

BS222 Practical 2. November 2019

Next Generation Sequencing (NGS) Analysis in Galaxy

Vladimir Teif (yteif@essex.ac.uk)

There are two ways of doing computational analysis of Next Generation Sequencing (NGS) data: the professional way and the easy way ☺. The professional way is to work in the command-line UNIX environment on a computer cluster, because the files are usually so large that it is not possible to operate with them on a personal computer. The easy way is to use an online software tool called Galaxy. Today we will be exploring the latter possibility. This “easy” way is sometimes also a very proper way, in particular if a friendly IT systems administrator has spent weeks to adjust this software, and your lecturers have double-checked that the computational tasks are doable during the practical ☺. Galaxy is intended to be the software of choice for learning and understanding how NGS analysis works, but it may have some glitches. If you encounter a glitch please keep patient and don’t panic – just wait for the lecturer who will save you. Importantly, the aim of the practical is to understand the main NGS concepts, so please try to see the forest for the trees.

It is also worth noting, that in this practical you will be working with real NGS data, and you can in principle make real scientific discoveries, in which case don’t forget to document them. If this does not interest you then I don’t know what else can make you interested in this practical. Oh, wait; may be also the fact that this is the only practical this year where you can get experience with NGS analysis to boost your employability and add “NGS data analysis skills” to your CV? ☺

Introduction. Our practical will be 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. The authors report both human and mouse ChIP-seq datasets, but mostly analyse the human data in the paper. Today we will perform analysis based on their mouse data. In this practical we will determine, where in the genome our protein of interest, called p53, is binding – because where it is binding determined which genes it is regulating. In the second practical we will be using this information to answer real biomedical questions, such as what happens with these cells as they respond to the anticancer drugs.

Plan of this practical:

1. Understand where to get NGS data online – follow the lecturer
2. Understand Galaxy – an online platform for NGS analysis – follow the lecturer
3. Understand ChIP-seq data formats – follow the lecturer
4. Understand how to map reads to the target genome in Galaxy
(*You do not need to perform the mapping step because I did the mapping for you already*)
5. Find peaks of p53 ChIP-seq (p53 binding sites) using MACS2 in Galaxy
6. Compare the peaks that we determined with the peaks reported by Younger et al.
7. Intersect p53 peaks with enhancers and promoters using BedTools in Galaxy
8. Find enrichment of p53 binding at enhancers and promoters using BedTools in Galaxy

Task 1. Understand where to get NGS data (GEO) – follow the lecturer. After carefully reading the paper’s abstract we scroll down to the bottom of the manuscript to find where the authors have deposited their data. We find the following:

ACCESSION NUMBERS

The Gene Expression Omnibus accession number for the RNA-Seq and ChIP-Seq data reported in this paper is **GSE55727**.

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

Using the Gene Expression Omnibus (GEO) accession number GSE55727 reported by the authors, we find their data at the following link:

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55727>

Opening this link in the browser, we can see the complete description of the experimental details of this study, and the list of the samples which they have deposited (you have to click on “more” next to the sample list):

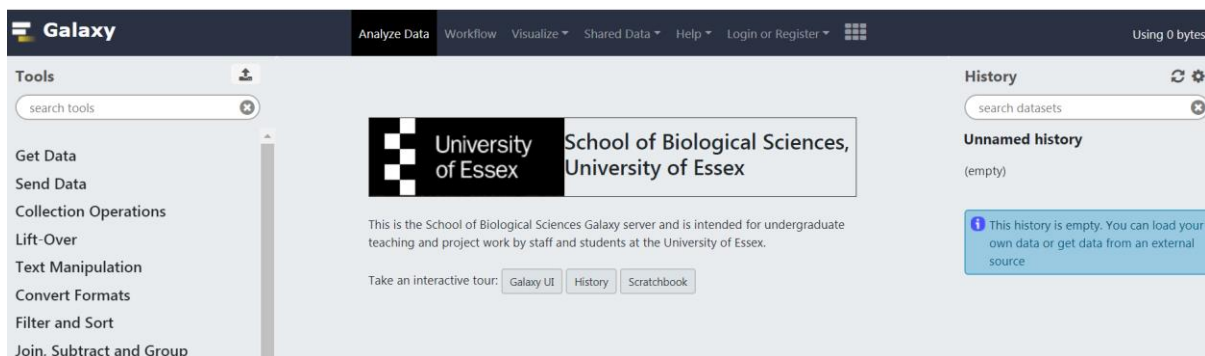
| | | |
|-------------------------|----------------------------|----------------------|
| Samples (24) | GSM1342483 | GM06170_RNA_unt_rep1 |
| Less... | GSM1342484 | GM06170_RNA_unt_rep2 |
| | GSM1342485 | GM06170_RNA_dox_rep1 |
| | GSM1342486 | GM06170_RNA_dox_rep2 |
| | GSM1342487 | GM06170_ChIP_input |
| | GSM1342488 | GM06170_ChIP_p53 |
| | GSM1342489 | GM00011_RNA_unt_rep1 |
| | GSM1342490 | GM00011_RNA_unt_rep2 |
| | GSM1342491 | GM00011_RNA_dox_rep1 |
| | GSM1342492 | GM00011_RNA_dox_rep2 |
| | GSM1342493 | GM00011_ChIP_input |
| | GSM1342494 | GM00011_ChIP_p53 |
| | GSM1342495 | MEF_WT_RNA_unt_rep1 |
| | GSM1342496 | MEF_WT_RNA_unt_rep2 |
| | GSM1342497 | MEF_WT_RNA_unt_rep3 |
| | GSM1342498 | MEF_WT_RNA_dox_rep1 |
| | GSM1342499 | MEF_WT_RNA_dox_rep2 |
| | GSM1342500 | MEF_WT_RNA_dox_rep3 |
| | GSM1342501 | MEF_ChIP_input |
| | GSM1342502 | MEF_ChIP_p53 |
| | GSM1375967 | MEF_KO_RNA_unt_rep1 |
| | GSM1375968 | MEF_KO_RNA_unt_rep2 |
| | GSM1375969 | MEF_KO_RNA_dox_rep1 |
| | GSM1375970 | MEF_KO_RNA_dox_rep2 |

We will be working with the samples MEF_ChIP_p53 and MEF_ChIP_Input. “MEF” stands for mouse embryonic fibroblasts. “p53” stands for the sample which has undergone ChIP-seq with antibody against p53 protein, and “Input” is the same sample, but sequenced without antibody. Our task for this practical will be to analyse these data: check whether the conclusions of the authors of the paper are correct (or may be suggest new scientific conclusions and make a scientific discovery!)

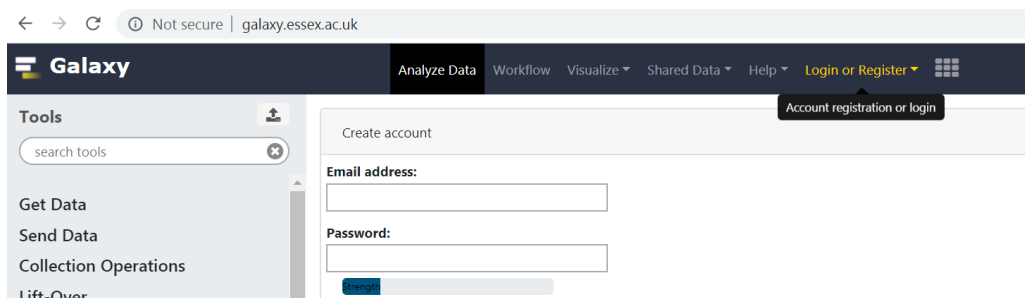
Task 2. Understand Galaxy – an online platform for NGS analysis – follow the lecturer.

Galaxy is open-source software arising from a large international project that aims to provide a user-friendly environment for all kinds of NGS analysis. Galaxy provides a web server that can be installed locally, and then the systems administrator has to take care of this server and install all the required software. We have a local version at <http://galaxy.essex.ac.uk>, which can be accessed only from a university computer. Almost any software tool that exists as a command line tool for UNIX can be also installed on Galaxy, where users do not need to struggle with the “unfriendly” UNIX environment. However, many serious programmers work in UNIX. The teaching materials about Galaxy are available here: <https://galaxyproject.org/learn/>

2.1. Let us open our Galaxy. Open an internet browser and type this address: galaxy.essex.ac.uk:



2.2. Create your account on Galaxy. Go to User > Register, and register an account using your university email. Then login as this user, and continue working under the same user name today.



Now we need to load our data to the Galaxy. Since the files that we want to use are quite large, there is no need that each of us uploads such files to the Galaxy. It is enough that the files have been uploaded once, and then we all can use them. In Galaxy there are several ways to share files between different users. One way is to share “history”. The history is what you see on your right side of the Galaxy. Currently your history is empty. I have previously uploaded to Galaxy the files that we need today, and made this history public. You now can find my shared history and import it, so that it will become your history. Let us do this.

2.3. Click Shared data/Histories:

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data' (highlighted), 'Help', and 'Login or Register'. On the left, a 'Tools' sidebar is visible. The main content area is titled 'Published Histories' and features a search bar with the placeholder text 'search name, annotation, ow'. Below the search bar is a table of published histories. The table has columns for 'Name', 'Annotation', 'Owner', 'Community Rating', 'Community Tags', and 'Last Updated!'. One history is listed: 'Vlad's BS222 history' by 'vlad' with a rating of 5 stars and updated '6 hours ago'.

| Name | Annotation | Owner | Community Rating | Community Tags | Last Updated! |
|----------------------|------------|-------|------------------|----------------|---------------|
| Vlad's BS222 history | | vlad | ★★★★★ | | 6 hours ago |

2.4. Select the history called “Vlad’s BS222 history” (which is also accessible at this link: <http://galaxy.essex.ac.uk/u/vlad/h/bs222>)

2.5. Click on the “Import history” button at the top right corner:

The screenshot shows the 'About this History' page for 'Vlad's BS222 history'. The top navigation bar is the same as in the previous screenshot. The main content area has a large empty space for the history's details. On the right, a sidebar titled 'About this History' contains the author 'vlad', a section for 'Related Histories' (listing 'All published histories' and 'Published histories by vlad'), and an 'Import history' button with a plus icon. The top right corner of the main area shows 'Using 0 bytes'.

After you clicked on the “import history”, you will see something like this:

The screenshot shows a dialog box titled 'Importing history "Vlad's BS222 history"'. The dialog has a warning message: 'As an anonymous user, unless you login or register, you will lose your current history after importing this history. You can login here or register here.' Below the warning, there is a text input field for 'Enter a title for the new history:' with the text 'imported: Vlad's BS222 history' entered. At the bottom right of the dialog are 'Cancel' and 'Import' buttons.

2.6. Click “Import”. This will replace your empty history with the history “Imported BS222 2017-18”:

The screenshot shows the Galaxy web interface at galaxy.essex.ac.uk. The main panel displays a welcome message: "Hello, Galaxy is running!" with instructions to customize the page or take an interactive tour. The left sidebar contains a list of tools categorized under "Tools". The right sidebar shows the "History" panel with a search bar and a list of datasets. The datasets listed are:

- imported: BS222 2017-18 (6 shown, 5.81 GB)
- 98: promoters_mm9_sorted.bed
- 87: mm9_genome.sorted
- 86: p53_published_MEF_ChIP_peaks_sorted.bed
- 72: p53.fastq
- 71: Input.fastq

Task 3. Understand ChIP-seq data formats – follow the lecturer

As you can see, your new history in Galaxy now includes several files. The most important are p53.fastq and Input.fastq – they contain the initial raw data as they were received from the sequencing machine. Let’s look at each of them and understand how they are structured. You can click on the eye pictogram to look inside each of these file. For example, here we have opened the p53.fastq file – this file contains the sequences of all DNA fragments determined in the ChIP-seq experiment with p53 antibody. The red rectangle shows an example of one read. In this case each read is 36 nucleotide long:

The screenshot shows the Galaxy web interface with the p53.fastq file open. A warning message at the top states: "This dataset is large and only the first megabyte is shown below. Show all | Save". The main panel displays the contents of the file, which are FASTQ format sequences. A red rectangle highlights a single read, showing the sequence: "NNCTCTCTTCANCTGNGCCCAACCAATGTGTTTA". The right sidebar shows the "History" panel with a search bar and a list of datasets. The datasets listed are:

- imported: BS222 2017-18 (6 shown, 5.81 GB)
- ks_sorted.bed
- 72: p53.fastq (3.0 GB, format: fastqsanger, database: mm9, uploaded fastq file)

Task 4. Understand how to map reads to the target genome in Galaxy.

The first step in ChIP-seq analysis is mapping (also known as “alignment”) of the reads. This is usually the most time-consuming and computationally demanding task in NGS analysis. Therefore, in order to save your time I have already performed this task for you – you do not need to perform the mapping, but you need to understand how the mapping was done by me. Before the mapping is performed we only know the DNA sequence of each read, but do not know yet where each read is positioned in the genome. After the mapping is performed, we know for each read its location in the genome (for some reads there could be potentially several locations – the lecturer will discuss this).




4.1. Locate in the left panel software Bowtie2, and click on it:

Bowtie2 - map reads against reference genome (Galaxy Version 2.3.4.3+galaxy0) Versions Options

Is this single or paired library

Single-end

FASTA/Q file



11: p53.fastq

Must be of datatype "fastqsanger" or "fasta"

Write unaligned reads (in fastq format) to separate file(s)

Yes

No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

Yes

No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

4.2. Select library type “single-end”, file name “p53.fastq”, and reference genome “Mus musculus (mm9)”. **Do NOT click “execute” (do NOT start the mapping), because if we all submit jobs for alignment (aka mapping) we all will have to wait very long.**

You do not need to do the alignment because I have already did it for you and placed the results in the history that you have imported. However, advanced students can try to do their own alignment if time remains at the end of the practical.

The history that I have shared with you contains files “p53 aligned reads” and “Input aligned reads”. These are the aligned data based on p53 and its control experiment “Input” correspondingly. Let us now look at them to understand the data structure.

Attributes updated

Attributes Convert Format Datatype Permissions

Edit Attributes

Name:
Input aligned reads (sorted BAM)

Info:
100: p53 aligned reads (sorted BAM)

History

search datasets

imported: BS222 2017-18
8 shown
6.89 GB

101: Input aligned reads (sorted BAM)

100: p53 aligned reads (sorted BAM)

4.3. Click on the files “p53 aligned reads” and “Input aligned reads”. Hint: Do not click on the eye pictogram; click directly on the file name.

Discuss in pairs: How many reads are aligned in the p53 sample? How many reads are aligned in the Input? How many reads did not map in p53? How many reads did not map in Input?

Task 5. Find peaks of p53 binding using MACS2 in Galaxy.

Now let us do some calculations.

If our local Galaxy installation at Essex is overloaded we will do the next steps at the central Galaxy installation at <https://usegalaxy.org>. In the latter case we will jump to Task 5* on page 9.

Let us determine the locations of bound p53 genome-wide. If you remember the lecture about NGS analysis, the genomic locations which are bound by proteins a visually seen as peaks on the protein binding occupancy landscape. We now need to locate the positions of all these peaks. At this step we will need both the p53 and Input mapped reads. Why? In order to consider only the peaks that appeared in the ChIP-seq experiments using antibody against protein p53, and not the peaks which appeared in the control experiment where antibody was not added (this control is called Input).

5.1. Search on the left panel of Galaxy for the software called MACS2, and click on it:

Tools

macs2

ChipSeq

MACS2.1.0 Model-based Analysis of ChIP-Seq: peak calling

Workflows

All workflows

MACS2.1.0 Model-based Analysis of ChIP-Seq: peak calling (Galaxy Version 2.1.0-6)

Experiment Name
MACS2.1.0 in Galaxy

Type of region to call
Narrow regions
Broad regions are formed by linking nearby enriched regions

Format of input read data
BAM (single-end)
Specify the format of the input data and whether or not it is paired end (--format)

ChIP-seq read file
100: p53 aligned reads (sorted BAM)

ChIP-seq control read file
101: Input aligned reads (sorted BAM)

Effective genome size
Mouse (1.87e9)
Either pre-defined (for common organisms), or user-defined (--gsz)

Band width
200

History

search datasets

imported: BS222 2017-18
8 shown
6.89 GB

101: Input aligned reads (sorted BAM)

100: p53 aligned reads (sorted BAM)

98: promoters_mm9_sorted.bed

87: mm9.genome.sorted

86: p53_published_MEF_ChIP_peaks_sorted.bed

72: p53.fastq

3.0 GB

5.2. Select Type of regions to call “Narrow regions”, Format of input read data “BAM (single-end)”, ChIP-seq read file “p53 aligned reads”, ChIP-seq control read file “Input aligned reads”, and Effective genome size “Mouse”. Then click “execute” at the bottom of the page:

This calculation will take about 15-20 minutes (if our server is in a good mood☺)

5.3. While the job is being executed listen to the lecturer’s explanations about the algorithm of peak calling, and read about the parameters of MACS2 at its Galaxy page, as well as at its own web page: <https://github.com/taoliu/MACS>

5.4. When the calculation is finished we can have a look at each of the four new files that are created:

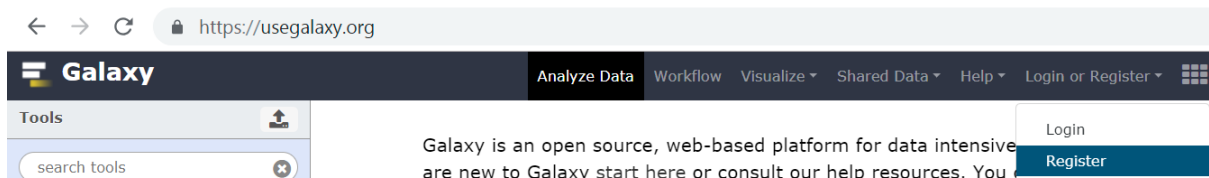
Which parameters determine the number of regions that are reported as peaks? How can we change these parameters to get more/less peaks?

5.5. How many regions did we get as p53 binding peaks?

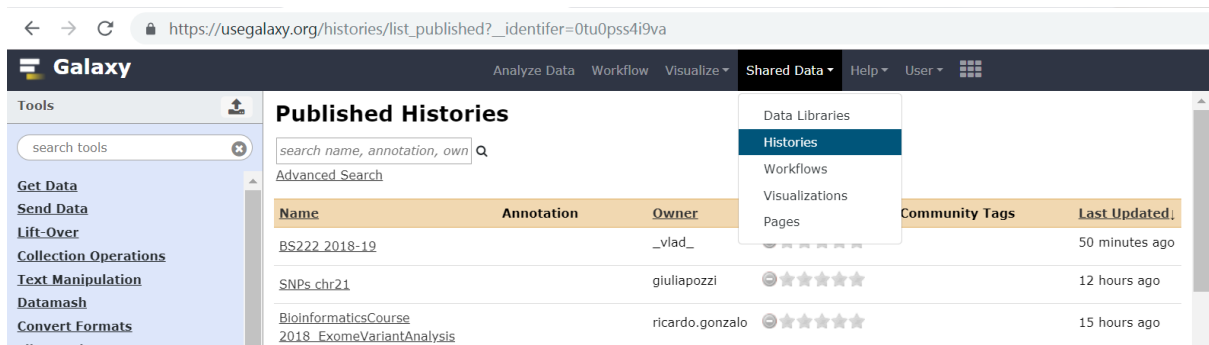
Task 5*. If our Galaxy installation at Essex is overloaded we will do a variation of step 5 at the central Galaxy installation at <https://usegalaxy.org>, as described below. Otherwise, skip to step 6 on page 12.

If we can't use our Essex Galaxy, we have a plan B! We can try use another installation of Galaxy at <https://usegalaxy.org> (or even better at <https://usegalaxy.eu/>), which are available for anyone, not just for students of our university. In fact, you can use them in your future projects or play individually.

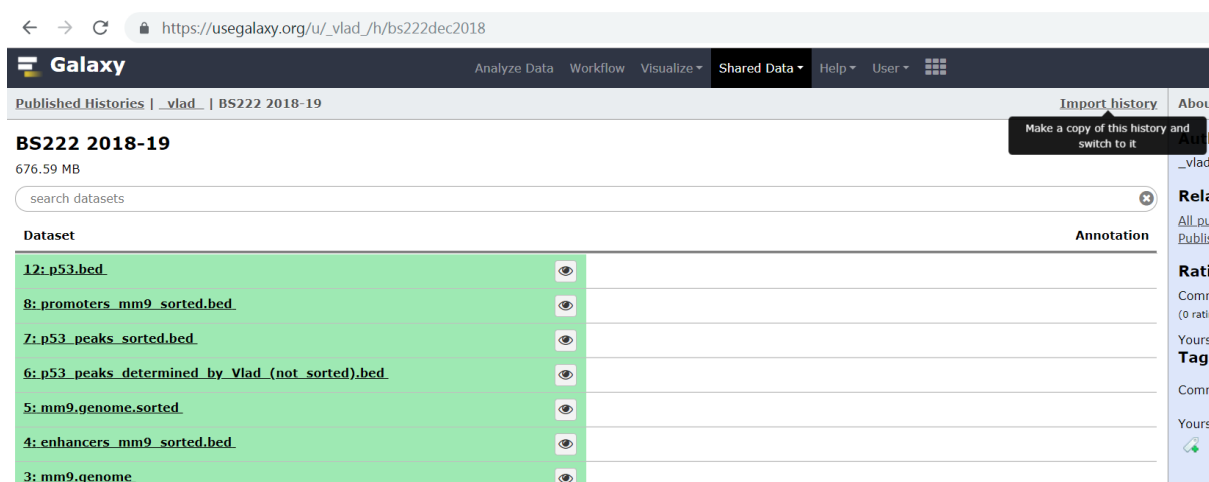
Go to <https://usegalaxy.org> and register an account in the same way as you previously did at our local Galaxy installation:



Then go to menu “Shared data”> “Histories”, and select history named “BS222 2018-19”:



Then open the history “BS222 2018-19” and import it to your current history:



This history contains files with the same names as in our local Galaxy installation. In addition, there are mapped DNA reads in a more compact BED format. In this format, only the genomic coordinates are provided. You can see these files by clicking on the eye icon on the files “p53.bed” and “Input.bed”:

The screenshot shows the Galaxy web interface at <https://usegalaxy.org>. The main panel displays a table with the following columns: Chrom, Start, End, Name, Score, Strand, ThickStart, ThickEnd, ItemRGB, BlockCount, BlockSizes, and BigWig. The table contains 12 rows of data for various chromosomes (chr5, chr13, chr18, chr2, chr15, chr11, chr6, chrX). The right-hand panel shows a history of datasets, including an imported dataset named "BS222 2018-19" and a dataset named "12: p53.bed" which is approximately 11,000,000 regions in BED format for the mm9 database.

Now let's determine p53 binding peaks (or in other words, do peak calling), using these two files "p53.bed" and "Input.bed". To do so, let's locate a program MACS2 that is doing peak calling. For example, we can find it by entering the name "MACS2" in the search field as shown below:

The screenshot shows the Galaxy web interface with the search results for "MACS2". The search bar on the left contains the text "MACS2". Below the search bar, a list of tools is displayed under the heading "NGS: Peak Calling". The tools listed are: DiffBind (differential binding analysis of ChIP-Seq peak data), MACS2 bdgdiff (Differential peak detection based on paired four bedgraph files), MACS2 bdgcmp (Deduct noise by comparing two signal tracks in bedGraph), MACS2 bdgbroadcall (Call broad peaks from bedGraph output), and MACS2 callpeak (Call peaks from alignment results). The main panel displays a table with the following columns: Chrom, Start, End, Name, Score, and Strand. The table contains 12 rows of data for various chromosomes (chr5, chr13, chr18, chr2, chr13, chr15, chr11, chr6, chrX, chr4, chr7, chr11, chr18, chr12, chr16, chr1, chr8).

Then let's select "MACS2 callpeak" (Call peaks from alignment results). In the MACS2 menu let us select the following options:

- ChIP-seq treatment file: p53.bed
- Do you have a control file: yes
- ChIP-seq control file: Input.bed
- Format of input files: single-end BED
- Effective genome size: m. musculus

Keep the rest parameters as they are by default (do not change).

Scroll to the end of the page and click the "Execute" button:

MACS2 callpeak Call peaks from alignment results (Galaxy Version 2.1.1.20160309.4) Versions Options

Are you pooling Treatment Files?

 For more information, see Help section below

ChIP-Seq Treatment File

 (-t)

Do you have a Control File?

Are you pooling Control Files?

 For more information, see Help section below

ChIP-Seq Control File

 (-c)

Format of Input Files

 For Paired-end BAM (BAMPE) the 'Build model step' will be ignored and the real fragments will be used for each template defined by leftmost and rightmost mapping positions (--format). Default: Single-end BAM

Effective genome size

If you did everything correct, the following kind of screen appears:

Executed **MACS2 callpeak** and successfully added 1 job to the queue.

The tool uses 2 inputs:

- 12: p53.bed**
- 2: Input.bed**

It produces this output:

- 16: MACS2 callpeak on data 2 and data 12 (narrow Peaks)**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

search datasets

imported: BS222 2018-19
10 shown
675.32 MB

16: MACS2 callpeak on data 2 and data 12 (narrow Peaks)

12: p53.bed
~11,000,000 regions
format: bed, database: mm9
uploaded bed file

The new dataset with peaks that is being created will be first grey (which means it is standing in a queue), then it will turn yellow (which means that this job is running), and finally when it will turn green the calculation will be finished. In my experience this calculation took about 10 minutes, but it may be very different if all of us will submit our jobs to the same server at the same time, so be prepared for longer waiting times. It is also a good time to go get a cup of tea ☺

When this calculation is finished you can have a look at the resulting file:

| | | | | | | | | | |
|------|---------|---------|-----------------|------|---|----------|-----------|-----------|-----|
| chr1 | 6476725 | 6477052 | p53_bed_peak_7 | 561 | . | 24.42301 | 60.59121 | 56.11501 | 182 |
| chr1 | 6988686 | 6988861 | p53_bed_peak_8 | 189 | . | 12.15005 | 22.90046 | 18.99440 | 85 |
| chr1 | 7387677 | 7387892 | p53_bed_peak_9 | 60 | . | 4.42201 | 9.50980 | 6.05489 | 104 |
| chr1 | 9628826 | 9628982 | p53_bed_peak_10 | 101 | . | 8.02496 | 13.76951 | 10.12431 | 49 |
| chr1 | 9629054 | 9629377 | p53_bed_peak_11 | 1836 | . | 44.85268 | 189.18132 | 183.60217 | 161 |

16: MACS2 callpeak on data 2 and data 12 (narrow Peaks)

9,148 regions
format: bed, database: mm9

The number of peaks that you obtain may be different from me if you changes some parameters (in which case you need to be able to explain what you changed and how it affected your peak calling ☺)

In the example above, the resulting file has 9.148 regions. Each region corresponds to one ChIP-seq peak, or in other words, to one p53-bound genomic location.

Task 6. Compare the peaks that we determined with the peaks reported by Younger et al.

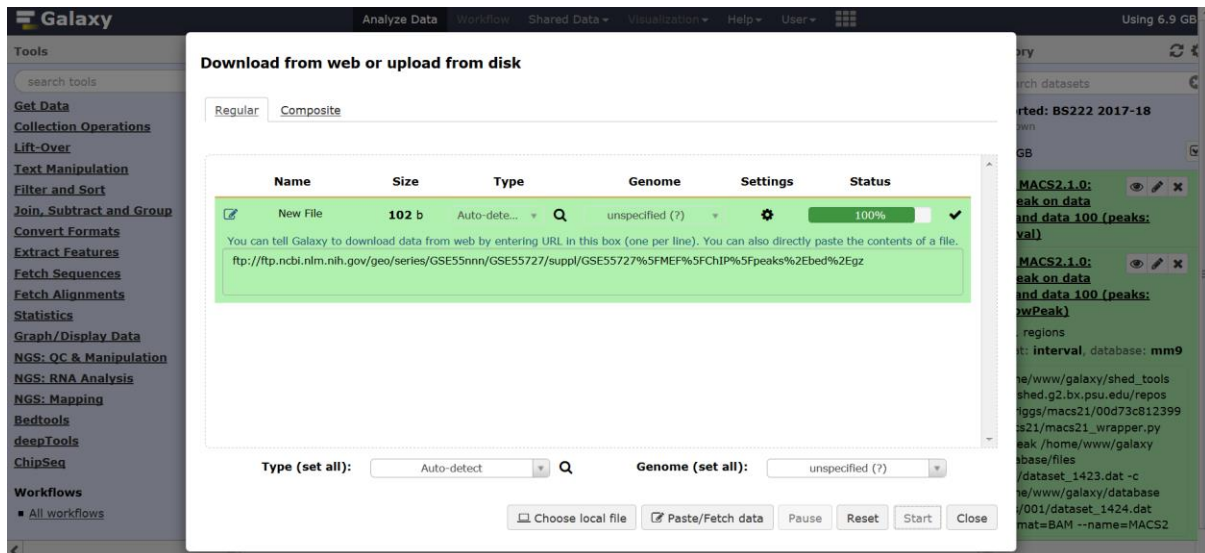
Now let's have a look at the peaks that have been reported by the authors of this study. Remember where the data came from? We can look in the GEO database, where we took the data from (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55727>). At the bottom of the entry, we can see the following:

| Relations | | | |
|---|-------------|--|--------------------|
| BioProject | PRJNA240784 | | |
| SRA | SRP039598 | | |
| Download family | | Format | |
| SOFT formatted family file(s) | | SOFT ? | |
| MINIML formatted family file(s) | | MINIML ? | |
| Series Matrix File(s) | | TXT ? | |
| Supplementary file | Size | Download | File type/resource |
| SRP/SRP039/SRP039598 | | (ftp) | SRA Study |
| GSE55727_Human_ChIP_peaks.bed.gz | 24.2 Kb | (ftp) (http) | BED |
| GSE55727_Human_RNA_Expression_Matrix.txt.gz | 1000.0 Kb | (ftp) (http) | TXT |
| GSE55727_MEF_ChIP_peaks.bed.gz | 27.6 Kb | (ftp) (http) | BED |
| GSE55727_MEF_KO_RNA_Expression_Matrix.txt.gz | 570.5 Kb | (ftp) (http) | TXT |
| GSE55727_MEF_WT_RNA_Expression_Matrix.txt.gz | 784.4 Kb | (ftp) (http) | TXT |

Raw data provided as supplementary file

Processed data is available on Series record

We are particularly interested in the file “GSE55727_MEF_ChIP_peaks.bed.gz”. This is the file with the peaks determined by the authors. I have already copied it to the Galaxy history shared with you so you need not to download it from the Internet. But if you wish to do so you can do this by selecting Get Data from the left menu and following the screenshot below that shows you how I did this:



In the Galaxy history shared with you I have renamed this file to “p53 peaks sorted”. We can view this file. We are mostly interested in the question how many regions (ChIP-seq peaks) are there.

The screenshot shows the Galaxy web interface. The main panel displays a table of genomic regions with columns: Chrom, Start, End, Name, Score, Strand, ThickStart, ThickEnd, ItemRGB, and BlockCount. The table lists 18 regions across chromosomes chr1 and chr2. The right panel shows the 'History' section with a search bar and a list of datasets. The top dataset is 'imported: BS222 2017-18' with 13 shown items and a size of 6.89 GB. Below it, a preview of a BED file is shown, indicating 3,100 regions in BED format.

| Chrom | Start | End | Name | Score | Strand | ThickStart | ThickEnd | ItemRGB | BlockCount |
|-------|-----------|-----------|------|-------|--------|------------|----------|---------|------------|
| chr2 | 52887755 | 52888422 | | | | | | | |
| chr2 | 53937582 | 53938135 | | | | | | | |
| chr2 | 54150598 | 54151013 | | | | | | | |
| chr2 | 54220708 | 54220897 | | | | | | | |
| chr1 | 158969084 | 158969473 | | | | | | | |
| chr2 | 54758085 | 54758420 | | | | | | | |
| chr2 | 57551055 | 57551601 | | | | | | | |
| chr2 | 58338055 | 58338467 | | | | | | | |
| chr2 | 59189751 | 59190155 | | | | | | | |
| chr2 | 60377479 | 60378021 | | | | | | | |
| chr2 | 60664931 | 60665649 | | | | | | | |
| chr2 | 60693112 | 60693579 | | | | | | | |
| chr2 | 64460446 | 64460692 | | | | | | | |
| chr2 | 64877894 | 64878420 | | | | | | | |
| chr2 | 65206758 | 65207040 | | | | | | | |
| chr1 | 159250521 | 159251028 | | | | | | | |

There is one peak per line (or “per region”).

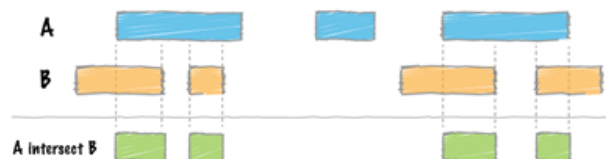
Discuss in pairs: How many regions are there in this file? Why is the number of peaks that we have found different from the number of peaks determined by the authors of this paper?

Task 7. Intersect p53 peaks with enhancers and promoters using BedTools in Galaxy

[7*. If our local Galaxy installation at Essex is overloaded we can do this step at the central Galaxy installation at <https://usegalaxy.org>. In the latter case instead of “Vlad’s BS222 history” on our local server <http://galaxy.essex.ac.uk>, you can use history “BS222 2018-19” on <https://usegalaxy.org>]

Peaks are genomic regions (defined by the chromosome, region start, region end, etc). In the BED format (the format typically used to store genomic regions after peak calling), we have columns in exactly this order (chromosome, region start, region end).

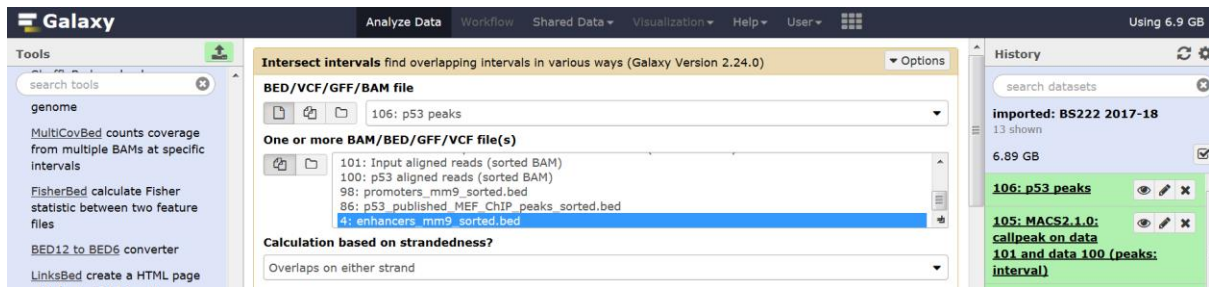
In the next task we want to intersect the genomic regions which are identified as p53 binding sites by the authors of the original paper with the regions corresponding to mouse enhancers and promoters. Here is a schematic picture which explains the “intersection” between two sets of genomic regions:



Intersection is one of the main concepts in ChIP-seq analysis. To do this we will use command [Intersect Intervals](#) from the software package [BedTools](#).

A detailed description of all parameters of this command is provided at the following link: <http://bedtools.readthedocs.io/en/latest/content/tools/intersect.html>

7.1. Locate on Galaxy software “BedTools”, and inside BedTools select “Intersect Intervals”. Open it. Then select the names of two files with genomic regions that you want to intersect. Select as the first file “*p53 peaks sorted*”, and as the second file “*enhancers_mm9.sorted*” (this is the file with mouse enhancers which I have prepared for you). Then click “Execute” at the bottom of the page. This calculations will take just several seconds is there is no queue on the server.



7.2. Click on the file with the intersection of p53 peaks and enhancers:

1 job has been successfully added to the queue - resulting in the following datasets:

107: Intersect intervals on data 4 and data 106

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

search datasets

imported: BS222 2017-18
13 shown
6.89 GB

107: Intersect intervals on data 4 and data 106

380 regions
format: bed, database: mm9

display in IGB View
display at Ensembl Current
display with IGV local Mouse mm9
display at UCSC main test

| 1.Chrom | 2.Start | 3.End |
|---------|----------|----------|
| chr2 | 60693112 | 60693579 |
| chr2 | 72833515 | 72834290 |
| chr2 | 75606183 | 75606697 |

How many regions are there in this file? Do you remember how many regions there were in total in the file with "*p53 peaks sorted*" reported by the authors that you have used in this intersection? Are there many regions intersecting with enhancers? How did you decide that this is "many"?

7.3. Repeat step 6.2., but now intersect p53 peaks with promoters (select for the intersection files "*p53 peaks sorted*" and "promoters_mm9_sorted.bed"). How many p53 peaks intersects with promoters? Is it a lot? How do you know that this is a lot or not?

8. Finding enrichment of p53 binding at enhancers and promoters using BedTools in Galaxy

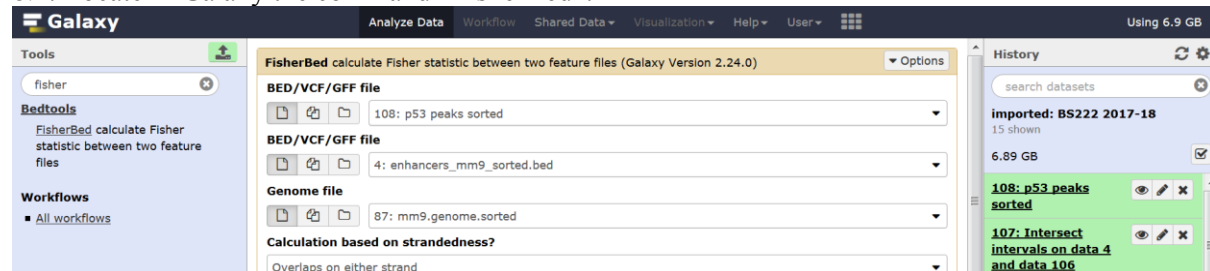
[8*. If our local Galaxy installation at Essex is overloaded we can do this step at the central Galaxy installation at <https://usegalaxy.org>. In the latter case instead of "Vlad's BS222 history" on our local server <http://galaxy.essex.ac.uk>, you can use history "BS222 2018-19" on <https://usegalaxy.org>]

At the previous steps you have noticed that the pure knowledge of how many p53 peaks intersects with promoters or enhancers does not tell us whether this is a lot or not. Indeed, the critical thing that we do not know is how many regions would intersect with promoters or enhancers by chance if we would randomly select the same number of regions as in the set of p53 peaks, just with random genomic locations. Basically, if our peaks have a higher proportion of regions that intersect with enhancers than what one would expect by chance, then we can say that p53 peaks are statistically enriched in enhancers. There are different ways to check for this statistical hypothesis. One of the simplest possibilities is to perform the Fisher test (remember the introduction to statistics from Year 1?) The Fisher test will give us a quantitative measure of the statistical significance of our hypothesis (our hypothesis is that p53 peaks are enriched in enhancers). The Fisher test will calculate for us a P-value, which is the probability that the same situation happens just by chance (randomly). Obviously, if it can happen by chance randomly, this is not a real biological effect. Only if our biological finding

has a very low probability to happen by chance (low P-value), only then our finding is statistically significant. So let us test the conclusion of the authors of the paper that p53 likes to bind in enhancers.

8.1. The Fisher text that is available on the Galaxy will only work on the sorted data, so we have to sort our peaks first. I did it for you already for the peaks reported by the authors of this paper, which are in the file named “*p53 peaks sorted*” (this is why it has the word sorted in its name). If you want to do it yourself later you can use the command “SortBed” in Galaxy to sort any BED-format file.

8.2. Locate in Galaxy the command “FisherBed”:



8.3. Select the names of the two files for which we want to perform the Fisher test: “p53 peaks sorted”, and “enhancers_mm9_sorted”. For the Genome file select “mm9.genome.sorted” (this is the file that contains the lengths of all mouse chromosomes – this information is needed to perform the statistical significance test). Then click “execute”:

8.4. The results of the Fisher test are reported in the following way:

| Chrom | Start | End | Name | Score | Strand | ThickStart | ThickEnd | ItemRGB |
|--|-------------|-------------|----------|-------|--------|------------|----------|---------|
| # 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 | | | | | |

In the table above, we need to look at the two-tail P-value. If the P-value is smaller than 0.05 the results are usually considered as significant. Are our results significant? The value indicated as “ratio” shows the enrichment of p53 peaks with enhancers. In the case above, for example, ratio=11.599. This means that p53 binding sites are more than 11-fold enriched with enhancers in comparison with what would be expected by chance.

8.5. Determine p53 enrichment at promoters using the file promoters_mm9_sorted.bed and following the steps 7.2-7.4.

Discuss in pairs: Is p53 also enriched at promoters? Where is it enriched stronger, at promoters or enhancers?

This is the end of our first computational practical. Please keep all your notes as these will be helpful for your next practical, as well as for answering the questions in the coursework. If you will forget any numbers obtained during this practical you will be able to see them again on the Galaxy. You will be also able to play with Galaxy later. Your Galaxy account will be retained for you.

The standard practical ends here. Below is an additional step for advanced students, that explains how to do any analysis in Galaxy later on (if you are interested).

9* Optional task for advanced students. Learn how to get the data from the GEO database to Galaxy.

To do so, scroll down to the bottom of the GEO entry and go to the link to the “SRA” entry. SRA stands for “Short Read Archive”, and it stores the raw data (while the initial GEO entry stores processed data):


Platforms (4)
[More...](#)

[GPL9115](#) Illumina Genome Analyzer II (Homo sapiens)
[GPL9250](#) Illumina Genome Analyzer II (Mus musculus)
[GPL11154](#) Illumina HiSeq 2000 (Homo sapiens)

Samples (24)
[More...](#)

[GSM1342483](#) GM06170_RNA_unt_rep1
[GSM1342484](#) GM06170_RNA_unt_rep2
[GSM1342485](#) GM06170_RNA_dox_rep1

Relations
 BioProject [PRJNA240784](#)
 SRA [SRP039598](#)



Download family
[SOFT formatted family file\(s\)](#)
[MINIML formatted family file\(s\)](#)
[Series Matrix File\(s\)](#)

Format
 SOFT [?](#)
 MINIML [?](#)
 TXT [?](#)

Then go to the SRA entry for the p53 ChIP-seq experiment:

Platform
 Illumina (24)

Strategy
 other (24)

Data in Cloud
 GS (24)
 S3 (24)

[Clear all](#)
[Show additional filters](#)

☐ [GSM1375970: MEF KO RNA dox rep2; Mus musculus; RNA-Seq](#)
 1. 1 ILLUMINA (Illumina HiSeq 2000) run: 30.6M spots, 6.2G bases, 4.2Gb downloads
 Accession: SRX529164

☐ [GSM1375969: MEF KO RNA dox rep1; Mus musculus; RNA-Seq](#)
 2. 1 ILLUMINA (Illumina HiSeq 2000) run: 36.4M spots, 7.3G bases, 4.9Gb downloads
 Accession: SRX529163

☐ [GSM1375968: MEF KO RNA unt rep2; Mus musculus; RNA-Seq](#)
 3. 1 ILLUMINA (Illumina HiSeq 2000) run: 35.6M spots, 7.2G bases, 4.9Gb downloads
 Accession: SRX529162

☐ [GSM1375967: MEF KO RNA unt rep1; Mus musculus; RNA-Seq](#)
 4. 1 ILLUMINA (Illumina HiSeq 2000) run: 38M spots, 7.7G bases, 5.2Gb downloads
 Accession: SRX529161

☒ [GSM1342502: MEF ChIP p53; Mus musculus; ChIP-Seq](#)
 5. 1 ILLUMINA (Illumina Genome Analyzer II) run: 16.7M spots, 600.7M bases, 299.5Mb downloads
 Accession: SRX483599

☐ [GSM1342501: MEF ChIP input; Mus musculus; ChIP-Seq](#)
 6. 1 ILLUMINA (Illumina Genome Analyzer II) run: 16M spots, 576.6M bases, 281.6Mb downloads
 Accession: SRX483598

Click of the “MEF ChIP p53” (Accession number SRX483599):

SRX483599: GSM1342502: MEF_ChIP_p53; Mus musculus; ChIP-Seq

1 ILLUMINA (Illumina Genome Analyzer II) run: 16.7M spots, 600.7M bases, 299.5Mb downloads

Submitted by: NCBI (GEO)

Study: Integrative Genomic Analysis Reveals Widespread Enhancer Regulation by p53 in Response to DNA Damage

[PRJNA240784](#) • [SRP039598](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

Sample: MEF_ChIP_p53

[SAMN02678396](#) • [SRS567918](#) • [All experiments](#) • [All runs](#)

Organism: [Mus musculus](#)

Library:

Instrument: Illumina Genome Analyzer II

Strategy: ChIP-Seq

Source: GENOMIC

Selection: ChIP

Layout: SINGLE

Construction protocol: RNA: TriZol / ChIP: Immunopurification of DNA followed by phenol:chloroform extraction Illumina TruSeq

Experiment attributes:

GEO Accession: GSM1342502

Links:

Runs: 1 run, 16.7M spots, 600.7M bases, [299.5Mb](#)

| Run | # of Spots | # of Bases | Size | Published |
|----------------------------|------------|------------|---------|------------|
| SRR1186260 | 16,685,745 | 600.7M | 299.5Mb | 2015-07-22 |

At the very bottom of this page we finally see the link to the raw data and it's accession number: SRR1186260. Let's copy this number.

The go to the Galaxy and click “get data”:

Galaxy Analyze Data Workflow Visualize Shared Data Help Login or Register

Tools search tools

Get Data

- [Download and Generate Pileup Format from NCBI SRA](#)
- [Faster Download and Extract Reads in FASTQ format from NCBI SRA](#)
- [Download and Extract Reads in BAM format from NCBI SRA](#)
- [Download and Extract Reads in FASTA/Q format from NCBI SRA](#)
- [Upload File from your computer](#)

University of Essex School of Biological Sciences, University of Essex

This is the School of Biological Sciences Galaxy server and is intended for undergraduate teaching and project work by staff and students at the University of Essex.

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

As you can see, there are many ways to get data to the Galaxy. In this case let's select "Download and Extract Reads in FASTA/Q format from NCBI SRA":

Download and Extract Reads in FASTA/Q format from NCBI SRA (Galaxy Version 2.10.4) ▼ Options

select input type

SRR accession ▼

Accession

Must start with SRR, DRR or ERR, e.g. SRR925743, ERR343809


Select output format


☐ gzip compressed fastq

☒ Uncompressed fastq

☐ bzip2 compressed fastq

Compression will greatly reduce the amount of space occupied by downloaded data. Downstream applications such as a short-read mappers will accept compressed data as input. Consider this example: an uncompressed 400 Mb fastq datasets compresses to 100 Mb or 80 Mb by gzip or bzip2, respectively. (--gzip --bzip2)

[Advanced Options](#) 

 **Execute**

Now we just need to paste the SRR accession number that we copied from the NCBI SRA database, select "uncompressed FASTQ" for the output format and click "execute".

Et voilà, you've got your dataset of interest to the Galaxy and can do any analysis with it!