

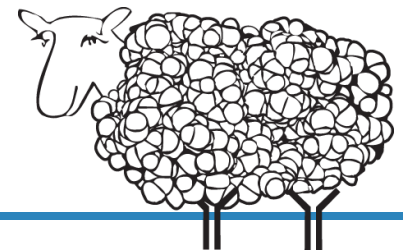
A step-by-step guide to ChIP-seq data analysis

December 03, 2014

Xi Chen, Ph.D.

EMBL-European Bioinformatics Institute

Wellcome Trust Sanger Institute



Target audience

- Wet-lab biologists with no experience of programming and Unix/Linux environment
- ChIP-seq beginners

Raw sequencing reads (fastq files)

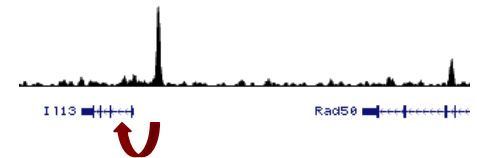
```
NTCACCTGCCTGAACCTTTGAACCACTCCCAACACA  
ACTTTTTCATGCTTATTATATCTGTTATGGTGATCT  
CTCCTTTTCTCTCCCTGTAATATGCCAAGGACTGTG  
GGTAGCAGGTGATTTACTTTAGCTAAACACTAATAA  
.  
.  
.
```



Alignments



Genes and GO terms



Peaks



Motifs



Webinar topics

QC of sequencing reads (FastQC)

Read alignment/mapping (Galaxy/bowtie)

Peak calling (Galaxy/macs)

Binding signal visualization (UCSC genome browser)

De novo motif discovery (MEME-ChIP)

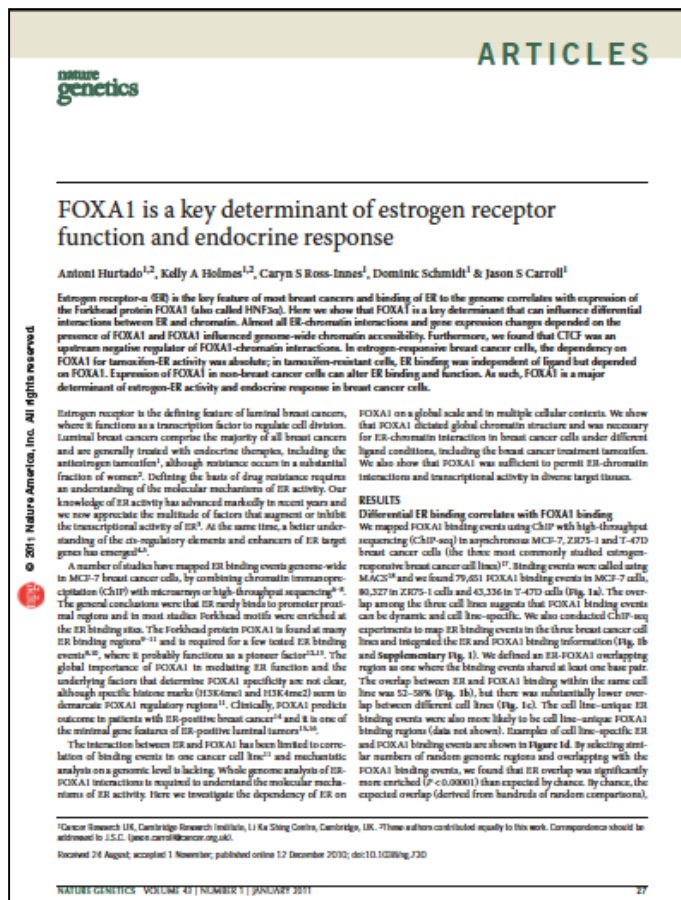
Gene ontology of binding sites (GREAT)

Heatmap representation of binding signals (seqMINER)

FOXA1 ChIP-seq data sets

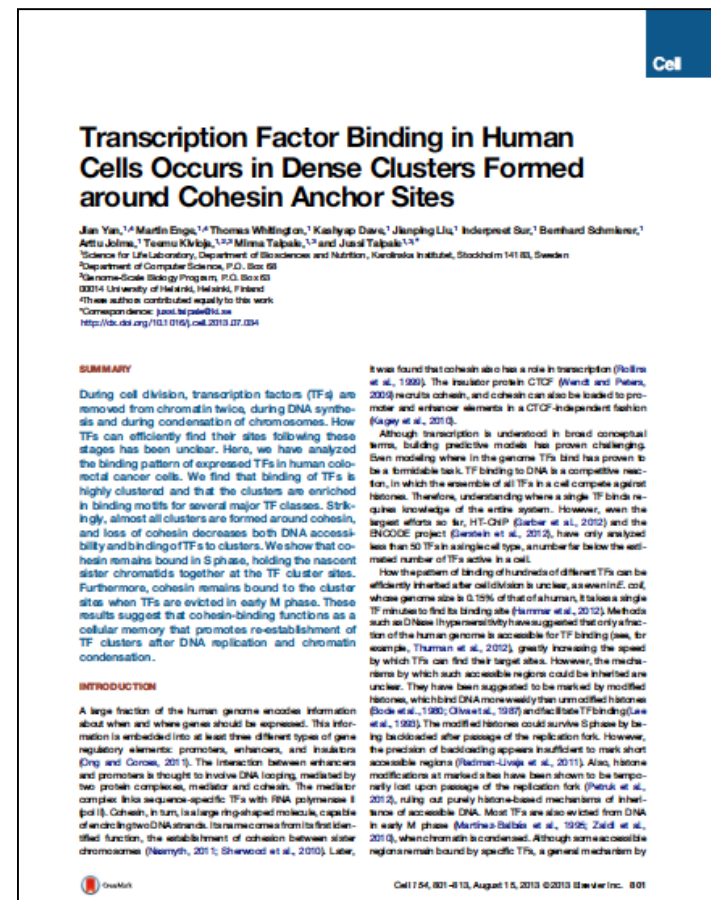
A successful experiment

α -FOXA1, ab5089 & ab23738, MCF7 cells




A failed experiment

α -FOXA1, other vendor, LoVo cells



QC of sequencing reads

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>



Babraham Bioinformatics

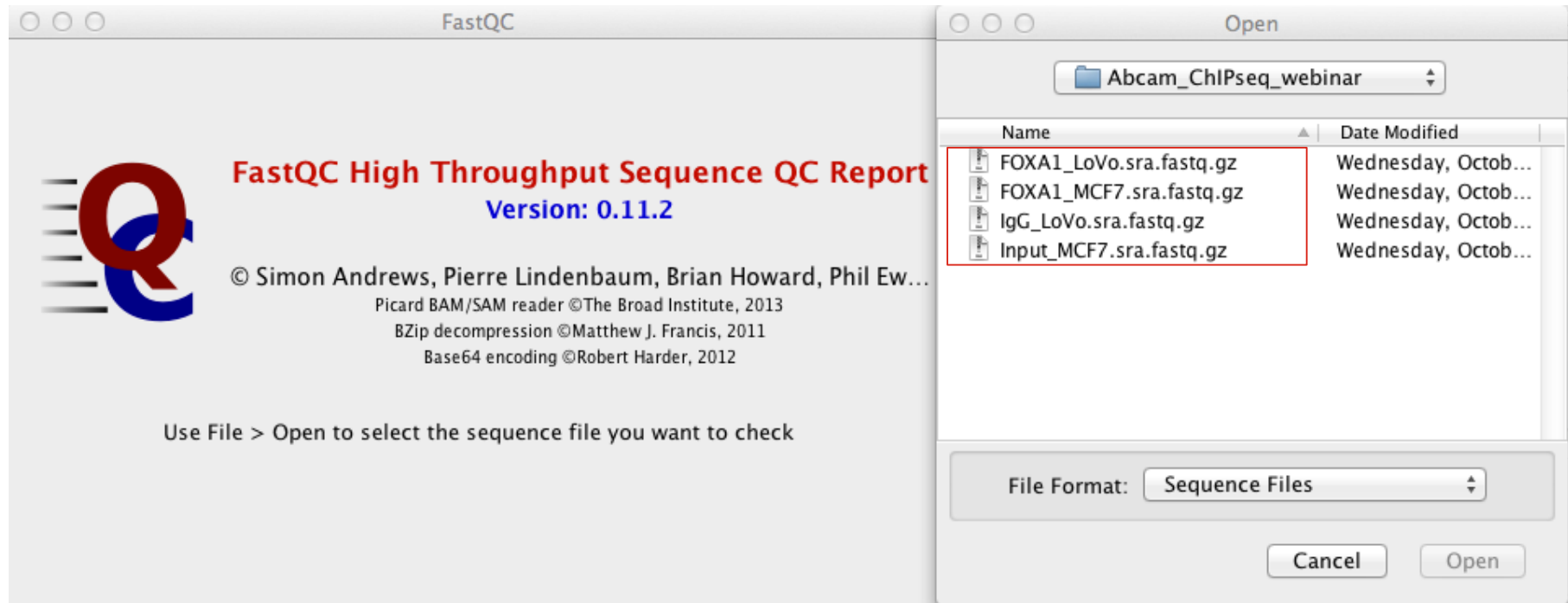
[About](#) | [People](#) | [Services](#) | [Projects](#) | [Training](#) | [Publications](#)

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews

[Download Now](#)

QC of sequencing reads



QC of sequencing reads

FastQC High Throughput Sequence QC Report
Version: 0.11.2

© Simon

Use File > O

Basic sequence stats

Measure	Value
Filename	FOXA1_MCF7.sra.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	26940662
Sequences flagged as poor quality	0
Sequence length	36
%GC	42

Metrics:

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Open Dialog:

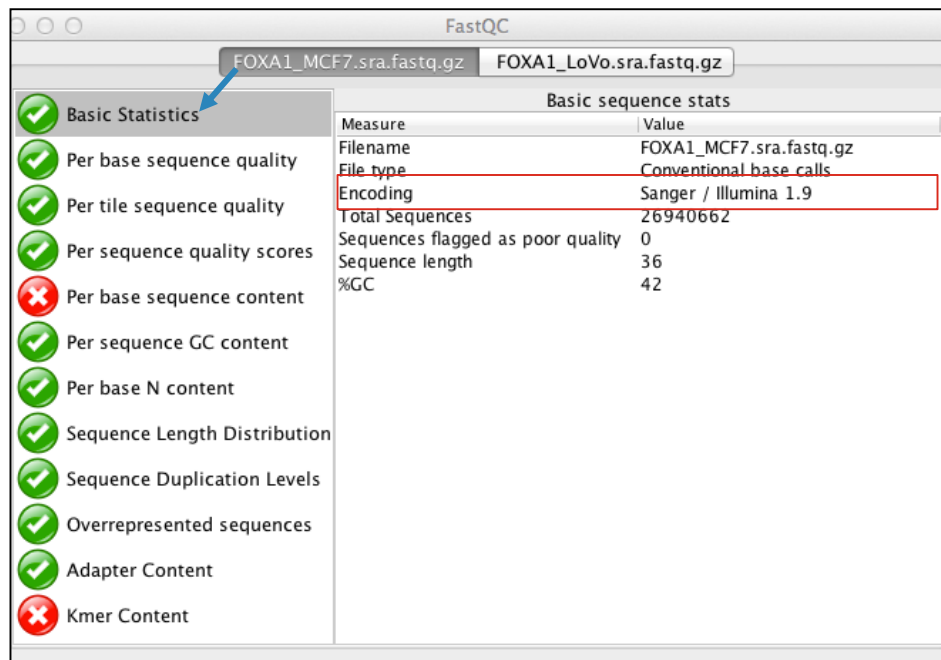
Abcam_ChIPseq_webinar

Name	Date Modified
FOXA1_LoVo.sra.fastq.gz	Wednesday, Octob...
FOXA1_MCF7.sra.fastq.gz	Wednesday, Octob...
IgG_LoVo.sra.fastq.gz	Wednesday, Octob...
	Wednesday, Octob...

Cancel Open

QC of sequencing reads

FOXA1, MCF7 cells



FastQC

FOXA1_MCF7.sra.fastq.gz FOXA1_LoVo.sra.fastq.gz

Basic Statistics

Basic sequence stats

Measure	Value
Filename	FOXA1_MCF7.sra.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	26940662
Sequences flagged as poor quality	0
Sequence length	36
%GC	42

Per base sequence quality

Per tile sequence quality

Per sequence quality scores

Per base sequence content

Per sequence GC content

Per base N content

Sequence Length Distribution

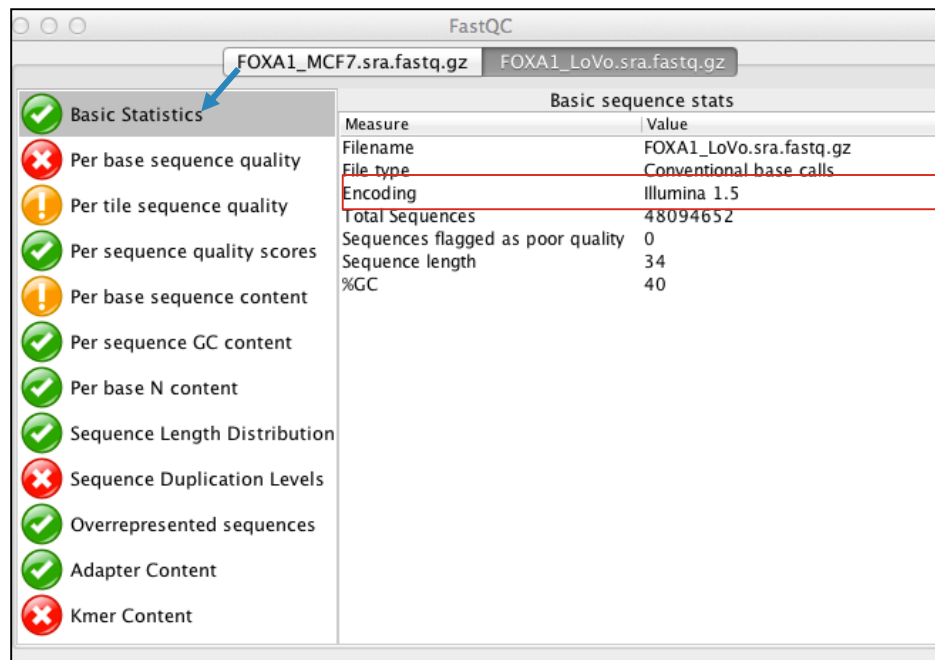
Sequence Duplication Levels

Overrepresented sequences

Adapter Content

Kmer Content

FOXA1, LoVo cells



FastQC

FOXA1_MCF7.sra.fastq.gz FOXA1_LoVo.sra.fastq.gz

Basic Statistics

Basic sequence stats

Measure	Value
Filename	FOXA1_LoVo.sra.fastq.gz
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	48094652
Sequences flagged as poor quality	0
Sequence length	34
%GC	40

Per base sequence quality

Per tile sequence quality

Per sequence quality scores

Per base sequence content

Per sequence GC content

Per base N content

Sequence Length Distribution

Sequence Duplication Levels

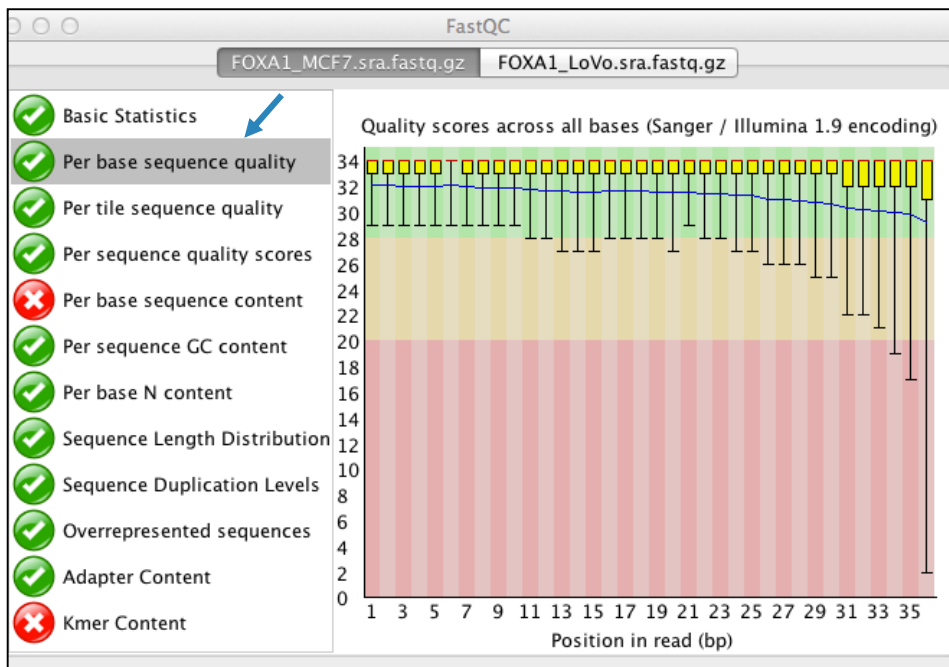
Overrepresented sequences

Adapter Content

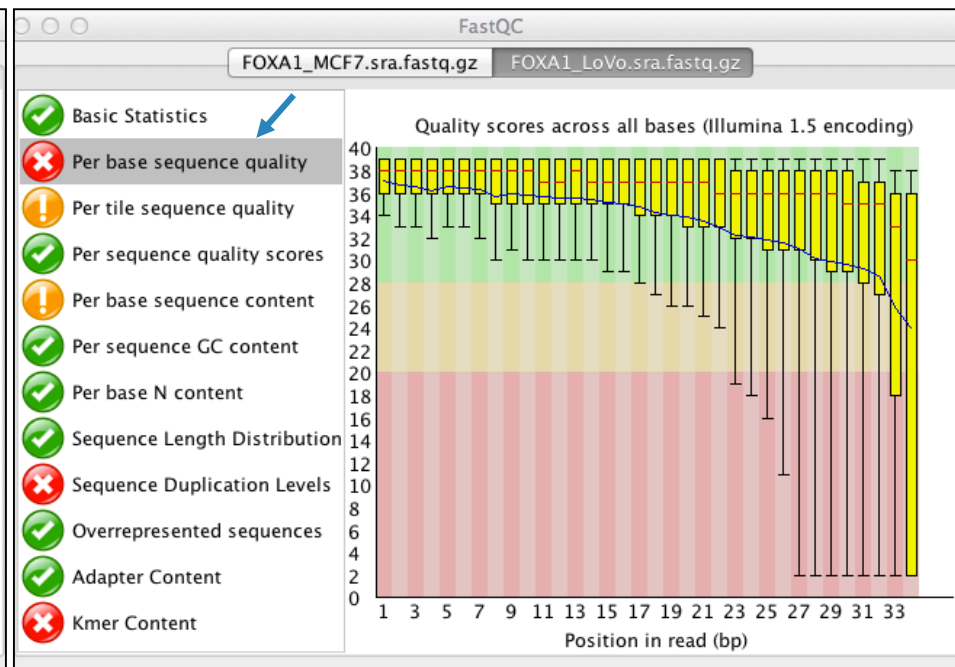
Kmer Content

QC of sequencing reads

FOXA1, MCF7 cells

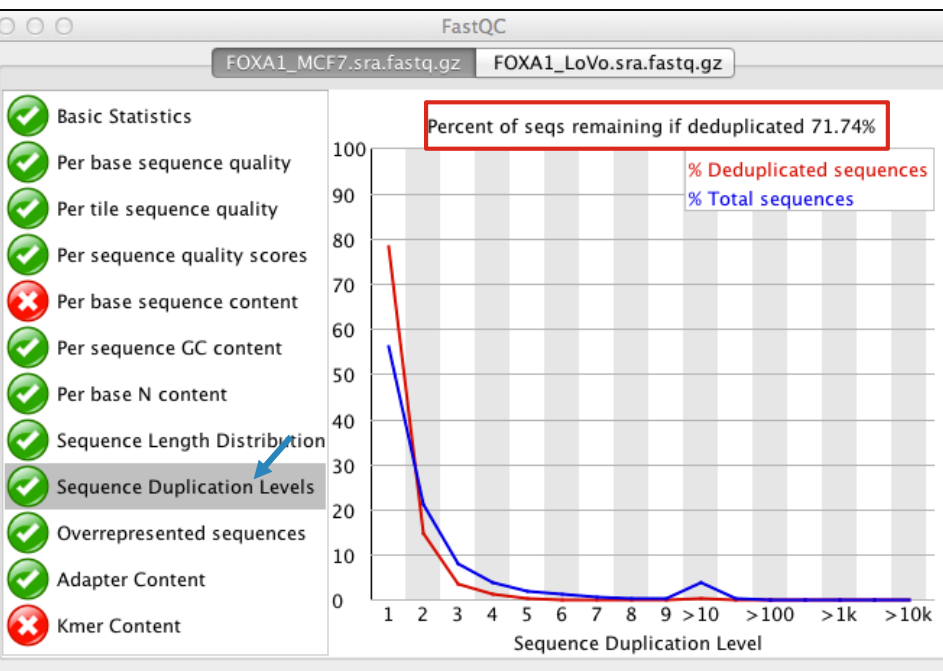


FOXA1, LoVo cells

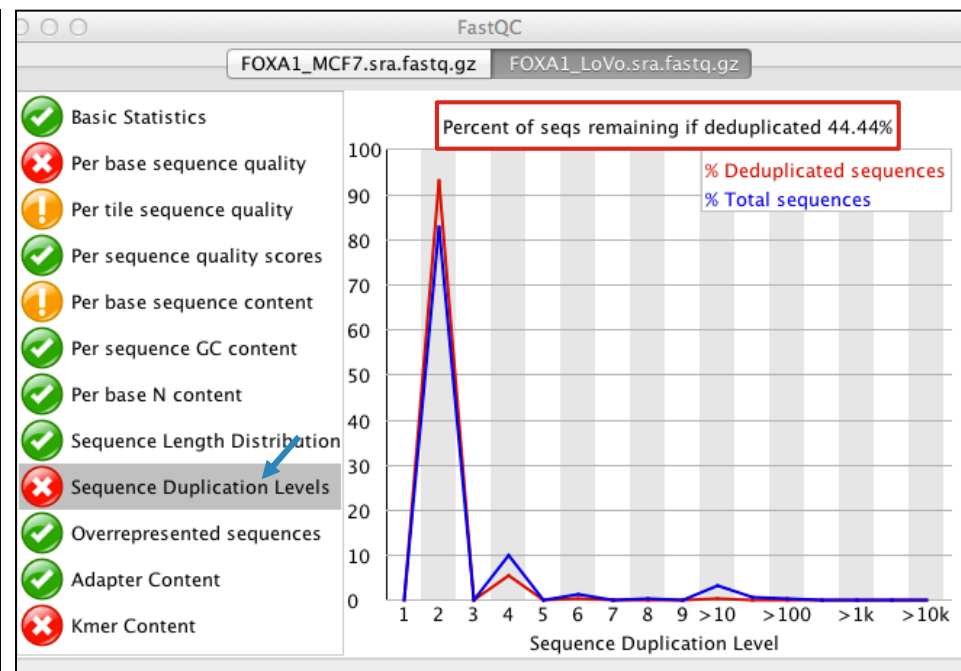


QC of sequencing reads

FOXA1, MCF7 cells



FOXA1, LoVo cells



Data upload (Galaxy)

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 24%

Tools search tools

Get Data

- [Upload File from your computer](#)
- [UCSC Main table browser](#)
- [UCSC Archaea table browser](#)
- [EBI SRA ENA SRA](#)
- [BioMart Central server](#)
- [GrameneMart Central server](#)
- [Flymine server](#)
- [modENCODE fly server](#)
- [modENCODE modMine server](#)
- [MouseMine server](#)
- [Ratmine server](#)
- [YeastMine server](#)
- [modENCODE worm server](#)
- [WormBase server](#)
- [EuPathDB server](#)

Upload File (version 1.1.4) Help from Biostar

File Format:
Auto-detect
Which format? See help below

File:
[Choose File](#) no file selected

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
Your FTP upload directory contains no files.		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at [usegalaxy.org](#) using your Galaxy credentials (email address and password).

Convert spaces to tabs:
☐ Yes


History search datasets

Abcam_ChIPseq_webinar
0 bytes

i This history is empty. You can [load your own data](#) or [get data from an external source](#)

Data upload (Galaxy)

xichen@ebi.ac.uk@usegalaxy.org - FileZilla

Host: Username: Password: Port: 

Command: PASV
Response: 227 Entering Passive Mode (129,114,60,179,199,208).
Command: LIST
Response: 150 Opening BINARY mode data connection for file list
Response: 226 Transfer complete
Status: Directory listing successful

Local site: Remote site:

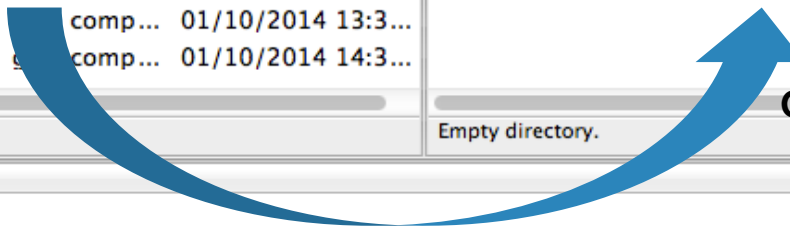
Filename	Filesize	Filetype	Last modified
..			
FOXA1_LoVo.sra.fastq.gz	2126876...	gzip comp...	01/10/2014 11:3...
FOXA1_MCF7.sra.fastq.gz	9705448...	gzip comp...	01/10/2014 11:1...
IgG_LoVo.sra.fastq.gz	1146432...	comp...	01/10/2014 13:3...
Input_MCF7.sra.fastq.gz	8789903...	comp...	01/10/2014 14:3...

4 files. Total size: 5122844366 bytes

Remote site:

Empty directory listing

Empty directory.

Local computer  **Galaxy FTP**

Server/Local file | Direction | Remote file

Queued files | Failed transfers | Successful transfers

Queue: empty

Data upload (Galaxy)

xichen@ebi.ac.uk@usegalaxy.org - FileZilla

Host: usegalaxy.org Username: xichen@ebi.ac. Password: Port: Quickconnect

Command: PASV
Response: 227 Entering Passive Mode (129,114,60,180,171,179).
Command: LIST
Response: 150 Opening BINARY mode data connection for file list
Response: 226 Transfer complete
Status: Directory listing successful

Local site: /Users/xichen/Documents/Abcam_ChIPseq_webinar/ Remote site: /

Filename	Filesize	Filetype	Last modified
..			
FOXA1_LoVo.sra.fastq.gz	2126876...	gzip comp...	01/10/2014 11:3...
FOXA1_MCF7.sra.fastq.gz	9705448...	gzip comp...	01/10/2014 11:1...
IgG_LoVo.sra.fastq.gz	1146432...	gzip comp...	01/10/2014 13:3...
Input_MCF7.sra.fastq.gz	8789903...	gzip comp...	01/10/2014 14:3...

4 files. Total size: 5122844366 bytes

Local computer

Filename	Filesize	Filetype
..		
FOXA1_LoVo.sra.fastq.gz	2126876671	gzip com ...
FOXA1_MCF7.sra.fastq.gz	970544837	gzip com ...
IgG_LoVo.sra.fastq.gz	1146432523	gzip com ...
Input_MCF7.sra.fastq.gz	878990335	gzip com ...

4 files. Total size: 5122844366 bytes

Galaxy FTP

Server/Local file | Direction | Remote file

Queued files Failed transfers Successful transfers (4)

Queue: empty

Data upload (Galaxy)

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 24%

Tools search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- MouseMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server

Upload File (version 1.1.4) Help from Biostar

File Format:
Auto-detect
Which format? See help below

File:
Choose File no file selected
TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:
Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
<input type="checkbox"/> FOXA1_LoVo.sra.fastq.gz	2.0 GB	10/08/2014 04:32:58 AM
<input type="checkbox"/> FOXA1_MCF7.sra.fastq.gz	925.6 MB	10/08/2014 04:07:10 AM
<input type="checkbox"/> IgG_LoVo.sra.fastq.gz	1.1 GB	10/08/2014 04:44:43 AM
<input type="checkbox"/> Input_MCF7.sra.fastq.gz	838.3 MB	10/08/2014 04:47:26 AM

History search datasets

Abcam_ChIPseq_webinar
0 bytes

This history is empty. You can [load your own data](#) or [get data from an external source](#)

Data upload (Galaxy)

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 24%

Tools search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- MouseMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server

Upload File (version 1.1.4) Help from Biostar

File Format:

Auto-detect

eset

fasta

fastq

fastqssanger

fastqillumina

fastqsanger

fastqsolexa

fped

Files larger than 2GB is guaranteed method (below) or FTP (if enabled by

fastqssanger (SOLiD)

fastqillumina

fastqsanger

fastqsolexa

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
<input checked="" type="checkbox"/> FOXA1_LoVo.sra.fastq.gz	2.0 GB	10/08/2014 04:32:58 AM
<input type="checkbox"/> FOXA1_MCF7.sra.fastq.gz	925.6 MB	10/08/2014 04:07:10 AM
<input type="checkbox"/> IgG_LoVo.sra.fastq.gz	1.1 GB	10/08/2014 04:44:43 AM
<input type="checkbox"/> Input_MCF7.sra.fastq.gz	838.3 MB	10/08/2014 04:47:26 AM

This Galaxy server allows you to upload files via FTP. To upload some

History search datasets

Abcam_ChIPseq_webinar

0 bytes

This history is empty. You can [load your own data](#) or [get data from an external source](#)

FastQ format

Format	Illumina pipelines	Quality encoding	Score range	ASCII range
Fastqsanger	Starting Illumina 1.8	Phred + 33	0 to 93	33 - 126
Fastqillumina	Starting Illumina 1.3 Before Illumina 1.8	Phred + 64	0 to 62	64 - 126
Fastqsolexa	Solexa/early Illumina	Solexa + 64	-5 to 62	59 - 126

- Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38:1767-71
- http://en.wikipedia.org/wiki/FASTQ_format

Data upload (Galaxy)

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A status bar on the right indicates 'Using 24%'. The left sidebar contains a 'Tools' menu with various options like 'Ratmine server', 'YeastMine server', 'modENCODE worm server', 'WormBase server', 'EuPathDB server', 'GenomeSpace import from file browser', 'Send Data', 'Lift-Over', 'Text Manipulation', 'Convert Formats', 'FASTA manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', and 'Statistics'. The main content area is titled 'fastqc' and shows a list of tools: 'fastqc', 'fastqc-sanger', 'fastqc-illumina', 'fastqc-sanger', and 'fastqc-solexa'. The 'fastqc-illumina' tool is selected. Below the tool list, there is a text input field for specifying URLs or pasting file contents. A table titled 'Files uploaded via FTP:' lists four files: 'FOXA1_LoVo.sra.fastq.gz' (2.0 GB, 10/08/2014 04:32:58 AM), 'FOXA1_MCF7.sra.fastq.gz' (925.6 MB, 10/08/2014 04:07:10 AM), 'IgG_LoVo.sra.fastq.gz' (1.1 GB, 10/08/2014 04:44:43 AM), and 'Input_MCF7.sra.fastq.gz' (838.3 MB, 10/08/2014 04:47:26 AM). Below the table, there is a section 'Convert spaces to tabs:' with a 'Yes' checkbox. A 'Genome:' dropdown menu is set to 'Additional Species Are Below'. A red arrow points to the 'Execute' button at the bottom of the main content area. The right sidebar shows a 'History' section with a search bar and a message: 'This history is empty. You can load your own data or get data from an external source'.

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 24%

Tools

- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- GenomeSpace import from file browser
- Send Data
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics

fastqc
fastqc-sanger
fastqc-illumina
fastqc-sanger
fastqc-solexa

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
<input checked="" type="checkbox"/> FOXA1_LoVo.sra.fastq.gz	2.0 GB	10/08/2014 04:32:58 AM
<input type="checkbox"/> FOXA1_MCF7.sra.fastq.gz	925.6 MB	10/08/2014 04:07:10 AM
<input type="checkbox"/> IgG_LoVo.sra.fastq.gz	1.1 GB	10/08/2014 04:44:43 AM
<input type="checkbox"/> Input_MCF7.sra.fastq.gz	838.3 MB	10/08/2014 04:47:26 AM

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at usegalaxy.org using your Galaxy credentials (email address and password).

Convert spaces to tabs:

☐ Yes
Use this option if you are entering intervals by hand.

Genome:
----- Additional Species Are Below -----

Execute

History

search datasets

Abcam_ChIPseq_webinar
0 bytes

i This history is empty. You can [load your own data](#) or [get data from an external source](#)

Data upload (Galaxy)

The screenshot displays the Galaxy web interface. At the top, the 'Galaxy' logo is on the left, and navigation tabs for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User' are in the center. A status bar on the right indicates 'Using 34%'. The left sidebar contains a 'Tools' section with a search bar and a list of data sources under 'Get Data', including 'Upload File from your computer', 'UCSC Main table browser', 'UCSC Archaea table browser', 'EBI SRA ENA SRA', 'BioMart Central server', 'GrameneMart Central server', 'Flymine server', 'modENCODE fly server', 'modENCODE modMine server', 'MouseMine server', 'Ratmine server', 'YeastMine server', 'modENCODE worm server', 'WormBase server', and 'EuPathDB server'. The main content area features a large green notification box with a checkmark icon, stating: 'A job has been successfully added to the queue – resulting in the following datasets: 2: FOXA1_MCF7.sra.fastq.gz, 3: IgG_LoVo.sra.fastq.gz, 4: Input_MCF7.sra.fastq.gz. You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.' To the right of the notification is the 'History' pane, which has a search bar and a list of datasets. The datasets listed are 'Abcam_ChIPseq_webinar' (24.0 GB) and four files: '4: Input_MCF7.sra.fastq', '3: IgG_LoVo.sra.fastq', '2: FOXA1_MCF7.sra.fastq', and '1: FOXA1_LoVo.sra.fastq'. Each dataset entry includes icons for viewing, editing, and deleting. A red rectangle highlights the bottom four dataset entries in the History pane.

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 34%

Tools

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- MouseMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server

History

search datasets

Abcam_ChIPseq_webinar
24.0 GB

- 4: Input_MCF7.sra.fastq
- 3: IgG_LoVo.sra.fastq
- 2: FOXA1_MCF7.sra.fastq
- 1: FOXA1_LoVo.sra.fastq

View FastQ files

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 34%

Tools

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- MouseMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server

History

search datasets

- Abcam_ChIPseq_webinar
24.0 GB
- 4: Input_MCF7.sra.fastq
- 3: IgG_LoVo.sra.fastq
- 2: FOXA1_MCF7.sra.fastq**
- 1: FOXA1_LoVo.sra.fastq

Warning: This dataset is large and only the first megabyte is shown below. [Show all](#) | [Save](#)

```
@FOXA1_MCF7.sra.1 HWI-EAS202_0334:3:1:1019:2155 length=36
NTCACCTGCCTGAACCTTTGAACCACTCCCAACACA
+FOXA1_MCF7.sra.1 HWI-EAS202_0334:3:1:1019:2155 length=36
!#####
@FOXA1_MCF7.sra.2 HWI-EAS202_0334:3:1:1020:20749 length=36
NATTAACAAGAAAGAACCTCAAGCTTTAGCCAAAC
+FOXA1_MCF7.sra.2 HWI-EAS202_0334:3:1:1020:20749 length=36
!#####
@FOXA1_MCF7.sra.3 HWI-EAS202_0334:3:1:1020:6178 length=36
NCCAGCCAATAAAGCCATTTAATTGGGACGAAAT
+FOXA1_MCF7.sra.3 HWI-EAS202_0334:3:1:1020:6178 length=36
!#####
@FOXA1_MCF7.sra.4 HWI-EAS202_0334:3:1:1021:19696 length=36
NGTTTGGAGATTACTGATTCATGAGTCACTGTTTAC
+FOXA1_MCF7.sra.4 HWI-EAS202_0334:3:1:1021:19696 length=36
!#####
@FOXA1_MCF7.sra.5 HWI-EAS202_0334:3:1:1022:10967 length=36
NAAAAACAACAAAAAACCCCTGCATGTCTACATT
+FOXA1_MCF7.sra.5 HWI-EAS202_0334:3:1:1022:10967 length=36
!#####
@FOXA1_MCF7.sra.6 HWI-EAS202_0334:3:1:1022:9556 length=36
NAGGTCAATGGTAGAAAAGGAAATATCTTCGTATAA
+FOXA1_MCF7.sra.6 HWI-EAS202_0334:3:1:1022:9556 length=36
!#####
@FOXA1_MCF7.sra.7 HWI-EAS202_0334:3:1:1022:12326 length=36
NCAGTACGAAATAGATTTCATTCGAGGCTGAAA
```

View FastQ files

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 34%

Tools search tools

Get Data Upload File from your computer UCSC Main table browser

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@
1	1	[START OF HEADING]	33	21	!	65	41	A
2	2	[START OF TEXT]	34	22	"	66	42	B
3	3	[END OF TEXT]	35	23	#	67	43	C
4	4	[END OF TRANSMISSION]	36	24	\$	68	44	D
5	5	[ENQUIRY]	37	25	%	69	45	E
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F
7	7	[BELL]	39	27	'	71	47	G
8	8	[BACKSPACE]	40	28	(72	48	H
9	9	[HORIZONTAL TAB]	41	29)	73	49	I
10	A	[LINE FEED]	42	2A	*	74	4A	J
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K
12	C	[FORM FEED]	44	2C	,	76	4C	L
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M
14	E	[SHIFT OUT]	46	2E	.	78	4E	N
15	F	[SHIFT IN]	47	2F	/	79	4F	O
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W
24	18	[CANCEL]	56	38	8	88	58	X
25	19	[END OF MEDIUM]	57	39	9	89	59	Y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z
27	1B	[ESCAPE]	59	3B	;	91	5B	[
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D]
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_

@FOXA1_MCF7.sra.178 HWI-EAS202_0334:3:1:1102:18899 length=36
 AGGATGGCAGGCATGCTAGAAACAGGTCTGGGGGTT
 +FOXA1_MCF7.sra.178 HWI-EAS202_0334:3:1:1102:18899 length=36
 B@A@BB:B@ABDBBB(7987BB<AA#####
 NTCACCTGCCTGAACCTTTGAACCACTCCCAACACA
 +FOXA1_MCF7.sra.1 HWI-EAS202_0334:3:1:1019:2155 length=36

Position	1	2
Base	A	G
Quality string	B	@
ASCII	66	64
Phred	66-33=33	64-33=31
Error Prob.	0.0005	0.0008

115 73 s
 116 74 t
 117 75 u
 118 76 v
 119 77 w
 120 78 x
 121 79 y
 122 7A z
 123 7B {
 124 7C |
 125 7D }
 126 7E ~
 127 7F [DEL]

4: Input_MCF7.sra.fastq
3: IgG_LoVo.sra.fastq
2: FOXA1_MCF7.sra.fastq
1: FOXA1_LoVo.sra.fastq

FastQ Groomer

Galaxy

Analyze DataWorkflowShared DataVisualizationCloudHelpUser

Tools

NGS: QC and manipulation

Select high quality segmentsBuild base quality distributionTabular to FASTQ converterFASTQ Quality Trimmer by sliding windowFASTQ Trimmer by columnFASTQ to Tabular converterFASTQ to FASTA converterFASTQ Summary Statistics by columnFASTQ splitter on joined paired end readsFASTQ joiner on paired end readsFASTQ Masker by quality scoreManipulate FASTQ reads on various attributesFASTQ Groomer convert between various FASTQ quality

FASTQ Groomer (version 1.0.4)Help from Biostar

File to groom:1: FOXA1_LoVo.sra.fastq

Input FASTQ quality scores type:ILLUMINA 1.3-1.7

Advanced Options:Hide Advanced Options

Execute

What it does

This tool offers several conversions options relating to the FASTQ format.

When using *Basic* options, the output will be *sanger* formatted or *cssanger* formatted (when the input is Color Space Sanger).

When converting, if a quality score falls outside of the target score range, it will be coerced to the closest available value (i.e. the minimum or maximum).

When converting between Solexa and the other formats, quality scores are mapped between Solexa and PHRED scales using the equations found in [Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009 Dec 16.](#)

When converting between color space (csSanger) and base/sequence space (Sanger, Illumina, Solexa) formats, adapter bases are lost or gained; if gained, the base 'C' is used as the adapter. You cannot convert a color space read to base

History

search datasets

Abcam_ChIPseq_webinar24.0 GB

4: Input_MCF7.sra.fastq

3: IgG_LoVo.sra.fastq

2: FOXA1_MCF7.sra.fastq

1: FOXA1_LoVo.sra.fastq

FastQ Groomer

Galaxy

Analyze DataWorkflowShared DataVisualizationCloudHelpUser

Using 34%

Tools

NGS: QC and manipulation

Select high quality segmentsBuild base quality distributionTabular to FASTQ converterFASTQ Quality Trimmer by sliding windowFASTQ Trimmer by columnFASTQ to Tabular converterFASTQ to FASTA converterFASTQ Summary Statistics by columnFASTQ splitter on joined paired end readsFASTQ joiner on paired end readsFASTQ Masker by quality scoreManipulate FASTQ reads on various attributesFASTQ Groomer convert between various FASTQ quality

FASTQ Groomer (version 1.0.4)Help from Biostar

File to groom:

1: FOXA1_LoVo.sra.fastq

Input FASTQ quality scores type:

Illumina 1.3-1.7

Advanced Options:

Hide Advanced Options

Execute

What it does

This tool offers several conversions options relating to the FASTQ format.

When using *Basic* options, the output will be *sanger* formatted or *cssanger* formatted (when the input is Color Space Sanger).

When converting, if a quality score falls outside of the target score range, it will be coerced to the closest available value (i.e. the minimum or maximum).

When converting between Solexa and the other formats, quality scores are mapped between Solexa and PHRED scales using the equations found in [Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009 Dec 16.](#)

When converting between color space (csSanger) and base/sequence space (Sanger, Illumina, Solexa) formats, adapter bases are lost or gained; if gained, the base 'G' is used as the adapter. You cannot convert a color space read to base

History

search datasets

Abcam_ChIPseq_webinar33.1 GB

5: FOXA1_LoVo_fastqsanger

4: Input_MCF7.sra.fastq

3: IgG_LoVo.sra.fastq

2: FOXA1_MCF7.sra.fastq

1: FOXA1_LoVo.sra.fastq

Read alignment (bowtie mapping)

The screenshot displays the Galaxy web interface for the Bowtie mapping tool. The left sidebar contains a 'Tools' menu with categories like 'Genome Diversity', 'NGS TOOLBOX BETA', 'Phenotype Association', 'NGS: QC and manipulation', and 'NGS: Mapping' (highlighted with a red box). A red arrow points to the 'Map with Bowtie for Illumina' option. The main panel shows the tool's configuration options, including 'Number of mismatches for SOAP-like alignment policy (-v)' set to -1, 'Whether or not to try as hard as possible to find valid alignments when they exist (-y)' set to 'Do not try hard', 'Report up to n valid alignments per read (-k)' set to 1, 'Whether or not to report all valid alignments per read (-a)' set to 'Do not report all valid alignments', and 'Suppress all alignments for a read if more than n reportable alignments exist (-m)' set to 1 (highlighted with a red box). The right sidebar shows a 'History' panel with a list of jobs, including '9: FOXA1_LoVo_bowtie_hg19' (highlighted with a red box), '8: Input_MCF7_bowtie_hg19', '7: IgG_LoVo_bowtie_hg19', '6: FOXA1_MCF7_bowtie_hg19', '5: FOXA1_LoVo_fastq_sanger', '4: Input_MCF7.sra.fastq', '3: IgG_LoVo.sra.fastq', '2: FOXA1_MCF7.sra.fastq', and '1: FOXA1_LoVo.sra.fastq'.

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 37%

Tools

- Genome Diversity
- NGS TOOLBOX BETA
- Phenotype Association
- NGS: QC and manipulation
- NGS: Mapping**
- Parse blast XML output
- Megablast compare short reads against htgs, nt, and wgs databases
- Bowtie2 is a short-read aligner
- Lastz map short reads against reference sequence
- Map with Bowtie for Illumina
- Map with BWA for Illumina
- NGS: SAM Tools
- NGS: GATK Tools (beta)
- NGS: Peak Calling
- NGS: RNA-seq
- NGS: Picard (beta)
- NGS: Variant Analysis
- NGS: VCF Manipulation

Round to nearest 10

Number of mismatches for SOAP-like alignment policy (-v):
-1
-1 for default MAQ-like alignment policy

Whether or not to try as hard as possible to find valid alignments when they exist (-y):
Do not try hard
Tryhard mode is much slower than regular mode

Report up to n valid alignments per read (-k):
1

Whether or not to report all valid alignments per read (-a):
Do not report all valid alignments

Suppress all alignments for a read if more than n reportable alignments exist (-m):
1
-1 for no limit

Write all reads with a number of valid alignments exceeding the limit set with the -m option to a file (--max):
☐

Write all reads that could not be aligned to a file (--un):
☐

Whether or not to make Bowtie guarantee that reported singleton alignments are 'best' in terms of stratum and in terms of the quality values

History

- Abcam_ChIPseq_webinar
44.2 GB
- 9: FOXA1_LoVo_bowtie_hg19
- 8: Input_MCF7_bowtie_hg19
- 7: IgG_LoVo_bowtie_hg19
- 6: FOXA1_MCF7_bowtie_hg19
- 5: FOXA1_LoVo_fastq_sanger
- 4: Input_MCF7.sra.fastq
- 3: IgG_LoVo.sra.fastq
- 2: FOXA1_MCF7.sra.fastq
- 1: FOXA1_LoVo.sra.fastq

View sam files

<http://samtools.github.io/hts-specs/SAMv1.pdf>

Tools

[Motif Tools](#)
[Multiple Alignments](#)
[Metagenomic analyses](#)
[Genome Diversity](#)

[NGS TOOLBOX BETA](#)
[Phenotype Association](#)
[NGS: QC and manipulation](#)
[NGS: Mapping](#)
[NGS: SAM Tools](#)

[Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region](#)

[Generate pileup from BAM dataset](#)

[SAM-to-BAM converts SAM format to BAM format](#)

[Slice BAM by provided regions](#)

[rmdup](#) remove PCR duplicates

[MPileup](#) SNP and indel caller

[flagstat](#) provides simple stats on BAM files

@SQ SN:chrX LN:155270560
@SQ SN:chrY LN:59373566
@PG ID:Bowtie VN:0.12.7 CL:"bowtie -q -p 8 -S -n 2 -e 70 -l 28 --maxbts 125 -k 1 -"
FOXA1_MCF7.sra.6 HWI-EAS202_0334:3:1:1022:9556 length=36 4 *
FOXA1_MCF7.sra.4 0 chr3
FOXA1_MCF7.sra.10 16 chr7
FOXA1_MCF7.sra.7 HWI-EAS202_0334:3:1:1022:12326 length=36 4 *
FOXA1_MCF7.sra.8 HWI-EAS202_0334:3:1:1023:20321 length=36 4 *
FOXA1_MCF7.sra.2 16 chr3
FOXA1_MCF7.sra.1 16 chr3
FOXA1_MCF7.sra.11 16 chr7
FOXA1_MCF7.sra.9 0 chr15
FOXA1_MCF7.sra.3 16 chr17
FOXA1_MCF7.sra.15 16 chr1
FOXA1_MCF7.sra.14 0 chr4
FOXA1_MCF7.sra.13 16 chr3
FOXA1_MCF7.sra.16 0 chr4
FOXA1_MCF7.sra.5 16 chr5
FOXA1_MCF7.sra.17 16 chr1
FOXA1_MCF7.sra.12 0 chr8
FOXA1_MCF7.sra.24 HWI-EAS202_0334:3:1:1031:19808 length=36 4 *
FOXA1_MCF7.sra.19 16 chr8
FOXA1_MCF7.sra.21 16 chr20
FOXA1_MCF7.sra.22 0 chr12

History

Abcam_ChIPseq_webinar
44.2 GB

9: FOXA1_LoVo_bowtie_hg19
8: Input_MCF7_bowtie_hg19
7: IgG_LoVo_bowtie_hg19
6: FOXA1_MCF7_bowtie_hg19
5: FOXA1_LoVo_fastq_anger
4: Input_MCF7.sra.fastq
3: IgG_LoVo.sra.fastq
2: FOXA1_MCF7.sra.fastq

Filter sam files

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 45%

Tools

- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- Genome Diversity
- NGS TOOLBOX BETA
- Phenotype Association
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools**
 - Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region
 - Generate pileup from BAM dataset
 - SAM-to-BAM converts SAM format to BAM format
 - Slice BAM by provided regions
 - rmdup remove PCR duplicates
 - MPileup SNP and indel caller
 - flagstat provides simple stats on BAM files

Filter SAM or BAM, output SAM or BAM (version 1.1.1)

Only output alignments with all of these flag bits set:

Select All Unselect All

- ☐ Read is paired
- ☐ Read is mapped in a proper pair
- ☐ The read is unmapped
- ☐ The mate is unmapped
- ☐ Read strand
- ☐ Mate strand
- ☐ Read is the first in a pair
- ☐ Read is the second in a pair
- ☐ The alignment or this read is not primary
- ☐ The read fails platform/vendor quality checks
- ☐ The read is a PCR or optical duplicate

Skip alignments with any of these flag bits set:

Select All Unselect All

- ☐ Read is paired
- ☐ Read is mapped in a proper pair
- ☒ The read is unmapped
- ☐ The mate is unmapped
- ☐ Read strand
- ☐ Mate strand
- ☐ Read is the first in a pair
- ☐ Read is the second in a pair
- ☐ The alignment or this read is not primary
- ☐ The read fails platform/vendor quality checks

History

search datasets

- Abcam_ChIPseq_webinar 44.2 GB
- 13: FOXA1_LoVo_aln.bam**
- 12: Input_MCF7_aln.bam**
- 11: IgG_LoVo_aln.bam**
- 10: FOXA1_MCF7_aln.bam**
- 9: FOXA1_LoVo_bowtie_hg19
- 8: Input_MCF7_bowtie_hg19
- 7: IgG_LoVo_bowtie_hg19
- 6: FOXA1_MCF7_bowtie_hg19

Peak calling (MACS)

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 46%

Tools

- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- Genome Diversity
- NGS TOOLBOX BETA
- Phenotype Association
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: GATK Tools (beta)
- NGS: Peak Calling
 - SICER Statistical approach for the Identification of ChIP-Enriched Regions
 - MACS Model-based Analysis of ChIP-Seq**
- NGS: RNA-seq
- NGS: Picard (beta)
- NGS: Variant Analysis
- NGS: VCF Manipulation
- snpEff

MACS (version 1.0.1) Help from Biostar

Experiment Name:
FOXA1_ChIPseq_MCF7

Paired End Sequencing:
Single End

ChIP-Seq Tag File:
10: FOXA1_MCF7_aln.bam

ChIP-Seq Control File:
12: Input_MCF7_aln.bam

Effective genome size:
270000000
default: 2.7e+9

Tag size:
36

Band width:
300

Pvalue cutoff for peak detection:
1e-05
default: 1e-5

Select the regions with MFOLD high-confidence enrichment ratio against background to build model:

History

- 19: MACS on data 12 and data 10 (html report)
- 18: MACS on data 12 and data 10 (control: wig)
- 17: MACS on data 12 and data 10 (treatment: wig)
- 16: MACS on data 12 and data 10 (negative peaks: interval)
- 15: MACS on data 12 and data 10 (peaks: interval)
- 14: MACS on data 12 and data 10 (peaks: bed)
- 13: FOXA1_LoVo_aln.bam
- 12: Input_MCF7_aln.bam
- 11: IgG_LoVo_aln.bam
- 10: FOXA1_MCF7_aln.bam

Peak calling (MACS)

Galaxy

Analyze DataWorkflowShared DataVisualizationCloudHelpUser

Using 46%

Tools

[Motif Tools](#)[Multiple Alignments](#)[Metagenomic analyses](#)[Genome Diversity](#)

[NGS TOOLBOX BETA](#)[Phenotype Association](#)[NGS: QC and manipulation](#)[NGS: Mapping](#)[NGS: SAM Tools](#)[NGS: GATK Tools \(beta\)](#)[NGS: Peak Calling](#)

[SICER Statistical approach for the Identification of ChIP-Enriched Regions](#)

[MACS Model-based Analysis of ChIP-Seq](#)

[NGS: RNA-seq](#)[NGS: Picard \(beta\)](#)[NGS: Variant Analysis](#)[NGS: VCF Manipulation](#)[snpEff](#)

Parse xls files into into distinct interval files:☒

Save shifted raw tag count at every bp into a wiggle file:

Save

Extend tag from its middle point to a wigextend size fragment.:

-1

Use value less than 0 for default (modeled d)

Resolution for saving wiggle files:

10

Use fixed background lambda as local lambda for every peak region:☐

up to 9X more time consuming

3 levels of regions around the peak region to calculate the maximum lambda as local lambda:

1000,5000,10000

Build Model:

Do not build the shifting model

Arbitrary shift size in bp:

100

Diagnosis report:

Do not produce report (faster)

up to 9X more time consuming

History

19: MACS on data 12 and data 10 (html report)

18: MACS on data 12 and data 10 (control: wig)

17: MACS on data 12 and data 10 (treatment: wig)

16: MACS on data 12 and data 10 (negative peaks: interval)

15: MACS on data 12 and data 10 (peaks: interval)

14: MACS on data 12 and data 10 (peaks: bed)

13: FOXA1_LoVo_aln.bam

12: Input_MCF7_aln.bam

11: IgG_LoVo_aln.bam

10: FOXA1_MCF7_aln.bam

Discover more at www.abcam.com

Copyright © 2014 Abcam plc.

Peak calling sanity check

Additional output created by MACS (FOXA1_ChIPseq_MCF7)

Additional Files:

- [FOXA1_ChIPseq_MCF7_negative_peaks.xls](#)
- [FOXA1_ChIPseq_MCF7_peaks.xls](#)

Messages from MACS:

```
INFO @ Wed, 08 Oct 2014 09:51:37:
# ARGUMENTS LIST:
# name = FOXA1_ChIPseq_MCF7
# format = BAM
# ChIP-seq file = /galaxy-repl/main/files/009/383/dataset_9383838.dat
# control file = /galaxy-repl/main/files/009/383/dataset_9383847.dat
# effective genome size = 2.70e+09
# tag size = 36
# band width = 300
# model fold = 32
# pvalue cutoff = 1.00e-05
# Ranges for calculating regional lambda are : peak_region,1000,50
INFO @ Wed, 08 Oct 2014 09:51:37: #1 read tag files...
INFO @ Wed, 08 Oct 2014 09:51:37: #1 read treatment tags...
INFO @ Wed, 08 Oct 2014 09:51:49: 1000000
INFO @ Wed, 08 Oct 2014 09:52:01: 2000000
INFO @ Wed, 08 Oct 2014 09:52:11: 3000000
INFO @ Wed, 08 Oct 2014 09:52:21: 4000000
INFO @ Wed, 08 Oct 2014 09:52:32: 5000000
INFO @ Wed, 08 Oct 2014 09:52:43: 6000000
INFO @ Wed, 08 Oct 2014 09:52:53: 7000000
INFO @ Wed, 08 Oct 2014 09:53:04: 8000000
INFO @ Wed, 08 Oct 2014 09:53:15: 9000000
```

ChIPseq_MCF7_treat_chr2" description="Shifted Merged MACS tag counts for every 10 bp"

11431	1
11441	1
11451	1
11461	1
11471	1
11481	1
11491	1
11501	1
11511	1
11521	1
11531	1
11541	1
11551	1
11561	1
11571	1
11581	1
11591	1
11601	1
11611	1
11621	1
11951	1
11961	3
11971	3

19: MACS on data 12 and data 10 (html report)

18: MACS on data 12 and data 10 (control: wig)

17: MACS on data 12 and data 10 (treatment: wig)

16: MACS on data 12 and data 10 (negative peaks: interval)

15: MACS on data 12 and data 10 (peaks: interval)

14: MACS on data 12 and data 10 (peaks: bed)

#peaks file

This file is generated by MACS

ARGUMENTS LIST:

name = FOXA1_ChIPseq_MCF7

format = BAM

ChIP-seq file = /galaxy-repl/main/files/009/383/dataset_9383838.dat

control file = /galaxy-repl/main/files/009/383/dataset_9383847.dat

effective genome size = 2.70e+09

tag size = 36

band width = 300

model fold = 32

pvalue cutoff = 1.00e-05

Ranges for calculating regional lambda are : peak_region,1000,5000,10000

unique tags in treatment: 14826089

total tags in treatment: 21188600

unique tags in control: 8228973

total tags in control: 8335100

d = 200

#chr	start	end	length	summit	tags	-10*log10(pvalue)	fo
chr1	780950	781339	389	195	14	110.23	
chr1	852480	853459	979	420	24	61.62	
chr1	1008944	1009947	1003	258	48	237.63	
chr1	1014777	1015550	773	183	19	68.97	
chr1	1059434	1061205	1771	850	104	359.87	
chr1	1067737	1068626	889	382	20	65.46	

track name="MACS peaks for FOXA1_CHIP_MCF7"

chr1	780950	781339	MACS_peak_1	110.23
chr1	852480	853459	MACS_peak_2	61.62
chr1	1008944	1009947	MACS_peak_3	237.63
chr1	1014777	1015550	MACS_peak_4	68.97
chr1	1059434	1061205	MACS_peak_5	359.87
chr1	1067737	1068626	MACS_peak_6	65.46
chr1	1071775	1072324	MACS_peak_7	154.33
chr1	1440160	1440574	MACS_peak_8	55.80
chr1	1611790	1612746	MACS_peak_9	55.53
chr1	1694316	1695236	MACS_peak_10	150.32
chr1	1724550	1725590	MACS_peak_11	214.50
chr1	1725625	1726254	MACS_peak_12	84.85
chr1	1784478	1785595	MACS_peak_13	528.68
chr1	1789576	1791759	MACS_peak_14	1161.26
chr1	1933232	1933725	MACS_peak_15	520.55
chr1	2035986	2037385	MACS_peak_16	60.77
chr1	2121024	2121820	MACS_peak_17	73.73
chr1	2139531	2140249	MACS_peak_18	82.17
chr1	2242849	2243468	MACS_peak_19	103.45
chr1	2454504	2455380	MACS_peak_20	62.60
chr1	2466235	2467162	MACS_peak_21	98.20

Peak calling sanity check

History

19: MACS on data 12 and data 10 (html report)

18: MACS on data 12 and data 10 (control: wig)

17: MACS on data 12 and data 10 (treatment: wig)

16: MACS on data 12 and data 10 (negative peaks: interval)

15: MACS on data 12 and data 10 (peaks: interval)

14: MACS on data 12 and data 10 (peaks: bed)

13: FOXA1 LoVo aln.bam

12: Input_MCF7_aln.bam

Download

at Ensembl

Current

display at RViewer

main

display at UCSC

main

1. Chrom

2. Start

3. End

#peaks file

This file is generated by MACS

ARGUMENTS LIST:

name = FOXA1_ChIPseq_MCF7

format = BAM

ChIP-seq file = /galaxy-repl/main/fil

	A	B	C	D	E	F	G	H	I
1	#peaks file								
2	# This file is generated by MACS								
3	# ARGUMENTS LIST:								
4	# name = FOXA1_ChIPseq_MCF7								
5	# format = BAM								
6	# ChIP-seq file = /galaxy-repl/main/files/009/383/dataset_9383838.dat								
7	# control file = /galaxy-repl/main/files/009/383/dataset_9383847.dat								
8	# effective genome size = 2.70e+09								
9	# tag size = 36								
10	# band width = 300								
11	# model fold = 32								
12	# pvalue cutoff = 1.00e-05								
13	# Ranges for calculating regional lambda are : peak_region,1000,5000,10000								
14	# unique tags in treatment: 14826089								
15	# total tags in treatment: 21188600								
16	# unique tags in control: 8228973								
17	# total tags in control: 8335100								
18	# d = 200								
19	#chr	start	end	length	summit	tags	#NAME?	fold_enrich	FDR(%)
20	chr1	780950	781339	389	195	14	110.23	12.75	1.08
21	chr1	852480	853459	979	420	24	61.62	11.35	2.62
22	chr1	1008944	1009947	1003	258	48	237.63	22.9	0.39
23	chr1	1014777	1015550	773	183	19	68.97	10.02	2.15
24	chr1	1059434	1061205	1771	850	104	359.87	27.36	0.21
25	chr1	1067737	1068626	889	382	20	65.46	12.34	2.37
26	chr1	1071775	1072324	549	370	24	154.33	19.12	0.63
27	chr1	1440160	1440574	414	194	16	55.8	5.36	3.15
28	chr1	1611790	1612746	956	557	18	55.53	10.02	3.19
29	chr1	1694316	1695236	920	279	31	150.32	24.59	0.66
30	chr1	1724550	1725590	1040	531	81	214.5	15.73	0.43
31	chr1	1725625	1726254	629	198	22	84.85	10.41	1.57
32	chr1	1784478	1785595	1117	774	82	528.68	38.16	0.15
33	chr1	1789576	1791759	2183	1383	158	1161.26	74.67	0.12
34	chr1	1933232	1933725	493	267	76	520.55	19.15	0.15
35	chr1	2035986	2037385	1399	703	44	60.77	9.88	2.69
36	chr1	2121024	2121820	796	597	19	73.73	8.2	1.93
37	chr1	2139531	2140249	718	392	25	82.17	11.1	1.62
38	chr1	2242849	2243468	619	324	30	103.45	10.74	1.19
39	chr1	2454504	2455380	876	414	22	62.6	10.57	2.55
40	chr1	2466235	2467162	927	478	28	98.2	12.49	1.29
41	chr1	2512229	2513235	1006	316	27	110.09	20.94	1.08
42	chr1	2718122	2718510	388	195	20	192.47	18.21	0.5
43	chr1	2782677	2783563	886	199	56	395.18	40.98	0.17
44	chr1	3001933	3003366	1433	690	50	150.18	22.2	0.66
45	chr1	3043579	3044573	994	353	55	252.29	19.86	0.36

Peak calling sanity check

History

- 19: MACS on data 12 and data 10 (html report)
- 18: MACS on data 12 and data 10 (control: wig)
- 17: MACS on data 12 and data 10 (treatment: wig)
- 16: MACS on data 12 and data 10 (negative peaks: interval)
- 15: MACS on data 12 and data 10 (peaks: interval)**
67,400 regions, 19 comments
format: interval, database: hg19
Download at Ensembl [Current](#)
display at RViewer [main](#)
display at UCSC [main](#)
- 14: MACS on data 12 and data 10 (peaks: bed)
- 13: FOXA1_LoVo_aln.bam
- 12: Input_MCF7_aln.bam

1. Chrom	2. Start	3. End
#peaks file		
# This file is generated by MACS		
# ARGUMENTS LIST:		
# name = FOXA1_ChIPseq_MCF7		
# format = BAM		
# ChIP-seq file = /galaxy-repl/main/fil		

Whether my ChIP-seq experiments work or not?

- Looking at interval files
- Visual inspection of binding signal
- Looking at motif discovery results

Peak calling sanity check

FOXA1, MCF7 cells

Smallest to largest

1	#chr	start	end	length	summit	tags	#NAME?	fold_enrich	FDR(%)
2	chr5	104951553	104954096	2543	968	711	3100	332.35	0
3	chr4	101699056	101703001	3945	2442	842	3100	315.05	0
4	chr5	19878819	19882069	3250	2235	682	3100	305.04	0
5	chr5	23493478	23496612	3134	1376	905	3100	303.22	0
6	chrX	16679646	16681811	2165	1052	669	3100	303.22	0
7	chrX	69439770	69440829	1059	618	485	3100	298.66	0
8	chr18	21794144	21796891	2747	1870	994	3100	297.75	0
9	chr5	57365345	57366720	1375	566	570	3100	296.84	0
10	chr22	29217442	29220242	2800	1644	617	3100	294.86	0
11	chr14	38014909	38017370	2461	1371	834	3100	291.39	0
12	chr18	52432497	52435727	3230	2002	637	3100	289.5	0
13	chr2	162727242	162729791	2549	1629	575	3100	283.18	0
14	chr10	68874080	68876479	2399	833	475	3100	282.27	0
15	chr2	189227945	189230610	2665	971	531	3100	280.45	0
16	chr4	68794026	68796123	2097	558	624	3100	280.45	0
17	chr1	22674381	22675898	1517	686	388	3100	279.54	0
18	chrX	82068375	82069481	1106	587	471	3100	277.72	0
19	chr4	162120738	162123716	2978	1471	950	3100	275.9	0
20	chr5	25254597	25256218	1621	843	371	3100	275.9	0
21	chr8	67625148	67627511	2363	1334	826	3100	273.17	0
22	chr1	66917118	66918987	1869	877	424	3100	270.44	0
23	chr12	97928917	97932273	3356	1373	538	3100	269.52	0
24	chr2	10626608	10628203	1595	836	396	3100	269.12	0
25	chr8	75684369	75686171	1802	933	417	3100	268.61	0
26	chr5	29905375	29908723	3348	914	596	3100	266.79	0
27	chr5	82560451	82562085	1634	775	444	3100	266.79	0
28	chr10	122012007	122013584	1577	877	483	3100	265.88	0
29	chr21	22610613	22612309	1696	796	481	3100	265.88	0
30	chr4	89331968	89333621	1653	848	372	3100	264.91	0
31	chr13	108614175	108616389	2214	643	553	3100	263.15	0
32	chr2	162700715	162702247	1532	801	429	3100	263.15	0
33	chr11	84898382	84899868	1486	873	486	3100	262.24	0
34	chr12	61097035	61099564	2529	1473	544	3100	262.24	0
35	chr6	125461794	125463647	1853	864	459	3100	260.51	0
36	chr2	213804218	213806420	2202	797	399	3100	259.63	0
37	chr6	136927574	136929981	2407	944	426	3100	259.53	0
38	chr5	21170344	21172340	1996	1080	497	3100	259.3	0
39	chr1	162843360	162844811	1451	646	373	3100	258.6	0
40	chr6	124051822	124055875	4053	3095	607	3100	258.6	0
41	chrX	11256815	11258383	1568	801	356	3100	258.6	0
42	chr2	198477539	198480531	2992	869	641	3100	258.16	0
43	chr4	125808108	125810040	1932	1022	498	3100	257.57	0
44	chr2	165548112	165552676	4564	2004	764	3100	256.78	0
45	chr8	94677448	94680000	2552	1164	904	3100	256.73	0
46	chr4	106506627	106508470	1843	784	423	3100	255.87	0
47	chr5	22762090	22763623	1533	799	443	3100	255.87	0
48	chr15	93288445	93289898	1453	852	368	3100	254.96	0
49	chr13	86289367	86291593	2226	1338	628	3100	253.11	0
50	chr3	140675149	140676928	1779	988	420	3100	252.22	0

of peaks with < 1% FDR: 42,041

Fold enrichment range: 3.8 ~ 332.35

Tags range: 14 ~ 4932

FOXA1, LoVo cells

Smallest to largest

1	#chr	start	end	length	summit	tags	#NAME?	fold_enrich	FDR(%)
2	chr5	116177834	116179259	1425	747	66	275.98	20.37	0
3	chr4	149874535	149876113	1578	766	67	238.31	15.93	0
4	chr17	15894112	15894996	884	567	31	143.25	15.65	0
5	chr17	76404816	76406095	1279	785	58	198.93	15.21	0
6	chr11	124693095	124694258	1163	618	36	131.89	14.64	0
7	chr19	55988312	55988773	461	185	20	148.25	13.74	0
8	chr4	113003730	113005494	1764	647	51	134.25	12.58	0
9	chr3	182326547	182328359	1812	1023	80	231.32	12.3	0
10	chr10	70976037	70977787	1750	1005	54	146.89	11.98	0
11	chr1	167051198	167053631	2433	741	92	149.84	11.79	0
12	chr3	124690675	124691866	1191	853	50	160.53	11.77	0
13	chr1	61606618	61608336	1718	733	60	132.7	10.98	0
14	chr13	41218459	41220190	1731	1014	59	167.81	10.98	0
15	chr15	29693139	29694540	1401	918	49	146.13	10.98	0
16	chr2	43861267	43862449	1182	708	57	151.81	10.98	0
17	chr2	208879431	208880975	1544	979	41	126.4	10.98	0
18	chr5	73388792	73390442	1650	807	67	129.52	10.75	0
19	chr18	77103220	77104798	1578	741	49	170.62	10.74	0
20	chr14	74599896	74601434	1538	648	40	136.94	10.71	0
21	chr17	64139390	64140549	1159	871	40	137.15	10.57	0
22	chr2	192507555	192508899	1344	781	50	159.17	10.43	0
23	chr10	69895168	69896783	1615	909	41	135.9	10.3	0
24	chr4	105887324	105888931	1607	937	41	143.13	10.3	0
25	chr17	28671108	28672535	1427	817	48	124.8	10.14	0
26	chr18	47653283	47656381	3098	1848	99	201.25	9.97	0
27	chr4	48579736	48581225	1489	793	46	147.13	9.88	0
28	chr14	75666996	75668487	1491	365	52	134.79	9.83	0
29	chr7	128980568	128983804	3236	1348	130	229.28	9.71	0
30	chr9	20382270	20383628	1358	683	48	132.47	9.63	0
31	chr11	114317990	114320346	2356	776	64	174.13	9.61	0
32	chr6	46949803	46951842	2039	539	55	167.27	9.52	0
33	chr10	115252379	115254684	2305	1681	58	153.74	9.51	0
34	chr5	174804394	174806984	2590	1281	128	206.45	9.49	0
35	chr4	113033655	113035260	1605	871	66	130.21	9.36	0
36	chr8	90623325	90624995	1670	943	54	145.61	9.34	0
37	chr11	110174970	110177126	2156	1480	81	156.82	9.3	0
38	chr17	42765829	42767833	2004	1767	61	134.06	9.29	0
39	chr1	192830624	192831986	1362	575	51	125.9	9.23	0
40	chr10	97387813	97389681	1868	977	86	180.98	9.23	0
41	chr18	20262322	20264372	2050	1189	87	162.19	9.23	0
42	chr6	70490080	70491679	1599	689	53	133.81	9.12	0
43	chr11	114653498	114655306	1808	1122	51	166.87	8.94	0
44	chr6	4319616	4322089	2473	743	82	177.59	8.85	0
45	chr2	46190988	46192447	1459	485	65	124.68	8.79	0
46	chr2	48340325	48342316	1991	1411	56	177.92	8.79	0
47	chr21	38159951	38161694	1743	721	54	138.2	8.79	0
48	chr6	136886254	136889902	3648	1164	130	191.09	8.67	0
49	chr17	39057862	39060389	2527	578	89	175.03	8.65	0
50	chr1	68663149	68664655	1506	243	56	173.19	8.64	0

of peaks with < 1% FDR: 430

Fold enrichment range: 3.05 ~ 20.37

Tags range: 15 ~ 224

Peak calling sanity check

FOXA1, MCF7 cells Largest to smallest

1	#chr	start	end	length	summit	tags	#NAME?	fold_enrichm	DR(%)
2	chr5	104951553	104954096	2543	968	711	3100	332.35	0
3	chr4	101699056	101703001	3945	2442	842	3100	315.05	0
4	chr5	19878819	19882069	3250	2235	682	3100	305.04	0
5	chr5	23493478	23496612	3134	1376	905	3100	303.22	0
6	chrX	16679646	16681811	2165	1052	669	3100	303.22	0
7	chrX	69439770	69440829	1059	618	485	3100	298.66	0
8	chr18	21794144	21796891	2747	1870	994	3100	297.75	0
9	chr5	57365345	57366720	1375	566	570	3100	296.84	0
10	chr22	29217442	29220242	2800	1644	617	3100	294.86	0
11	chr14	38014909	38017370	2461	1371	834	3100	291.39	0
12	chr18	52432497	52435727	3230	2002	637	3100	285	0
13	chr2	162727242	162729791	2549	1629	575	3100	283.18	0
14	chr10	68874080	68876479	2399	833	475	3100	282.27	0
15	chr2	189227945	189230610	2665	971	531	3100	280.45	0
16	chr4	68794026	68796123	2097	558	624	3100	280.45	0
17	chr1	22674381	22675898	1517	686	388	3100	279.54	0
18	chrX	82068375	82069481	1106	587	471	3100	277.72	0
19	chr4	162120738	162123716	2978	1471	950	3100	275.9	0
20	chr5	25254597	25256218	1621	843	371	3100	275.9	0
21	chr8	67625148	67627511	2363	1334	826	3100	273.17	0
22	chr1	66917118	66918987	1869	877	424	3100	270.44	0
23	chr12	97928917	97932273	3356	1373	538	3100	269.52	0
24	chr2	10626608	10628203	1595	836	396	3100	269.12	0
25	chr8	75684369	75686171	1802	933	417	3100	268.61	0
26	chr5	29905375	29908723	3348	914	596	3100	266.79	0
27	chr5	82560451	82562085	1634	775	444	3100	266.79	0
28	chr10	122012007	122013584	1577	877	483	3100	265.88	0
29	chr21	22610613	22612309	1696	796	481	3100	265.88	0
30	chr4	89331968	89333621	1653	848	372	3100	264.97	0
31	chr13	108614175	108616389	2214	643	553	3100	263.15	0
32	chr2	162700715	162702247	1532	801	429	3100	263.15	0
33	chr11	84898382	84899868	1486	873	486	3100	262.24	0
34	chr12	61097035	61099564	2529	1473	544	3100	262.24	0
35	chr6	125461794	125463647	1853	864	459	3100	260.51	0
36	chr2	213804218	213806420	2202	797	399	3100	259.61	0
37	chr6	136927574	136929981	2407	944	426	3100	259.51	0
38	chr5	21170344	21172340	1996	1080	497	3100	259.3	0
39	chr1	162843360	162844811	1451	646	373	3100	258.6	0
40	chr6	124051822	124055875	4053	3095	607	3100	258.6	0
41	chrX	11256815	11258383	1568	801	356	3100	258.6	0
42	chr2	198477539	198480531	2992	869	641	3100	258.16	0
43	chr4	125808108	125810040	1932	1022	498	3100	257.57	0
44	chr2	165548112	165552676	4564	2004	764	3100	256.78	0
45	chr8	94677448	94680000	2552	1164	904	3100	256.73	0
46	chr4	106506627	106508470	1843	784	423	3100	255.87	0
47	chr5	22762090	22763623	1533	799	443	3100	255.87	0
48	chr15	93288445	93289898	1453	852	368	3100	254.96	0
49	chr13	86289367	86291593	2226	1338	628	3100	253.13	0
50	chr3	140675149	140676928	1779	988	420	3100	252.22	0

of peaks with < 1% FDR: 42,041

Fold enrichment range: 3.8 ~ 332.35

Tags range: 14 ~ 4932

FOXA1, LoVo cells Largest to smallest

1	#chr	start	end	length	summit	tags	#NAME?	fold_enrichm	DR(%)
2	chr5	116177834	116179259	1425	747	66	275.98	20.37	0
3	chr4	149874535	149876113	1578	766	67	238.31	15.93	0
4	chr17	15894112	15894996	884	567	31	143.25	15.65	0
5	chr17	76404816	76406095	1279	785	58	198.93	15.21	0
6	chr11	124693095	124694258	1163	618	36	131.89	14.64	0
7	chr19	55988312	55988773	461	185	20	148.25	13.74	0
8	chr4	113003730	113005494	1764	647	51	134.25	12.58	0
9	chr3	182326547	182328359	1812	1023	80	231.32	12.3	0
10	chr10	70976037	70977787	1750	1005	54	146.89	11.98	0
11	chr1	167051198	167053631	2433	741	92	149.84	11.79	0
12	chr3	124690675	124691866	1191	853	50	160.53	11.77	0
13	chr1	61606618	61608336	1718	733	60	132.7	10.98	0
14	chr13	41218459	41220190	1731	1014	59	167.81	10.98	0
15	chr15	29693139	29694540	1401	918	49	146.13	10.98	0
16	chr2	43861267	43862449	1182	708	57	151.81	10.98	0
17	chr2	208879431	208880975	1544	979	41	126.4	10.98	0
18	chr5	73388792	73390442	1650	807	67	129.52	10.75	0
19	chr18	77103220	77104798	1578	741	49	170.62	10.74	0
20	chr14	74599896	74601434	1538	648	40	136.94	10.71	0
21	chr17	64139390	64140549	1159	871	40	137.15	10.57	0
22	chr2	192507555	192508899	1344	781	50	159.17	10.43	0
23	chr10	69895168	69896783	1615	909	41	135.9	10.3	0
24	chr4	105887324	105888931	1607	937	41	143.13	10.3	0
25	chr17	28671108	28672535	1427	817	48	124.8	10.14	0
26	chr18	47653283	47656381	3098	1848	99	201.25	9.97	0
27	chr4	48579736	48581225	1489	793	46	147.13	9.88	0
28	chr14	75666996	75668487	1491	365	52	134.79	9.83	0
29	chr7	128980568	128983804	3236	1348	130	229.28	9.71	0
30	chr9	20382270	20383628	1358	683	48	132.47	9.63	0
31	chr11	114317990	114320346	2356	776	64	174.13	9.61	0
32	chr6	46949803	46951842	2039	539	55	167.27	9.52	0
33	chr10	115252379	115254684	2305	1681	58	153.74	9.51	0
34	chr5	174804394	174806984	2590	1281	128	206.45	9.49	0
35	chr4	113033655	113035260	1605	871	66	130.21	9.36	0
36	chr8	90623325	90624995	1670	943	54	145.61	9.34	0
37	chr11	110174970	110177126	2156	1480	81	156.82	9.3	0
38	chr17	42765829	42767833	2004	1767	61	134.06	9.29	0
39	chr1	192830624	192831986	1362	575	51	125.9	9.23	0
40	chr10	97387813	97389681	1868	977	86	180.98	9.23	0
41	chr18	20262322	20264372	2050	1189	87	162.19	9.23	0
42	chr6	70490080	70491679	1599	689	53	133.81	9.12	0
43	chr11	114653498	114655306	1808	1122	51	166.87	8.94	0
44	chr6	4319616	4322089	2473	743	82	177.59	8.85	0
45	chr2	46190988	46192447	1459	485	65	124.68	8.79	0
46	chr2	48340325	48342316	1991	1411	56	177.92	8.79	0
47	chr21	38159951	38161694	1743	721	54	138.2	8.79	0
48	chr6	136886254	136889902	3648	1164	130	191.09	8.67	0
49	chr17	39057862	39060389	2527	578	89	175.03	8.65	0
50	chr1	68663149	68664655	1506	243	56	173.19	8.64	0

of peaks with < 1% FDR: 430

Fold enrichment range: 3.05 ~ 20.37

Tags range: 15 ~ 224

Signal visualization (UCSC)

The screenshot displays the Galaxy web interface. On the left, the 'Tools' sidebar is visible, with 'Convert Formats' and 'Wig/BedGraph-to-bigWig converter' highlighted with red boxes. The main panel shows the configuration for the 'Wig/BedGraph-to-bigWig (version 1.1.0)' tool. The 'Convert' section, also highlighted with a red box, shows a file named '17: MACS on data 12 and data 10 (treatment: wig)' selected. Below this, the 'Converter settings to use' are set to 'Default'. An 'Execute' button is present. The 'Syntax' section provides information about the tool's input formats: 'Wiggle format' and 'BED format'. The right sidebar shows a 'History' list with 10 entries, with the last four entries (17-24) highlighted with red boxes. These entries are related to MACS signal processing and visualization.

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 48%

Tools

Convert Formats

- Tabular-to-FASTA converts tabular file to FASTA format
- FASTA-to-Tabular converter
- FASTQ to FASTA converter
- FASTQ to FASTA converter
- BED-to-GFF converter
- GFF-to-BED converter
- MAF to BED Converts a MAF formatted file to the BED format
- MAF to Interval Converts a MAF formatted file to the Interval format
- MAF to FASTA Converts a MAF formatted file to FASTA format
- SFF converter
- Wig/BedGraph-to-bigWig converter**
- BED-to-bigBed converter

FASTA manipulation

Filter and Sort

Wig/BedGraph-to-bigWig (version 1.1.0) Help from Biostar

Convert: 17: MACS on data 12 and data 10 (treatment: wig)

Converter settings to use: Default

Default settings should usually be used.

Execute

Syntax

This tool converts bedgraph or wiggle data into bigWig type.

Wiggle format: The .wig format is line-oriented. Wiggle data is preceded by a UCSC track definition line. Following the track definition line is the track data, which can be entered in three different formats described below.

BED format with no declaration line and four columns of data:

```
chromA chromStartA chromEndA dataValueA
chromB chromStartB chromEndB dataValueB
```

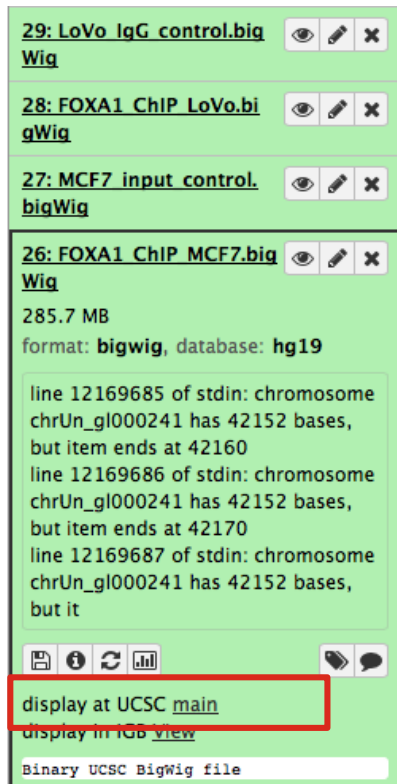
variableStep two column data; started by a declaration line and followed with chromosome positions and data values:

```
variableStep chrom=chrN [span=windowSize]
chromStartA dataValueA
chromStartB dataValueB
```

fixedStep single column data; started by a declaration line and followed with data

History

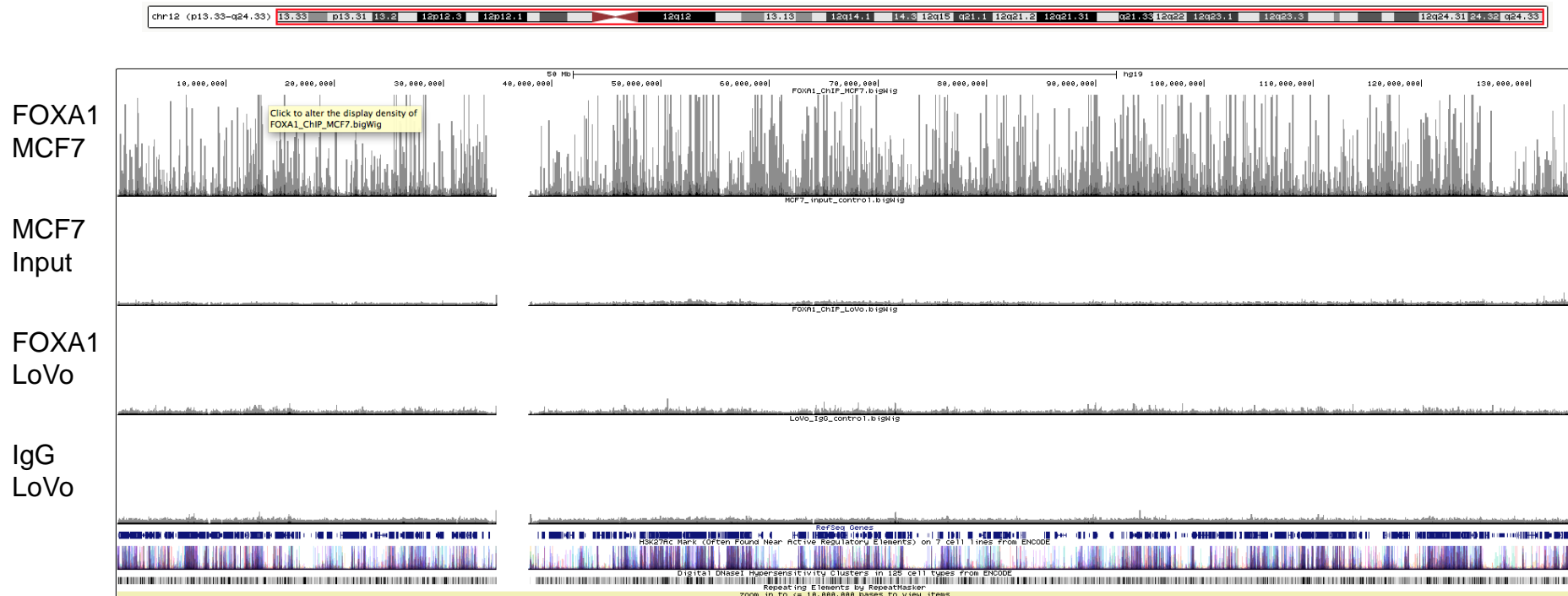
- 25: MACS on data 11 and data 13 (html report)
- 24: MACS on data 11 and data 13 (control: wig)
- 23: MACS on data 11 and data 13 (treatment: wig)
- 22: MACS on data 11 and data 13 (negative peaks: interval)
- 21: MACS on data 11 and data 13 (peaks: interval)
- 20: MACS on data 11 and data 13 (peaks: bed)
- 19: MACS on data 12 and data 10 (html report)
- 18: MACS on data 12 and data 10 (control: wig)
- 17: MACS on data 12 and data 10 (treatment: wig)
- 16: MACS on data 12 and data 10 (negative peaks: interval)



- Set “Display mode” to “full”
- Set “Vertical viewing range” to “min: 0, max: 200”
- Set “Data viewing scaling” to “use vertical viewing range setting”

- Peak shape
- Known target genes
- Chromosome-wide view

Signal visualization (UCSC)



De novo motif finding (MEME-ChIP)



Create coordinates of peak summit ± 50 bp

= Column A = Column B + Column E + 50
= Column B + Column E - 50 Name

FOXA1_summit_100bp.txt

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	#chr	start	end	length	summit	tags	#NAME?	fold_enrichm	FDR(%)					
2	chr5	104951553	104954096	2543	968	711	3100	332.35	0	chr5	104952471	104952571	FOXA1_MCF7_00001	
3	chr4	101699056	101703001	3945	2442	842	3100	315.05	0	chr4	101701448	101701548	FOXA1_MCF7_00002	
4	chr5	19878819	19882069	3250	2235	682	3100	305.04	0	chr5	19881004	19881104	FOXA1_MCF7_00003	
5	chr5	23493478	23496612	3134	1376	905	3100	303.22	0	chr5	23494804	23494904	FOXA1_MCF7_00004	
6	chrX	16679646	16681811	2165	1052	669	3100	303.22	0	chrX	16680648	16680748	FOXA1_MCF7_00005	
7	chrX	69439770	69440829	1059	618	485	3100	298.66	0	chrX	69440338	69440438	FOXA1_MCF7_00006	
8	chr18	21794144	21796891	2747	1870	994	3100	297.75	0	chr18	21795964	21796064	FOXA1_MCF7_00007	
9	chr5	57365345	57366720	1375	566	570	3100	296.84	0	chr5	57365861	57365961	FOXA1_MCF7_00008	
10	chr22	29217442	29220242	2800	1644	617	3100	294.86	0	chr22	29219036	29219136	FOXA1_MCF7_00009	
11	chr14	38014909	38017370	2461	1371	834	3100	291.39	0	chr14	38016230	38016330	FOXA1_MCF7_00010	
12	chr18	52432497	52435727	3230	2002	637	3100	285	0	chr18	52434449	52434549	FOXA1_MCF7_00011	
13	chr2	162727242	162729791	2549	1629	575	3100	283.18	0	chr2	162728821	162728921	FOXA1_MCF7_00012	
14	chr10	68874080	68876479	2399	833	475	3100	282.27	0	chr10	68874863	68874963	FOXA1_MCF7_00013	
15	chr2	189227945	189230610	2665	971	531	3100	280.45	0	chr2	189228866	189228966	FOXA1_MCF7_00014	
16	chr4	68794026	68796123	2097	558	624	3100	280.45	0	chr4	68794534	68794634	FOXA1_MCF7_00015	
17	chr1	22674381	22675898	1517	686	388	3100	279.54	0	chr1	22675017	22675117	FOXA1_MCF7_00016	
18	chrX	82068375	82069481	1106	587	471	3100	277.72	0	chrX	82068912	82069012	FOXA1_MCF7_00017	
19	chr4	162120738	162123716	2978	1471	950	3100	275.9	0	chr4	162122159	162122259	FOXA1_MCF7_00018	
20	chr5	25254597	25256218	1621	843	371	3100	275.9	0	chr5	25255390	25255490	FOXA1_MCF7_00019	
21	chr8	67625148	67627511	2363	1334	826	3100	273.17	0	chr8	67626432	67626532	FOXA1_MCF7_00020	
22	chr1	66917118	66918987	1869	877	424	3100	270.44	0	chr1	66917945	66918045	FOXA1_MCF7_00021	
23	chr12	97928917	97932273	3356	1373	538	3100	269.52	0	chr12	97930240	97930340	FOXA1_MCF7_00022	
24	chr2	10626608	10628203	1595	836	396	3100	269.12	0	chr2	10627394	10627494	FOXA1_MCF7_00023	
25	chr8	75684369	75686171	1802	933	417	3100	268.61	0	chr8	75685252	75685352	FOXA1_MCF7_00024	
26	chr5	29905375	29908723	3348	914	596	3100	266.79	0	chr5	29906239	29906339	FOXA1_MCF7_00025	
27	chr5	82560451	82562085	1634	775	444	3100	266.79	0	chr5	82561176	82561276	FOXA1_MCF7_00026	
28	chr10	122012007	122013584	1577	877	483	3100	265.88	0	chr10	122012834	122012934	FOXA1_MCF7_00027	
29	chr21	22610613	22612309	1696	796	481	3100	265.88	0	chr21	22611359	22611459	FOXA1_MCF7_00028	
30	chr4	89331968	89333621	1653	848	372	3100	264.97	0	chr4	89332766	89332866	FOXA1_MCF7_00029	
31	chr13	108614175	108616389	2214	643	553	3100	263.15	0	chr13	108614768	108614868	FOXA1_MCF7_00030	

De novo motif finding (MEME-ChIP)

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 49%

Tools

search tools

Get Data
Send Data
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Extract Genomic DNA using coordinates from assembled/unassembled genomes
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses

```
>hg19_chr5_104952471_104952571_+ FOXA1_MCF7_00001
AAGCTGTGTCATATTTGACTTGTGAATCAATGTACCCCTGTGATCAAGT
GATTAAGCTATTTACTGCTCTTGACAGAAAGCTATTAATAGATCGTAA
>hg19_chr4_101701448_101701548_+ FOXA1_MCF7_00002
CCTCAAACTTGCTGATCACATCTGTGCTGTGATATTTGAAATGTGCTCT
TTCAAGGGCATCTGTCAACATGTCACAAAGAATAAAGCAAAATGTCAGGTC
>hg19_chr5_19881004_19881104_+ FOXA1_MCF7_00003
TAATCTTATCTCTAAATCTTATTGACTCTGTTGGCTGTAATTGTTGACA
CTGCAAAACAATTTCTATTACTCTTACTTAGCATGAAGCCATATAGCCCTG
>hg19_chr5_23494804_23494904_+ FOXA1_MCF7_00004
TATAACAATATTTACATAGAGATAAAGAAAGTAATAACTCTATATATG
CAGCATTCAAATTTCTGCCCTTAGGAAGCTAACATGATTGGTGAAGGT
>hg19_chrX_16680648_16680748_+ FOXA1_MCF7_00005
GATCTTAGGAGTAATATTCCTATAGAAGGCTCCTGGTTTCCAAACTGT
TTACTTTCTAGCTAGAACAGCAGCAATGACCAGTCCCCACAAAGTAACT
>hg19_chrX_69440338_69440438_+ FOXA1_MCF7_00006
CATCTGTGCAACAAGTCTCCTCGAAGACAAGCAGAGTCTGGGAGATGC
CTCACTCTTTATCTCCGAGATCTGTACACCATATGACATATACAGGTCA
>hg19_chr18_21795964_21796064_+ FOXA1_MCF7_00007
GAAATAGAAATTGGATATTTACCTAAGATATTGAAAAAGTCCAGGGATCA
AAAGGAGAAAAATAACGAGACCTCTGAATGGGGAAGTAGTATCAAGGGAC
>hg19_chr5_57365861_57365961_+ FOXA1_MCF7_00008
GAATGGTAACATAAGAGAAAGAATATTACTTTGTGTAATATAGGTGAG
GCCAATGCAAAACATGACAGGCCAAATGTAATCCTGTGTGATGGAAATTA
>hg19_chr22_29219036_29219136_+ FOXA1_MCF7_00009
TGGCTTTTCTCTTCAACATTGCCACTGAATTGTGCTTTGTTGACAGAA
ATTAAACACTGCTCAGTAGATAGCATCAGGGCAGCCTTCAAGACTGGTT
>hg19_chr14_38016230_38016330_+ FOXA1_MCF7_00010
CATTCACACTGTGTAAACACCAAGCCCTTAACCTAGCAACAGTTGTTA
GAAGTGGGACACTCCAACCACATTCAGAGCTGAGATAAAATCAAATCACA
>hg19_chr18_52434449_52434549_+ FOXA1_MCF7_00011
TGCTCTAATACTCCTAAACAGGTTTGTGCTTGGGTAAATAACATGGAAA
GGCAAAATAACTCCACATACAACCTGTTTACCCAGTGGCTCATGCCAGTGTG
>hg19_chr2_162728821_162728921_+ FOXA1_MCF7_00012
TCAGTCTGCTCCAGCCTGTGTGAAAAATAAAAAATGATGTTAGCAGAGAAA
ACAGCACTGTGTACTTTAGCAAGGTGAGCTTTCTCCCTCTCATCTAAGT
>hg19_chr10_68874863_68874963_+ FOXA1_MCF7_00013
AGATTCCATCTCTGGGCTAATATTTGACTAACTCAAGATTATGTCACACA
CTGCTATGCCTCATGCAAAAGCCTCATGGAAGCAATTAATCTAACACA
>hg19_chr2_189228866_189228966_+ FOXA1_MCF7_00014
```

History

search datasets

Abcam_ChIPseq_webinar
61.8 GB

31: Extract Genomic DNA on data 30

1,000 sequences
format: fasta, database: hg19

>hg19_chr5_104952471_104952571_+ FOXA1
AAGCTGTGTCATATTTGACTTGTGAATCAATGTACCC
GATTAAGCTATTTACTGCTCTTGACAGAAAGCTATTA
>hg19_chr4_101701448_101701548_+ FOXA1
CCTCAAACTTGCTGATCACATCTGTGCTGTGATATTT
TTCAAGGGCATCTGTCAACATGTCACAAAGAATAAAGC

30: FOXA1_summit_100b p.txt

29: LoVo IgG_control.big Wig

28: FOXA1_ChIP_LoVo.bi gWig

27: MCF_input_control.bi gWig

26: FOXA1_ChIP_MCF7.bi gWig

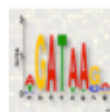
FOXA1_summit_100bp.fasta

De novo motif finding (MEME-ChIP)

<http://meme.nbcr.net/meme/cgi-bin/meme-chip.cgi>

MEME Suite Menu

- Submit A Job
- Documentation
- Downloads
- User Support
- Alternate Servers
- Authors
- Citing



MEME-ChIP

Motif Analysis of Large DNA Datasets

Version 4.9.1

Use this form to submit DNA sequences to MEME-ChIP. MEME-ChIP is designed especially for discovering motifs in **LARGE** (50MB maximum) sets of short (around 500bp) DNA sequences centered on locations of interest such as those produced by ChIP-seq experiments.

Data Submission Form

Perform motif discovery and enrichment on large DNA datasets.

Input the sequences

Enter DNA sequences in which you want to find motifs ?

or ☐ paste the sequences

Input the motif database

?

Input job queue details

Enter your email address. ?

Re-enter your email address.

Optionally enter a job description. ?

► Universal options

► MEME options

► DREME options

► CentriMo options

Version 4.9.1

Please send comments and questions to: meme@sdsc.edu

Powered by Opal





De novo motif finding (MEME-ChIP)

DESCRIPTION

FOXA1_top1000_peak_motifs

MOTIFS

The motifs found by the programs MEME, DREME and CentriMo; clustered by similarity and ordered by E-value.

<p>Motif Found</p>  <p>Reverse Complement ↔</p>	<p>Discovery/Enrichment Program ?</p> <p>MEME</p>	<p>E-value ?</p> <p>4.5e-378</p>	<p>Known or Similar Motifs ?</p> <p>FOXA1 (MA0148.3) Foxa2 (MA0047.2) Foxa2_primary (UP00073_1)</p>	<p>Distribution ?</p> 
<p>Motif Found</p>  <p>Reverse Complement ↔</p>	<p>Discovery/Enrichment Program ?</p> <p>DREME</p>	<p>E-value ?</p> <p>5.4e-010</p>	<p>Known or Similar Motifs ?</p> <p>Nr5a2 (MA0505.1) RORA_1 (MA0071.1) NR4A2 (MA0160.1)</p>	<p>Distribution ?</p> <p>Not Centrally Enriched</p>
<p>Motif Found</p>  <p>Reverse Complement ↔</p>	<p>Discovery/Enrichment Program ?</p> <p>DREME</p>	<p>E-value ?</p> <p>1.3e-006</p>	<p>Known or Similar Motifs ?</p>	<p>Distribution ?</p> <p>Not Centrally Enriched</p>


abcam[®]
discover more

FOXA1_LoVo

The motifs found by the programs MEME, DREME and CentriMo; clustered by similarity and ordered by E-value.

Reverse Complement \Rightarrow

Motif Found

Discovery/Enrichment Program 

DREME

E-value

4.1e-004

Known or Similar Motifs ⓘ

NFATC2 (MA0152.1)
Sp100_primary
(UP00049_1)
FOXD1 (MA0031.1)

Distribution

Not Centrally Enriched

Reverse Complement \Rightarrow

Motif Found

Discovery/Enrichment Program

DREME

E-value

□

Known or Similar Motifs ⓘ

Cdx2_4272.1
(UP00133_1)
CDX2 (MA0465.1)
Cdx1_2245.1
(UP00240_1)

Distribution ^[7]

Not Centrally
Enriched

Reverse Complement \Rightarrow

Motif Found

Discovery/Enrichment Program

MEME

E-value



Known or Similar Motifs ⓘ


ZNF263 (MA0528.1)
EWSR1-FLI1
(MA0149.1)
IRF1 (MA0050.2)

Distribution

Not Centrally
Enriched

Reverse Complement \Rightarrow

Motif Found

Discovery/Enrichment Program 

MEME

E-value

□

Known or Similar Motifs ⓘ

EGR2 (MA0472.1)
RREB1 (MA0073.1)

Distribution

Not Centrally Enriched

Gene ontology (GREAT)

_computational
BIOLOGY

ANALYSIS

GREAT improves functional interpretation of *cis*-regulatory regions

Cory Y McLean¹, Dave Bristor^{1,2}, Shoa I Clarke³, Bruce T Schaar², Craig B Lowe⁴, Aaron M Wenger¹ & Gill Bejerano^{1,2}

We developed the Genomic Regions Enrichment of Annotations Tool (GREAT) to analyze the functional significance of *cis*-regulatory regions identified by localized measurements of DNA binding events across an entire genome. Whereas previous methods look into account only binding proximal to genes, GREAT is able to properly incorporate distal binding sites and control for false positives using a binomial test over the input genomic regions. GREAT incorporates annotations from 20 ontologies and is available as a web application. Applying GREAT to data sets from chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-seq) of multiple transcription-associated factors, including SRP, NRSF, GABP, Stat3 and p300 in different developmental contexts, we recover many functions of these factors that are missed by existing gene-based tools, and we generate testable hypotheses. The utility of GREAT is not limited to ChIP-seq, as it could also be applied to open chromatin, localized epigenomic markers and similar functional data sets, as well as comparative genomics sets.

The coupling of chromatin immunoprecipitation with massively parallel sequencing, ChIP-seq, is ushering in a new era of genome-wide functional analysis¹⁻³. Thus far, computational efforts have focused on pinpointing the genomic locations of binding events from the deluge of reads produced by deep sequencing⁴⁻⁸. Functional interpretation is then performed using gene-based tools developed in the wake of the preceding microarray revolution⁹⁻¹¹. In a typical analysis, one compares the total fraction of genes annotated for a given ontology term with the fraction of annotated genes picked by proximal binding events to obtain a gene-based *P* value for enrichment (Fig. 1 and Online Methods).

This procedure has a fundamental drawback: associating only proximal binding events (for example, under 2–5 kb from the transcription start site) typically discards over half of the observed binding events (Fig. 2a). However, the standard approach to capturing distal events—associating each binding site with the one or two nearest

genes—introduces a strong bias toward genes that are flanked by large intergenic regions^{12,13}. For example, though the Gene Ontology¹⁴ (GO) term ‘multicellular organismal development’ is associated with 14% of human genes, the ‘nearest genes’ approach associates over 33% of the genome with these genes. This biological bias results in numerous false positive enrichments, particularly for the input set sizes typical of a ChIP-seq experiment (Fig. 2b and Supplementary Fig. 1). Building on our experience in addressing these pitfalls^{12,15,16}, we have developed a tool that robustly integrates distal binding events while eliminating the bias that leads to false positive enrichments.

RESULTS

Here we describe GREAT, which analyzes the functional significance of sets of *cis*-regulatory regions by explicitly modeling the vertebrate genome regulatory landscape and using many rich information sources.

A binomial test for long-range gene regulatory domains

GREAT associates genomic regions with genes by defining a ‘regulatory domain’ for each gene in the genome. Each genomic region is associated with all genes in whose regulatory domains it lies (Fig. 1b). High-throughput chromosomal conformation capture (3C) approaches such as 5C (ref. 17), Hi-C (ref. 18) or enhanced ChIP-4C (ref. 19) are providing first glimpses of actual gene regulatory domains. Because we still lack precise empirical maps, however, GREAT assigns each gene a regulatory domain consisting of a basal domain that extends 5 kb upstream and 1 kb downstream from its transcription start site (denoted below as 5+1 kb), and an extension up to the basal regulatory domain of the nearest upstream and downstream genes within 1 Mb (GREAT allows the user to modify the rule and distances). GREAT further refines the regulatory domains of a handful of genes, including several global control regions²⁰, by using their experimentally determined regulatory domains. Our tool can also incorporate additional locus-based and genome-wide data as they become available (Supplementary Fig. 2 and Online Methods).

Given a set of input genomic regions and an ontology of gene annotations, GREAT computes ontology term enrichments using a binomial test that explicitly accounts for variability in gene regulatory domain size by measuring the total fraction of the genome annotated for any given ontology term and counting how many input genomic regions fall into those areas (Fig. 1b and Online Methods). In the example above, GREAT expects 33% of all input elements to be associated with ‘multicellular organismal development’ by chance, rather than the 14% of input elements that a gene-based test assumes. The

<http://bejerano.stanford.edu/great/public/html/>

GREAT

Overview News Use GREAT Demo Video How to Cite Help Forum

Species Assembly

☒ Human: GRCh37 (UCSC hg19, Feb/2009)

☐ Human: NCBI build 36.1 (UCSC hg18, Mar/2006)

☐ Mouse: NCBI build 37 (UCSC mm9, Jul/2007)

☐ Zebrafish: Wellcome Trust Zv9 (danRer7, Jul/2010)

 Zebrafish CNE set

[Can I use a different species or assembly?](#)

Test regions

☒ BED file: FOXA1_s...0bp.txt

FOXA1_summit_100bp.txt

☐ BED data:

[What should my test regions file contain?](#)

[How can I create a test set from a UCSC Genome Browser annotation track?](#)

Background regions

☒ Whole genome

☐ BED file: no file selected

☐ BED data:

[When should I use a background set?](#)

[What should my background regions file contain?](#)

Association rule settings

Gene ontology (GREAT)

[Overview](#)[News](#)[Use GREAT](#)[Demo](#)[Video](#)[How to Cite](#)[Help](#)[Forum](#)

All genomic region-gene association tables (1000 regions, 1252 genes)

Job ID: 20141009-public-2.0.2-EKbLHw

Display name: FOXA1_summit_100bp.txt

What do these tables show?

Genomic region -> gene association table [Download table as text.](#)

Gene -> genomic region association table [Download table as text.](#)

Region	Gene (distance to TSS)	Gene	Region (distance to TSS)
FOXA1_MCF7_00001	NONE	AADAC	FOXA1_MCF7_00133 (-23,820)
FOXA1_MCF7_00002	EMCN (-262,248), PPP3CA (+567,130)	AADACL2	FOXA1_MCF7_00133 (+56,337)
FOXA1_MCF7_00003	CDH18 (+107,299)	ABCA12	FOXA1_MCF7_00065 (-78,574)
FOXA1_MCF7_00004	CDH12 (-641,123), PRDM9 (-12,870)	ABCC13	FOXA1_MCF7_00976 (+32,534)
FOXA1_MCF7_00005	S100G (+12,417), CTPS2 (+50,361)	ABHD2	FOXA1_MCF7_00778 (+27,658)
FOXA1_MCF7_00006	AWAT1 (-14,117), DGAT2L6 (+43,052)	ABLIM1	FOXA1_MCF7_00981 (-161,376)
FOXA1_MCF7_00007	CABYR (+77,059), OSBPL1A (+181,819)	ACMSD	FOXA1_MCF7_00198 (-14,984)
FOXA1_MCF7_00008	ACTBL2 (-587,275), PLK2 (+390,055)	ACP1	FOXA1_MCF7_00197 (-154,658)
FOXA1_MCF7_00009	ZNRF3 (-60,669), XBP1 (-22,526)	ACSL1	FOXA1_MCF7_00897 (-67,033)
FOXA1_MCF7_00010	SLC25A21 (-374,415), FOXA1 (+48,209)	ACSL3	FOXA1_MCF7_00185 (-64,157), FOXA1_MCF7_00314 (+165,515)
FOXA1_MCF7_00011	RAB27B (-61,209), C18orf26 (+176,109)	ACSL4	FOXA1_MCF7_00291 (-6,836)
FOXA1_MCF7_00012	DPP4 (+202,181), SLC4A10 (+248,026)		

Heatmap representation (seqMINER)

<http://sourceforge.net/projects/seqminer/>

History

12: Input_MCF7_aln.bam
423.3 MB
format: **bam**, database: **hg19**

[samopen] SAM header is present: 93 sequences.
[bam_index_core] the alignment is not sorted (ERR022055.sra.1): 30-th chr > 24-th chr
[bam_index_build2] fail to index the BAM file.

display at UCSC [main](#)
display at Ensembl [Current](#)
display with IGV [web](#) [current](#) [local](#)
display in IGB [View](#)

Binary bam alignments file

11: IgG_LoVo_aln.bam

10: FOXA1_MCF7_aln.bam
773.8 MB
format: **bam**, database: **hg19**

[samopen] SAM header is present: 93 sequences.
[bam_index_core] the alignment is not sorted (FOXA1_MCF7.sra.2): 42-th chr > 28-th chr
[bam_index_build2] fail to index the BAM file.

seqMINER_1.3.3e

Name	Date Modified
ensembl67_hg19.seqminer	20 July 2012 13:
ensembl67_mm9.seqminer	20 July 2012 13:
lib	24 October 2011
README.TXT	24 October 2011
refGene_hg19_genebody.bed	20 July 2012 13:
refGene_hg19_TSS.bed	20 July 2012 13:
refGene_mm9_genebody.bed	20 July 2012 13:
refGene_mm9_TSS.bed	20 July 2012 13:
run_in_linux.sh	24 October 2011
run_in_mac.command	24 October 2011
run_in_windows.bat	7 March 2011 1
seqMINER.jar	24 October 2011

Heatmap representation (seqMINER)

<http://sourceforge.net/projects/seqminer/>

History

12: Input_MCF7_aln.bam
423.3 MB
format: **bam**, database: **hg19**

[samopen] SAM header is present: 93 sequences.
[bam_index_core] the alignment is not sorted (ERR022055.sra.1): 30-th chr > 24-th chr
[bam_index_build2] fail to index the BAM file.

display at UCSC [main](#)
display at Ensembl [Current](#)
display with IGV [web](#) [current](#) [local](#)
display in IGB [View](#)

Binary bam alignments file

11: IgG_LoVo_aln.bam

10: FOXA1_MCF7_aln.bam
773.8 MB
format: **bam**, database: **hg19**

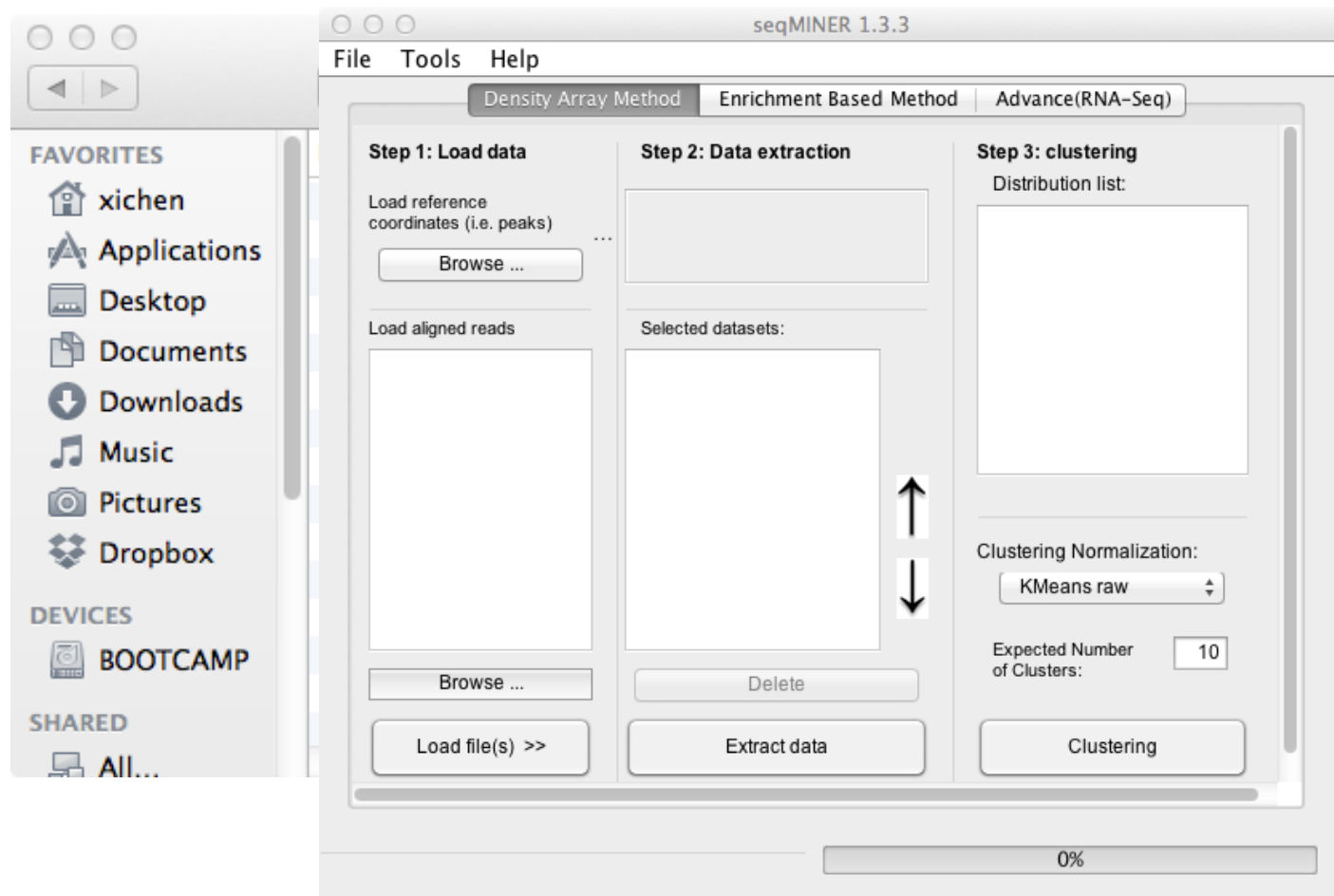
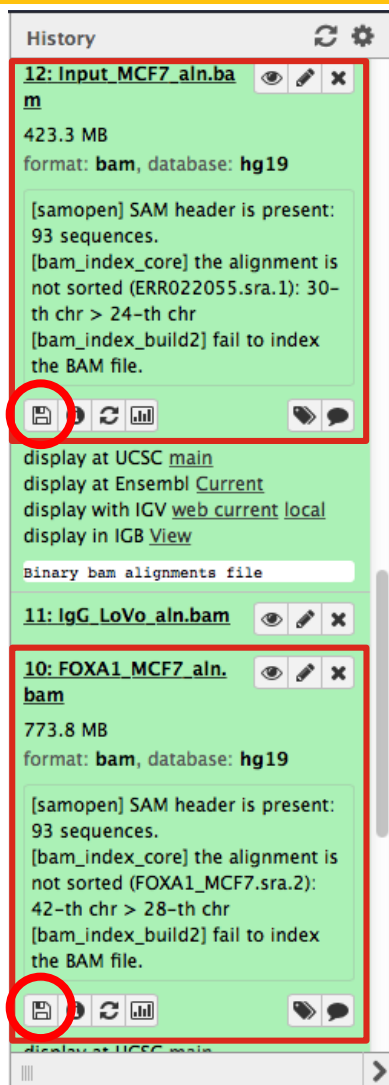
[samopen] SAM header is present: 93 sequences.
[bam_index_core] the alignment is not sorted (FOXA1_MCF7.sra.2): 42-th chr > 28-th chr
[bam_index_build2] fail to index the BAM file.

seqMINER_1.3.3e

Name	Date Modified
ensembl67_hg19.seqminer	20 July 2012 13:
ensembl67_mm9.seqminer	20 July 2012 13:
lib	24 October 2011
README.TXT	24 October 2011
refGene_hg19_genebody.bed	20 July 2012 13:
refGene_hg19_TSS.bed	20 July 2012 13:
refGene_mm9_genebody.bed	20 July 2012 13:
refGene_mm9_TSS.bed	20 July 2012 13:
run_in_linux.sh	24 October 2011
run_in_mac.command	24 October 2011
run_in_windows.bat	7 March 2011 1
seqMINER.jar	24 October 2011

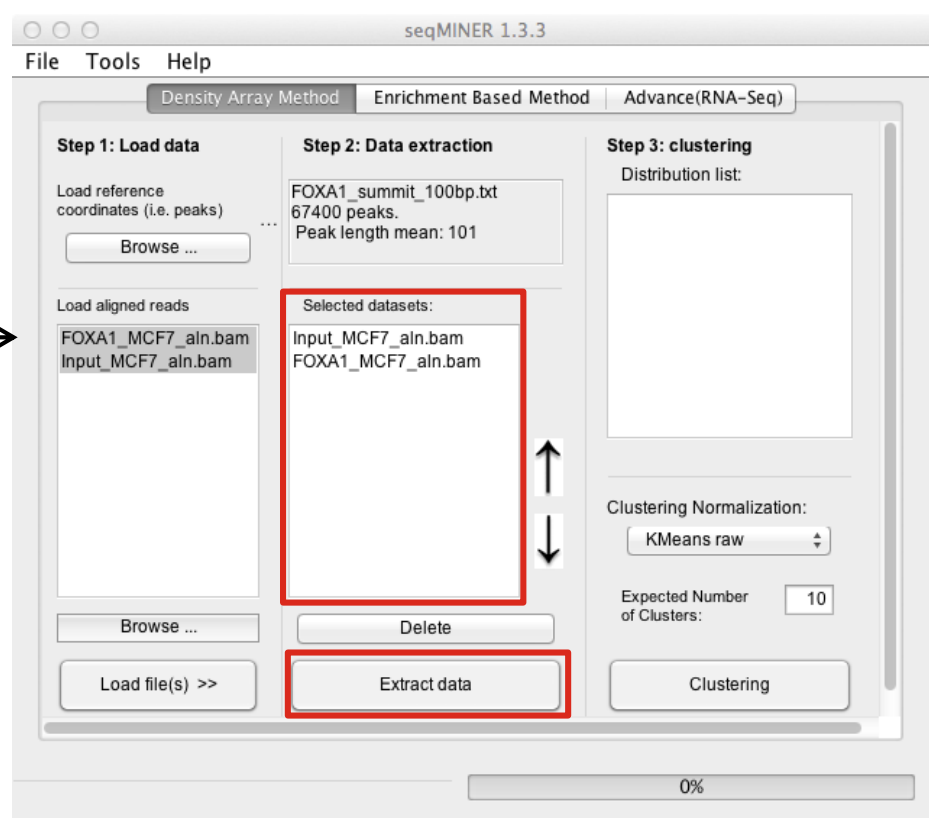
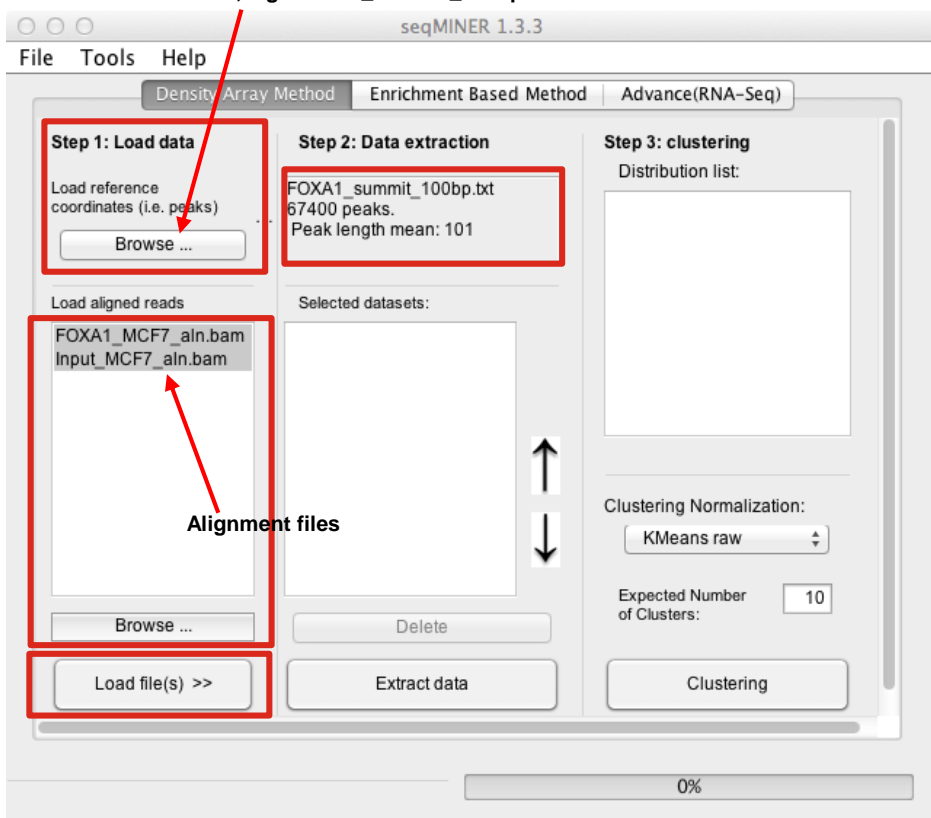
Heatmap representation (seqMINER)

<http://sourceforge.net/projects/seqminer/>

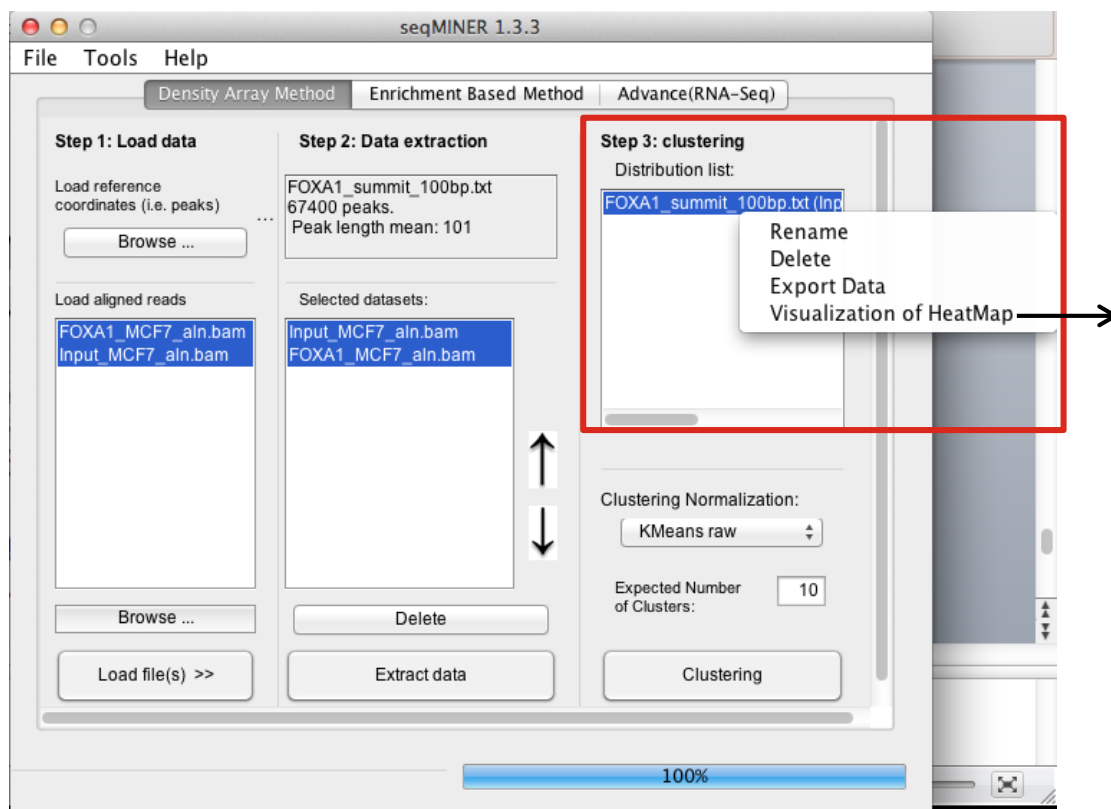


Heatmap representation (seqMINER)

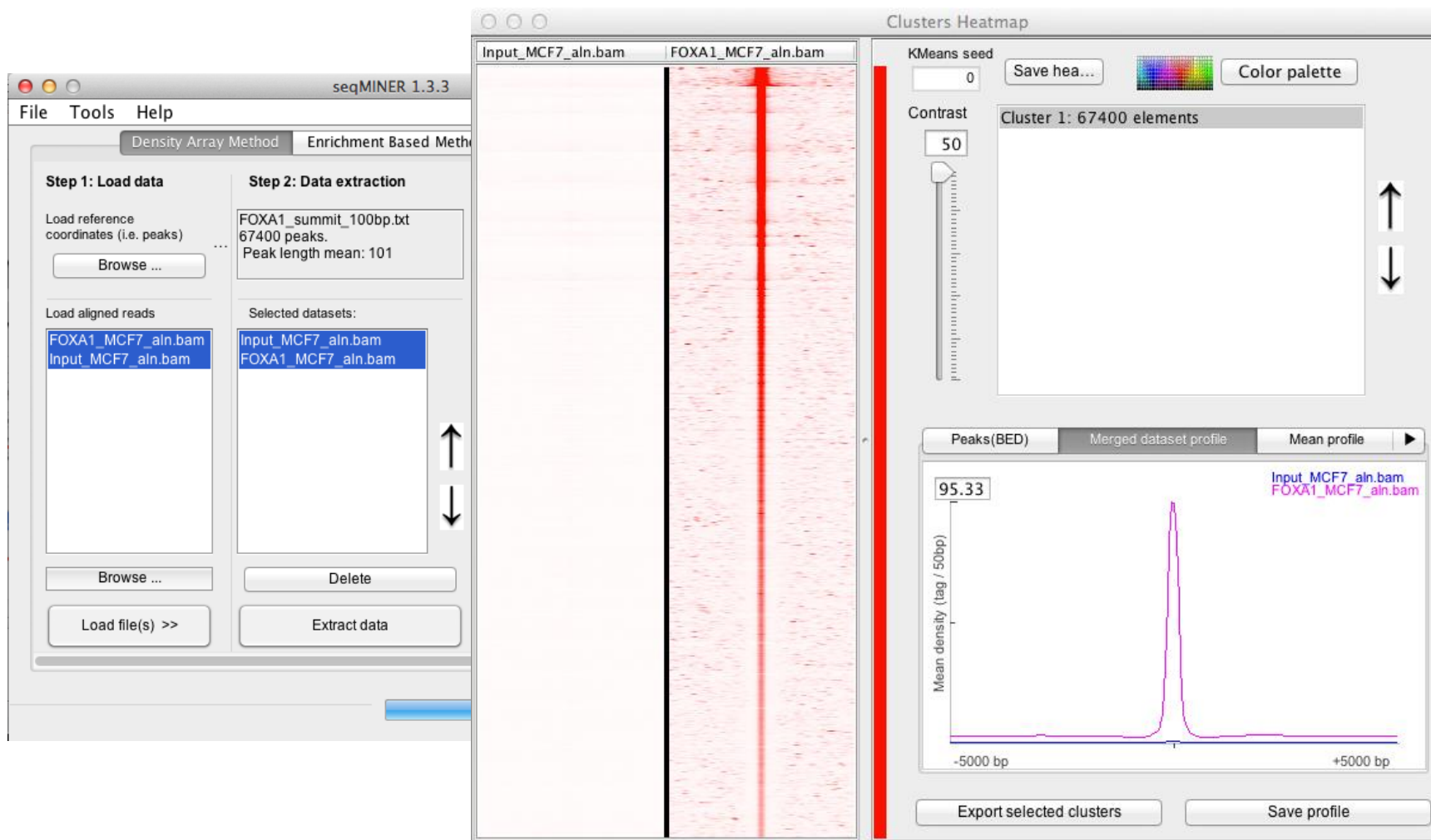
Peak file, e.g. FOXA1_summit_100bp.txt



Heatmap representation (seqMINER)



Heatmap representation (seqMINER)

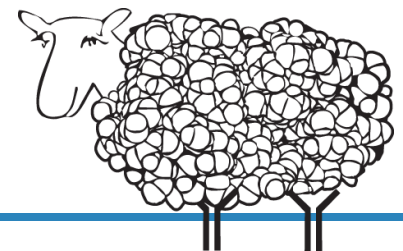


Conclusions & Questions

ChIP-seq pipeline	<ul style="list-style-type: none">• Bardet AF, He Q, Zeitlinger J, Stark A. (2011) A computational pipeline for comparative ChIP-seq analyses. <i>Nat Protoc.</i> 7:45-61.
Peak calling	<ul style="list-style-type: none">• Feng J, Liu T, Qin B, Zhang Y, Liu XS. (2012) Identifying ChIP-seq enrichment using MACS. <i>Nat Protoc.</i> 7:1728-40.
Motif discovery	<ul style="list-style-type: none">• Ma W, Noble WS, Bailey TL. (2014) Motif-based analysis of large nucleotide data sets using MEME-ChIP. <i>Nat Protoc.</i> 9:1428-50.• Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J. (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. <i>Nat Protoc.</i> 7:1551-68.
Gene ontology	<ul style="list-style-type: none">• McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. (2010) GREAT improves functional interpretation of cis-regulatory regions. <i>Nat Biotechnol.</i> 28:495-501.• Huang da W, Sherman BT, Lempicki RA. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. <i>Nat Protoc.</i> 4:44-57.
HOMER suite	<ul style="list-style-type: none">• http://homer.salk.edu/homer/

Products for your research

abcam[®]
discover more



Abcam Events – 10 Year Anniversary Giveaway

Four all-inclusive conference registrations up for grabs!



2014/15 Events and Webinars



It's our birthday!

Enter our giveaway to win all-inclusive registration to either:

Chromatin: Structure & Function 2015
November 16-19, 2015 - Grand Caymans

Organizer: Tony Kourzarides

Confirmed speakers include:

Shelley Berger, Laurie Boyer, Anne Brunet, Luciano Di Croce, Robert Kingston, Rob Martienssen, Danny Reinberg, Ramin Shiekhata, Ali Shilatifard, Ken Zaret

or

Maintenance of Genome Stability 2016
March 7-10, 2016 - Panama

Organizer: Steve Jackson

Enter giveaway @ www.abcam.com/10YearGiveaway

Competition closing date: December 12, 2014

High-sensitivity ChIP kit (ab185913)

Low input

Start from as low as 2,000 cells

Quick

Only 5 hours to perform ChIP (compared to 2 days using conventional method)

Compatible

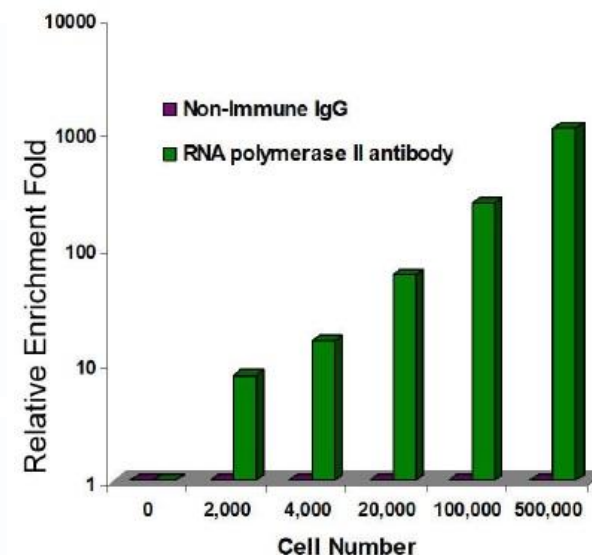
Eluted DNA can be processed straight away for DNA sequencing, microarrays or qPCR

Simple to use

Reaction takes place in the wells of the 96-wp: easy for standardization and high-throughput

Inclusive

Kit contains all necessary reagents for ChIP reaction (excluding cross-linking step)



ChIP-qPCR analysis of RNAPII enrichment in GAPDH promoters in MBD-231 cells chromatin using High-sensitivity ChIP Kit (ab185913).

High-sensitivity solutions for ChIP/ChIPseq

ChIP

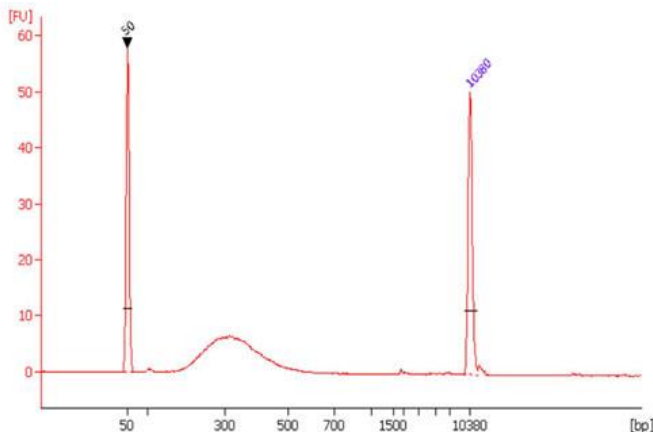
High Sensitivity ChIP Kit
(ab185913)

**Library
prep**

High Sensitivity DNA Library
Preparation Kit (for Illumina)
(ab185905)

**ChIP +
Library
prep**

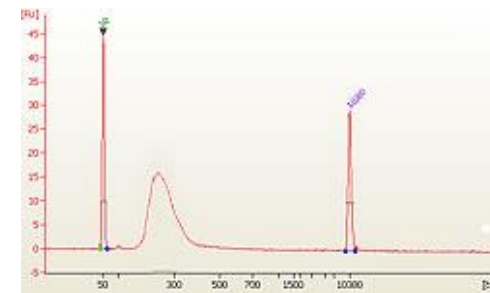
ChIP-Seq High Sensitivity Kit
(ab185908)



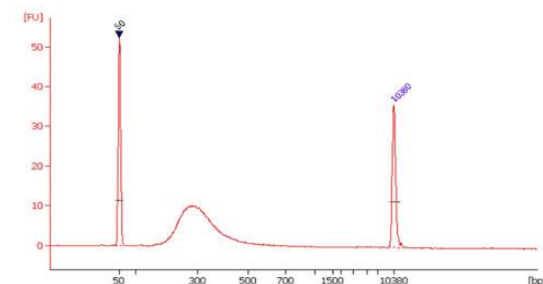
Library obtained from 0.2 ng of human placenta DNA using High Sensitivity DNA Library Preparation Kit (ab185905).

Bisulfite Sequencing

- **Post-Bisulfite DNA Library Preparation Kit (For Illumina[®]) (ab185906)**
 - High-sensitivity (1 µg)
 - Post-bisulfite
 - Compatible with WGBS, oxBS-seq, RRBS
 - Quick (5 hrs from bisulfite DNA to library)
- **Bisulfite-Seq High Sensitivity Kit (For Illumina[®]) (ab185907)**
 - High-sensitivity (< 0.5 µg)
 - Bisulfite conversion
 - Compatible with WGBS, oxBS-seq, RRBS
 - Quick (6 hrs from DNA to library)



Size distribution of library fragments. Post-bisulfite DNA library was prepared from 10 ng of input DNA using ab185906.



Size distribution of library fragments. Post-bisulfite DNA library was prepared from 10 ng of input DNA using ab185907.