

# BS312: Practical 6: ChIP-seq, motif analysis, DNA methylation

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

*In this practical session we will continue the p53 binding story started last week, adding to the analysis a new layer of gene regulation through DNA methylation. The P53 ChIP-seq and DNA methylation data will be used together. We will need the files created during our previous practical.*

**Summary of the previous practical.** Our previous practical was based 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 determine genome-wide p53 binding profiles in human and mouse cells. Their main finding is that p53 binding occurs predominantly within transcriptional enhancers. Last week we have 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. We have found that p53 binding sites are enriched at enhancers as judged by the comparison with a random dataset of genomic regions of the same size as p53 peaks.

## Plan for this practical:

**Task 1.** [For those who did not close their interactive sessions after the last practical]. Close previous interactive sessions if you did not do so yet.

**Task 2.** Copy files required for this practical to your home directory and understand these data.

**2\*** [Optional task for advanced students] Visualise these files in the UCSC Genome Browser

**Task 3.** Start HOMER’s script in interactive mode to find DNA sequence motifs for p53 peaks

**3\*** [Variation of Task 3 for advanced students] Instead submit this job using bash file.

**Task 4.** Perform Fisher’s test for enrichment of p53 sites in enhancers

**4\*** [Optional task for advanced students] Do the same also for promoters

**Task 5.** Calculate the average p53 binding profile around CpG islands using HOMER

**Task 6.** Visualise p53 binding profiles around CpG islands in Excel

**Tasks 7-9.** Calculate and visualise the average p53 binding profile around LMR, UMR, FMR

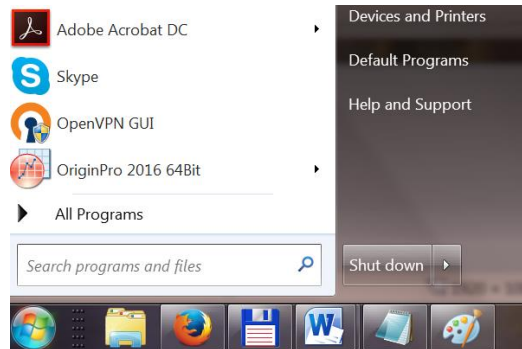
**Task 10.** Prepare HOMER tag directory for reduced RRBS DNA methylation in MEFs

**Task 11.** Calculate the average profile of 5mCs around p53 sites in MEFs

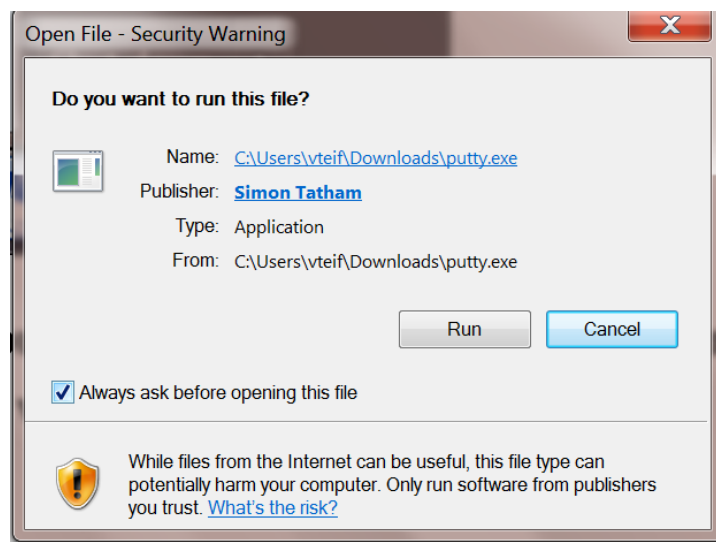
**Task 12.** Analyse results of the calculation of p53 motifs (which should be ready by this time)

**Task 13\*** [for advanced students] Calculate and visualise the average p53 binding profiles around promoters and enhancers

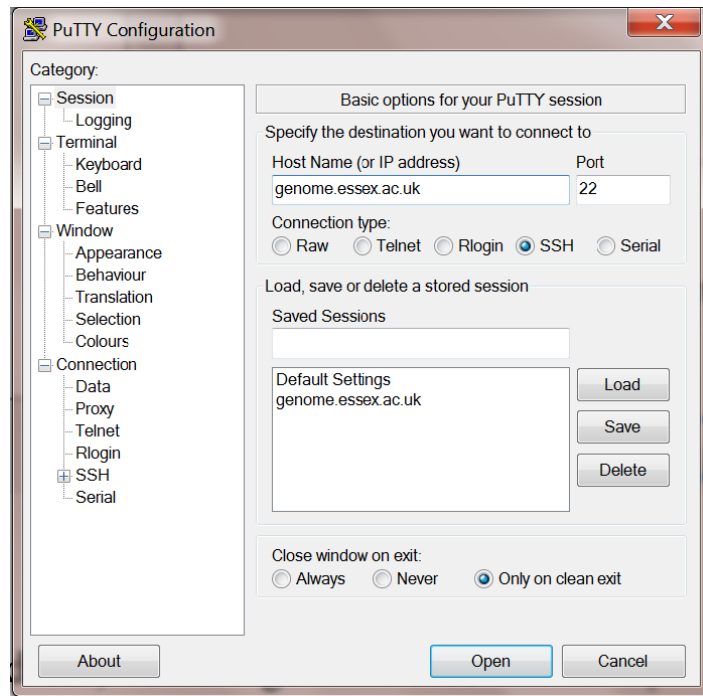
**A reminder how to connect to the computer cluster using Putty.** Our calculations deal with large files, and therefore have to be performed on the computer cluster. This is exactly how most serious sequencing analysis is being performed nowadays. Firstly, we need to connect from your computers to the cluster. We will do this using a program called **Putty**. A detailed description of this program can be found here: <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>. Let's open the "Start" menu of your Windows computers and type "Putty" in the "Search programs and files" field:



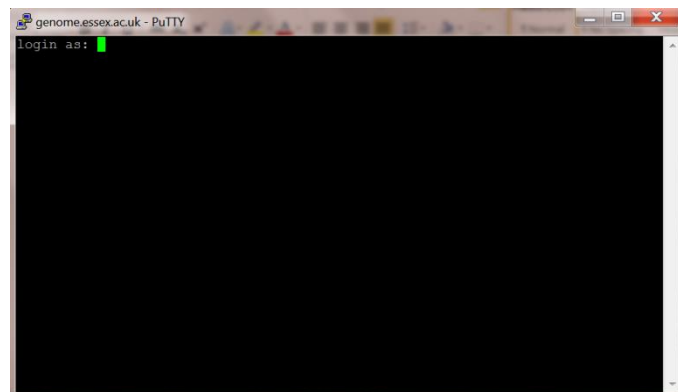
If Putty already exists on your computer, it will show up in the search results, and after clicking on this program it will open the following window:



Click "Run" and then agree to add the security key when it asks for it. Then Putty opens like this:



In the field “host name (or IP address)” instead of “genome.essex.ac.uk” as shown in the picture enter “ceres.essex.ac.uk”. Then click “Open”. It will open the black terminal screen:



Enter your university user name (the same as you use for your email), and press Enter. Then enter your university password and again click Enter. Note that when you are typing your password the cursor will not move on the screen, but this is fine, the computer is reading what you are writing. If you have correctly entered your password the welcome screen appears:

```

vteif@genome:~
Password:
Last login: Wed Dec 7 17:59:48 2016 from isslx029.essex.ac.uk
Rocks 6.2 (SideWinder)
Profile built 14:18 16-Nov-2015

Kickstarted 14:35 16-Nov-2015

=====
WELCOME TO THE SCHOOL OF BIOLOGICAL SCIENCES GENOME CLUSTER
=====

Users are advised that none of the file systems on the genome cluster
are backed up. Users are responsible for their own backup.

Users are also reminded that the command node (ie the server you have just logge
d in to)
is not intended for computation use. Non-root processes will be terminated autom
atically after
5 mins. If you need to run something outside a batch job, use qcrsh to start an i
nteractive
session on one of the compute nodes.

=====
Cluster guide website can be found on http://genomics.essex.ac.uk/cluster
=====
[vteif@genome ~]$

```

You are now located in your home directory on the cluster. Today we will be working in the “interactive mode”, that is, everything typed in the terminal will be executed immediately as we type. To switch to the interactive mode type the following command:

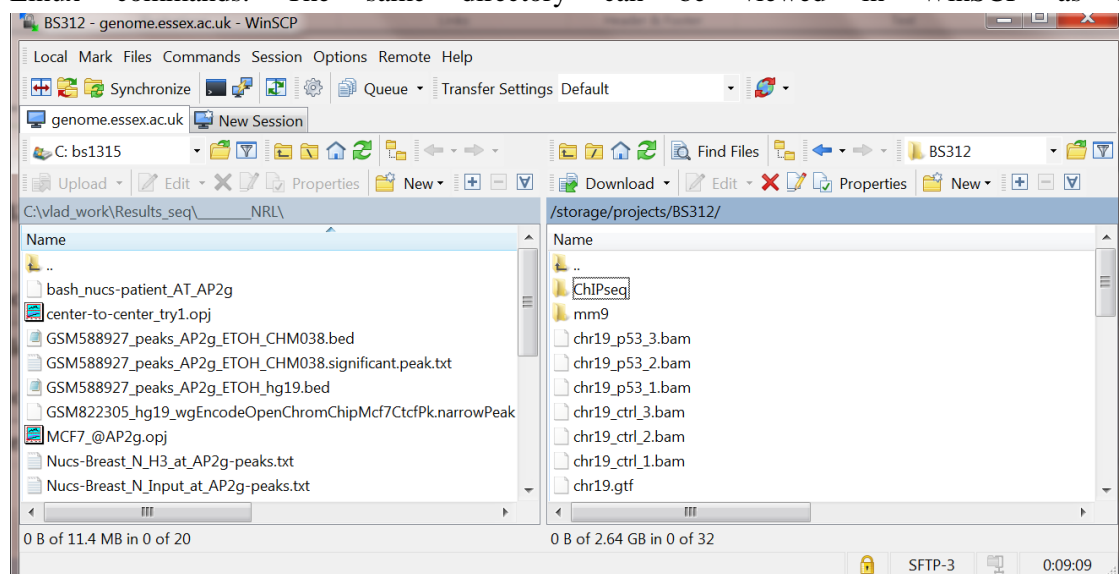
`qcrsh`

We have now entered the interactive mode.

A simple command “`ls`” will show you the content of the current directory. Just type “`ls`” and press [Enter]. We will be doing all the calculations in the home directory. The module materials including the data that we need for the analysis are stored in another directory which is situated at the following path: `/storage/projects/BS312/ChIPseq`. In order to go to any directory we can type the command `cd` followed by the path to the desired deirectory.

In order to go back to your home directory you can type the command “`cd`” without any parameters, or alternatively “`cd ~`”. In both cases it will bring you home.

**Connecting to the computer cluster using WinSCP.** As you have seen previously, there is also user-friendly software called WinSCP that allows to manage files on the cluster without typing any Linux commands. The same directory can be viewed in WinSCP as follows:



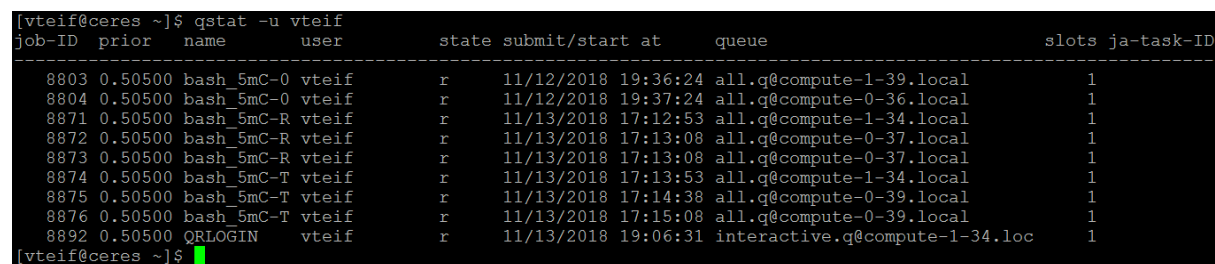
Here the right panel in WinSCP shows the directory on the cluster, and the left panel shows the directory on your local computer. You can copy files between the cluster and your computer by just dragging them by mouse. You can also view the content of the files by double-clicking on them. The files will then open in a text editor, where you can view and edit them. This can be only done for small files. Please do not attempt to do this for large files, as their opening on your computer can take ages. Large files can be viewed in Putty using the command “`less`”.

### Task 1. [For those who did not close their interactive sessions after the last practical].

**Close previous interactive sessions if you did not do so yet.**

Last time we have opened many interactive sessions and some of you forgot to close them. Our computer cluster can have only a limited number of interactive sessions. Therefore, first let's close interactive sessions that remained open from the previous practical(s). To see all your running sessions, type the following command by substituting the word “username” with your user name:

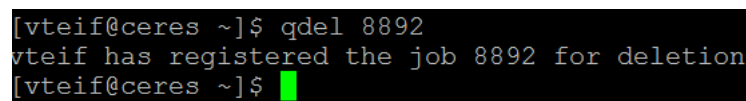
```
qstat -u username
```



job-ID	prior	name	user	state	submit/start at	queue	slots	ja-task-ID
8803	0.50500	bash_5mC-0	vteif	r	11/12/2018 19:36:24	all.q@compute-1-39.local	1	
8804	0.50500	bash_5mC-0	vteif	r	11/12/2018 19:37:24	all.q@compute-0-36.local	1	
8871	0.50500	bash_5mC-R	vteif	r	11/13/2018 17:12:53	all.q@compute-1-34.local	1	
8872	0.50500	bash_5mC-R	vteif	r	11/13/2018 17:13:08	all.q@compute-0-37.local	1	
8873	0.50500	bash_5mC-R	vteif	r	11/13/2018 17:13:08	all.q@compute-0-37.local	1	
8874	0.50500	bash_5mC-T	vteif	r	11/13/2018 17:13:53	all.q@compute-1-34.local	1	
8875	0.50500	bash_5mC-T	vteif	r	11/13/2018 17:14:38	all.q@compute-0-39.local	1	
8876	0.50500	bash_5mC-T	vteif	r	11/13/2018 17:15:08	all.q@compute-0-39.local	1	
8892	0.50500	QLOGIN	vteif	r	11/13/2018 19:06:31	interactive.q@compute-1-34.loc	1	

For example, the picture above shows that I have 10 active jobs running on the cluster. Now delete all your running jobs that you do not need. Use the job number from the left column. For example, I want to delete the job number 8892. I type the following command:

```
qdel 8892
```



```
[vteif@ceres ~]$ qdel 8892
vteif has registered the job 8892 for deletion
[vteif@ceres ~]$
```

### Task 2. Copy files required for this practical to your home directory and understand these data

Those who are no registered officially for BS312 will be provided with the USB key from which you copy files to your home directory on the cluster using the WinSCP file manager.

Those who are officially registered to this module BS312 have access to the directory on the cluster in which all the files are located. We will now copy these files to your home directory using command “`cp`” and will view them using command “`less`”:

Bisulfite sequencing in mouse embryonic stem cells (ESCs):

```
cp /storage/st10d/BS312/5mC/GSE30202_BisSeq_ES_CpGmeth.bed ~/
less GSE30202_BisSeq_ES_CpGmeth.bed
```

**Reduced Representation Bisulfite Sequencing (RRBS) in MEFs:**

```
cp /storage/st10d/BS312/5mC/Messner2008_5mC_MEF_meth.bed ~/
less Messner2008_5mC_MEF_meth.bed
```

**What is common and what is different between these two files?**

Now let's look at the coordinates of CpG islands in the mouse genome

```
cp /storage/st10d/BS312/5mC/CpGislands_mm9.bed ~/
less /storage/st10d/BS312/5mC/CpGislands_mm9.bed
```

**Coordinates of Fully Methylated Regions (FMR) in mouse embryonic stem cells**

```
cp /storage/st10d/BS312/5mC/Stadtler_ESC_FMR_corrected_v3.bed ~/
```

**Coordinates of Low Methylated Regions (LMR) in mouse embryonic stem cells**

```
cp /storage/st10d/BS312/5mC/Stadtler_ESC_LMR_corrected_v3.bed ~/
```

**Coordinates of Unmethylated Regions (UMR) in mouse embryonic stem cells**

```
cp /storage/st10d/BS312/5mC/Stadtler_ESC_UMR_corrected_v3.bed ~/
```

**You will also need the files with the numbers of nucleotides in each mouse chromosome:**

```
cp /storage/st10d/BS312/5mC/mm9.genome ~/
cp /storage/st10d/BS312/5mC/mm9.genome.sorted ~/
```

**Finally, let's copy the files with coordinates of promoters and enhancers:**

```
cp /storage/st10d/BS312/ChIPseq/promoters_mm9.bed ~/
cp /storage/st10d/BS312/ChIPseq/enhancers_mm9.bed ~/
```

In addition to these files, those who have been to the previous practical already have in their home directory the files with p53 peaks that we obtained in the previous practical: `GSE55727_MEF_ChIP_peaks.bed` (supplied by the authors of the paper Young et al.) and `peaks_formatted.bed` (p53 peaks that we determined ourselves).

**2\* [Optional task for advanced students]**

Visualise bed files with p53 peaks, CpG islands and low- and unmethylated regions in the UCSC Genome Browser (see instructions of our previous practical about how to create custom tracks in the UCSC Genome Browser).

Is there overlapping between p53 sites and CpG islands? Is there overlapping between CpG islands and unmethylated regions? Why not all CpG islands are unmethylated?

**Task 3. Start HOMER's script in interactive mode to find DNA sequence motifs for p53 peaks**

As we discussed during the lecture, there are two possible ways how DNA methylation can affect transcription factor (TF) binding: (a) directly, in the case if DNA binding motif recognised by a given TF contains a CpG that can be methylated, and (b) indirectly, in the case of DNA methylation affects binding of other proteins or overall chromatin structure that in turn affects the binding of TF of interest. To begin with, we need to determine the DNA sequence motif recognised by or TF. During the last practical we have determined p53 binding locations (peaks) based on ChIP-seq in mouse embryonic fibroblasts (MEFs). Now we will use this file with the peak coordinates to determine the DNA sequence features characteristic for p53 binding sites.

**3\* [Variation of Task 1 for advanced students]** Instead of finding p53 binding motifs in the interactive mode, we can start it using the bash file. There are many advantages of the job submission using bash files:

- 1) You do not need to keep open interactive sessions;
- 2) You can use more resources, and can submit more time-consuming jobs.
- 3) If many users submit jobs their jobs will be automatically handled by the queuing system.
- 4) You can create a bash file with several commands to be executed in a certain order.

The bash file has been already created for you, named "`bash_BS312_practical_6.sh`"

To run this bash file you just need to type the following command:

```
qsub bash_BS312_practical_6.sh
```

To check whether the job has been submitted to the cluster you can type the following:

```
qstat -u username
```

Obviously, instead of the word "username" you need to substitute your real user name.

If you have submitted the task to find motifs using bash file you can skip to task 2 now.

If you did not do the variation of the task 1 for advanced students explained above then let's do the following. In Putty switch to the interactive mode:

qrsh

Then run the following HOMER command:

```
findMotifsGenome.pl GSE55727_MEF_ChIP_peaks.bed mm9  
motifs_p53_GSE55727 -preparedDir ./Prepared
```

This calculation will take about 40 minutes. Meanwhile we can let it running and try something else. Let's open a new Putty window, and switch to the interactive mode (qrsh).

[Please do not open more than two Putty windows to not overload the computer cluster!]

#### Task 4. Perform Fisher's test for enrichment of p53 sites in promoters and enhancers

Remember on the last practical we have calculated the fold enrichment of p53 sites in regulatory regions by preparing random datasets, one random dataset per student and some students have asked whether there is a more systematic and statistically powered way of doing this? Now we will explore one possibility of doing this. The BedTools software suite contains a script called "fisher", which performs the Fisher exact test of statistical significance. It is doing so by calculating the fold enrichment in comparison with that expected by chance (as we did last week), and provides a P-value for this. The P value determines the statistical significance of the result (the smaller P value the better). The details about this function are available here:

<http://bedtools.readthedocs.io/en/latest/content/tools/fisher.html>

The general structure of the fisher command is like this:

```
bedtools fisher -a a.bed -b b.bed -g t.genome
```

The "fisher" command works only with sorted BED files, so we have to sort our files:

```
sort -k1,1 -k2,2n peaks_formatted.bed > peaks_formatted_sorted.bed  
  
sort -k1,1 -k2,2n GSE55727_MEF_ChIP_peaks.bed >  
GSE55727_MEF_ChIP_peaks_sorted.bed
```

The command "sort -k1,1 -k2,2n" sorts the BED file by chromosome number, and within each chromosome from the smallest region to the largest. Now after we have sorted our BED files we can apply the fisher command. Let's start with the Fisher test for the enrichment of our p53 peaks at enhancers:

```
bedtools fisher -a peaks_formatted_sorted.bed -b  
/storage/projects/BS312/ChIPseq/enhancers_mm9.bed -g  
/storage/projects/BS312/ChIPseq/mm9.genome.sorted
```

Here is what we will get as a result:



```
# Number of query intervals: 15377
# Number of db intervals: 44459
# Number of overlaps: 1075
# Number of possible intervals (estimated): 6048897
# phyper(1075 - 1, 15377, 6048897 - 15377, 44459, lower.tail=F)
# Contingency Table Of Counts
#
#      | in -b | not in -b |
# in -a | 1075  | 14302     |
# not in -a | 43384 | 5990136   |
#
# p-values for fisher's exact test
left   right   two-tail   ratio
0      1       0         10.378
```

In this case, the two-tail P value is 0 (aka “very small”), which means that the result is statistically significant, and the enrichment of our p53 peaks at enhancers is 10.378-fold.

Now we can repeat the same for p53 peaks obtained by the authors of the manuscript:

```
bedtools fisher -a GSE55727_MEF_ChIP_peaks_sorted.bed -b
/storage/projects/BS312/ChIPseq/enhancers_mm9.bed -g
/storage/projects/BS312/ChIPseq/mm9.genome.sorted
```

Here is the result that we get:

```
# Number of query intervals: 3100
# Number of db intervals: 44459
# Number of overlaps: 380
# Number of possible intervals (estimated): 3706962
# phyper(380 - 1, 3100, 3706962 - 3100, 44459, lower.tail=F)
# Contingency Table Of Counts
#
#      | in -b | not in -b |
# in -a | 380   | 2720      |
# not in -a | 44079 | 3659783   |
#
# p-values for fisher's exact test
left   right   two-tail   ratio
1      2.6176e-246  2.6176e-246  11.599
```

What is similar and what is different in the results of the Fisher test obtained with our p53 peaks versus the p53 peaks reported by the authors of the paper?

#### 4\* [Optional task for advanced students]

Calculate Fisher’s test for p53 enrichment at promoters. You need to substitute the file name of enhancers used in the previous command with the file name with promoters:

```
bedtools fisher -a GSE55727_MEF_ChIP_peaks_sorted.bed -b
/storage/projects/BS312/ChIPseq/promoters_mm9.bed -g
/storage/projects/BS312/ChIPseq/mm9.genome.sorted
```

Is p53 also enriched at promoters? Where it is more enriched, at promoters or at enhancers?

## Task 5. Calculate the average p53 binding profile around CpG islands using HOMER

If you remember On our previous practical we have calculated genome-wide occupancy landscapes of p53 binding using HOMER (also known as the HOMER tag directories). Now let us use these to perform a new type of analysis. One of the typical types of ChIP-seq analysis is the calculation of the average (also called aggregate) occupancy profiles of the ChIP-seq signal some genomic features. Let's practice this type of analysis using our new data. Let us start with CpG islands.

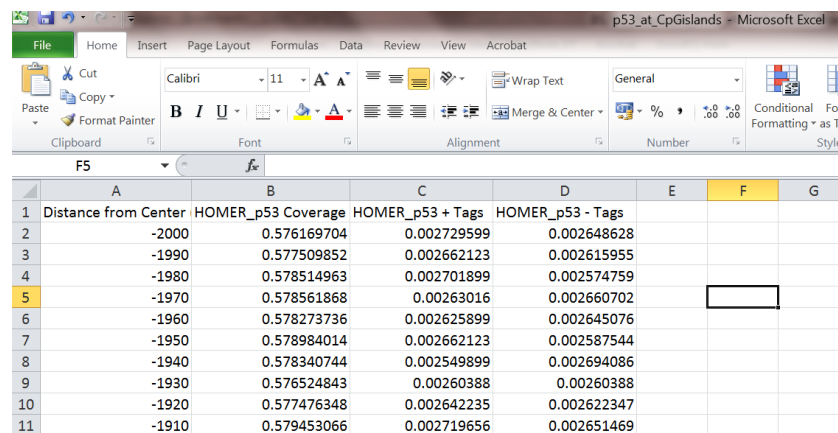
Read the HOMER's web site description to understand how this command works and what each parameter means in the command below: <http://homer.ucsd.edu/homer/ngs/quantification.html>

Now run this HOMER's command to calculate the average p53 profile at CpG islands:

```
annotatePeaks.pl CpGislands_mm9.bed mm9 -size 4000 -hist 10 -d  
HOMER_p53 > p53_at_CpGislands.txt
```

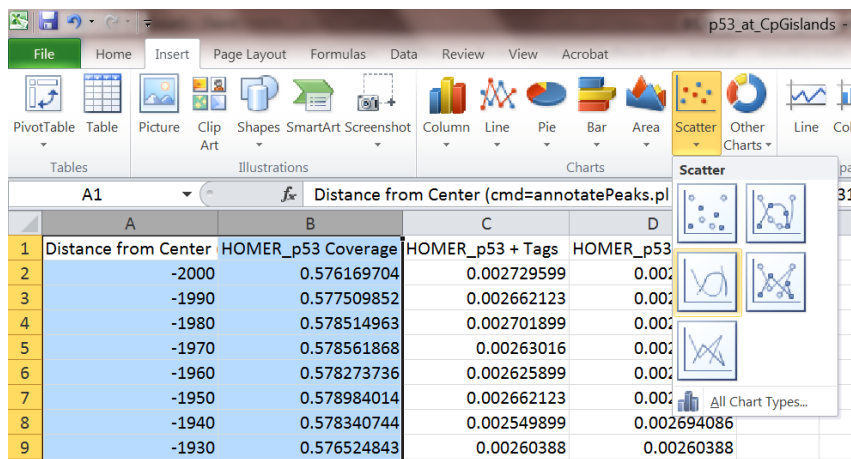
## Task 6. Visualise p53 binding profiles around CpG islands in Excel

Use WinSCP to copy the results of the previous calculation (`p53_at_CpGislands.txt`) to your local computer, then open this file in Excel, select the first two columns and plot the graph:

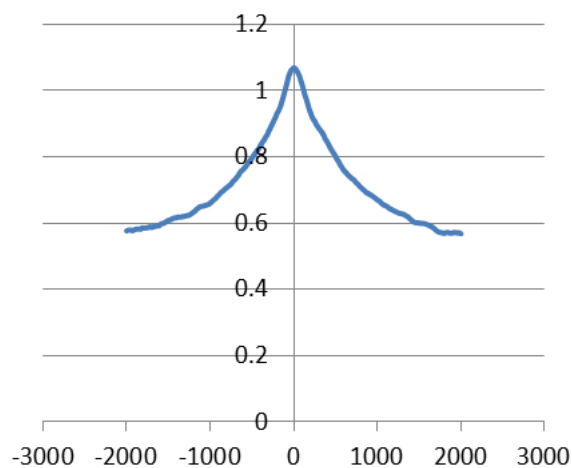


	A	B	C	D	E	F	G
1	Distance from Center	HOMER_p53 Coverage	HOMER_p53 + Tags	HOMER_p53 - Tags			
2	-2000	0.576169704	0.002729599	0.002648628			
3	-1990	0.577509852	0.002662123	0.002615955			
4	-1980	0.578514963	0.002701899	0.002574759			
5	-1970	0.578561868	0.00263016	0.002660702			
6	-1960	0.578273736	0.002625899	0.002645076			
7	-1950	0.578984014	0.002662123	0.002587544			
8	-1940	0.578340744	0.002549899	0.002694086			
9	-1930	0.576524843	0.00260388	0.00260388			
10	-1920	0.577476348	0.002642235	0.002622347			
11	-1910	0.579453066	0.002719656	0.002651469			

Inserting the graph:



Your graph will look like this:



The X axis here is the distance from the centre of CpG domain and Y axis is p53 occupancy.

What does this graph demonstrate?

**Task 7. Calculate and visualise the average p53 binding profile around LMR regions.**

Repeat the calculations from tasks 5 and 6 using the file [Stadtler\\_ESC\\_LMR\\_corrected\\_v3.bed](#).

**Task 8. Calculate and visualise the average p53 binding profile around UMR regions.**

Repeat the calculations from tasks 5 and 6 using the file [Stadtler\\_ESC\\_UMR\\_corrected\\_v3.bed](#).

**Task 9. Calculate and visualise the average p53 binding profile around FMR regions.**

Repeat the calculations from tasks 5 and 6 using the file [Stadtler\\_ESC\\_FMR\\_corrected\\_v3.bed](#).

Now compare the figures obtained at steps 6-8 and discuss what happens with p53 binding at LMR, UMR and FMR regions. What can you say about the relation between p53 binding and DNA methylation?

### Step 10. Prepare HOMER tag directory for reduced RRBS DNA methylation in MEFs

At this step we will take the mapped reads for the Reduced Representation Bisulfite Sequencing (RRBS) in MEFs and will treat them similarly to how we previously have treated ChIP-seq data. Indeed, any genomic signal including DNA methylation can be represented in the form of genomic “occupancy landscapes”. For DNA methylation, the situation is complicated by the fact that methylation values for each nucleotide are continuous. In the file provided to you ([Messner2008\\_5mC\\_MEF\\_meth.bed](#)) I have already filtered only those CpGs, which have a probability to be methylated above 50%. If you carefully look at the values in the last column you will see that they are all above 0.5. In all other respects this file has a structure of a usual BED file format, and it can be treated as a BED file. In particular, we can now create HOMER’s tag directory with chromosome-wide occupancies for this dataset using the following command:

```
makeTagDirectory HOMER_5mC-MEF Messner2008_5mC_MEF_meth.bed -  
genome mm9
```

### Task 11. Calculate the average profile of 5mCs around p53 sites in MEFs

Similarly to steps 4 and 5, we can calculate and visualise the average DNA methylation profile around p53 binding sites. This is the command to perform the calculation:

```
annotatePeaks.pl peaks_formatted.bed mm9 -size 4000 -hist 10  
-d HOMER_5mC-MEF > 5mC-MEF_at_p53.txt
```

After the calculation has finished (which is usually less than a minute), plot the graph as in step 5.

What can you say now about the relation of DNA methylation and p53 binding? Did this calculation change your previous conclusions? Do you notice some controversy between p53 profiles at LMRs, UMRs, FMRs and the DNA methylation profile around p53 sites? How can you explain this controversy?

### Task 12. Analyse results of the calculation of p53 motifs (which should be ready by this time)

Now let’s look at the DNA motifs inside the p53 ChIP-seq peaks. ChIP-seq peaks represent regions of about 1kb protein binding sites, but the protein binding sites themselves are just several nucleotides. Binding sites of a transcription factor such as p53 are usually defined by a distinct DNA sequence motif. Remember, that we have already started the motif calculation of our first analysis pipeline? Now let us look at the results.

Using WinSCP, copy the folders “[motifs\\_p53\\_our\\_peaks](#)” or “[motifs\\_p53\\_GSE55727](#)” from the cluster to your local computer (drag and drop it from the right panel to the left panel). Now you

can open the folder “motifs” on your computer locate the file [homerResults.html](#) and open it. To open a file right-click on it in WinSCP, and then select “open”. (You may also open files in the same way by right-clicking on them in the right WinSCP panel and then selecting “open”). The results will look like this:

## Homer *de novo* Motif Results (p53motifs\_GSE55727/)

[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 = 3100

Total background sequences = 44227

\* - possible false positive

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-1650	-3.801e+03	54.26%	3.01%	44.0bp (71.6bp)	p53(p53)/mES-cMyc-ChIP-Seq(GSE11431) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>
2		1e-800	-1.843e+03	10.87%	0.02%	52.2bp (31.0bp)	Gata1/MA0035.3/Jaspar(0.661) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>
3		1e-748	-1.724e+03	10.97%	0.03%	52.7bp (50.3bp)	Sox4(HMG)/proB-Sox4-ChIP-Seq(GSE500) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>
4		1e-670	-1.543e+03	10.39%	0.03%	62.7bp (61.1bp)	FOS/MA0476.1/Jaspar(0.686) <a href="#">More Information</a>   <a href="#">Similar Motifs Found</a>

What does this table tell us? What is the consensus motif of p53?

Now you can open the corresponding matrix and the consensus motif, and answer the following questions:

What is the consensus p53 motif? Does the consensus p53 motif contain CpGs? How does this new information change our understanding of the interplay of p53 & DNA methylation?

**Task 13\*** [For advanced students] Calculate and visualise the average p53 binding profiles around promoters and enhancers similar to tasks 5-6, using the files [promoters\\_mm9.bed](#) and [enhancers\\_mm9.bed](#)