An introduction to ChIP-seq analysis with Galaxy

Sebastian Schmeier

Institute of Natural and Mathematical Sciences Massey University Auckland, New Zealand http://sschmeier.com s.schmeier@gmail.com

2016-03-23

Contents

An introduction to ChIP-seq analysis with Galaxy	3
1.0 Preface	3
1.1 Overview	3
1.2 The task at hand	3
1.3 Log into Galaxy	4
1.4 Create a new history	6
1.5 Loading the data	7
1.6 Investigate the data	9
1.7 Quality assessment	11
1.7.1 FastQC	11
1.7.2 Read filtering	12
1.7.3 Quality trimming	14
1.8 Mapping reads	15
1.8.1 Bowtie2	16
1.8.2 Post-mapping processing	18
1.9 Peak calling	21
1.9.1 MACS	21
2.0 Post-processing	24
2.1 Overlap peaks with promoter regions	24
2.1.1 Get genes	24
2.1.2 Get promoter	25
2.1.3 Join	26
2.2 Enrichment analysis (genes) with Enrichr	27
2.3 Enrichment analysis (peaks) with GREAT	31
2.4 Visualisation	33
2.5 Motif finding	35
2.5.1 Find the peak center	36
2.5.2 Get flanking regions	38
2.5.3 Extract fasta-sequence	39
2.5.4 Run MEME-ChIP	39
2.6 References	41
2.7 Web links	41

An introduction to ChIP-seq analysis with Galaxy

1.0 Preface

In this brief tutorial we will learn how to use the excellent tool Galaxy (http://galaxyproject.org/) to analyse data from a chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiment. It is part of a series of introductory tutorials that can be found at http://sschmeier.github.io/bioinf-workshop/.

A PDF-version of this tutorial can be downloaded here or at http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/doc/Galaxy-ChIPseq-Introduction_sschmeier.pdf

Two accompanying lectures for this tutorial are available:

- 1. ChIP-seq introduction (http://dx.doi.org/10.6084/m9.figshare.1545468).
- 2. ChIP-seq data processing (http://dx.doi.org/10.6084/m9.figshare.1554130).

1.1 Overview

In this brief tutorial we will learn how to use the excellent tool Galaxy to analyse ChIP-seq data. If you are new to Galaxy, you might want to start with the Galaxy introductory tutorial (http://sschmeier.github.io/bioinf-workshop/ #!galaxy-intro/).

1.2 The task at hand

The overall purpose in this tutorial is to:

- Understand better the Galaxy system (1.3-1.4).
- Understand how to get your data of interest into the system (1.5).
- Understand how to quality control your sequencing data (1.7).
- Understand how to map sequence reads to a reference genome (1.8).
- Post-mapping processing to clean the data before peak calling (1.8.2)
- Understand how to call ChIP-peaks based on the mapped reads (1.9).
- Understand how to gather additional information about you data (2.0).

In order to develop an understanding of the points above, you will run through the workflow to analyse ChIP-seq data (see *Figure 1*):



Figure 1: ChIP-workflow (adjusted from Park2009).

The individual tasks are:

- $1. \ \ \text{Load the dataset}.$
- $2. \ \ Quality \ assess \ the \ reads.$
- 3. Map the reads to the genome using Bowtie2.
- 4. Call peaks using MACS.
- 5. Run Enrichr with genes and GREAT with the peak regions to find enriched annotations.
- 6. Visualise the peaks in UCSC browser.
- 7. Prepare peak data and use MEME to find TFBS motifs.

1.3 Log into Galaxy

First, go to https://usegalaxy.org/ and log into your Galaxy account (see Figure 2 and Figure 3).



Figure 2: Log into your Galaxy account.

ng Galaxy	Analyze Data Workflow Shared Data - Visualization Cloud - Help - Use	- Using 0 bytes
Tools	Login	History C 🌣
search tools	Username / Email Address:	search datasets
<u>Get Data</u> <u>Lift-Over</u>	Password:	Unnamed history 0 bytes
Inte-Over Text Manipulation Convert Formats Filter and Sort Join, Subtract and Group NGS: QC and manipulation NGS: Mapping NGS: RNA-seq NGS: SAMtools NGS: BAM Tools NGS: Picard NGS: VCF Manipulation Extract Features Fetch Sequences Fetch Alignments Get Genomic Scores	Forgot password? Reset here Login OpenID Login OpenID URL:	This history is empty. You can load your own data or get data from an external source
	Or, authenticate with your Google + account. Login Terms and Conditions for use of this service	

Figure 3: Log into your Galaxy account with your credentials.

1.4 Create a new history

Create a new history (see Figure 4) and rename it to something useful (see Figure 5).



Figure 4: Log into your Galaxy account.



Figure 5: Log into your Galaxy account with your credentials.

1.5 Loading the data

We are going to use some Shared Data from the Galaxy Demonstration dataset.

- 1. Click on the *Shared Data* tab (see *Figure 6*).
- 2. Search for the Demonstration Datasets. (see Figure 7)



Figure 6: Load shared data tab.

🗧 Galaxy	Analyze Data	Workflow	Shared Data +	Visualization •	Cloud -	Help -	User +		Using 0%
Data Libraries									
search dataset name, info, message,	dbkey Q								
Advanced Search									
Data library name↓		Data li	brary descriptio	n					
1000 Genomes		Data fr	om the 1000 Ger	nomes Project FTP	site				
AC-exome									
Bushman		Data fo	or two papers abo	out the Khoisan an	d other pop	oulations.			
Charts Example Data									
ChIP-Seq Mouse Example		Data u	sed in examples	that demonstrate	analysis of	ChIP-Seq	data		
Chobi									
CloudMap		Contai	ns userguide, ref	erence files, and c	onfiguratio	n files for	the Cloudm	ap WGS analysis	s pipeline
Codon Usage Frequencies									
Coleman		IonPGN	4						
Datalmport-00107dec-86f4-44f6-at cb20d69c2fe9@createprivatelibrarye	f87– xample.com								
Datalmport-ff2b1cbd-ded0-41b8-at 741f1534ceb8@createprivatelibrarye	f <u>db-</u> xample.com								
Demonstration Datasets		Demor	nstration datasets	collected from va	rious Galax	y tutorials	5		
Denisovan sequences		Files fr 2012 a	om 'A high-cove and basic process	rage genome sequ ed data.	ence from	an archaic	Denisovan	Individual" Mey	er et al. Science
Erythroid Epigenetic Landscape		Dynam	ics of the epigen	etic landscape dur	ing erythro	id differer	ntiation afte	r GATA1 restora	ation

Figure 7: Look for the Demonstration Datasets.

Load the following 4 files: G1E CTCF, G1E_ER4 CTCF, G1E ER4 input, G1E input (see Figure 8).

📮 Galaxy 🛛 🗛	nalyze Data Workfic	w Shared Data -	Visualization -	Cloud +	Help +	User -	200	Using 0%
Data Library "Demons	tration Data	sets"						
🖸 Name	Message				1	Data type	Date uploaded	File size
Human RNA-seq: CHB ENCODE Exercise	■ Data on h1-hESC Sequencing Works	and CD20 produced hops	by ENCODE and u	sed by the (СНВ			
Mouse ChIP-seq: G1E CTCF binding	Sample datasets f	om Hardison lab for	ChIP-seq analysis					
GIE CTCF (chr19) -					f	astqsanger	Sun Jan 11 18:38:58 2015 (UTC)	29.2 MB
G1E_ER4 CTCF (chr19) →					f	astqsanger	Sun Jan 11 18:38:57 2015 (UTC)	29.2 MB
G1E_ER4 input (chr19) -	None				ſ	astqsanger	Sun Jan 11 18:38:58 2015 (UTC)	3 17.1 MB
G1E input (chr19) 🗸					f	astqsanger	Sun Jan 11 18:38:59 2015 (UTC)	28.5 MB
For selected datasets: Import to	current history 🗘	Go						



Once the files are loaded we can switch back to the analysis window by clicking *Analyze Data* tab (see *Figure 9*). We should find four datasets in the history panel (see *Figure 10*).

n Galaxy An	alyze Data Workflow Shared Data	 Visualization - Cloud - 	Help 🗸 User 🕶	888	Using 0%
Data Library "Demonst 4 datasets imported into 1 history: Ch	ration Datasets"				
Name	Message		Data type	Date uploaded	File size
Human RNA-seq: CHB ENCODE Exercise	Data on h1-hESC and CD20 produce Sequencing Workshops	ed by ENCODE and used by the	СНВ		
Mouse ChIP-seq: G1E CTCF .	Sample datasets from Hardison lab	for ChIP-seq analysis			
G1E CTCF (chr19) 🗸			fastqsanger	Sun Jan 11 18:38:58 2015 (UTC)	29.2 MB
G1E_ER4 CTCF (chr19) -			fastqsanger	Sun Jan 11 18:38:57 2015 (UTC)	29.2 MB
G1E_ER4 input (chr19) -	None		fastqsanger	Sun Jan 11 18:38:58 2015 (UTC)	17.1 MB
G1E input (chr19) 🕶			fastqsanger	Sun Jan 11 18:38:59 2015 (UTC)	28.5 MB
For selected datasets: Import to c	urrent history 💠 Go				

Figure 9: Load the datasets.

🗧 Galaxy		Analyze Data	Workflow	Shared Data -	Visualization -	Cloud - Help -	Jser v		Using 0%
Tools	1	Calannia		ourse web bas	ad platform for	data intensiva	History		200
search tools	0	biomedica	al research	. If you are new	to Galaxy start	here or consult our	search	n datasets	(
Get Data Upload File from your comp	uter	help reso	urces.				ChIP 4 shown 103.9 M	в	S B 9
UCSC Archaea table browser EBI SRA ENA SRA	e.						4: G1E in	n <u>put (chr19)</u> R4 input (chr19)	• / :
<u>BioMart</u> Central server <u>GrameneMart</u> Central server				Try G	ialaxy		2: G1E_E	R4 CTCF (chr19)	• *
Elymine server modENCODE fly server				onthe	Cloud		<u>1: G1E C</u>	<u>TCF (chr19)</u>	• / :
modENCODE modMine serve	er								

Figure 10: Loaded data in history panel

Alternatively, you can download the data chipdata.zip or (~40MB) from http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/data/chipdata.zip, unzip it and upload the files to the Galaxy history.

Hint! Should you need to refresh how to upload data to Galaxy, have a look at the Galaxy introductory tutorial (http://sschmeier.github.io/bioinf-workshop/#!galaxy-intro/)).

1.6 Investigate the data

The four files that we have now in our history are: G1E CTCF, G1E_ER4 CTCF, G1E ER4 input, G1E input. A closer look reveals that they are in fastq-sanger format (see *Figure 11* and *Figure 12*).

🗧 Galaxy	Analyze Data Workflow Shared Data + Visualization + Cloud + Help + User	-	Using 0%
Tools		History	200
search tools	This dataset is large and only the first megabyte is shown below. Show all Save	search datasets	0
Get Data		ChIP	
Upload File from your computer	@HWUSI-EAS610:2:1:4:959#0/1	4 shown	
UCSC Main table browser	GATTAATTGGCTGGCGGGCACATACAGGGGCTC +	103.9 MB	8 % 9
UCSC Archaea table browser	>1A@98BBABA@@@<3=):?0928967779;:&-9; @HWUSI-EAS610:2:1:4:677#0/1	4: G1E input (chr19)	• * ×
EBI SRA ENA SRA	AGGAACTCGAGGCAGNAACTGAAGCAGGAACACAGC	28.5 Mb	View data
BioMart Central server	+ BCCBBCCBCCCCBA=\$9AABCBBCA7BBBE>E>AAA	format: fastqsanger	database: mm9
GrameneMart Central server	<pre>@HWUSI-EAS610:2:1:4:1092#0/1</pre>	802	• •
Elymine server	+	@HWUSI-EAS610:2:1:4:9	59#0/1
modENCODE fly server	29@B;749@=1.5<4==/)/7?BB@5>5<=9027## @HWUSI-EAS610:2:1:5:1009#0/1	GATTAATTGGCTGGCTGTGGG	GACATACAGGGGCTC
modENCODE modMine server	AAACGTTGTCTTTCTGTCATAGTACTGTGGACTTAA	>1A098BBABA000<3=):?0	928967779;:&-9;
MouseMine server	+ =>CB?ABB@BA <abbb==8@<>>36>:@5@=?=<3?@</abbb==8@<>	@HWUSI-EAS610:2:1:4:6 AGGAACTCGAGGCAGNAACTG	77#0/1 AAGCAGGAACACAGC
<u>Ratmine</u> server	@HWUSI-EAS610:2:1:5:166#0/1 GATTGGAGATATGGTTGGTGGCCCAGCTGT	3: C1E ER4 input (chr	19)
YeastMine server	+	J. GAL_ERY INPACTON	
modENCODE worm server	BB??BBAAAA?BB>>7=?AA?9?==??<7>4?;1>5 @HWUSI-EAS610:2:1:5:1797#0/1	2: GIE_ER4 CTCF (chr	19) 👁 🖋 🗙
WormBase server	GAACATTTGGAGAAAAGGGTTTATTTCAAATTATAC	1: G1E CTCF (chr19)	• / ×
ZebrafishMine server	+		

Figure 11: Information about the data.

FastQ-form	nat		
			Sequence id
@EAS139:136:FC GATTTGGGGTTCAA	706VJ:2:2104:15 AGCAGTATCGATCA	5343:197393 1:Y:18:ATC	Sequence
+ ''*((((***+))%	%%++)(%%%%).1**	**-+*''))**55CCF>>>>	
Phred quality			Phred quality of the corresponding nucleotide (ASCII code)
One ASCII	character per nu	ucletide.	
Encodes for	r a quality $Q = -$	10*log ₁₀ (P), where P is t	he error probability
	The Relationship Betweer	n Quality Score and Base Call Accuracy	
	Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
-10*log ₁₀ (0.1) =	10 (Q10)	1 in 10	90%
	20 (Q20)	1 in 100	99%
	30 (Q30)	1 in 1000	99.9%
			Sebastian Schmeier



What we are looking at is data from the G1E mouse cell-line from Gata1-null mouse embryonic stem cells. We are looking at two conditions, normal G1E cell-lines and G1E-ER4 cell-lines, where the effect of Gata1 deletion is restored. Under both conditions Ctcf has been ChIP'ed and sequenced. The "input" samples denote samples where the DNA was fragmented but before the immunoprecipitation against Ctcf and can thus be used as controls. We

are also only looking at a subset of the full dataset, only chr19. Thus, we can compare the CTCF occupancy between G1E and G1E-ER4 cell-lines.

Note! **TODO**:

- 1. Find out what **Ctcf** is.
- 2. Find out why studying Gata1 in mouse embryonic stem cells is of interest?

Hint! You can use NCBI gene or wikigenes or even wikipedia to find out about Ctcf and Gata1.

Attention! **Before you go further**: Step 1.7 up to step 1.9 is the same process for each fastq-file individually. Thus, it might be a good idea to do the steps for **ONE** file and then create a workflow out of the steps that allows you to do all of the steps for each of the other three files in one go. This will significantly speed up the processing time. Should you not remember how to create a workflow, see Galaxy introductory tutorial (http://sschmeier.com/bioinf-workshop/#!galaxy-intro/)..

1.7 Quality assessment

Now we need to assess the quality of the reads in each sample and filter and quality trim the reads if necessary.

1.7.1 FastQC

First, we run FastQC on each sample to get a feel for the overall quality of the data (see *Figure 13*).





Have a look at the HTML result page. Depending on what the results are you might want to do some filtering and quality trimming.

Note!

TODO: Run FastQC on all four files and investigate the quality. Note for each sample the nucleotide number where the quality markedly drops.

```
Sebastian Schmeier
```

1.7.2 Read filtering

Here, we want to get rid of all reads that are of low quality. This strongly depends on your definition for "low quality". In the figure below the default values are used (see *Figure 14*). The *Quality cut-off* value is 20 and 90% of all nucleotides of the read need to be equal or above this cut-off value to be accepted.

📮 Galaxy	Analyze Data Workflow Shared Data - Visualization - Cloud - Help -	User -
Tools	Filter by quality (Galaxy Tool Version 1.0.0)	ptions
NGS: QC and manipulation FastQC Read Quality reports Select high quality segments Build base quality distribution Draw quality score boxplot Quality format converter (ASCII- Numeric) Filter by quality FASTQ to FASTA converter Remove sequencing artifacts	Library to filter Library to filter 1 1 2 Percent of bases in sequence that must have quality equal to / higher than cutvalue 90 V Execute	-off
Barcode Splitter Clip adapter sequences Collapse sequences Draw nucleotides distribution chart Compute quality statistics Rename sequences	 What it does This tool filters reads based on quality scores. Using percent = 100 requires all cycles of all reads to be at least the quality cut-or value. Using percent = 50 requires the median quality of the cycles (in each read) to be least the quality cut-off value. 	off at

Figure 14: Filtering reads of bad quality.

Furthermore, I edited the dataset name (1) to keep track the kind of data (see *Figure 15*). The original name I copied into the notes field (2), however it is not strictly necessary as the information from which dataset this one was derived is still available when clicking the info button (see *Figure 16*). Finally, I renamed the dataset to something useful (3, see *Figure 15*).

📲 Galaxy	Analyze Data Workflow Shared Data + Visualization + Cloud + Help + User+		Using 0%
Tools 1	Attributes Convert Format Datatype Permissions	History	C 🕈 🗆
NGS: Mapping		search datasets	0
<u>Bowtie2</u> – map reads against reference genome <u>BWA</u> – map short reads (< 100 bp) against reference genome	Edit Attributes Name: C1E input filtered Info:	Copy of 'ChIP' (active it 6 shown, 4 <u>deleted</u> 138.6 MB	ems only)
<u>BWA-MEM</u> - map medium and long reads (> 100 bp) against reference genome <u>Parse blast XML output</u> <u>Megablast</u> compare short reads against htgs, nt, and wgs	Quality cut-off: 20 Minimum percentage: 90 Annotation / Notes: Filter by quality on data 4	10: Filter by quality on 4 21.9 MB format: fastqsanger, da Quality cut-off: 20 Minimum percentage: 9	data' @ / X Edit attributes
databases <u>Map with BWA for Illumina</u> Map with Bowtie for Illumina	Add an annotation or notes to a dataset, annotations are available when a history is viewed. Database/Build: Mouse July 2007 (NCBI37/mm9) (mm9)	Input: 276733 reads. Output: 212756 reads. discarded 63977 (23%) reads.	low-quality
Lastz map short reads against reference sequence	Save	802	
NGS: RNA-seq NGS: SAMtools NGS: BAM Tools NGS: Picard NGS: VCF Manipulation	Auto-detect This will inspect the dataset and attempt to correct the above column values if they are not accurate.	<pre>etmUSI-EAS610:2:1:4:677 AGGAACTCGAGGCAGNAACTGAA + BCCBBCCBCCCCBA-N9AABCGBB eHMUSI-EAS610:2:1:5:100 AAACGTTGTCTTCTGTCATAGT</pre>	#0/1 GCAGGAACACAGC CA?BBB®>®>AAA 9#0/1 ACTGTGGACTTAA

Figure 15: Rename the dataset to keep track.

📲 Galaxy	Ana	lyze Data Workflow Shared Data - Visualization - Cloud - Help -	- User-		Using 0%
Tools	Tool: Filter by qua	lity		History	2 ¢ 🗆
NGS: Mapping	Name:	G1E input filtered		Canada data anda	0
Bowtie2 - map reads against	Created:	Tue Sep 22 01:23:37 2015 (UTC)		.search datasets	0
reference genome	Filesize:	21.9 MB		Copy of 'ChIP' (active iten	is only)
RWA - man short reads (< 100	Dbkey:	mm9		6 shown, 4 deleted	
bp) against reference genome	Format:	fastqsanger		138.6 MB	
op) against reference genome	Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/devteam/fastq_quality_filter/cshl_fastq_qua	lity_filter/1.0.0		
BWA-MEM - map medium and	Galaxy Tool Version:	1.0.0		10: G1E input filtered	● / ×
long reads (> 100 bp) against	Tool Version:			21.0 MR	
reference genome	Tool Standard Output:	stdout		format: fastosander data	hase mm9
Parse blast XML output	Tool Standard Error:	stderr		ionnat insidsninger, cara	oure mins
Megablast compare short reads	Tool Exit Code:	0		Quality cut-off: 20	
against htgs, nt, and wgs	API ID:	bbd44e69cb8906b505edda71e814f4ab		Minimum percentage: 90	
databases	History ID:	62fc6065d03a7359		Input: 276733 reads.	
Map with BWA for Illumina	UUID:	6d8cd290-e71f-4f23-979e-b7de3bc731e4		discarded 63977 (23%) lov	w-quality
Map with Bowtie for Illumina	Input Parameter	Value	Note for rerun	reads.	
Lastz man short reads against	Library to filter	4: G1E inpu	ut (chr19)	2 0 C	۰ ک
reference sequence	Quality cut-off value	20		PUNCT FIFEID. 3. 1. 1. 67740	
NGS: RNA-seq	Percent of bases in seq	uence that must have quality equal to / higher than cut-off value 90		AGGAACTCGAGGCAGNAACTGAAGC	AGGAACACAGC
NGS: SAMtools	Inheritance Ch	ain		+	
NGS: BAM Tools			1	BUCBBUCBUCCCBA-#9AABCBBCA	ruuue>e>AAA
NGS: Picard		G1E input filtered		PHWU51-EA5610:2:1:5:1009#	5/1
the second s				AAACGITGICITTCIGICATAGIACI	GIOGACITAA

Figure 16: Detailed information about a dataset can be gathered by clicking the info button.

Note!

TODO: Run the filtering on all four files and note how many reads got excluded for each sample (see the next section on how to speed this process up by re-running analyses).

Re-running an analysis

Click on the re-run button of the analysis (1) you woud like to re-run (see *Figure 17*). The parameter window pops up with all the original parameters used. Now you can select a different dataset (2) and run the original analysis with the same parameters (see *Figure 17*).

🗧 Galaxy	Analyze Data Wor	rkflow Shared Data + Visualization + Cloud + Help + User +		Using 0%
Tools	Filter by quality	(Galaxy Tool Version 1.0.0)	History	200
Clip adapter sequences	Library to filter		search datasets	0
Collapse sequences	0 2 0	4: G1E input (chr19) 2	ChIP	
Draw nucleotides distribution chart	Quality cut-off v	Q	6 shown, 4 <u>deleted</u> 149.0 MB	C 🗞 🗩
Compute quality statistics	20	10: G1E input trimmed	10: G1E input trimmed	
Rename sequences	Percent of bases	9: GIE input filtered 4: GIE input (chr19)	9: G1E input filtered	
Reverse-Complement	90	3: G1E_ER4 input (chr19)	21.9 MB	
Trim sequences		2: G1E_ER4 CTCF (chr19)	format: fastqsanger, dat	tabase: mm9
Combine FASTA and QUAL into FASTQ	✓ Execute	1: G1E CTCF (chr19)	Filter by quality on data	4
Filter FASTQ reads by quality score and length <u>Manipulate FASTQ</u> reads on various attributes FASTQ Groomer convert between	What it does This tool filters read Using percent = value.	s based on quality scores. 100 requires all cycles of all reads to be at least the quality cut-off 50 requires the median quality of the cycles (in each read) to be at	Quality cut-off: 20 Minimum percentage: 9 Input: 276733 reads. Output: 212756 reads. discarded 63977 (23%) I reads.	0 Iow-quality
FASTO Masker by quality score	least the quality	cut-off value.	B 0 C 1	•
FASTQ joiner on paired end reads FASTQ splitter on joined paired	Quality score distrib quality cut-off value Example:	ution (of all cycles) is calculated for each read. If it is lower than the - the read is discarded.	<pre>@HMUSI-EAS610:2:1:4:677# AGGAACTCGAGGCAGNAACTGAAC + BCCBBCCBCCCCCBA=%9AABCBBC</pre>	10/1 GCAGGAACACAGC CA?BBB®>®>AAA
end reads	BCSHL 4 FC042AGO	DTT+1+2+214+584	@HWUSI-EAS610:2:1:5:1009	##0/1 ACTGTGGACTTAA
column	GACAATAAAC		A: C1E input (chr10)	
FASTQ to FASTA converter	+CSHL_4_FC042AGOO 30 30 30 30 30 30 30	0 30 30 20 10	H. GIL INPUT (CIT 15)	• * *
FASTQ to Tabular converter	Using percent = 50	and cut-off = 30 - This read will not be discarded (the median quality	3: G1E_ER4 input (chr19) • # ×
FASTQ Trimmer by column	is higher than 30).		2: G1E_ER4 CTCF (chr19) • # ×
FASTQ Quality Trimmer by sliding window	Using percent = 90 no have quality equa	and cut-off = 30 – This read will be discarded (90% of the cycles do al to / higher than 30).	1: G1E CTCF (chr19)	• # ×

Figure 17: Re-run button to re-run the same analysis.

1.7.3 Quality trimming

Finally, we can use a quality trimmer to get rid of bad starts and ends of reads (see *Figure 19*). To do so, select the *FASTQ Quality Trimmer* (1). Choose the **filtered** dataset from the step before (2). In *Figure 19* I use a simple window size of 1 (3) and a quality score of 20 (4) to just trim of the ends on both sides.

🚍 Galaxy	Analyze Data Workflow Shared Data + Visualization + Cloud + Help + User +		Using 0%
Tools	FASTQ Quality Trimmer by sliding window (Galaxy Tool Version 1.0.0) • Options	History	2 � □
Clip adapter sequences	FASTQ File	search datasets	0
Collapse sequences	🖸 🔁 🗅 9: G1E input filtered 🔸	ChiP	
Draw nucleotides distribution chart	Keep reads with zero length	6 shown, 3 <u>deleted</u> 127.1 MB	8 % 9
Compute quality statistics	Trim and	9: G1E input filtered	• / ×
Rename sequences Reverse-Complement	S' and 3'	5: FastQC on data 4: Web page	• / ×
Trim sequences	Window size	4: G1E input (chr19)	• / ×
Combine FASTA and QUAL into FASTQ	1	3: G1E_ER4 input (chr19)	• / ×
Filter FASTQ reads by quality score and length	Step Size	2: G1E_ER4 CTCF (chr19)	• # ×
Manipulate FASTQ reads on various attributes	Maximum number of bases to exclude from the window during aggregation	1: G1E CTCF (chr19)	⊛ # ×
FASTQ Groomer convert between	0		
various FASTQ quality formats	Aggregate action for window		
FASTQ Masker by quality score	min score 🔹		
reads	Trim until aggregate score is		
FASTQ splitter on joined paired end reads			
FASTO Summary Statistics by column	Quality Score		
FASTQ to FASTA converter			
FASTQ to Tabular converter	▼ Execute		
FASTQ Trimmer by column FASTQ Quality Trimmer by sliding window	This tool allows you to trim the ends of reads based upon the aggregate value of quality scores found within a sliding window; a sliding window of size 1 is equivalent to 'simple' trimming of the ends.		

Figure 19: Filtering reads of bad quality.

Note!

TODO: Run the quality trimmer on all *filtered* datasets and rename the sets to something meaningful.

Attention! Trimming reads is not always necessary or desired. Here, we do it to see how the trimming process works in Galaxy. However, in a real situation we might decide not to trim at all.

1.8 Mapping reads

By know we should have 4 sets of filtered and trimmed reads with a meaningful name (see *Figure 19*). These form the basis for the subsequent analyses. Now we are going to map the reads to the reference genome.



Figure 19: The datasets for mapping.

1.8.1 Bowtie2

We can now map the trimmed data to the reference genome using Bowtie2. Select Bowtie2 in the tools panel under section NGS: Mapping (1, see Figure 19). We select the trimmed dataset we want to map (2) and select an appropriate reference genome (3).

Tots Percentiz2 - map reads against reference genome (Galaxy Tool Version 04) Version * Options* version search against reference genome NGS. Manghilding Bandust reference genome Is this single or paired library Single-end NGS. Manghilding Bandust reference genome Is this single or paired library Single-end NGS. Manghilding Bandust reference genome Is charact statis NGS. Manghilding Press Is charact statis NGS. MANG Colliminal Is charact statis Maas with Row Ka Colliminal Seccre Centeroce genome Seccre Centeroce genome Seccre Centeroce Seccre Centeroce genome	🚍 Galaxy	Analyze Data Workflow Shared Data + Visualization + Cloud + Help + User +		Using 0%
NGX-Mapping BMX-mapping BMX-mapping </td <td>Tools 📩</td> <td>Bowtie2 - map reads against reference genome (Galaxy Tool Version 0.4) & Versions • Options</td> <td>History</td> <td>€ ♥ 🗆</td>	Tools 📩	Bowtie2 - map reads against reference genome (Galaxy Tool Version 0.4) & Versions • Options	History	€ ♥ 🗆
Bottle2 - map reads against reads (a 100 bb) against reference genome Singleend BXMA - map notit meads (a 100 bb) against reference genome Do CIE input trimmed BXMA - map notit meads (b) against reference genome Nucleating against reference genome Base Maax MM, output Nucleating against reference genome Base Maax MM, output Mapasitis RMA, for, illumina Mapasitis RMA, for, and way: databases Mapasitis RMA, for, illumina Mapasitis RMA, for, illumina </td <td>NGS: Mapping</td> <td>Is this single or paired library</td> <td>search datasets</td> <td>0</td>	NGS: Mapping	Is this single or paired library	search datasets	0
BXAmap short reads (x 100 bb) against reference genome reference genome gainst htps, nt, and wegs databases I :: Gil: Erget trimmed integress (x 100 bp) against reference genome intervences gainst htps, nt, and wegs databases I :: Gil: Cil: input trimmed integress (x 100 bp) against reference genome intervences intervences gainst htps, nt, and wegs databases I :: Gil: Erget input trimmed intervences gainst htps, nt, and wegs databases I :: Gil: Cil: input trimmed intervences genome index I :: Gil: Erget input trimmed intervences genome index I :: Gil: Erget input trimmed intervences genome index Map with BWA for Illumina Map with BWA for Illumina Lit Status sterim of with status Di Status to use presets? Deceare Deceare Consent Intervals BWA for Illumina Devery sensitite end -to-end (enstity) Destatus end -to-	Bowtie2 - map reads against reference genome	Single-end •	ChIP	
BWA-MMM - map medium and long reads (>> 100 bp) splainst reference genome Wite unaligned reads (in fastq format) to separate file(s) Parse blast XML suituit megabilist on parse short reads against htps, m, and was databases Wite unaligned reads (in fastq format) to separate file(s) Map with BWA for Illumina Lasz map short reads against reference sequence Wite value a built-in genome index Map with BWA for Illumina Lasz map short reads against reference sequence Wile value interest is not listed, contact the Galaxy team NGS. RMA-reads CS. SAMIOSIS Select reference genome 3 NGS. RMA-reads CS. SAMIOSIS Select reference genome 3 NGS. RMA-reads CS. SAMIOSIS Select reference genome 3 NGS. RMA-reads CS. SAMIOSIS Specifying read group for this file? 2 NGS. RMA-reads CS. SAMIOSIS Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets: Se help below more details 3 NGS. RMA Tools Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple datasets: Se help below more details 3 NGS. RMA Tools Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple datasets: Se help below more details 3 Call ER4 ctrCF chrlig? ? X Scille ER4 input thrimsed	<u>BWA</u> – map short reads (< 100 bp) against reference genome	FASTQ file	12 shown, 4 <u>deleted</u> 264.6 MB	2 % 9
Parse blait XML output No Magabilits Compare short reads against hos, n.r. and was databases No Magabilits tompare short reads against hos, n.r. and was databases Will you select a treference genome from your history or use a built-in index? Magabilits hos, n.r. and was databases Will you select a treference genome from your history or use a built-in index? Magabilits hos, n.r. and was databases Will you select a treference genome index Magabilits hos, n.r. and was databases Will you select a treference genome index Magabilits hos, n.r. and was databases Will you select a treference genome index Magabilits hos, n.r. and was databases Will you select a treference genome index Mata with Boxel for illimina Mouse Mus musculus): mm9 Canonical Moss: BMA Teols Mouse Mus musculus): mm9 Canonical Mos: Extract Features Select reference genome Extract Features Select hole reads on greatly simplify your downstream analyses by allowing combining multiple datasets. See help below for more details Statistics Do you want to use prests? Operators of commic Intervals Operators of commit Intervals Statistics Overy sate that use prests parameter settings. Choosing between these will result in dramatic changes in rusting. Setting of (-very-sensitive local) Operstaes of commic Inte	<u>BWA-MEM</u> - map medium and long reads (> 100 bp) against	Nucleotide-space. Must nave sanger-scaled quality values with ASLII offset 3.3	16: GIE ER4 input trimmed	● / X
Image Nation Coupon Megability Course, Not, and was databases Image Number Not, not, and was databases Image Number Not, not, mits triggersum parameter for single reads andum-conc for paired reads Map with BVA for Illumina Will you select a reference genome Irlow, See 'Indexes' section of help below 3 Map with BVA for Illumina Select reference genome 3 Map with BVA for Illumina Select reference genome 3 Map with BVA for Illumina Select reference genome 3 Most RMA-seq Mouse (Mus musculus): mm9 Canonical If your genome of interest is not listed, contact the Galaxy team 5 NGS: BMI Tools Specify the read group for this file? 5 NGS: Pleard No Specify the read group for the file? 5 NGS: Pleard Specify the read group for the file? 5 NGS: Pleard Specify the read group information can greatly simplify your downstream analyses by allowing combining multiple datasets. See thelp below for more details 5 Cast Genomic Corres On you want to use presets? On you want to use presets? 5 Op you want to use presets? On you want to use presets? Sect analysis mode 2 Cast Genomic Chineraals Sectify analysis mode 2 2 2 2	Parra blact VML output	Ves No	15: GIE_ER4 CTCF trimmed	• / ×
sapalist higs, nf, and wgs databases Will you select a reference genome from your history or use a built-in index? Max with BWA for Illumina Use a built-in genome index Max with Boatie for Illumina Select reference genome Max with Boatie for Illumina Select reference genome Select reference squaree Mouse (Mus muculus): mm9 Canonical If your genome of interest is not listed, contact the Galaxy team SS 25Mitools NGS: RMA-seq NGS: Plant Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple databases Secta Genomic Scores Op you want to use presets? Op you want to use presets? Op you want to use presets? Overy fast thet-local (fast) Orvery fast theta (enstitive) Overy fast theta (enstitive) Overy fast theta (enstitive) Overy fast theta (enstitive) Overy fast theta (enstitive)-tota) Overy fast theta (enstitive)-tota) Overy fast theta	Megablast compare short reads	un/un-conc; This triggersun parameter for single reads andun-conc for paired reads	14: G1E CTCF trimmed	@/×
Map. with BWA for Illumina Use a built-in genome index Lastz map short reads against reference sequence Built-ins were indexed using default options. See 'Indexes' section of help below 3 Lastz map short reads against reference sequence Mowse (Mus musculus): mm9 Cannical It If your genome of interest is not listed, contact the Galaxy team 9. GIE input filtered * X NGS. BMA tools Specify the read group for this file? * * NGS. Pricard Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple * X KSX EPK mainpulation Select analysis mode * X * X Select analysis mode 1: Default setting only 4 * X: GIE ER4 CTCF (chr.19) * X No. Select analysis mode 1: Default setting only * X X: GIE ER4 CTCF (chr.19) * X Op you want to use presets? Ono, just use defaults Ono, just use defaults Sensitive and-to-end (sensitive) Sensitive and-to-end (sensitive) * X: GIE ER4 CTCF (chr.19) * X Sensitive local (verysensitive) Overy use sensitive end-to-end (verysensitive) Sensitive and-to-end (verysensiti	against htgs, nt, and wgs	Will you select a reference genome from your history or use a built-in index?	13: GIE CTCF filtered	• * ×
Max with Bouxter Munitive Control Minima Lastz map short reads against reference sequence MGS: MALE sequences Sector reference sequence MGS: SANAL reads Mouse (Mus musculus): mm9 Canonical If your genome of interest is not listed, contact the Galaxy team Specify the read group for this file? NGS: RAM. Tools NGS: YCH Manipulation Castats: See help below for more details Sector reads of the sector sector Sector reads of the sector Sector reads of the sector sector	Man with BWA for Illumina	Use a built-in genome index -	12: GIE_ER4 CTCF filtered	@ / X
Lastz map short reads against reference sequence Mouse (Mus musculus): mo9 Canonical Image: Musculus): mo9 Canonical MGS: RMA-seq Mouse (Mus musculus): mo9 Canonical Image: Musculus): mo9 Canonical MGS: RMA-seq Specify the read group for this file? NGS: RAM Tools Specify the read group information can greatly simplify your downstream analyses by allowing combining multiple MGS: VCF Manipulation Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple MGS: VCF Manipulation Select analysis mode 1: Default setting only Image: Calority information Cal Canonic Intervals Oxy us want to use presets? Operate on Genomic Intervals Oxy y tast indu-to-end (verysensitive) Statistics Overy y sensitive end-to-end (verysensitive) Sector (-very- sensitive) Overy sensitive end-to-end (verysensitive) Overy y sensitive end-to-end (verysensitive) Overy fast local (resensitive) Sector (-very- sensitive end-to-end) Overy sensitive end-to-end (verysensitive) Overy y sensitive end-to-end (verysensitive) Overy fast local (resensitive) Overy y sensitive local (very-fast-local) Overy sensitive end-to-end (verysensitive) Overy y sensitive local (very-fast-local)	Map with Bowtle for Illumina	Built-ins were indexed using default options. See 'Indexes' section of help below Select reference genome	11: G1E ER4 input filtered	* / ×
NGS: RNA-seq If your genome of interest is not listed, contact the Galaxy team NGS: RAA-seq Specify the read group for this file? NGS: RAM Tools No NGS: PLANT Tools Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple datasets. See help below for more details Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple datasets. See help below for more details Stract Features Select analysis mode 1: Default setting only Image: Comparison of the end (very-tast) Opparate on Genomic Intervals O you want to use presets? Oparate on Genomic Intervals Orey tast the -deta(very-tast) Graph/Display Data Osensitive end-to-end (very-tast) Phenotype Association Sensitive end-to-end (very-sensitive) Setor fail Overy fast tool-(-cer) -sensitive) Very sensitive coal (very-sensitive) Overy sensitive end-to-end (very-sensitive) Overy fast tool (very-sensitive) Overy sensitive end-to-end (very-sensitive) Overy fast tool (very-sensitive) Overy sensitive end-to-end (very-sensitive) Overy fast tool (very-sensitive-local) Sensitive local (very-sensitive-local) Genome Diversity Overy sensitive end (very-se	Lastz map short reads against reference sequence	Mouse (Mus musculus): mm9 Canonical +	10: G1E input trimmed	• / ×
NGS: SAMtools Specify the read group for this file? NGS: Pland No NGS: Pland Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple NGS: VET Manipulation datasets. See help below for more details Stract Features Select analysis mode Fetch Alignments 1: Default setting only Operate on Genomic Intervals 0 No, just use defaults Overy last end-to-end (fast) Graph/Display Data Overy fast local (mast-local) Overy sensitive end-to-end (sensitive-local) Genome Diversity Osensitive-local (sensitive-local) MBOSS Overy sensitive end-to-cand MSDS S Overy sensitive-local (sensitive-local) Matiple Alignments Job Resource Parameters	NGS: RNA-seq	If your genome of interest is not listed, contact the Galaxy team	9: GIE input filtered	
NGS: EAM Tools No NGS: Pland Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple datasets. See help below for more details NGS: VCF Manipulation Select analysis mode Attact Features. Select analysis mode Fetch Alignments So you want to use presets? Operate on Genomic Intervals No, just use defaults Statistics Overly tast end-to-end (fast) Graph/Display Data Sensitive end-to-end (fast) Phenotyce Association Sensitive end-to-end (very-sensitive) Super Diversity Sensitive local (very-fast-local) Genome Diversity Sensitive local (very-fast-local) Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Volution Job Resource Parameters	NGS: SAMtools	Specify the read group for this file?	A CIE Innut (sha10)	
NSS: Pleard Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple 3: G1E ER4 input (chr19) 2: X MGS: VCF Manipulation Select analysis mode 2: G1E ER4 input (chr19) 2: X Extract Features Select analysis mode 2: G1E ER4 input (chr19) 2: X Fetch Asiguments Do you want to use presets? 0 you want to use presets? 0 you want to use presets? Operate on Genomic Intervals O sensitive end-to-end (fast) Fast end-to-end (fast) 0 Yery sastitive end-to-end (reast) Phenotype Association Sensitive end-to-end (reast) 0 Yery fast local (-very-fast) 0 Yery sastitive end-to-end (reast) Genome Diversity O Sensitive local (-fast-local) 0 Yery sensitive-local) 0 Yery sensitive-local) Genome Diversity O Sensitive local (-very-fast-local) 0 Yery sensitive-local) 0 Yery sensitive-local) Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Job Resource Parameters Job Resource Parameters Job Resource Parameters Job Resource Parameters Yery sensitive local (-very-sensitive-local)	NGS: BAM Tools	No	4: GTE input (cnr19)	• / ×
MGS: VCF: Manipulation datasets. See help below for more details Extract Features Select analysis mode 1: Default setting only 4 2: GLE ER4 CTCF (chr19) × × Fetch Sequences Do you want to use presets? Operate on Genomic Intervals Overy tast end-to-end (test) Statistics Overy fast end-to-end (ensitive) Graph/Diplay Data Overy fast end-to-end (very-resultive) Overy fast end-to-end (sensitive) Overy fast local (very-fast-local) Genome Diversity Sensitive end-to-end (very-ensitive) EMBOSS Overy sensitive local (very-sensitive) Allow selecting among several presets. Job Resource Parameters Job Resource Parameters Job Resource Parameters	NGS: Picard	Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple	3: G1E_ER4 input (chr19)	@ / X
Extract Features Fetch Alignments Cet Cenomic Scores Operate on Genomic Intervals Statistics Operate on Genomic Intervals Object of Commic Scores Operate on Genomic Intervals Statistics Overy tast end-to-end (fast) Overy sensitive end-to-end (fast) Overy sensitive end-to-end (fast) Overy sensitive end-to-end (sensitive) Statistics Overy sensitive end-to-end (fast) Overy sensitive end-to-end (sensitive) Sensitive end-to-end (fast) Overy sensitive end-to-end (sensitive) Sensitive ind-to-end (sensitive) Overy sensitive end-to-end (sensitive) Overy sensitive end-to-end (very-sensitive) Overy sensitive local (very-sensitive) Overy sensitive local (very-sensitive) Overy sensitive local (very-sensitive) Osensitive local (sensitive-local) Overy sensitive local (very-sensitive) Overy sensitive loc	NGS: VCF Manipulation	datasets. See help below for more details	2. CIE ERA CTCE (aba10)	
Fetch Alignments 1: Default setting only 4 • Fetch Alignments Do you want to use presets? • • Operate on Genomic Intervals • • • • Statistics • • • • • • Operate on Genomic Intervals •	Extract Features	Select analysis mode	2. GIE EN4 CICF (CIRIS)	• / ×
Fetch Alignments Gat Genomic Scores Do you want to use presets? Operate on Genomic Intervals ONO, just use defaults Statistics Overy tast end-to-end (fast) Oraph/Display Data Osensitive end-to-end (restrict) Phenotype Association Osensitive end-to-end (represensitive) Statistics Overy fast end-to-end (represensitive) Statistics Overy fast local (represensitive) Statistics Overy fast local (fast-local) Genome Diversity Osensitive end-to-cal) Genome Diversity Osensitive local (very-sensitive-local) EMBOSS Osensitive local (very-sensitive-local) Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Evolution Job Resource Parameters	Fetch Sequences	1: Default setting only 4	1: GIE CTCF (chr19)	@ / X
Value Control Scores Operate on Genomic Intervals Operate on Genomic Intervals No, just use defaults Overy tast end-to-end (very-fast) Oraph/Display Data Overy fast end-to-end (very-fast) Overy fast end-to-end (very-fast) Overy fast local (very-fast-local) Overy fast local (fast) Overy fast local (very-fast-local) Overy fast local (very-fast-local) Overy fast local (very-fast-local) Overy ensitive end (very-fast-local) Overy sensitive local (very-fast-local) Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Valutiple Alignments	Fetch Alignments	Do you want to use presents?		
Substitute O No., just use defaults O Ro, just use defaults Graph/Display_Data O Fast end-to-end (fast) Phenotype Association snpEff O Very snsitive end-to-end (very-ensitive) BEDTools Genome Diversity EMBOSS Multiple Alignments Job Resource Parameters	Get Genomic Scores	bo you want to use presets:		
Statistics Overy fast end-to-end (very-fast) Graph/Display Data OFast end-to-end (fast) Phenotype Association Sensitive end-to-end (very-sensitive) snpEff Overy fast local (very-fast-local) BEDTools Overy fast local (very-fast-local) Graph/Display Data Overy sensitive end-to-end (very-sensitive) BEDTools Overy fast local (very-fast-local) Genome Diversity Osensitive local (sensitive-local) EMBOSS Overy sensitive local (very-sensitive-local) Regional Variation Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Evolution Job Resource Parameters	Statistics	ONo, just use defaults		
Mitability State Orast end-u-end (tast) Phenotype Association O Sensitive end-u-end (tast) SinpEff O Very sensitive end-u-end (very-sensitive) BEDTools O Very fast local (very-fast-local) Genome Diversity O Sensitive (ocal) EMBOSS O Very sensitive local (sensitive-local) Regional Variation Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Evolution Job Resource Parameters	Granh / Dienlay Data	O very fast end-to-end (very-fast)		
Instruction Association Overy sensitive end-to-end (very-sensitive) SupEff Overy sensitive end-to-end (very-sensitive) BEDTools Overy fast local (fast-local) Genome Diversity O Sensitive local (fast-local) EMBOSS Osensitive local (very-sensitive-local) Overy sensitive local (very-sensitive-local) Overy sensitive local (very-sensitive-local) Regional Variation Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Koulution Job Resource Parameters	Phenotype Accoriation	Sensitive end-to-end (sensitive)		
Sinberti Overy fast local (very-fast-local) Genome Diversity Offast local (fast-local) Sensitive local (fast-local) Overy fast local (fast-local) MB0255 Osensitive local (very-fast-local) Regional Variation Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Evolution Job Resource Parameters	rienotype Association	Very sensitive end-to-end (very-sensitive)		
BED Tools O Fast local (fast-local) Genome Diversity O Fast local (fast-local) EMBOSS O Sensitive local (sensitive-local) Regional Variation O Very sensitive local (very-sensitive-local) FASTA manipulation Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Evolution Job Resource Parameters	Shperi	Overy fast local (very-fast-local)		
Cenome Diversity O Sensitive local (sensitive-local) EMBOSS O Very sensitive local (very-sensitive-local) Regional Variation Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Evolution Job Resource Parameters	BED TOOIS	() Fast local (fast-local)		
EMBOSS Overy sensitive local (very-sensitive-local) Regional Variation Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Koultiple Alignments Job Resource Parameters	Genome Diversity	O Sensitive local (sensitive-local)		
Regional Variation Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets. Evolution Job Resource Parameters	EMBOSS	Overy sensitive local (very-sensitive-local)		
EASTA manipulation runtime. See help below to understand effects of these presets. Evolution Job Resource Parameters	Regional Variation	Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in		
Evolution Multiple Alignments Job Resource Parameters	FASTA manipulation	runtime. See help below to understand effects of these presets.		
Multiple Alignments Job Kesource Parameters	Evolution			
	Multiple Alignments	Job Resource Parameters		
Metagenomic analyses Use default job resource parameters -	Metagenomic analyses	Use default job resource parameters 🔹		
Motif Tools	Motif Tools	✓ Execute		

Figure 19: Bowtie2.

Using an inbuilt index choose the same genome built as your data is (here we are looking at mouse mm9 data). Choose the canonical index, **mm9 Canonical**.

From the Galaxy Bowtie tool description:

A Note on Built-in Reference Genomes

The default variant for all genomes is "Full", defined as all primary chromosomes (or scaffolds/contigs) including mitochondrial plus associated unmapped, plasmid, and other segments. When only one version of a genome is available in this tool, it represents the default "Full" variant. Some genomes will have more than one variant available. The "Canonical Male" or sometimes simply "Canonical" variant contains the primary chromosomes for a genome. For example a human "Canonical" variant contains chr1-chr22, chrX, chrY, and chrM. The "Canonical Female" variant contains the primary chromosomes excluding chrY.

Finally, we just use the default parameters of Bowtie2 (4) and execute the analysis (see Figure 19).

We can not look at the resulting data in detail, as the output is in a format called *bam* which is a binary version of the Sequence Alignment/Map (SAM) format (see http://genome.ucsc.edu/goldenpath/help/bam.html and <a href="http://genome.ucsc.edu/goldenpath/help/bam.html"



Figure 20: Bowtie2 mapping information.

Note!

TODO: Run Bowtie2 on each of the four trimmed datasets. Note for each sample the number of reads that could be aligned exactly once to the genome and the overall alignment percentage.

1.8.2 Post-mapping processing

First, we need to filter out multi-mapping reads. We will use samtools to do this. The important parameter here is the **Minimum MAPQ quality score** which should be set to 1, which will remove multi-mapping reads, as reads that multi-map will get a score of 0 (see *Figure 20b*).

= Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User -
Tools	Eilter SAM or PAM output SAM or PAM Size on ELAC MADO PC LN or by region (Colony Version 1.1.1)
NGS: DeepTools	
NGS: Mapping	SAM or BAM file to filter
NGS: RNA Analysis	□ 2 C 18: G1E CTCF bowtie2
NGS: SAMtools	
<u>Filter mapped reads</u> on MD tag string	Include Header -
Merge BAM Files merges BAM files together	Minimum MAPQ quality score
<u>SAM-to-BAM</u> convert SAM to BAM	1 (-q)
<u>Pileup-to-Interval</u> condenses	Filter on bitwise flag
bases	no
<u>MPileup</u> call variants	Select alignments from Library
<u>bcftools view</u> Converts BCF format to VCF format	(-1) Requires headers in the input SAM or BAM, otherwise no alignments will be output
<u>Reheader</u> copy SAM/BAM header between datasets	Select alignments from Read Group
Split BAM dataset on readgroups	(-r) Requires headers in the input SAM or BAM, otherwise no alignments will be output
<u>Stats</u> generate statistics for BAM dataset	Output alignments overlapping the regions in the BED FILE
<u>BAM-to-SAM</u> convert BAM to SAM	(-L) Nothing selected
Sort BAM dataset	Select regions (only used when the input is in BAM format)
CalMD recalculate MD/NM tags	
<u>BedCov</u> calculate read depth for a set of genomic intervals	region should be presented in one of the following formats: `chr1', `chr2:1,000' and `chr3:1000-2,000'
IdxStats tabulate mapping	BAM (-b)
Flagstat tabulate descriptive	✓ Execute
Slice BAM by genomic regions	What it does
<u>RmDup</u> remove PCR duplicates	This tool uses the samtools view command in <u>SAMtools</u> toolkit to filter a SAM or BAM file on the MAPQ (mapping quality), FLAG bits, Read
Filter pileup on coverage and SNPs	Input
Convert SAM to interval	Input is either a SAM or BAM file.
Filter SAM on hitwise flag values	Output
Filter SAM on bitwise hag values	The output file will be SAM or BAM (depending on the chosen option), filtered by the selected options.
<u>Generate pileup</u> from BAM dataset	Options
Filter SAM or BAM, output SAM	Filtering by read group or library requires headers in the input SAM or BAM file.
or BAM files on FLAG MAPQ RG LN or by region	If regions are specified, only alignments overlapping the specified regions will be output. An alignment may be given multiple times if it is overlapping several regions. A region can be presented, for example, in the following format:
Figure 20b: Samtools filte	ering.

Second, sort the output from the former step (see $\it Figure~20c$).

Tools ▲ Sort Sort BAM dataset (Galaxy Version 2.0) ● Options Filter and Sont BAM File Search datasets Sort data in ascending or descending order Sort data in ascending or descending order Sort data in ascending or descending order Since Status What it does This tool uses searcols sort command to sort BAM datasets in coordinate or read name order. ChP Since Status What it does This tool uses searcols sort command to sort BAM datasets in coordinate or read name order. ChP Since Status Chains: G Show BbTeX Definition of SAM/BAM format. [Link] Upto SAM or BAM on data 23 Image: Status CFF Extract features from CFF data Up A classet of Chain as a contigone sort (Link) Image: Status Image: Status Image: Status Link (C) CF data by fature court using simple expressions Link (2011). Impriving SNP discovery by base alignment quality. In Bioinformatics, 27 (B, p. 2078-2079. [doi:10.1093/Bioinformatics/Btri232III.nh] Image: Status Zie Extract Genomic D Image: X Stort SAM dataset NGS-Elcard Durbin, R Segregation based metric for variant call QC. [Link] Image: X Zie Compute on data 5 Zie Compute on data 5 X Stort SAM dataset by coordinate NGS-Elcard Sont SAM	🚍 Galaxy	Analyze Data Workflow Shared Data + Visualization + Help + User +	sort	
Sort ● Text Manipulation Unique lines assuming metassuming metasuming metassuming metassum	Tools	Sort BAM dataset (Galaxy Version 2.0) Options	History	S¢I
Text. Manipulation Unique lines assuming imaging in the sassuming imaging in the sassuming imaging or descending or descesending descending or descending or descendin	sort O	BAM File	search datasets	0
Values_list Durbin, R Segregation based metric for variant call QC. [Link] 70: Compute on data 6 3 Sort BAM dataset Li, H Mathematical Notes on SAMtools Algorithms. [Link] 3 63: Cut on data 62 * X NGS: Picard SontSam sort SAM/BAM updated dataset SontSam sort SAM/BAM updated dataset 60: Build custom track on data 27, data 25, an data 25, an data 55 an data 35 an data	sort C Text Manipulation Unique lines assuming new sorted input file Sort data in ascending or descending order Filter and Sort Sort data in ascending or descending order Select lines that match an expression Eilter data on any column using simple expressions GFF Extract features from GFF data Eilter GFF data by attribute using simple expressions Eilter GFF data by feature count using simple expressions Eilter GFF data by attribute	BAM File	Search datasets ChIP 34 shown, 40 deleted 310.33 MB C 74: Sort on data 73 73: Filter SAM or BAM. output SAM or BAM. IO.1 MB format: bam, database: r D C C Idd display at UCSC main display at Ensembl Currer display at Ensembl Currer display at Ensembl Currer display at Ensembl Currer display in IGB View. Binary ban alignments fi 72: Extract Genomic D NA on data 71 71: Get flanks on data 70	mm9
NGS: Picard SAMTools GitHub page. [Link] SortSam sort SAM/BAM updated dataset 60: Build custom track on data 27, data 25, an d others NGS: VCF Manipulation s8: Compare two Datas ets on data 55 and dat a 57	NGS: SAMtools Sort BAM dataset	Durbin, R Segregation based metric for variant call QC. [Link] Li, H Mathematical Notes on SAMtools Algorithms. [Link]	70: Compute on data 6 3 63: Cut on data 62	• / ×
NOS: VCF Manipulation 58: Compare two Datas VCFsort, Sort VCF dataset by coordinate 58: Compare two Datas BEDTable a 57	NGS: Picard SortSam sort SAM/BAM updated dataset	SAM LOOIS GITHUD Page, [LINK]	60: Build custom track on data 27, data 25, an d others	• / ×
DEL/10015	NGS: VCF Manipulation VCFsort: Sort VCF dataset by coordinate BEDTools		58: Compare two Datas ets on data 55 and dat a 57	• / ×

Figure 20c: Samtools filtering.

Third, remove duplicate reads with samtools. Here you need to specify that we are dealing with single-end reads (see *Figure 20d*).

🔁 Galaxy	Analyze Data Workflow Shared Data + Visualization + Help + User +		Using 0%
Tools	RmDup remove PCR duplicates (Galaxy Version 2.0) & Versions Options	History	2 ¢ []
rmdup	BAM File	search datasets	0
NGS: SAMtools <u>RmDup</u> remove PCR duplicate	5 This paired-end or single end data	ChIP 34 shown, 40 <u>deleted</u> 310.33 MB	8 8 9
Workflows All workflows	BAM is single-end (-s)	74: Sort on data 73	• # ×
	Execute What it does Remove potential PCR duplicates: if multiple read pairs have identical external coordinates, only retail the pair with highest mapping quality. In the paired-end mode, this command ONLY works with FR orientation and requires ISIZE is correctly set. It does not work for unpaired reads (e.g. two ends manned to different chromosomes or ormaban reads). This tool has the following narameters:	73: Filter SAM or BAM, output SAM or BAM on data 18: bam 1 10.1 MB format: bam, database: n	• * ×
	-s rmdup for SE reads -S treat PE reads as SE in rmdup (force -s)	display at UCSC <u>main</u> display at Ensembl <u>Curre</u> display with IGV <u>local Mo</u> display in IGB <u>View</u>	nt juse mm9
	Definition of SAM/BAM format [] ink]	Binary bam alignments fi	le
	Li, H. and Handsaker, B. and Wysoker, A. and Fennell, T. and Ruan, J. and Homer, N. and Marth, G. and Abecasis, G. and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. In	72: Extract Genomic D NA on data 71	● / ×
	Bioinformatics, 25 (16), pp. 2078-2079. [doi:10.1093/bioinformatics/btp352][Link] Li, H. (2011). Improving SNP discovery by base alignment quality. In <i>Bioinformatics, 27 (8), pp.</i>	71: Get flanks on data 70	• # ×
	1157-1158. [doi:10.1093/bioinformatics/btr076][Link] Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and	70: Compute on data 6 3	• / ×
	population genetical parameter estimation from sequencing data. In <i>Bioinformatics, 27 (21), pp. 2987–2993.</i> [doi:10.1093/bioinformatics/btr509][Link]	63: Cut on data 62	• / ×
	Danecek, P., Schiffels, S., Durbin, R Multiallelic calling model in bcftools (-m). [Link] Durbin, R Segregation based metric for variant call QC. [Link]	60: Build custom track on data 27, data 25, an d others	• # ×
	LI, H., MALTERNATICAL NOTES ON SAMIDOIS ALGORITIMS. [LINK] SAMTools GitHub page. [Link]	58: Compare two Datas ets on data 55 and dat a 57	• / ×

Figure 20d: Samtools rmdup.

1.9 Peak calling

Hint! You should have 4 bowtie2-generated bam-files in your history. If Galaxy did not run your bowtie2 tasks it could be that the queues are full. In this case, please download the Bowtie2 bam-files here or at http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/data/bowtie2-results-bam.zip. Unzip the files and upload all files to your Galaxy history and go to 1.8.2 and finally, continue to 1.9.1.

1.9.1 MACS

Select the MACS tool in the NGS Peak Calling section:

- 1. Once you have the tool open (see *Figure 21*), give it a useful name.
- 2. We are dealing with single-end reads, so select this option.
- 3. We give it the Bowtie mapped file of the CTCF-ChIP'ed experiment and the "input" of the same cell-line as a control-file.
- 4. We need to adjust the genome size to that of mm9 Canonical which is 1.87e+9.
- 5. We also change the tag-size to 36.
- 6. Finally, we adjust the peak detection method to the "new" one.

🔁 Galaxy		Analyze Data Workflow Shared Data - Visualization - Cloud - Help - Us			Using 0%
Tools	<u>±</u>	F MACS Model-based Analysis of ChIP-Seq (Galaxy Tool Version 1.0.1)	Options	History	C O 🗆
search tools	0	Experiment Name		search datasets	0
Get Data Send Data		MACS GIE CTCF		ChIP 16 shown, 6 <u>deleted</u>	
Lift-Over		Paired End Sequencing		322.4 MB	8 \$ 9
Text Manipulation				22: G1E input bowtie2	@ / X
Filter and Sort		ChiP-Seq Tag File	-	20: G1E ER4 input bowtie2	@ / X
Join, Subtract and Group		Chill See Control Elle 3		19: G1E_ER4 CTCF bowtie2	x
NGS: Mapping		Chi P set Control Pile 22: G1E input bowtie2	-	18: G1E CTCE howtie?	
NGS: RNA-seq					
NGS: SAMtools NGS: 8AM Tools		Effective genome size		16: GIE EK4 input trimmed	* / X
NGS: Picard		default: 2.7#+9		15: G1E_ER4 CTCF trimmed	@ / X
NGS: VCF Manipulation		Tag size		14: G1E CTCF trimmed	@ / X
Fetch Sequences		36 5		13: G1E CTCF filtered	@ / X
Fetch Alignments		Band width		12: GIE ER4 CTCF filtered	@ # X
Operate on Genomic Interval	s	300		11: G1E_ER4 input filtered	@ / X
Statistics		Pvalue cutoff for peak detection		10: G1E input trimmed	
Graph/Display Data Phenotype Association		0.00001		0. C1E locat filmed	
snpEff		default: 1e-5		9. GTE INPUT HITERED	* / ×
BEDTools Conome Divertity		Select the regions with MFOLD high-confidence enrichment ratio against background to build model		4: G1E input (chr19)	@ / X
EMBOSS		32		3: G1E_ER4 input (chr19)	* / ×
Regional Variation		Parse xis files into into distinct interval files		2: G1E ER4 CTCF (chr19)	@ / X
Evolution		Yes No		1: GIE CTCF (chr19)	@ / X
Multiple Alignments		Save shifted raw tag count at every bp into a wiggle file			
Metagenomic analyses		Do not create wig hie (raster)			
MIGHT TOOLS	_	Use fixed background lambda as local lambda for every peak region			
NGS: Peak Calling		up to 9X more time consuming			
MACS Model-based Analysis	of	3 levels of regions around the peak region to calculate the maximum lambda as local lambda			
SICER Statistical approach fo	ir the	1000,5000,10000			
Identification of ChIP-Enrich Regions	ed	Build Model			
NGS: Variant Analysis		Build the shifting model			
NGS: GATK Tools (beta)		Diagnosis report			
NGS: Picard (beta) RNA Structure Prediction		Do not produce report (faster) up to 9X more time consuming			
Workflows		Perform the new peak detection method (futurefdr)			
All workflows		Yes No 6			
		The default menode only consider the peak location, 14, 36, and 10k regions in the control data; whereas the new future method also consider the 5k, 10k regions in treatment data to calculate local bias.	e		
		✓ Execute			
2					

Figure 21: MACS peak calling.

The results of the MACS run are two datasets (see *Figure 22*). One bed-file that contains the enriched regions and a html-file that provides more information about the MACS run, e.g. we can have a look at the estimated peak model (see *Figure 23*) or get more information about the peaks in the created xls-file (see *Figure 24*).

ng Galaxy	Analyze Data Workflow Shared Data + Visualization + Cloud + Help + Us	er-	Using 0%
Tools	Additional output created by MACS (MACS_G1E_CTCF)	History	2 ¢ []
search tools	Additional Files:	search datasets	0
Get Data Send Data Lift-Over	MACS GIE CTCF model.pdf MACS GIE CTCF modelr MACS GIE CTCF modelr.log	ChIP 18 shown, 6 <u>deleted</u> 322.6 MB	* •
Convert Formats	MACS GIE CTCF negative peaks.xls MACS GIE CTCF peaks.xls	24: MACS on data 22 and data 18 (html report)	• / ×
Filter and Sort Join, Subtract and Group	Messages from MACS:	3.3 KB format: html, database: mm9	
NGS: QC and manipulation	INFO @ Mon, 21 Sep 2015 23:36:23:	B 0 2	• •
NGS: RNA-seq	<pre># ARGUMENTS LIST: # name = MACS_GIE_CTCF</pre>	HTML file	
NGS: SAMtools	<pre># format = BAM # ChIP-seq file = /galaxy-repl/main/files/012/526/dataset_12526748.dat # control file = /galaxy-repl/main/files/012/526/dataset_12526910.dat</pre>	23: MACS on data 22 and data 18 (peaks: bed)	● / ×
NGS: Picard	<pre># effective genome size = 1.87e+09 # tag size = 36</pre>	405 regions, a comments format: peg, database: mm9	
NGS: VCF Manipulation	<pre># band width = 300 # model fold = 32 # model fold = 32</pre>	B 0 2 M	*
Fetch Sequences	<pre># Poalde Celori = 1.000-05 # Ranges for calculating regional lambda are : peak_region,1000,5000,10000 INFO @ Mon, 21 Sep 2015 23:36:23: #1 read tag files</pre>	display in IGB <u>View</u> display at Ensembl Current	
Fetch Alignments Get Genomic Scores	INFO @ Mon, 21 Sep 2015 23:36:23: #1 read treatment tags WARNING @ Mon, 21 Sep 2015 23:36:26: NO records for chromosome chr13, minus stra	display at UCSC main	
Operate on Genomic Intervals	WARNING & Mon, 21 Sep 2015 23:36:26: NO records for chromosome chr10, minus stra WARNING & Mon, 21 Sep 2015 23:36:26: NO records for chromosome chr18, minus stra	1.Chrom 2.Start 3.End 4.Name 5 track name="MACS peaks for MACS_GIE_CTCF"	
Statistics	INFO @ Mon, 21 Sep 2015 23:36:26: #1.2 read input tags WARNING @ Mon, 21 Sep 2015 23:36:28: NO records for chromosome chr16, minus stra	chr19 3204402 3204776 MACS_peak_1 92.61	
Graph/Display Data Phenotype Association	WARNING @ Mon, 21 Sep 2015 23:36:28: NO records for chromosome chr15, plus stran INFO @ Mon, 21 Sep 2015 23:36:28: #1 Background Redundant rate: 0.01	chr19 3291823 3292396 MACS_peak_2 132.00 chr19 3450651 3452121 MACS_peak_3 64.64	
snpEff BEDTools	INFO @ Mon, 21 Sep 2015 23:36:28: #: finished1 INFO @ Mon, 21 Sep 2015 23:36:28: #2 Build Peak Model INFO @ Mon, 21 Sep 2015 23:36:34: #2 number of paired peaks: 9733 INFO @ Mon, 21 Sep 2015 23:36:34: #2 finished1	chr19 3587686 3588189 MACS_peak_4 103.41 chr19 3623513 3624226 MACS_peak_5 71.88	





Figure 23: MACS peak model.

_	A	B	C	D	E	F	G	Н	1	
1	# This file is a	generated by	MACS							
2	# ARGUMEN	TS LIST:								
3	# name = MA	ACS_G1E_CTC	F							
4	# format = B	AM								
5	# ChIP-seq fi	le = /galaxy-re	epl/main/files	/012/526/dat	aset_125267	48.dat				
6	# control file	= /galaxy-rep	l/main/files/0	012/526/data	set_12526910).dat				
7	# effective g	enome size = :	1.87e+09							
8	# tag size = 3	6								
9	# band width	n = 300								
10	# model fold	= 32								
11	# pvalue cut	off = 1.00e-05								
12	# Ranges for	calculating re	gional lambd	a are : peak_r	egion,1000,5	000,10000				
13	# unique tag	s in treatment	t: 213711							
14	# total tags i	n treatment: 2	214878							
15	# unique tag	s in control: 2	10164							
16	# total tags i	n control: 210	781							
17	# d = 141									
18	chr	start	end	length	summit	tags	-10*LOG10(pvalue)	fold_enrichment	FDR(%)	
19	chr19	3204403	3204776	374	266	12	92.61	17.73	0	
20	chr19	3291824	3292396	573	337	35	132	10.3	0	
21	chr19	3450652	3452121	1470	765	44	64.64	10.69	0.34	
22	chr19	3587687	3588189	503	257	29	103.41	12.95	0	
23	chr19	3623514	3624226	713	381	29	71.88	11.44	0	
24	chr19	3723759	3725525	1767	1270	40	53.23	12.07	1.32	
25	chr19	3946485	3947973	1489	438	33	51.4	14.51	2.27	
26	chr19	3980149	3981146	998	393	49	156.53	12.41	0	
27	chr19	4012707	4013391	685	242	28	95.26	11.32	0	
28	chr19	4047685	4048485	801	365	22	56.6	6.3	0.84	
29	chr19	4098871	4099526	656	340	30	120	16 92	0	

Figure 24: MACS peak details.

Note! **TODO**:

- 1. Do the MACS peak calling for both cell-lines. Look at both peak models and note the distance and differences between the models.
- 2. Rename the peak-files to something meaningful and while you are doing it change the score-column to 5.
- 3. What do you expect in terms of called peaks if you would run G1E-CTCF without a control (the "input"-file)?
- 4. RUN G1E-CTCF without the input control. Note the differences.

2.0 Post-processing

Now that we established the peaks, we can do several different analyses to gain information about the genes they regulate or differences in peak abundance as well as functional association.

2.1 Overlap peaks with promoter regions

2.1.1 Get genes

Let's upload some genes and extract promoter information for them. Please download the following file (mm9_chr19_NCBIgenes.bed or from http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/data/mm9_chr19_NCBIgenes.bed) and upload to your Galaxy history (see *Figure 25*). the file contains 1428 gene regions in bed-format.



Figure 25: Upload the gene bed-file.

🚍 Galaxy	Analyze Data	Workflow	Shared Dat	a - Visualization -	۳.	Cloud - He	lp 👻	User - Using	0%
Tools	1 2	3	3(4	5	6	1	History CH	
search tools	chr19 3	065710	3197714	AK077035 AK006563	1	-		search datasets	0
Get Data	chr19 3	153798	3197714	AK007025	1	-		ChIP 21 shown, 30 deleted	
Upload File from your computer	chr19 3 chr19 3	259075 264810	3283010 3283010	Ighmbp2 Ighmbp2	1	-		323.9 MB	,
UCSC Archaea table browser	chr19 3 chr19 3	272720 283046	3283010 3291197	lghmbp2 Mrpl21	1	-+		51: mm9_chr19_NCBIgenes.bed @	×
EBI SRA ENA SRA	chr19 3	283046	3292837	Mrpl21	1	+		1,428 regions format: bed, database: mm9	
BioMart Central server	chr19 3	388868	3398168	Mtl5	1	+		uploaded bed file	
Elymine server	chr19 3 chr19 3	388868 389400	3407785 3407785	MtI5 MtI5	1	++		802M 🔊	9
modENCODE fly server	chr19 3 chr19 3	409916 454927	3414457 3575749	Gal Ppp6r3	1	-		display in IGB <u>View</u> display at Ensembl <u>Current</u>	
MouseMine server	chr19 3	477775	3575749	Ppp6r3	1	-		display at UCSC main	
Ratmine server	chr19 3 chr19 3	483527 483527	3494038 3575749	Ppp6r3	1	-		chr19 3065710 3197714 AK077035 1 -	
YeastMine server	chr19 3 chr19 3	510945 584824	3575749 3615879	mKIAA1558 Lrp5	1	-		chr19 3153210 3197714 AK006563 1 - chr19 3153798 3197714 AK007025 1 -	
WormBase server	chr19 3	584824	3686564	Lrp5	1	-		chr19 3259075 3283010 Ighmbp2 1 - chr19 3264810 3283010 Ighmbp2 1 -	
ZebrafishMine server	chr19 3	689686	3708168	AK144662	1	-		chr19 3272720 3283010 Ighmbp2 1 -	

Figure 26: The file is in bed-format.

2.1.2 Get promoter

Get the promoter regions by using **Operate on Genomic Intervals** => **Get flanks**. Choose the upstream regions and 10,000 bases (see *Figure 27*). Rename the promoter-set to something meaningful.

🗧 Galaxy	Analyze Data Workflow Shared Data - Visualization - Cloud - Help	- User -	Using 0%	
Tools	Get flanks returns flanking region/s for every gene Options	History	C 🕸 🗆	
Extract Features	(Galaxy Tool Version 1.0.0)	search datasets	0	
Fetch Alignments	Select data	Chip		
Get Genomic Scores	The Children Stimmer christen NCRicenes hed	21 shown, 30 <u>deleted</u>		
Operate on Genomic Intervals	St. mins_cirits_recigenes.ord	323.9 MB		
Profile Annotations for a set of	Region	F1 mm0 sha10 NCPleases had		
genomic intervals	Around Start -	51: mm9_cnr19_NCBigenes.bed	• # ×	
Merge the overlapping intervals of a dataset	Location of the flanking region/s	1,428 regions format: bed , database: mm9		
Fetch closest non-overlapping	Upstream	uploaded bed file		
feature for every interval	Offset	B 0 2 M		
Concatenate two datasets into	0	display in IGB View		
Subtract the intervals of two	Use positive values to offset co-ordinates in the direction of transcription and negative values to offset in the opposite direction.	display at Ensembl <u>Current</u> display at UCSC <u>main</u>		
	Length of the flanking region(s)	1.Chrom 2.Start 3.End 4.Name	5 6.Strand	
side-by-side	10000	chr19 3065710 3197714 AK077035 3	1 -	
Intersect the intervals of two	Use non-negative value for length	chr19 3153210 3197714 AK006563 1 chr19 3153798 3197714 AK007025 1	l - 1 -	
Get flanks returns flanking	✓ Execute	chr19 3259075 3283010 Ighmbp2 1 chr19 3264810 3283010 Ighmbp2 1	l - 1 -	
region/s for every gene	This seal finds the unstances and (as descentions finalize sealar(s) of all	chr19 3272720 3283010 Ighmbp2 3	1 -	
<u>Coverage</u> of a set of intervals on second set of intervals	the selected regions in the input file.	28: MACS G1E_ER4 CTCF + input (html)	• / ×	

Figure 27: Get upstream flanking regions of the TSS of genes.

2.1.3 Join

Now we are going to join (overlap) the peaks with the promoter regions by choosing the tool: **Operate on Genomic Intervals** => **Join** (see *Figure 28*). Again rename the resulting dataset to something useful.

🚍 Galaxy	Analyze Data Workflow Shared Data - Visualization - Cloud - Help	- User-	Using 0%
Tools	Join the intervals of two datasets side-by-side (Galaxy Options)	History	2 ¢ []
Convert Formats	Tool Version 1.0.0)	search datasets	0
Filter and Sort	Join	ChiP	
NGS: QC and manipulation	🗋 🙆 🖸 52: Gene promoters chr19 👻	22 shown, 30 deleted	
NGS: Mapping	First dataset	324.0 MB	8 9 9
NGS: RNA-seq	with	52: Gene promoters chr19	• / ×
NGS: SAMtools	🖸 🖄 🗀 25: MACS GIE CTCF + input (peaks) 👻	1,428 regions	
NGS: BAM Tools	Second dataset	format: interval, database: mm9	
NGS: VCF Manipulation	with min overlap	Location: Upstream, Region: start	t, Flank-
Extract Features	1	length: 10000, Offset: 0	
Fetch Sequences	(bp)	802W	۲
Fetch Alignments	Return	display at Ensembl Current	
Get Genomic Scores	Only records that are joined (INNER JOIN)	display at UCSC main	
Operate on Genomic Intervals		1.Chrom 2.Start 3.End 4.Name	5 6.Strand
genomic intervals	✓ Execute	chr19 3197714 3207714 AK006563	1 -
Merge the overlapping intervals	B TIP If your dataset does not annear in the pulldown menu it means	chr19 3197714 3207714 AK007025	1 -
of a dataset	that it is not in interval format. Use "edit attributes" to set chromosome,	chr19 3283010 3293010 Ighmbp2	1 -
Fetch closest non-overlapping feature for every interval	start, end, and strand columns.	chr19 3283010 3293010 Ighmbp2	1 -
Concatenate two datasets into	Screencasts!	51: mm9_chr19_NCBIgenes.bed	• / ×
one dataset	See Galaxy Interval Operation Screencasts (right click to open this link in	28: MACS GIE_ER4 CTCF + input	x
Subtract the intervals of two datasets	another window).	(html)	
Join the intervals of two datasets side-by-side	Syntax	27: MACS GIE_ER4 CTCF + input (peaks)	. • # ×

Figure 28: Overlap promoter and peaks with the join tool.

Sebastian Schmeier

Note!

TODO: Join the peak file for G1E CTCF and G1E_ER4 CTCF with the gene promoter regions. Note the numbers and differences in promoter numbers that overlap Ctcf peaks for both peak-files.

2.2 Enrichment analysis (genes) with Enrichr

Now lets take the genes with Ctcf in their promoter regions and do some functional annotation. To do this, we need the unique genes from the overlap of peaks and promteors form the step before. We will be using the tool: **Join, Subtract and Group** => **Group** to do this. **Group** aggregates data in a certain column. We will use it to aggregate column 4, the gene symbol column (see *Figure 29*). Copy the resulting genes symbol (see *Figure 30*).



Figure 29: Aggregate the gene symbol column.

ng Galaxy	Analyze Data Workflow Shared Data - Visualization - Cloud -	Help + User +	Using 0%
Tools	1	History	2 4 []
search tools	1810006K21Rik 4930579J09Rik	search datasets	0
Get Data	AK006563	ChIP	
Send Data	AK007025	24 shown, 31 deleted	
Lift-Over	AK008826	324.0 MB	S D
Text Manipulation	AK016444	EE: Crown on data E2	
Convert Formats	AK028012	33. Group on data 35	• * *
Filter and Sort	AK029443	156 lines	
Join, Subtract and Group	AK036616	format: tabular, database: mm	9
Subtract Whole Dataset from	AK040197	Group by c4:	
another dataset	AK040231	E A CIN	
loin two Datasets side by side on	AK048429		• •
a specified field	AK049068	1	
Compare two Datasets to find	AK052572	1810006K21Rik	
common or distinct rows	AK052812	4930579J09Rik	
	AK076978	AK006563	
Group data by a column and	AK077035	AK007025	
other columns.	AK082813	AK008826	
	AK148054	AK016444	
NGS: QC and manipulation	AK158434	53: Overlap promoter + G1E C	Tarr
NGS: Mapping	AK168860	CE	
NGS: RNA-seq	44107072		

Figure 30: The aggregated gene symbols.

Now, go to the online tool Enrichr (http://amp.pharm.mssm.edu/Enrichr/). Enrichr provides a way to analyse mammalian gene lists to find enriched annotation terms to get a better understanding of the functions of the gene list under investigation. Go to **Analyze** tab and paste your gene list into the field (see *Figure 31*). Click on the arrow.

✤Enrichr	Login Regi 413,746 lists analy
alyze What's New? Libraries Find A Gene	About Help
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. Choose File No file chosen	Or paste in a list of gene symbols optionally followed by a comma and levels of membership. Try a regular example or an example of a quantitative set. Ifit2 Ighmbp2 Kazald1 Kcnip2 Kcnk4 Lbx1 Lcor Lrp5 Lrrn4cl Map4k2 Mark2
Enter a brief description for the list in case you	156 gene(s) entered
Enter a brief description for the list in case you	want to share it. (Optional)
Contribute Please acknowledge Enrichr in your publications by citing the followi Chen EV Tan CM Kou X Duan O Wang 7 Meirelles GV Clark NR Ma	ng reference:

Figure 31: The Enrichr tool.

On the result pages (see *Figure 32*) you will find several different categories (e.g. *Transcription, Pathways*, etc.) of with different databases where term-gene association information was extracted. *Figure 32* for example shows the enriched pathways from the Reactome (http://www.reactome.org/) database.

βÂ	nric	hr						Login Regi
anscriptio	n Pathways	Ontologies	Disease/Drugs	Cell	Types	Misc	Legacy	Crowd
scription	No descriptior	n available (15	6 genes)					
KEGG	2015							
WikiPa	athways 20)15						
Reacto	ome 2015			Bar	Graph	Table	Grid	Network 🗳
Hover ead	ch row to see th	ne overlappin	g genes.					
10 🛟	entries per pag	e				Search:		
Index	Name				♦ P-	value	Z- score	Combined Score
1	Signaling by \	Nnt			0.	.03101	-2.24	1.74
2	Organelle bio	genesis and i	maintenance		0.0	04934	-2.20	1.71
3	misspliced LR catenin-depe	RP5 mutants h ndent signaliı	lave enhanced be າg	ta-	0.	.02128	-2.19	1.70
4	RNF mutants proliferation	show enhand	ed WNT signaling	and	0.	.02128	-2.16	1.68
5	XAV939 inhib	its tankyrase,	stabilizing AXIN		0.	.02128	-2.16	1.68
6	TCF depende	nt signaling ir	response to WN	Г	0.	.02128	-2.16	1.68
7	Signaling by \	WNT in cance	r		0.	.03189	-2.13	1.65
8	Polymerase s telomere*	witching on t	he C-strand of the		0.0	07373	-2.07	1.61
9	Telomere C-s	trand (Laggin	g Strand) Synthesi	s*	0.	.01758	-2.07	1.61
10	Lagging Stran	nd Synthesis*			0.	.01499	-2.03	1.57
Showing 1 Terms mar	to 10 of 415 entr ked with an * hav	ies <mark>Export ent</mark> ve an overlap o	ries to table f less than 5				< Pi	revious Next 🕨

Figure 32: The Enrichr results show enriched term associations to the input gene list.

Note! **TODO**:

- 1. Find and note the top 5 enriched Gene Ontology process terms for both the G1E and G1E_ER4 genes that have Ctcf in their promoters.
- 2. Now that you have unique gene lists for G1E and G1E_ER4, how many genes are in common, e.g. which genes in both cases have Ctcf in their promtoer region?

Hint! For point 2. you can use the Join, Subtract and Group => Compare two Datasets tool.

2.3 Enrichment analysis (peaks) with GREAT

Here we are going to use another tool called GREAT (http://bejerano.stanford.edu/great/public/html/). Great as opposed to Enrichr excepts bed-regions directly, thus we do not need to get the genes that overlap our peak regions. Take the results from MACS, cut out the first 4 columns with **Text Manipulation** => **Cut** (as GREAT does not except floats as scores and will produce errors), copy the regions and paste them into the GREAT interface.

GREAT version 3.0.0	current (02/15/	2015 to now)			\$		
GREAT pre	dicts fur	nctions	of ci	s-regula	atory re	egions.	
any cooling genes are blogical meaning to a udying cis functions of IP-seq) and by comp	set of non-codin f sets of non-cod utational method	g genomic reg ding genomic r s (e.g. compared	ions by a egions. C rative ger	inalyzing the an is-regulatory re nomics). For mo	notations of t egions can be ore see our Na	the nearby genes. Thus, identified via both expe ature Biotech Paper.	it is particularly useful ir rimental methods (e.g.
ews							
 Feb 15, 2015: Apr 3, 2012: GREAT Feb 18, 2012: The C 	GREAT version Version 2.0 add	3.0 switches to is new annotat re released, alle	Ensemb ions to h	ol genes, adds uman and mou creased user-to	the mouse mi se ontologies -user interact	m10 assembly, and add and visualization tools 1 ion	s new ontologies. or data exploration.
ore news items							
Species Assembly	 Human: GF Mouse: NC Mouse: NC Zebrafish: ' Can I use a diffe 	RCh37 (UCSC BI build 37 (UCSC BI build 38 (UC Wellcome Trus event species or	hg19, Fe CSC mm CSC mm t Zv9 (da assembly	b/2009) 9, Jul/2007) 10, Dec/2011) nRer7, Jul/201 ?	0) Zebrafisl	n CNE set	
Test regions	BED file:	Choose File	alaxy25-	·bed)].bed			
	BED data:	chr19 609 chr19 611 chr19 611 chr19 611 chr19 612	68112 60078 73687 85950 75219	60968799 61161098 61174023 61186399 61276078	MACS_pea MACS_pea MACS_pea MACS_pea MACS_pea	ak_401 ak_402 ak_403 ak_404 ak_404 ak_405	
	What should my How can I creat	test regions file a test set from	e contain? a UCSC	Genome Brows	er annotation t	rack?	
Background	Whole gen	ome					
regions	BED file:	Choose File	lo file cho	osen			
	BED data:						
	Million should be		1				
	What should my	/ background re	gions file	contain?			

Figure 33: The GREAT website.

,GREAT,	Overview	News	Use GREAT	Demo	Video How	to Cite Hel	p Forum					Bejera	ino Lab, Star	nford Uni	iversity
GREAT versi	ion 3.0.0 curre	ent (02/15/	2015 to now)			٠									
+ Job De	scription														
+ Region	-Gene Asso	ociatio	n Graphs												
+ Global	Controls	Global E	kport	÷	🗯 wл	ich data is exp	orted by each o	ption?							
GO Mo	lecular Fun	ction (4 terms)		-									G	lobal cont
Table controls:	Export	•	Shown top ro	ws in this	table: 20	Set	Term annotatio	n count: Min: 1	Max: Inf	Set	Visuali	ze this table: 💥	select o	ne]	•
	Term Name			Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Se Coverag
oxidoreducta donors, with resulting in the to two molection	ase activity, acti oxidation of a p the reduction of cules of water	ing on pa pair of do molecula	ired nors ir oxygen	1	2.6326e-16	9.1693e-13	78.0167	10	2.47%	1	8.8353e-6	37.2489	6	9	1.58%
iron ion bind	ling			4	2.2764e-12	1.9822e-9	5.2559	28	6.91%	3	3.7623e-3	3.9732	16	225	4.22%
stearoyl-CoA	A 9-desaturase a	activity		6	1.7961e-11	1.0426e-8	69.6114	7	1.73%	2	2.5647e-3	37.2489	4	6	1.06%
oxidoreducta donors, with molecular ox	ase activity, acti Incorporation c cygen	ing on pa or reducti	ired on of	17	6.7018e-8	1.3731e-5	3.9978	22	5.43%	4	2.1452e-2	3.7607	14	208	3.69%
The test set of GO Molecular 3,483 ontology	f 405 genomic re <i>Function</i> has 3,4 y terms (100%) v	egions picł 483 terms were teste	(ed 379 (2%) of covering 15,73 d using an anno	all 21,176 5 (74%) o ptation co	5 genes. f all 21,176 ge unt range of [enes, and 181, 1, Inf].	165 term - gene	associations.							
o GO Bio	logical Pro	cess (1	term)											GI	lobal cont
Table controls:	Export	\$	Shown top ro	ws in this	table: 20	Set	Term annotatio	n count: Min: 1	Max: Inf	Set	Visuali	ze this table: 🍀	select o	ne]	÷
	Term Name			Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Se Coverag
Wnt receptor	r signaling path	way		2	1.4738e-14	7.4104e-11	4.1678	42	10.37%	3	4.7691e-2	3.7249	15	225	3.96%

Figure 34: GREAT result page.

Note!

TODO: Run **GREAT** for both MACS result-files and note the top 5 **GO Biological processes**. Are they different to the ones from **Enrichr**?

2.4 Visualisation

Let us now create a visualisation track of the promoters that overlap G1E CTCF peaks and G1E_ER4 CTCF peaks. Use **Graph/Display Data** => **Build custom track** (see *Figure 33*). Also add the two MACS peak bed-files. Look at the track at UCSC (see *Figure 36* and *Figure 37*).

🚍 Galaxy	Analyze Data Workflow Shared Data - Visualization - Cloud - Help -	User 🕶 🚺	Using 0%
Tools 🔹	Build custom track for UCSC genome browser (Galaxy Tool Version Option 1.0.0)	15 History search datasets	2 ¢ [] 0
<u>Get Data</u> Send Data Lift-Over	Track 1: Track	ChIP 27 shown, 31 <u>deleted</u> 324.0 MB	S D
Text Manipulation Convert Formats Filter and Sort	Dataset 56: Overlap promoter + G1E_ER4 CTCF	 58: Compare two Datas ets on data 55 and dat a 57 	• / ×
Join, Subtract and Group NGS: OC and manipulation	Gene promoter overlapping G1E_ER4 CTCF	57: Group on data 56	• / ×
NGS: Mapping NGS: RNA-seq NGS: SAMtools	description User Supplied Track (from Galaxy)	+ G1E_ER4 CTCF	* / X
NGS: BAM Tools NGS: Picard	Color Green	53: Overlap promoter + G1E CTCF	• # ×
NGS: VCF Manipulation Extract Features Fetch Sequences	Visibility Dense	52: Gene promoters ch r19	• / ×
Fetch Alignments Get Genomic Scores	2: Track	51: mm9_chr19_NCBlg enes.bed	• / ×
Operate on Genomic Intervals Statistics Granh/Display Data	S3: Overlap promoter + G1E CTCF	 28: MACS GIE_ER4 CTC F + input (html) 27: MACS GIE_ER4 CTC 	• / ×
Build custom track for UCSC genome browser	Gene promoter overlapping G1E CTCF	F + input (peaks) 26: MACS on data 22 a	• / ×
Scatterplot of two numeric columns	description User Supplied Track (from Galaxy)	nd data 18 (html repor t)	
<u>Plotting tool</u> for multiple series and graph types	Color Black	25: MACS GIE CTCF + I nput (peaks) 22: GIE input bowtie2	* * ×
GMAI Multiple Alignment Viewer Boxplot of quality statistics	Visibility Dense	20: G1E_ER4 input bow tie2	• # ×
VCF to MAF Custom Track for display at UCSC	3: Track	19: G1E_ER4 CTCF bo wtie2	• / ×
Phenotype Association snpEff	25: MACS G1E CTCF + input (peaks)	18: G1E CTCF bowtie2	* / ×
DED TOOIS	name	med	

Figure 35: Building a custom UCSC track.

= Galaxy		Analyze Data	Workflow	Shared Data -	Visualization 👻	Cloud -	Help - User -		Using 0%
Tools	±.	1	2		3	2	4	History	2 0 []
search tools	Θ	track name	="Gene prom	noter overlapping	G1E_ER4 CTCF" de	scription	="User Supplied Tr		-
Get Data		chr19		3197714	320	07714	0	search datasets	G
Send Data		chr19		3197714	320	07714	1	ChIP	
Lift-Over		chr19		3197714	320	07714	2	28 shown, 33 deleted	
Text Manipulation		chr19		3283010	329	93010	3	324.2 MB	2 > >
Convert Formats		chr19		3283010	329	93010	4	CO. Build sustain treak	Include the
Convert Pormats		chr19		3283010	329	93010	5	on data 27. data 25. a	• / ×
Filter and Sort		chr19		3313300	332	23300	6	nd others	
Join, Subtract and Group		chr19		3575749	35	85749	7	1,641 lines, 5 comments	
NGS: QC and manipulation		chr19		3575749	358	85749	8	format: customtrack, da	tabase:
NGS: Mapping		chr19		3575749	35	85749	9	mm9	
NGS: RNA-seq		chr19		3575749	358	85749	10	Constated a sustam trac	-k
NGS: SAMtools		chr19		3615879	36	25879	11	containing 4 subtracks.	
NGS: BAM Tools		chr19		3615879	362	25879	12		
NGS: Picard		chr19		3698332	370	08332	13	B O C M	•
NGS: VCF Manipulation		chr19		3698410	370	08410	14	display at UCSC main	
Extract Features		chr19		3976570	398	36570	15	1	2
Fetch Sequences		chr19		3976660	398	86660	16	track name="Gene promote	er overlappir
Fetch Alignments		chr19		3990579	400	00579	17	chr19	3197714
Get Genomic Scores		chr19		3990743	400	00743	18	chr19	3197714
Operate on Genomic Intervals		chr19		3993384	400	03384	19	chr19	3197714
Statistics		chr19		4012725	402	22725	20	chr19	3283010
Graph/Display Data		chr19		4037912	404	47912	21	chr19	3283010
an apart a reprint Sealing		chr19		4042221	405	\$2221	22		2

Figure 36: Visualising a Galaxy dataset/track.

Â	Genomes	Genome	Browser	Tools	Mirrors	Dov	vnloads	My Data	View
JCSC	Genom	e Brow	ser on	Mouse	July 2 3x 10x	2007 base	(NCB	137/mm	9) Assen
chr19:3	,152,709-3,	252,718 10	00,010 bp.	enter position	n, gene symbol	l or search	terms		go <u>hg38</u>
c	nr19 (qA)	19qA		19q8	19qC1	19962	19qC3	19qD1 19qD2	qD3
ping GIE_ER User r lapping GI for MACS_GI MACS_GIE_ER ENCFF STS N AH AH	Scale chr19: * CTCF * Track ECTCF *E_CTCF *4.CTCF ************************************	3,170,000	50 kb 3,180,000 ch ucsc	3, 198, 868) User MR MRCS IP-seq of MEL select or click to zoom § Genes (RefSe	3,200,000 3 Supplied Tr User Supp CS peaks for cell line - on Genetic a rd, GenBank,	8,210,000 ack (fro ack (fro ack (fro MACS_G1 MACS_G1E_ IACS_G1E_ I I ENCSR70 and Radia tRNAS &) 3,229,000 m Galaxy) ck m Galaxy) E_CTCF ER4_CTCF SHGT(Target sHGT(Target ition Hybrid Comparative	mm9 a 3,230,000 - USF1) Maps - Genomics)	3,240,000 3,25

Figure 37: Custom UCSC track at the UCSC genome browser website.

2.5 Motif finding

Here we want to establish enriched sequence motifs in the peak regions to hypothesise on the acctual binding site of Ctcf. We are going to use MEME-ChIP (http://meme.ebi.edu.au/meme/tools/meme-chip) for this. However, MEME-ChIP expects fasta-sequence data as an input, not bed-files. So, we need to extract for our peak bed-files the actual sequence. Another restriction is, that MEME-ChIP expects regions of similar size, this is also not a given in the MACS results. The workflow for this analysis looks like this:

- 1. Find the center of each MACS peak region.
- 2. Get the flanking region +-250 bases (as the recommended region size for MEME-ChIP is 500bp).
- 3. Extract the fasta-sequence for the regions.
- 4. Download the fasta-file and upload to MEME-ChIP.
- 5. Run MEME-ChIP.

2.5.1 Find the peak center

We have the start and stop position in our bed-files, thus we can calculate the center point with start + ((stop-start)/2). The tool we need is: Text Manipulation => Compute (see *Figure 38*). Make sure you round the results.

📮 Galaxy	Analyze Data	Workflow	Shared Data -	Visualization -	Cloud +	Help +	User
Tools Search tools Get Data Send Data Lift-Over Text Manipulation Compute an expression on every row Add column to an existing dataset	Analyze Data Computer Add express c2+((c3-c2) as a new co C2+(c3-c2) as a new co C2+(c3-c2) Add express C2+((c3-c2) C2+(c3-c2)	Workflow an expressi asion)/2) blumn to 25: M sing? See TIP alt?	Shared Data + on on every row (IACS G1E CTCF + below	Visualization - Galaxy Tool Version input (peaks)	Cloud -	Help →	User
<u>Add column</u> to an existing dataset <u>Concatenate datasets</u> tail-to- head	Round resu YES ✓ Execute	ılt?					•



Now we cut out the first column (chromosome) and last column (center) with **Text Manipulation** => **Cut** (see *Figure 39*).

🗧 Galaxy	Analyze Data	Workflow	Shared Data -	Visualization -	Cloud -	Help -	User
Tools	Eut colun	nns from a ta	ble (Galaxy Tool	Version 1.0.2)		▼ Optic	ons
Text Manipulation Compute an expression on every row	Cut column	s					
<u>Add column</u> to an existing dataset	Delimited b	y					
<u>Concatenate datasets</u> tail-to- head	Tab						•
<u>Condense</u> consecutive characters		G4: C	ompute on data 6	53			•
Convert delimiters to TAB	✓ Execute						
Merge Columns together		. This tool b	areaks column a	ssianments. To re	-establish	column	
dataset	assignme	nts run the to	ools and click on	the pencil icon in	the latest h	nistory iter	n.
Cut columns from a table	The outpu are comm	ut of this too as. they will	l is always in tabu be replaced with	ılar format (e.g., il tabs). For exampl	f your origi e:	nal delimit	ters

Figure 39: Cut columns.

We add another column to the result that will reppresent the stop-postion with **Text Manipulation** => **Compute**. Make this c2 + 1 (see *Figure*).

ng Galaxy	Analyze Data	Workflow	Shared Data -	Visualization -	Cloud -	Help +	User
Tools 1 Text Manipulation Compute an expression on every row Add column to an existing dataset Concatenate datasets tail-to-head Condense consecutive characters Convert delimiters to TAB Merge Columns together Merge Columns together	Compute Add express c2+1 as a new co as a new co Dataset mis Round resu YES VES	e an expression column to 63: C sing? See TIP alt?	on on every row (ut on data 62 below	Galaxy Tool Versi	on 1.1.0)	✓ Optio	v v

Figure 40: Compute the stop-position.

Now, the resulting data is in *tabular* format (see *Figure 41*) and we need to change it first to *bed-format* (see *Figure 42*).

Analyze Dat	a Workflow	Shared Data -	Visualization -	Cloud -	Help -	User 🔫		Using 0%	
1	2	3				1	History	C 🕈 [П
chr19	3204589	3204590							
chr19	3292109	3292110					search datasets	8	り
chr19	3451386	3451387					ChIP		
chr19	3587937	3587938					30 shown, 34 deleted		
chr19	3623869	3623870					324.2 MB	🗹 📎 🗩	
chr19	3724641	3724642					64. Compute on data (
chr19	3947228	3947229					3		
chr19	3980647	3980648					= 405 regions		
chr19	4013048	4013049					format: tabular, databa	se: mm9	
chr19	4048084	4048085							
chr19	4099198	4099199					Creating column 3 with	n expression	
chr19	4206891	4206892					kept 100.00% of 405 li	nes.	
chr19	4228453	4228454							
chr19	4309935	4309936					₿ 0 2 ш	۲	•
chr19	4338322	4338323					1 2 3		
chr19	4442144	4442145					chr19 3204589 3204590		
chr19	4455946	4455947					chr19 3292109 3292110		
chr19	4755852	4755853					chr19 3451386 3451387		
chr19	4847634	4847635					chr19 3587937 3587938		
chr19	4908202	4908203					chr19 3623869 3623870		
chr19	4961967	4961968					chr19 3724641 3724642		
1 10	4000122	4000122							

Figure 41: Center peaks.

Analyze Data Workfl	ow Shared Data -	Visualization -	Cloud -	Help -	User -		Using 0%
Attributes Conve	rt Format Datatype	Permissions				History	2 ¢ []
Change data type						search datasets	8
New Type:		-				ChIP	
bed						324.2 MB	
this if Galaxy has inc	correctly guessed the	type of your datas	et.	contents.	Use	64: Compute on data 6	
Save						<u>3</u> 405 regions	
						format: tabular , databas	e: mm9
						Creating column 3 with c2+1 kept 100.00% of 405 lir	expression nes.
						882	۲
						1 2 3 cbr19 3204589 3204590	
						chr19 3292109 3292110	
						chr19 3451386 3451387	
						chr19 3587937 3587938	
						chr19 3623869 3623870 chr19 3724641 3724642	

Figure 42: Change the data-format to bed.

2.5.2 Get flanking regions

Use **Operate on Genomic Intervals** => **Get flanks**. Extend **both** sides of the start position by **500** bases (see *Figure 43*).



Figure 43: Get flanking regions.

2.5.3 Extract fasta-sequence

Use Fetch Sequences => Extract Genomic DNA to extract for the regions the genomic DNA (see Figure 44).



Figure 44: Extract DNA for regions.

🗧 Galaxy	Analyze Data Workflow Shared Data + Visualization + Cloud + Help + User +		Using 0%
Tools	>mm9_chr19_3204089_3204589_+ 3204589	History	2 � □
Filter and Sort	TTTAANTACCAGGGGAGTGAAAAATCTCACTATTCCACATCTGTTATATA	0	
Join, Subtract and Group	AAACTGTACAAGAAGGGCTTAGCACAAAGCCTTGGGTACCACAGTTGCTT AACACCTGTTCTTGTTTTTCTCTGCTGATACCACAACACCCCTATTCCCAG	search datasets	0
NGS: QC and manipulation	TAAAGATCTCTAGAACCAGGTTCTCTACCAGACTAGGTTTATCATACCTT	ChIP	
NGS: Mapping	ACACGATAAAGGCACAGCAGTCTGAGTTCAAACGAAGGACAGAATTCACA	32 shown, 40 deleted	
NCS: BNA see	TGGCCAGGTAAACACAGGGTGTAAAAATAAAAAGGGAACTCTGGTAAAGG	224 7 40	
NGS. KNA-Sey	AGAGTATAAGATTTAGACACCCTGAAACTGAGAAGGCATAGCCCAGGGAG	324.7 MB	
NGS: SAMtools	TCCTTALCATTCCCCACCCCCCCCCCCCCCCCCCCCCCC	72: Extract Conomic D	0.04
NGS: BAM Tools	TTGCTACTAGACAAGTTTGCTTGGGGGAAGGGAACGTTTGCTCAGATTTGA	NA on data 71	
NGS: Picard	>mm9 chr19 3204589 3205089 + 3205089	In on wata / A	
NCC NCC Manipulation	GGTTGCACAGGGTCTTACAAACTTGAACTTGAATTCTACAGTCTTGTAGT	810 sequences	
NGS. VCP Manipulation	CTGTAGTGACTGGCTGAAGCAAAGGGCTCAAAGGCGCCACCTTCTGGAAG	format: fasta, database: mm9	
Extract Features	GATGCTCCCCTGCATCCCTACATCTTCTCTATCCTCTGAATGCCTAGTTT	B 0 3	
Fetch Sequences	TCTTATCTTGTGTCTGATAAGAAAACTAGTTCCCTTTTCTTGTTCAGCTG		
Extract Genomic DNA using	CCTCATTCCTCAACCCTTCAGAACCACATTATTCCTGCTGATTTCTTTAA	>mm9_chr19_3204089_3204589_+ 3204589 TTTAAATACCAGGGGAGTGAAAAATCTCACTATTCC	
coordinates from	gattcccccgagdtctcccatagatagttcaaaataatcttcccacatggggctcttc		
assembled/unassembled	catacatcctcacagtccctcttgctgtgtTCAGGTCAGCAGAGCTAGAT	AAACTGTACAAGAAGGGCTTAGG	ACAAAGCCTTGGC
genomes	AGTGGGTGTCTCTGGGGGTCACCATCTGTCCACCATATCTCTGTTCTGGGG	AACACCTGTTCTTGTTTTCTCT	GETGATACCACA
	TAATTCCAGCACTGTCTCTGCCATCCATTCATAGCCATGAAGAAAGA	TAAACATCTCTACAACCACCTT	TCTACCACACTAC
Fetch Alignments	>mm9_chr19_3291609_3292109_+ 3292109		TCTACCAGACTAC
Get Genomic Scores	CTTCCTTTAGCCTCCATGGTTCTATGACAGGGGTCACAGTACAGGGTCAG	ALALGATAAAGGLALAGLAGTLI	GAGTICAAALGAA
Operate on Genomic Intervals	TGCTCCAGCAGCCCCCCAGGGGGTGGGGGGGGGGGGGGG		
Statistics	GAGGGTGAGGCAGGAGTTCTGATCCTAGGAACGGTTTGGGGCCTCAAGGGC	71: Get flanks on data	• / ×
Statistics	ACACAGGCTCTAAGGATCAAGCCAGAAGCATCGAATCAGAGGAGGAGGAGGAGGAG	<u>70</u>	
Graph/Display Data	ACACAGO I CIARGONI CARGO CAGANGONI CAGAGONG CAC		

Figure 45: Region in fasta-format.

2.5.4 Run MEME-ChIP

Go to MEME-ChIP (http://meme.ebi.edu.au/meme/tools/meme-chip) and copy the fasta-sequences into the field and run the application (see *Figure 45*). This may result in enriched sequence motifs that were found in the uploaded sequences (see *Figure 46*).

MEME Suite 4.10.1	MEME-ChIP performs comprehensive motif analysis (including motif discovery) on LARGE (50MB maximum) sets of nucleotide batasets Version 4.10.1						
Motif Discovery							
Motif Enrichment	Data Submission Form						
Motif Scanning	Perform motif discovery, motif enrichment analysis and clustering on large nucleotide datasets.						
Motif Comparison	Input the sequences						
Manual	Enter the (equal-length) nucleotide sequences to be analyzed.						
► Guides & Tutorials							
Sample Outputs							
► File Format Reference							
▶Databases							
Download & Install	Input the motifs						
► Help	Select, upload or enter a set of known motifs.						
Alternate Servers	Vertebrates (In vivo and in silico)						
Authors & Citing	Input job details						
Recent John	(Optional) Enter your email address. ?						
P Recent 0003							
← Previous version	(Optional) Enter a job description ?						
4.10.0							
	► Universal options						
	► MEME options						
	► DREME options						
	► CentriMo options						
	Note: if the combined form inputs exceed 80MB the job will be rejected.						
	Start Search Clear Input						

Figure 45: MEME-ChIP interface.

Name 🙎	Alt. Name 🙎	Preview 🕐		Matches 🕅	List 🙎	
1	MEME		₳ ∓₽ ₳₽ ₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽	1	UP00082_2 (Zfp187_secondary)	
2	MEME	CCARGAC TERRARGACT	<mark>င့င္ ၊</mark> ငိုင္ခင္	٥		
3	MEME	^ŀ <mark>]</mark> ⊊Į _{₮♀} ₽ <mark>₽</mark> Į <mark>₽Į</mark> ₽ ₽ ₽ <mark>₽</mark>	REFERENCE	2	MA0080.3 (Spi1), UP00031_2 (Zbtb3_secondary)	
Farget Da	TABASES					Previous Next Top
Database	?	Number of Motifs ? Mo	otifs Matched 🔞			
jolma2013. JASPAR_CO uniprobe_m	meme RE_2014_vertebrat oouse.meme	843 es.meme 205 386	0 1 2			
MATCHES TO	O QUERY MOTH	F 1 (MEME)				Previous Next Top
Summary	?	Alignment <table-cell></table-cell>				
Name Alt. Nam Database	e Zfp187_sec e uniprobe_m	ondary ouse.meme 措 1-	00		c	
p-value E-value	0.00051472	25	- SEVV	ŢŦÄlħ	Y SE	
q-value	1	د ن	- 0 0 4 0 0	10 110	12 14 15 16	
Overlap	16	2				
Offset	-6			C	<u> </u>	
Orientatio	on Normal	ੵੵ <mark>ੑੑੵੑਲ਼ੑੑ</mark> ਸ਼ੵਫ਼ੵਫ਼ੵ				
Cr	eate custom LOGO	I				[Query Top]

Figure 46: MEME-ChIP results.

Note!

TODO: Note the enriched motif for the G1E CTCF and G1E_ER4 peak regions. Are there any differences?

2.6 References

Hawkins RD, Hon GC & Ren B. Next-generation genomics: an integrative approach. *Nature Reviews Genetics. 2010; 11, 476-486*

Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics. 2009; 10, 669-680*

2.7 Web links

Galaxy: https://usegalaxy.org

Enrichr: http://amp.pharm.mssm.edu/Enrichr/

GREAT: http://bejerano.stanford.edu/great/public/html/

Gene Ontology: http://amigo.geneontology.org/

MEME-ChIP: http://meme.ebi.edu.au/meme/tools/meme-chip

This tutorial: http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/

Sebastian Schmeier