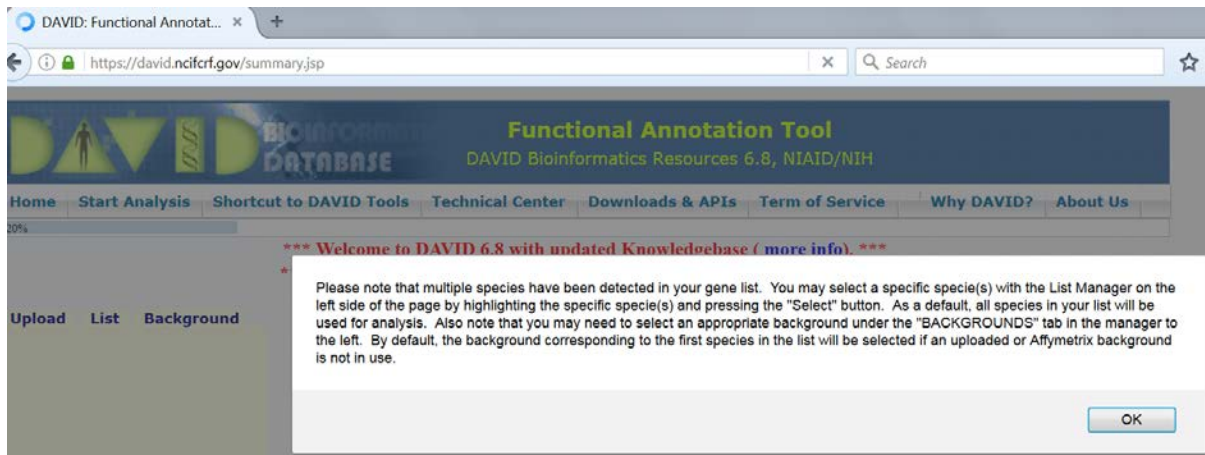
What's Important in DAVID?

- New requirement to cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina

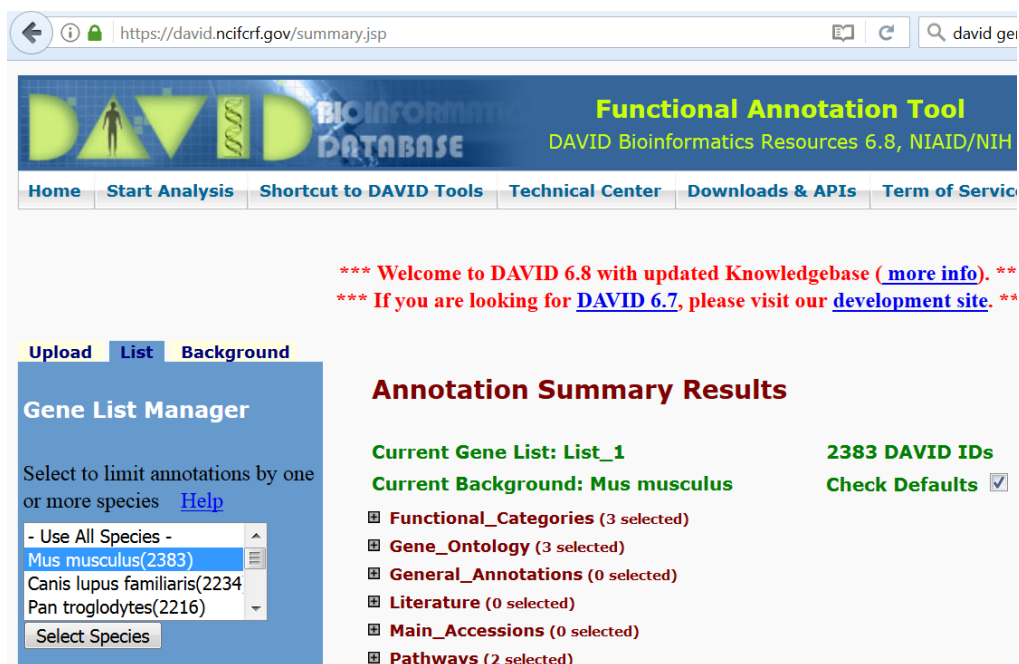
Select “Functional annotation”:

Select the “upload” link, then under “step 1” paste your list of genes from Excel in the gene list manager, under “step 2” select “official gene name”, and under “step 3” select “gene list”:

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



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



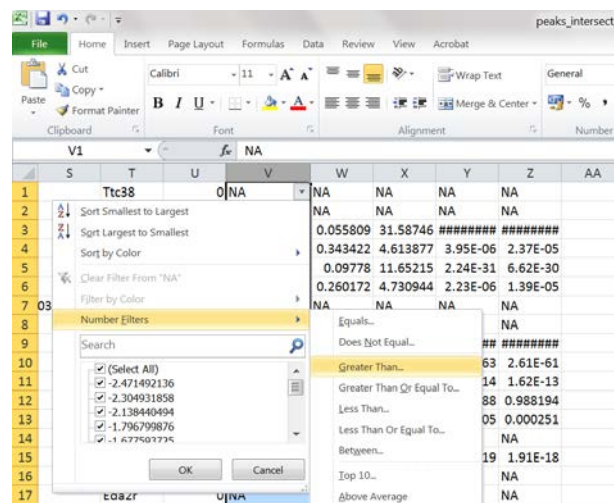
Then click “Functional annotation clustering”:

**281 Cluster(s)** [Download File](#)

Annotation Cluster	Enrichment Score	Count	P_Value	Benjamini
<b>Annotation Cluster 1</b>	<b>Enrichment Score: 12.97</b>			
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
<b>Annotation Cluster 2</b>	<b>Enrichment Score: 9.9</b>			
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
<b>Annotation Cluster 3</b>	<b>Enrichment Score: 9.61</b>			
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



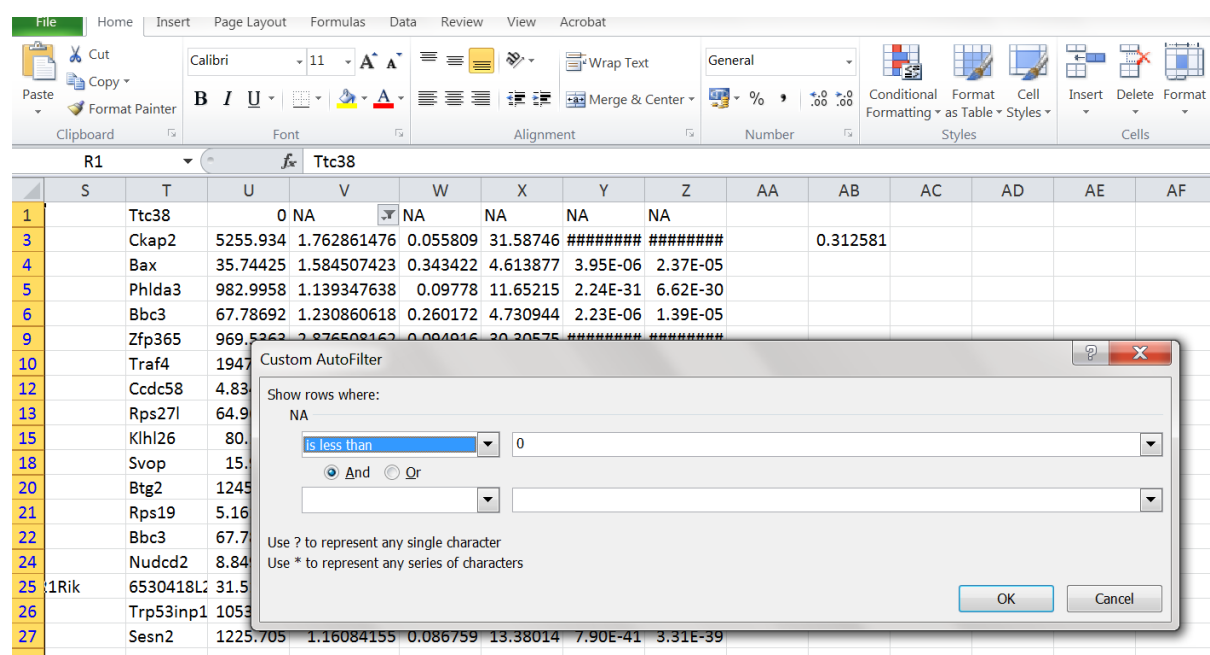
Now let's go back to the Excel file and select only those genes which have p53 at their promoters and their expression was upregulated upon treatment (log2 fold change >0):



Now submit them again to DAVID and do all the steps in DAVID as above:

122 Cluster(s)				Download File		
Annotation Cluster 1	Enrichment Score: 6.94			Count	P-Value	Benjamini
<input type="checkbox"/> GOTERM_CC_DIRECT	mitochondrion	RI		123	4.2E-10	2.2E-7
<input type="checkbox"/> UP_KEYWORDS	Mitochondrion	RI		81	3.6E-9	4.1E-7
<input type="checkbox"/> UP_KEYWORDS	Transit peptide	RI		45	4.3E-7	2.1E-5
<input type="checkbox"/> UP_SEQ_FEATURE	transit peptide:Mitochondrion	RI		40	2.8E-4	4.0E-1
Annotation Cluster 2	Enrichment Score: 6.01			Count	P-Value	Benjamini
<input type="checkbox"/> UP_KEYWORDS	Lysosome	RI		32	7.1E-9	4.9E-7
<input type="checkbox"/> GOTERM_CC_DIRECT	lysosome	RI		38	2.2E-8	3.0E-6
<input type="checkbox"/> KEGG_PATHWAY	Lysosome	RI		19	1.1E-5	2.7E-3
<input type="checkbox"/> GOTERM_CC_DIRECT	lysosomal membrane	RI		22	5.5E-4	2.6E-2
Annotation Cluster 3	Enrichment Score: 3.85			Count	P-Value	Benjamini
<input type="checkbox"/> UP_KEYWORDS	Metal-binding	RI		181	1.4E-6	5.3E-5
<input type="checkbox"/> GOTERM_MF_DIRECT	metal ion binding	RI		180	4.5E-5	2.0E-2
<input type="checkbox"/> UP_KEYWORDS	Zinc	RI		109	7.3E-4	1.6E-2
<input type="checkbox"/> UP_KEYWORDS	Zinc-finger	RI		79	8.9E-3	8.0E-2

Now let's repeat this only for the genes which contain p53 at their promoters and are downregulated upon treatment:




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

Annotation Cluster 1		Enrichment Score: 4.82			Count	P_Value	Benjamini
<input type="checkbox"/>	UP_KEYWORDS	Cell cycle	RT		43	1.4E-7	7.5E-6
<input type="checkbox"/>	GOTERM_BP_DIRECT	cell cycle	RT		43	7.0E-7	1.7E-3
<input type="checkbox"/>	UP_KEYWORDS	Cell division	RT		26	4.8E-5	1.4E-3
<input type="checkbox"/>	GOTERM_BP_DIRECT	mitotic nuclear division	RT		22	1.1E-4	8.4E-2
<input type="checkbox"/>	UP_KEYWORDS	Mitosis	RT		20	1.2E-4	3.2E-3
<input type="checkbox"/>	GOTERM_BP_DIRECT	cell division	RT		26	1.8E-4	7.2E-2
Annotation Cluster 2		Enrichment Score: 4.33			Count	P_Value	Benjamini
<input type="checkbox"/>	UP_KEYWORDS	Mitochondrion	RT		63	2.4E-8	1.5E-6
<input type="checkbox"/>	UP_KEYWORDS	Transit peptide	RT		30	2.3E-4	4.3E-3
<input type="checkbox"/>	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
<input type="checkbox"/>	UP_KEYWORDS	Protein transport	RT		39	9.0E-7	3.1E-5
<input type="checkbox"/>	GOTERM_BP_DIRECT	protein transport	RT		39	1.0E-5	1.3E-2
<input type="checkbox"/>	UP_KEYWORDS	Transport	RT		71	9.0E-3	7.5E-2
<input type="checkbox"/>	GOTERM_BP_DIRECT	transport	RT		70	2.8E-2	7.9E-1
Annotation Cluster 4		Enrichment Score: 2.75			Count	P_Value	Benjamini
<input type="checkbox"/>	UP_KEYWORDS	Endoplasmic reticulum	RT		49	1.6E-4	3.5E-3
<input type="checkbox"/>	GOTERM_CC_DIRECT	endoplasmic reticulum	RT		57	4.5E-3	1.1E-1
<input type="checkbox"/>	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. What does this mean? It means that the cells are struggling with doxorubicin-induced DNA damage and cannot enter the cell cycle. This is quite consistent with doxorubicin action leading to cell apoptosis.

A similar GO analysis can be also performed in another GO software called GOrilla

<http://cbl-gorilla.cs.technion.ac.il>



**GORILLA**  
*Gene Ontology enRichment anaLysis and visualiZation tool*

GORilla is a tool for identifying and visualizing enriched GO terms in ranked lists of genes. It can be run in one of two modes:

1. Searching for enriched GO terms that appear densely at the top of a ranked list of genes or
2. Searching for enriched GO terms in a target list of genes compared to a background list of genes.

For further details see [References](#).

[Running example](#)
[Usage instructions](#)
[GORilla News\(Updated March 8th 2013\)](#)
[References](#)
[Contact](#)

**Step 1: Choose organism**

Homo sapiens

**Step 2: Choose running mode**

☒ Single ranked list of genes ☐ Two unranked lists of genes (target and background lists)


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

Names should be separated by an <ENTER>. The preferred format is gene symbol. Other supported formats are: gene and protein RefSeq, Uniprot, Unigene



## Gene Ontology enrichment analysis using EnrichR

Open <http://amp.pharm.mssm.edu/Enrichr/>. Then upload your BED file with all p53 peaks (`peaks_formatted.bed`) using the “Browse” button, select “mouse mm9”, then click “submit”:

 **Enrichr** Login | Register  
5,576,889 lists analyzed

Analyze | What's New? | Libraries | Find a Gene | About | Help

### Input data

Choose an input file to upload. Either in BED format or a list of genes. For a quantitative set, add a comma and the level of membership of that gene. The membership level is a number between 0.0 and 1.0 to represent a weight for each gene, where the weight of 0.0 will completely discard the gene from the enrichment analysis and the weight of 1.0 is the maximum.

Try an example [BED file](#).

peaks\_formatted.bed

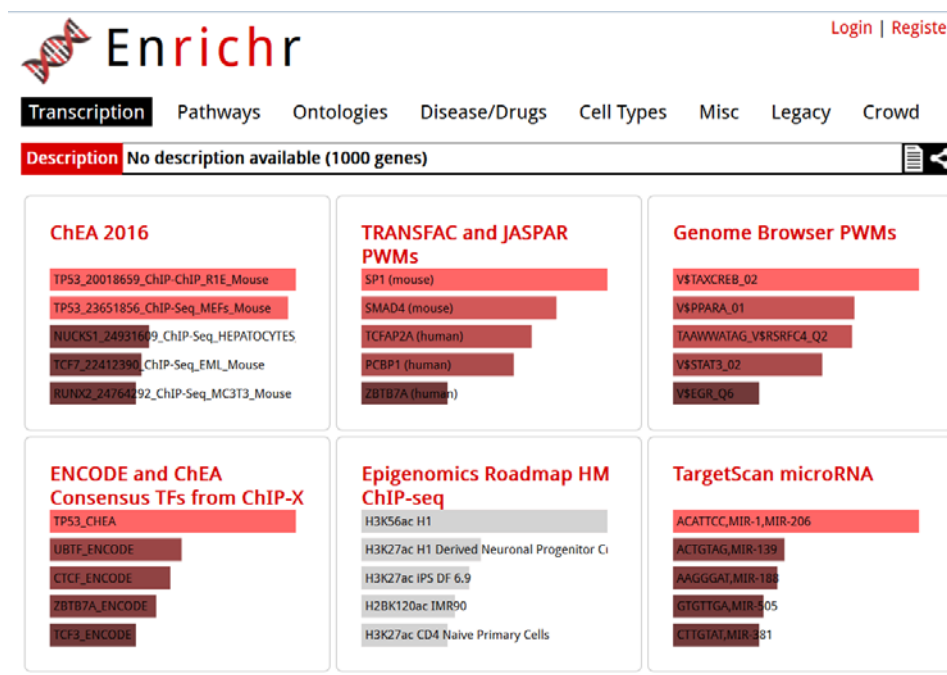
Select parameters for bed file to gene list conversion.

Species:

Max number of genes:

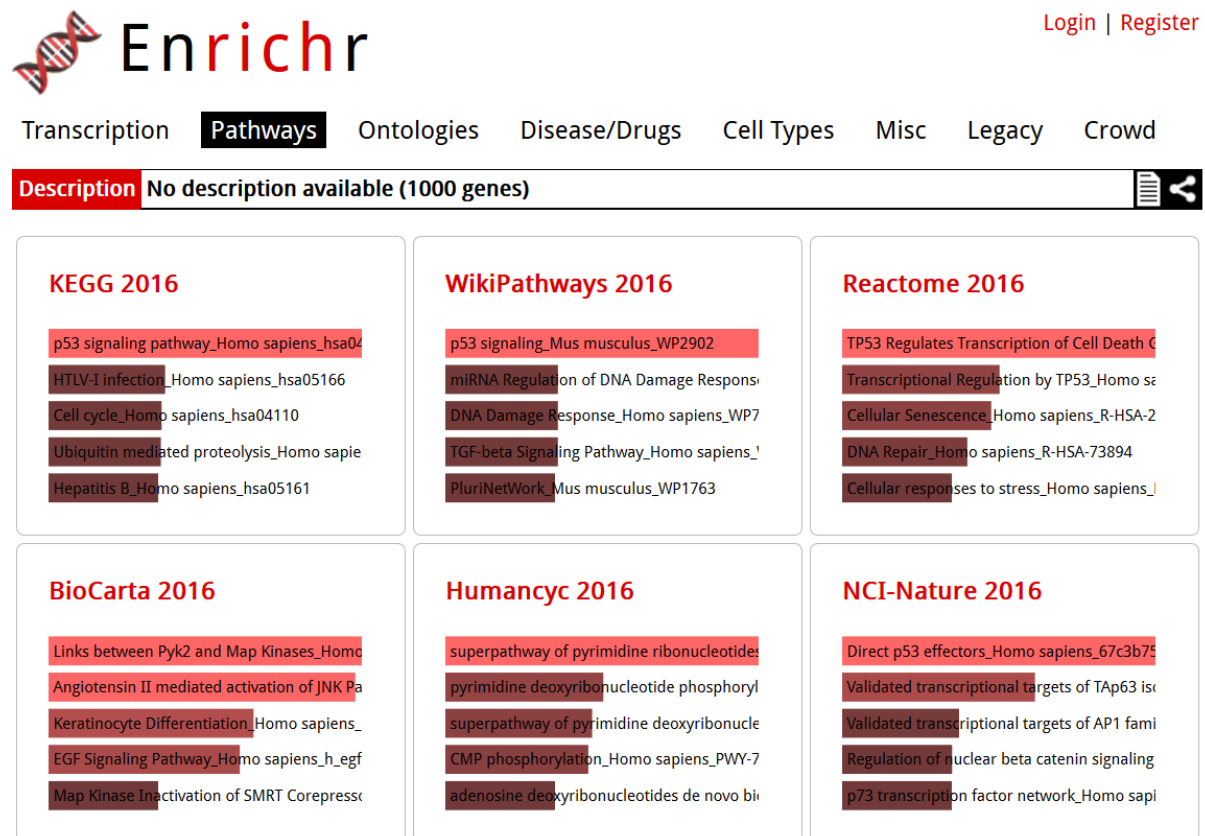
0 gene(s) entered

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 that I’ve got:



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:



**Enrichr** [Login](#) | [Register](#)

Transcription **Pathways** Ontologies Disease/Drugs Cell Types Misc Legacy Crowd

**Description** No description available (1000 genes)

KEGG 2016	WikiPathways 2016	Reactome 2016
p53 signaling pathway_Homo sapiens_hsa04110	p53 signaling_Mus musculus_WP2902	TP53 Regulates Transcription of Cell Death C
HTLV-I infection_Homo sapiens_hsa05166	miRNA Regulation of DNA Damage Respons	Transcriptional Regulation by TP53_Homo se
Cell cycle_Homo sapiens_hsa04110	DNA Damage Response_Homo sapiens_WP7	Cellular Senescence_Homo sapiens_R-HSA-2
Ubiquitin mediated proteolysis_Homo sapie	TGF-beta Signaling Pathway_Homo sapiens_	DNA Repair_Homo sapiens_R-HSA-73894
Hepatitis B_Homo sapiens_hsa05161	PluriNetWork_Mus musculus_WP1763	Cellular responses to stress_Homo sapiens_

BioCarta 2016	Humancyc 2016	NCI-Nature 2016
Links between Pyk2 and Map Kinases_Homo	superpathway of pyrimidine ribonucleotides	Direct p53 effectors_Homo sapiens_67c3b75
Angiotensin II mediated activation of JNK Pa	pyrimidine deoxyribonucleotide phosphoryl	Validated transcriptional targets of TP63 is
Keratinocyte Differentiation_Homo sapiens_	superpathway of pyrimidine deoxyribonucle	Validated transcriptional targets of AP1 fami
EGF Signaling Pathway_Homo sapiens_h_egf	CMP phosphorylation_Homo sapiens_PWY-7	Regulation of nuclear beta catenin signaling
Map Kinase Inactivation of SMRT Corepress	adenosine deoxyribonucleotides de novo bi	p73 transcription factor network_Homo sapi

We can let you play with EnrichR a bit more on your own. This is the end of the practical.

If time remains at the end, please feel free to suggest for discussion your own directions for integrative analysis, or ask the lecturers how to do the analysis for your experimental system.

We hope you enjoyed the course!