



ChIP-seq data analysis using command line tools

Endre Barta^{1,2} and Gergely Nagy¹, Hungary

¹University of Debrecen, Department of Biochemistry
and Molecular Biology, barta.endre@unideb.hu

²Agricultural Biotechnology Center, Gödöllő, Agricultural Genomics
and Bioinformatics Group, barta@abc.hu

Helsinki, Finland, September 19, 2014

Tasks

- Running ChIP-seq_analyze script on the cluster from different sources (SRA, fastq files, bam files)
- Visualizing ChIP-seq results in IGV genome browser
- Defining and visualizing overlapping peak sets using intersectBed and VennMaster
- Creating heat maps with HOMER, Cluster and TreeView
- Mapping *de novo* motifs to the peaks
- Summarizing the numbers of the analysis

Nagy G, Dániel B, Jónás D, Nagy L, Barta E. A novel method to predict regulatory regions based on histone mark landscapes in macrophages. *Immunobiology*. 2013 Nov;218(11):1416-27. doi: 10.1016/j.imbio.2013.07.006. Epub 2013 Jul 26. PubMed PMID: 23973299.

Cuaranta-Monroy I, Simandi Z, Kolostyak Z, Doan-Xuan QM, Poliska S, Horvath A, Nagy G, Bacso Z, Nagy L. Highly efficient differentiation of embryonic stem cells into adipocytes by ascorbic acid. *Stem Cell Res*. 2014 Jul;13(1):88-97. doi: 10.1016/j.scr.2014.04.015. Epub 2014 May 5. PubMed PMID: 24858493.

Daniel B, Nagy G, Nagy L. The intriguing complexities of mammalian gene regulation: how to link enhancers to regulated genes. Are we there yet? *FEBS Lett*. 2014 Aug 1;588(15):2379-91. doi: 10.1016/j.febslet.2014.05.041. Epub 2014 Jun 16. PubMed PMID: 24945732.

Daniel B, Nagy G, Hah N, Horvath A, Czimmerer Z, Poliska S, Gyuris T, Keirsse J, Gysemans C, Van Ginderachter JA, Balint BL, Evans RM, Barta E, Nagy L. The active enhancer network operated by liganded RXR supports angiogenic activity in macrophages. *Genes Dev*. 2014 Jul 15;28(14):1562-77. doi: 10.1101/gad.242685.114. PubMed PMID: 25030696; PubMed Central PMCID: PMC4102764.

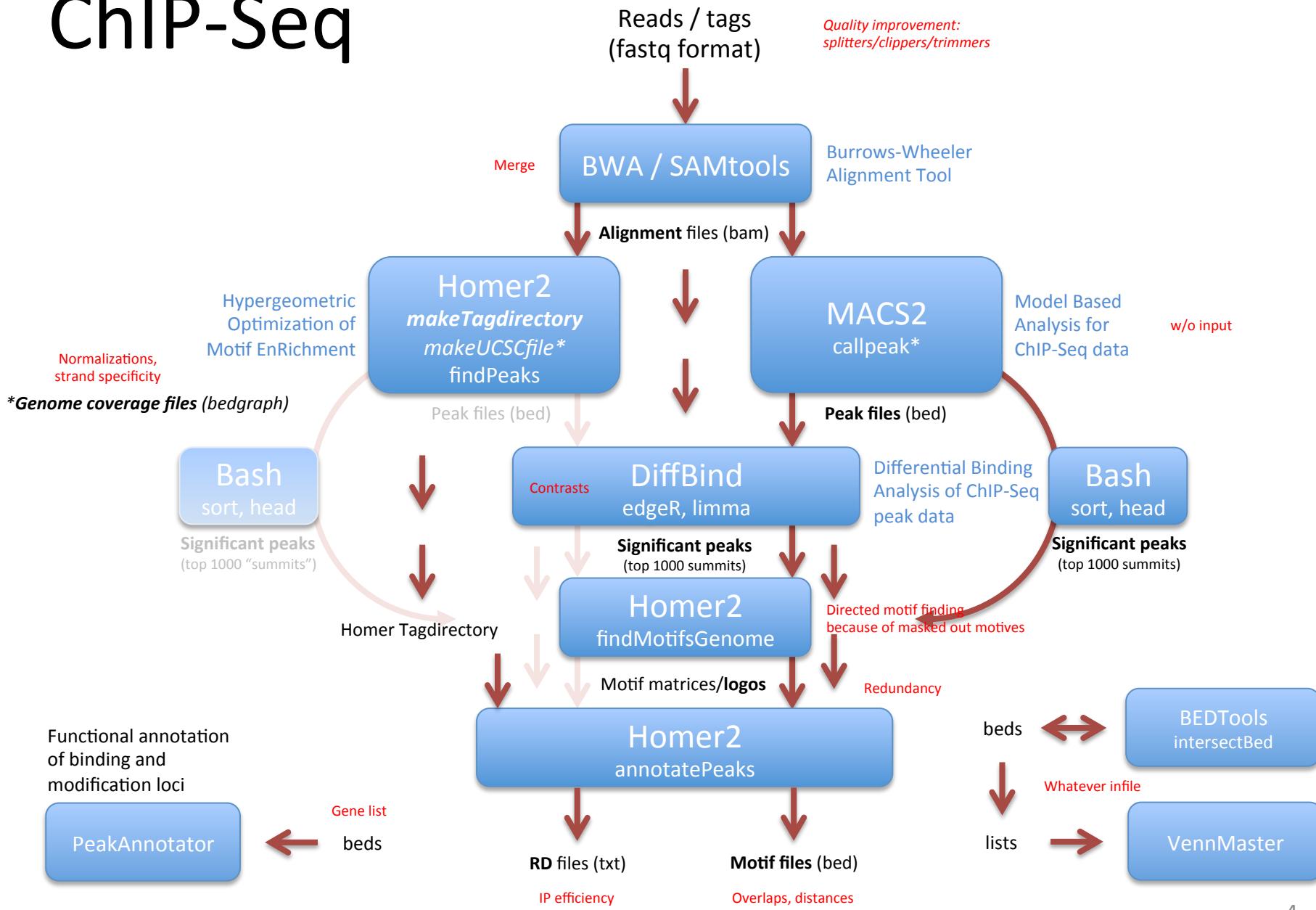
ChIP-seq analyze script

- Aim: To have a simple script that can be used either to analyze local ChIP-seq sequencing data or to do meta-analysis of ChIP-seq experiments stored on the NCBI SRA database
- Command line tool
- Input: SRA (NCBI reads), fastq (reads), bam (alignments)
 - Downloads SRA format files NCBI
 - Maps fastq format reads
 - Peak calling by HOMER and MACS
 - Peak annotation by HOMER
 - Known and *denovo* motif finding by HOMER
 - GO enrichment analysis by HOMER
 - Generates bedgraph and bed files for visualization

Barta E

Command line analysis of ChIP-seq results.
EMBNET JOURNAL 17:(1) pp. 13-17. (2011)

ChIP-Seq



Reads (fastq format)

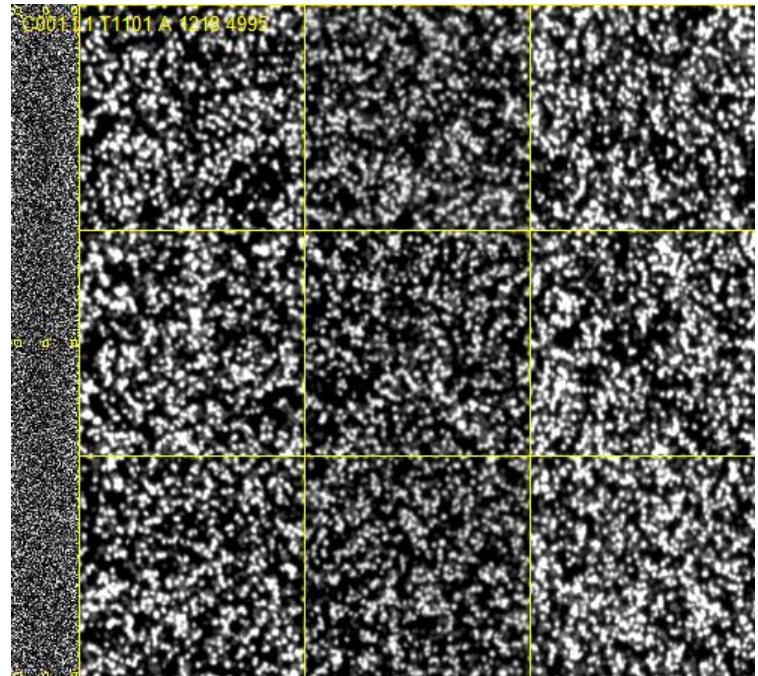
- Raw data from pictures

illumina®



- Separating individual samples
(`fastx_barcode_splitter.pl`)

- FastQC
- Adaptors, linkers, barcodes, low QC bases – removing with splitters, trimmers, clippers



fastq files

- Large text files
 - Identifier
 - Instrument name
 - Lane number
 - Tile number
 - X-coordinate
 - Y-coordinate
 - Member of the pair
 - Passing filter
 - Control bits
 - *Index sequence*
 - Raw sequence
 - Optional description
 - Quality values
- Can be splitted and concatenated

```
@HWI-H148:96:C0FTRACXX:1:1101:1634:2070 1:N:0:  
TAGTCTCTTAAGAGCCATGGCTACTAGAAAATTGATAGATCTGGATACACN  
+  
+1:B1BBDFDBCF>GD@9A9;@?:B+9A+9AG4?:C**1::11:9?B*?#  
@HWI-H148:96:C0FTRACXX:1:1101:2201:2016 1:N:0:  
GAGTTTCAGGATCTGTTATGTCTCCCTTCATTTCTGATTNNNNNNNN  
+  
@@;DDDDAHH<FHHIH:CEGEDFFEHIIGEHGEGHEGHIFHI#####  
@HWI-H148:96:C0FTRACXX:1:1101:2165:2024 1:N:0:  
GAGCTTTTTTCCTCGCCATATTCACGTCTAAAGTGTGTATTNNNNNNNN  
+  
=?1+ADDDHF>FHIIHBDGHGGDCB@GH?FDGG9BDG<BF?FFH#####  
@HWI-H148:96:C0FTRACXX:1:1101:2314:2026 1:N:0:  
GATTTAGGTAACAGCCAGGGCAGGATAATCAAGGACATTTTNNNNNNN  
+  
11144AABDADHDIEGIIIHDGHDEFHGHIIGGEGCE<GHII#####
```

SRA toolkit

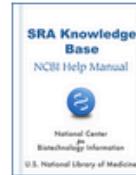
Contents ▾

Bookshelf ID: NBK56560

[Print View](#)

< Prev

Next >



SRA Knowledge Base [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-.

[Table of Contents Page](#) | [Cite this Page](#)

Other titles in this collection

[NCBI Help Manual](#)

Recent activity

[Turn Off](#) [Clear](#)

[Using the SRA Toolkit - SRA Knowledge Base](#)

Bookshelf

[See more...](#)

Using the SRA Toolkit

What is the purpose of the SRA toolkit?

Is there documentation that will show me how to convert the files I downloaded from SRA into a format I'm familiar with?

I'm having problems using the toolkit, and the documentation doesn't cover the problem I'm having. Who do I contact for help?

What is the purpose of the SRA toolkit?

The [SRA Toolkit](#), also known as the SRA System Development Kit (SDK), will allow you to programmatically access data housed within SRA and convert it from the SRA format to any of the following formats:

- AB SOLiD native
- FASTQ
- SFF (Roche 454)
- Illumina native

You can also use the toolkit to convert from the formats listed below into the SRA format:

- FASTQ
- AB SOLiD-SRF
- AB SOLiD-Native
- Illumina SRF
- Illumina Native
- SFF

The SRA toolkit is available in versions compatible with Linux, Windows and Mac operating systems.

Finding ChIP-seq experiments in the NCBI's SRA database

NCBI Resources How To Sign in to NCBI

SRA SRA Incap AR foxn1 knockdown Search Save search Limits Advanced Help

Display Settings: Summary, 20 per page Send to: Filter your results:

The following term was not found in SRA: foxn1.

See the search details.

Results: 6

- GSM862357: LNCaP_abl_AR_KD_rep3; Homo sapiens; RNA-Seq
 - 1 ILLUMINA (Illumina HiSeq 2000) run: 4M spots, 143.8M bases, 80.4Mb downloads
Accession: SRX116038
- GSM862356: LNCaP_abl_AR_KD_rep2; Homo sapiens; RNA-Seq
 - 1 ILLUMINA (Illumina HiSeq 2000) run: 7.3M spots, 263.1M bases, 162.6Mb downloads
Accession: SRX116037
- GSM862355: LNCaP_abl_AR_KD_rep1; Homo sapiens; RNA-Seq
 - 1 ILLUMINA (Illumina HiSeq 2000) run: 3.1M spots, 110.8M bases, 62.5Mb downloads
Accession: SRX116036
- GSM916522: LNCaP_shFoxA1_regular_medium_AR; Homo sapiens; ChIP-Seq
 - 1 ILLUMINA (Illumina HiSeq 2000) run: 8.8M spots, 440.8M bases, 256.7Mb downloads
Accession: SRX142907
- GSM923509: AR_ChIP-seq_shFoxA1_hormone-deprived; Homo sapiens; ChIP-Seq
 - 1 ILLUMINA (Illumina HiSeq 2000) run: 9.6M spots, 482.1M bases, 277.7Mb downloads
Accession: SRX147605
- GSM775365: LNspip53-AR ChIPSeq
 - 1 ILLUMINA (Illumina Genome Analyzer IIx) run: 13.6M spots, 559.4M bases, 314.2Mb downloads
Accession: SRX092448

All (6)
access: Controlled (0)
access: Public (6)
aligned data (0)
source: DNA (3)
source: metagenomic (0)
source: RNA (3)
type: exome (0)
type: genome (0)

Manage Filters

Search in related databases

Database	Access		all
	public	controlled	
BioSample	6		6
BioProject			
dbGaP			
GEO Datasets			

Find related data

Database: Select

8

Getting the link to the FTP site for downloading sra format reads

SRA Help

Display Settings: Full

GSM916522: LNCaP_shFoxA1_regular_medium_AR; Homo sapiens; ChIP-Seq

Accession: SRX142907
Experiment design: n/a
Submission: SRA051824 by GEO
Study summary: GSE37345: FoxA1 inhibits androgen receptor expression and suppresses prostate cancer metastasis [ChIP-seq] (SRP012266) • Study • All experiments (more...)
Sample: GSM916522: AR_ChIP-seq_shFoxA1_R1881 (SRS309547) (less...)

Organism: [Homo sapiens](#)
Attributes:
GEO Accession: GSM916522
cell line: LNCaP
genotype/variation: FoxA1 knockdown
chip antibody: anti-AR
chip antibody vendor: Millipore
External link: [GEO Sample](#)

Library: GSM916522: LNCaP_shFoxA1_regular_medium_AR (more...)
Platform: Illumina (more...)
Spot descriptor:


Experiment attributes:
GEO Accession: GSM916522
instrument model: Illumina HiScanSQ
Total: 1 run, 8.8M spots, 440.8M bases, [256.7Mb](#)

Run # of Spots # of Bases Size

1.	SRR488262	8,816,568	440.8M	256.7Mb
----	---------------------------	-----------	--------	-------------------------

Send to:
BioProject
BioSample
GEO DataSets
Taxonomy

Recent activity

Turn Off Clear

Inc AR foxn1 knockdown (6) SRA
srx142906 (1) SRA
Inc AR foxa1 (34) SRA
Inc AR foxa1ko (53) SRA
Inc AR foxn (53) SRA
See more...

ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX142/SRX142907

Introduction

Burrows-Wheeler Aligner (BWA) is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome. It implements two algorithms, bwa-short and BWA-SW. The former works for query sequences shorter than 200bp and the latter for longer sequences up to around 100kbp. Both algorithms do gapped alignment. They are usually more accurate and faster on queries with low error rates. Please see the [BWA manual page](#) for more information.

FAQ

How can I cite BWA?

The short read alignment component (bwa-short) has been published:

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754–60. [PMID: [19451168](#)]

If you use BWA-SW, please cite:

Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, Epub. [PMID: [20080505](#)]

(See also Errata below for a minor correction to the formulae in these papers.)

BWA:

[SF project page](#)
[SF download page](#)
[Mailing list](#)
[BWA manual page](#)
[Repository](#)

Links:

[SAMtools](#)
[MAQ](#)

Introduction

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

General Information

[SAM Spec v1.4](#)

[SF Project Page](#)

[SF Download Page](#)

[Mailing Lists](#)

[SVN Browse](#)

[Related Software](#)

[FAQ](#)

SAMtools in C

[General Introduction](#)

[Manual Page \(0.1.17\)](#)

[Variant Calling \(mpileup\)](#)

[Text Alignment Viewer](#)

[API Documentation](#)

[Example C Program](#)

BAMTOOLS

BamTools provides both a programmer's API and an end-user's toolkit for handling BAM files.

I. Learn More

II. License

III. Acknowledgements

IV. Contact

I. Learn More:

Installation steps, tutorial, API documentation, etc. are all now available through the BamTools project wiki:

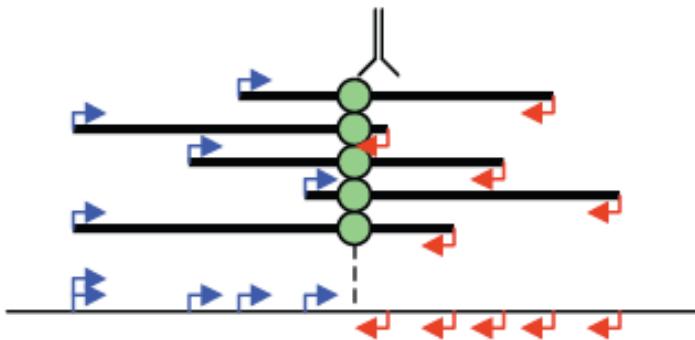
<https://github.com/pezmaster31/bamtools/wiki>

Join the mailing list(s) to stay informed of updates or get involved with contributing:

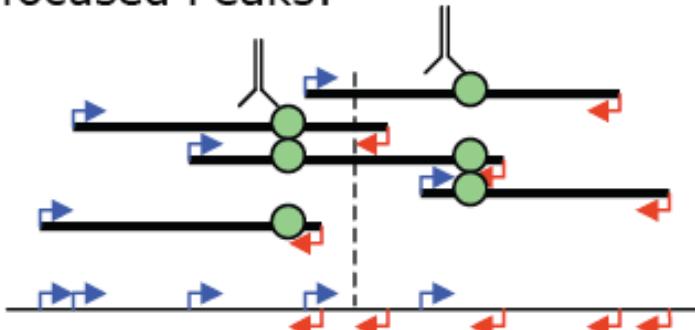
<https://github.com/pezmaster31/bamtools/wiki/Mailing-lists>

- The number of peaks depends on the methods used and the cutoff values applied.
- More reads doesn't mean necessarily more peaks
- Different methods give only 60-80% similar peaks!

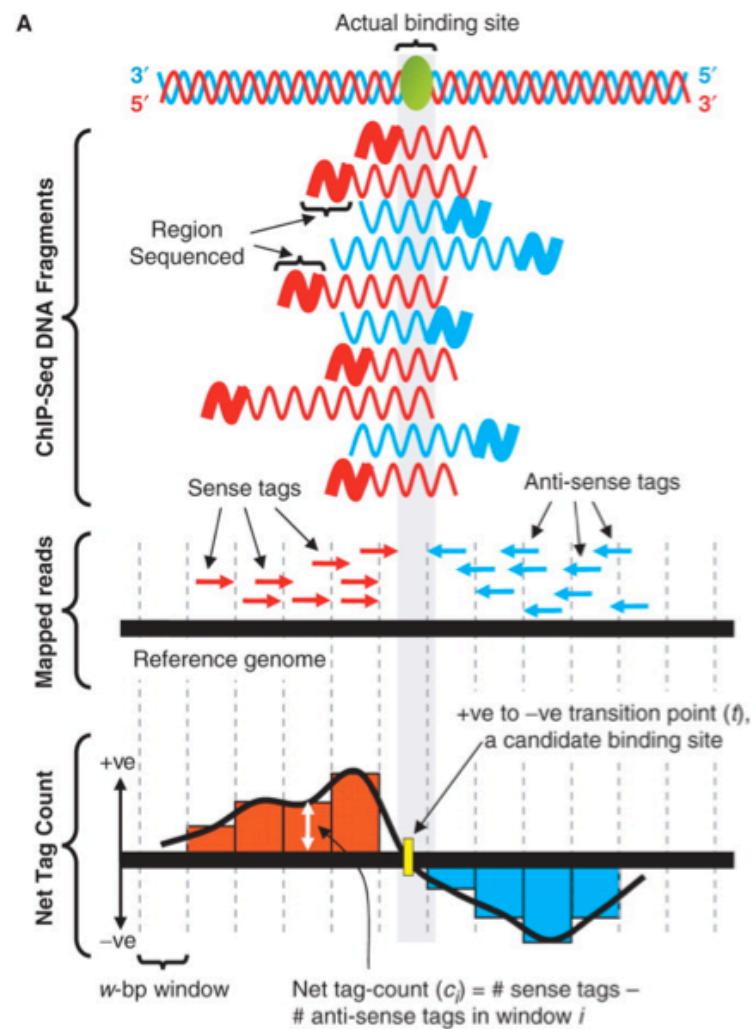
Focused Peaks:



Unfocused Peaks:



Finding peaks





Model-based Analysis for ChIP-Seq

[Readme](#)
[Install](#)
[Download](#)
[Contributions](#)
[FAQ](#)
[ChangeLog](#)

About

Next generation parallel sequencing technologies made chromatin immunoprecipitation followed by sequencing (ChIP-Seq) a popular strategy to study genome-wide protein-DNA interactions, while creating challenges for analysis algorithms. We present Model-based Analysis of ChIP-Seq ([MACS](#)) on short reads sequencers such as Genome Analyzer (Illumina / Solexa). [MACS](#) empirically models the length of the sequenced ChIP fragments, which tends to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites. [MACS](#) also uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence, allowing for more sensitive and robust prediction. [MACS](#) compares favorably to existing ChIP-Seq peak-finding algorithms, is publicly available open source, and can be used for ChIP-Seq with or without control samples.

Now, the newest version is [version 1.4.2](#)

Author

[MACS](#) is written by Yong Zhang and [Tao Liu](#) from Xiaole Shirley Liu's Lab.

Source Code

[On Github](#)

Citation

Our paper has been [published](#) in Genome Biology. Please cite "Zhang et al. Model-based Analysis of ChIP-Seq ([MACS](#)). *Genome Biol* (2008) vol. 9 (9) pp. R137".

Peak finding with MACS

- MACS is a PYTHON script developed and maintained by Tao Liu
- It is the most widely used and cited method now
- There are a lot of switch to fine tune the analysis
- Single experiments and control-treated pairs can be analyzed as well
- It provides a BED format file for the peaks (ChIP regions) and for the summits and an XLS file for the peaks.
- It also provides bedgraph format coverage files

HOMER



HOMER

Software for motif discovery and next-gen sequencing analysis

Next-Generation Sequencing Analysis

ChIP-Seq is the best thing that happened to ChIP since the antibody. It is 100x better than ChIP-Chip since it escapes most of the problems of microarray probe hybridization. Plus it is cheaper, and genome wide. But ChIP-Seq is only the tip of the iceberg - there are many inventive ways to use a sequencer. Below are a list of the more popular methods that will be covered below:

ChIP-Seq: Isolation and sequencing of genomic DNA "bound" by a specific transcription factor, covalently modified histone, or other nuclear protein. This methodology provides genome-wide maps of factor binding. Most of HOMER's routines cater to the analysis of ChIP-Seq data.

DNase-Seq: Treatment of nuclei with a restriction enzyme such as DNase I will result in cleavage of DNA at accessible regions. Isolation of these regions and their detection by sequencing allows the creation of DNase hypersensitivity maps, providing information about which regulatory elements are accessible in the genome.

MNase-Seq: Micrococcal Nuclease (MNase) is a restriction enzyme that degrades genomic DNA not wrapped around histones. The remaining DNA represents nucleosomal DNA, and can be sequencing to reveal nucleosome positions along the genome. This method can also be combined with ChIP to map nucleosomes that contain specific histone modifications.

RNA-Seq: Extraction, fragmentation, and sequencing of RNA populations within a sample. The replacement for gene expression measurements by microarray. There are many variants on this, such as Ribo-Seq (isolation of ribosomes translating RNA), small RNA-Seq (to identify miRNAs), etc.

GRO-Seq: RNA-Seq of nascent RNA. Transcription is halted, nuclei are isolated, labeled nucleotides are added back, and transcription briefly restarted resulting in labeled RNA molecules. These newly created, nascent RNAs are isolated and sequenced to reveal "rates of transcription" as opposed to the total number of stable transcripts measured by normal RNA-seq.

Hi-C: Genomic interaction assay for understanding genome 3D structure. This assay is much more specialized - For more information about how to use HOMER to analyze Hi-C data, check out the [Hi-C analysis section](#).

List of HOMER utilities

HOMER Program Index

Below is a quick introduction to the different programs included in HOMER. Running each program without any arguments will provide basic instructions and a list of command line options.

FASTA file Motif Discovery

[findMotifs.pl](#) - performs motif analysis with lists of Gene Identifiers or FASTA files (See [FASTA file analysis](#))

[homer2](#) - core component of motif finding (Called by everything else , See [FASTA file analysis](#))

Gene/Promoter-based Analysis

[findMotifs.pl](#) - performs motif and gene ontology analysis with lists of Gene Identifiers, both promoter and mRNA motifs (See [Gene ID Analysis Tutorial](#))

[findGO.pl](#) - performs only gene ontology analysis with lists of Gene Identifiers (Called by findMotifs.pl, See [Gene Ontology Analysis](#))

[loadPromoters.pl](#) - setup custom promoter sets for specialized analysis (See [Customization](#))

Next-Gen Sequencing/Genomic Position Analysis

[findMotifsGenome.pl](#) - performs motif analysis from genomic positions (See [Finding Motifs from Peaks](#))

[makeTagDirectory](#) - creates a "tag directory" from high-throughput sequencing alignment files, performs quality control (See [Creating a Tag Directory](#))

[makeUCSCfile & makeBigWig.pl](#) - create bedGraph file for visualization with the UCSC Genome Browser (See [Creating UCSC file](#))

[findPeaks](#) - find peaks in ChIP-Seq data, regions in histone data, de novo transcripts from GRO-Seq (See [Finding ChIP-Seq Peaks](#))

[analyzeChIP-Seq.pl](#) - automation of programs found above (See [Automation of ChIP-Seq analysis](#))

[annotatePeaks.pl](#) - annotation of genomic positions, organization of motif and sequencing data, histograms, heatmaps, and more... (See [Annotating Peaks, Quantification](#))

[analyzeRNA.pl](#) - quantification of RNA levels across transcripts (See [RNA quantification](#))

[mergePeaks](#) - find overlapping peak positions (See [Comparing ChIP-Seq Peaks](#))

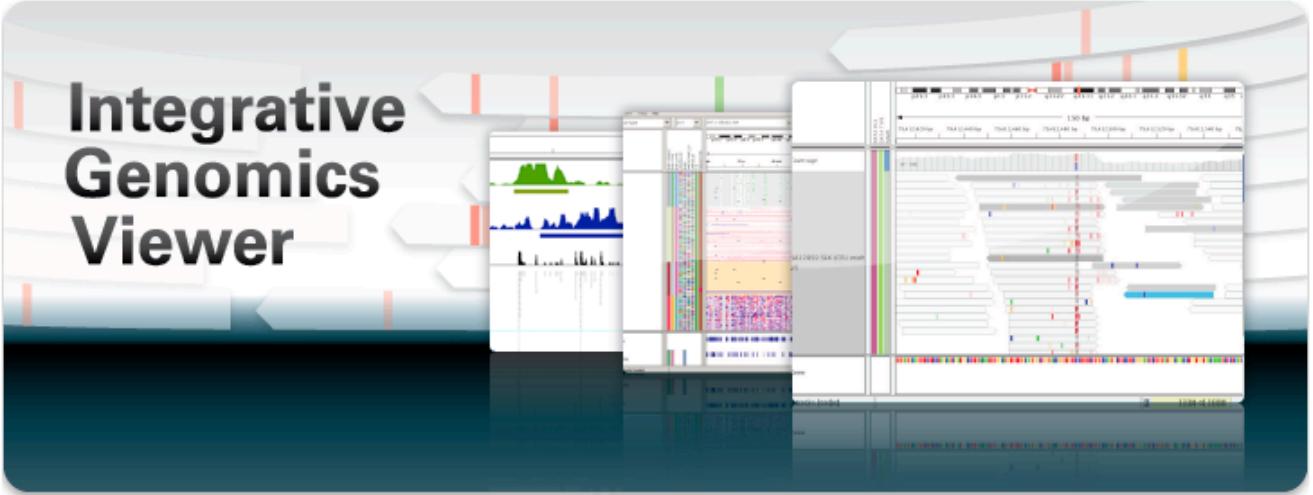
[homerTools](#) - basic sequence manipulation (See [Sequence Manipulation](#))

[tagDir2bed.pl](#) - output tag directory as an alignment BED file (See [Miscellaneous](#))

[bed2pos.pl, pos2bed.pl](#) - convert between HOMER peak file format and BED file format (See [Miscellaneous](#))

[checkPeakFile.pl](#) - use this to see if your peak file is in the correct format

The IGV genome browser (for visualization of the genomic data)



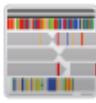
What's New



April 20, 2012. IGV 2.1 has been released. See the [release notes](#) for more details.

April 19, 2012. See our new [IGV paper](#) in *Briefings in Bioinformatics*.

Overview



The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

Downloads



Please [register](#) to download IGV. After registering, you can log in at any time using your email address. Permission to use IGV is granted under the GNU [LGPL license](#).

Citing IGV

To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. *Nature Biotechnology* 29, 24–26 (2011), or

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. *Briefings in Bioinformatics* 2012.

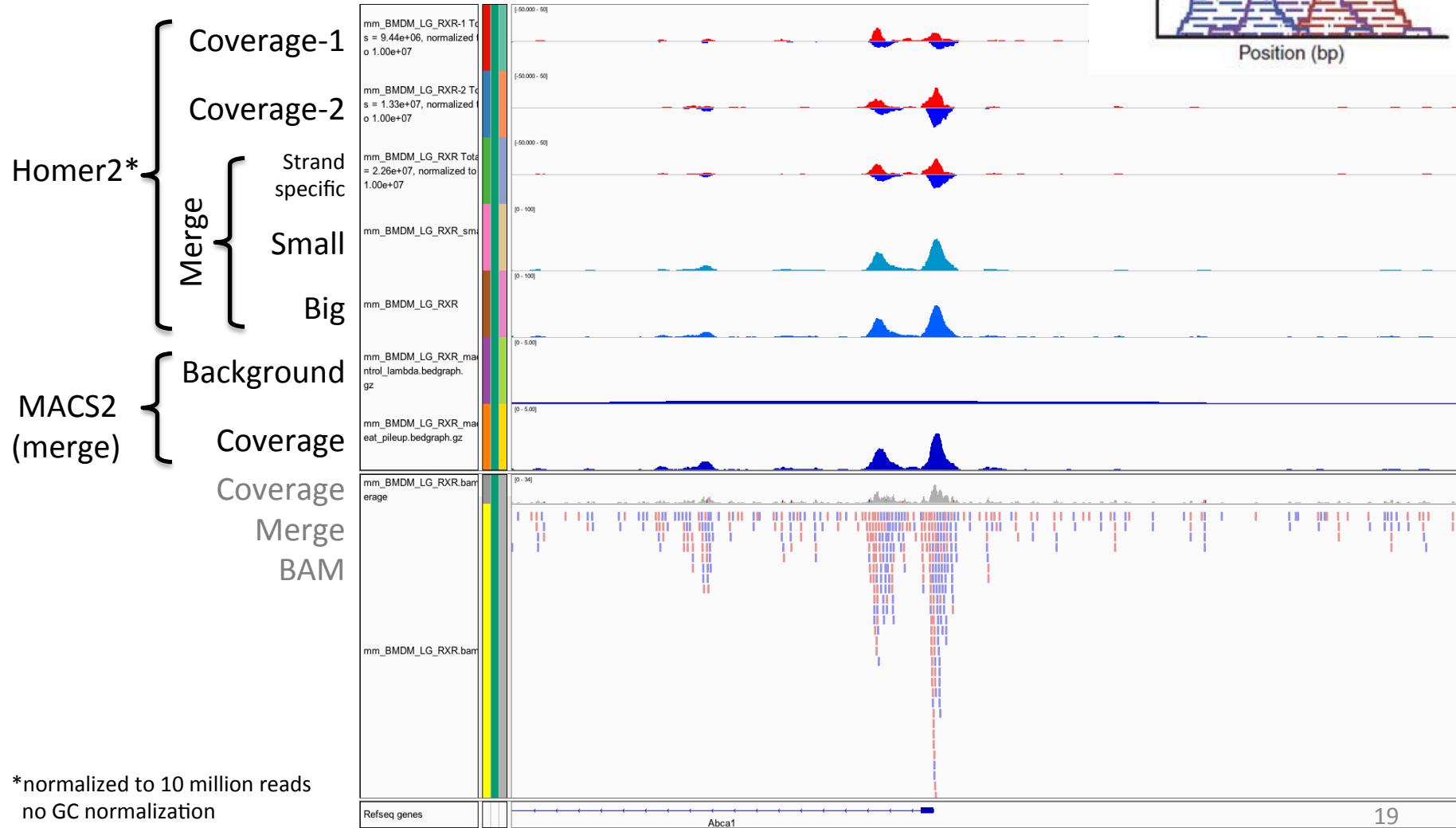
Funding

Development of IGV is made possible by funding from the [National Cancer Institute](#), the [National Institute of General Medical Sciences](#) of the [National Institutes of Health](#), and the [Starr Cancer Consortium](#).



IGV is participating in the [GenomeSpace](#) initiative.

BAM and bedgraph files



*normalized to 10 million reads
no GC normalization

Peak finding using HOMER (findPeaks)

```
# HOMER Peaks
# Peak finding parameters:
# tag directory = Sox2-ChIP-Seq
#
# total peaks = 10280
# peak size = 137
# peaks found using tags on both strands
# minimum distance between peaks = 342
# fragment length = 132
# genome size = 4000000000
# Total tags = 9908245.0
# Total tags in peaks = 156820.0
# Approximate IP efficiency = 1.58%
# tags per bp = 0.001907
# expected tags per peak = 0.523
# maximum tags considered per bp = 1.0
# effective number of tags used for normalization = 10000000.0
# Peaks have been centered at maximum tag pile-up
# FDR rate threshold = 0.001000
# FDR effective poisson threshold = 0.000000
# FDR tag threshold = 8.0
# number of putative peaks = 10800
#
# size of region used for local filtering = 10000
# Fold over local region required = 4.00
# Poisson p-value over local region required = 1.00e-04
# Putative peaks filtered by local signal = 484
#
# Maximum fold under expected unique positions for tags = 2.00
# Putative peaks filtered for being too clonal = 36
#
# cmd = findPeaks Sox2-ChIP-Seq -style factor -o auto
#
# Column Headers:
```

Genome size represents the total effective number of mappable bases in the genome (remember each base could be mapped in each direction)

Approximate IP efficiency describes the fraction of tags found in peaks versus. genomic background. This provides an estimate of how well the ChIP worked. Certain antibodies like H3K4me3, ERα, or PU.1 will yield very high IP efficiencies (>20%), while most fall in the 1-20% range. Once this number dips below 1% it's a good sign the ChIP didn't work very well and should probably be optimized.

- Column 1: PeakID - a unique name for each peak (very important that peaks have unique names...)
- Column 2: chr - chromosome where peak is located
- Column 3: starting position of peak
- Column 4: ending position of peak
- Column 5: Strand (+/-)
- Column 6: Normalized Tag Counts - number of tags found at the peak, normalized to 10 million total mapped tags (or defined by the user)
- Column 7: (-style factor): Focus Ratio - fraction of tags found appropriately upstream and downstream of the peak center. (see below)
 - (-style histone/-style groseq): Region Size - length of enriched region
- Columns 8+: Statistics and Data from filtering

De novo motif finding with HOMER

How `findMotifsGenome.pl` works:

1. Verify peak/BED file
2. Extract sequences from the genome corresponding to the regions in the input file, filtering sequences that are >70% "N"
3. Calculate GC/CpG content of peak sequences.
4. Preparse the genomic sequences of the selected size to serve as background sequences
5. Randomly select background regions for motif discovery
6. Auto normalization of sequence bias.
7. Check enrichment of known motifs
8. de novo motif finding

HOMER *denovo* motif finding result

Total target sequences = 4219

Total background sequences = 45719

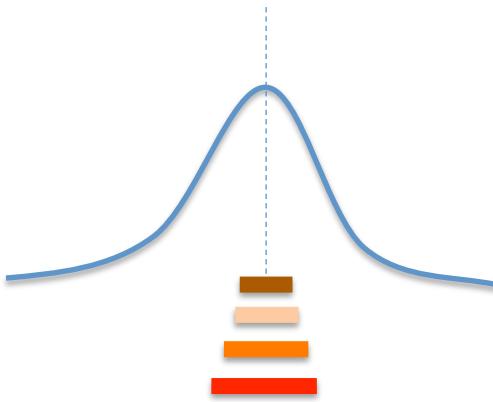
* - possible false positive

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-658	-1.516e+03	54.23%	16.93%	414.8bp (300.1bp)	PB0058.1_Sfpi1_1 More Information Similar Motifs Found
2		1e-131	-3.036e+02	46.43%	28.60%	465.7bp (296.4bp)	PB0049.1_Nr2f2_1 More Information Similar Motifs Found
3		1e-131	-3.025e+02	32.80%	17.26%	489.8bp (298.2bp)	Jun-AP1(bZIP)/K562- cJun-ChIP-Seq/Homer More Information Similar Motifs Found
4		1e-70	-1.632e+02	29.18%	17.91%	483.4bp (300.2bp)	MA0102.2_CEBPA More Information Similar Motifs Found

- ▶ RXR peaks overlapping with GRO-seq paired peaks
- ▶ Enrichment = % of Targets / % of Background
- ▶ The P-value depends on the size of the sample (not comparable between different samples)
- ▶ Best match (HOMER has its own motif library coming from the JASPAR database and from ChIP-seq analyses) does not mean perfect match!

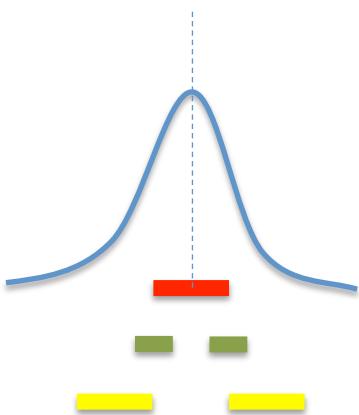
Homer

Motives are close to peak summits



	Width	p-value	target %	bg %	fold
PU.1	50	1E-74	23.17	4.66	4.97
	60	1E-91	28.62	5.94	4.82
	80	1E-142	34.02	5.32	6.39
	100	1E-169	44.24	8.28	5.34

	Width	p-value	target %	bg %	fold
NRhalf	50	1E-151	39.27	6.96	5.64
	60	1E-184	41.05	6.02	6.82
	80	1E-254	52.78	7.63	6.92
	100	1E-258	59.23	10.39	5.70



	p-value	target %	bg %	fold
PU.1	1E-86	18.97	5.34	3.55
	1E-135	21.85	4.62	4.73
	1E-169	44.24	8.28	5.34

	p-value	target %	bg %	fold
NRhalf	1E-18	0.64	0.01	64.00
	1E-52	9.52	2.18	4.37
	1E-258	59.23	10.39	5.70

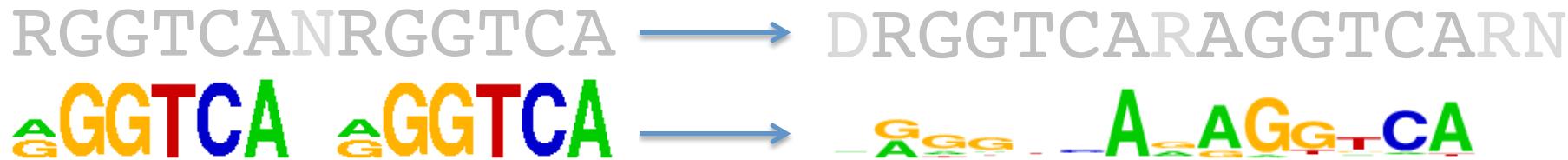


Homer

Searching for motif enrichments

- Functions that make impossible complex motif search if a simple (component) motif is highly enriched:
 - cyclic re-optimization of motives
 - masking out found motives from further searching
- Parameters:
 - quickMask: omits “rank masking”
 - maskMotif: masks given motives
 - len: 10,12,14,16 instead of 8,12,16,20
 - size: 200, 100, given (summit +/-50)
 - opt: optimizing a given motif
 - mis: (3 instead of) **2**

Motif matrix files and logos



>Consensus sequence				Name	Score threshold
>RGGTCANRGGTCA	DR1	0.499	0.001	0.499	15.2051888767745
R 0.499	0.001	0.499	0.001		
G 0.001	0.001	0.997	0.001		
G 0.001	0.001	0.997	0.001		
T 0.001	0.001	0.001	0.997		
C 0.001	0.997	0.001	0.001		
A 0.997	0.001	0.001	0.001		
N 0.25	0.25	0.25	0.25		
R 0.499	0.001	0.499	0.001		
G 0.001	0.001	0.997	0.001		
G 0.001	0.001	0.997	0.001		
T 0.001	0.001	0.001	0.997		
C 0.001	0.997	0.001	0.001		
A 0.997	0.001	0.001	0.001		

>Consensus sequence				Name	Score threshold
>DRGGTCARAGGTCARN	1-DRGGTCARAGGTCARN	0.367	0.067	0.311	0.256
D 0.367	0.001	0.466	0.001	0.532	0.001
R 0.189	0.022	0.111	0.100	0.656	0.133
G 0.111	0.001	0.133	0.256	0.667	0.122
T 0.166	0.512	0.133	0.256	0.156	0.455
C 0.066	0.222	0.166	0.512	0.011	0.100
A 0.966	0.011	0.966	0.001	0.011	0.022
R 0.378	0.477	0.378	0.067	0.078	
A 0.821	0.177	0.821	0.001	0.001	
G 0.066	0.922	0.066	0.011	0.001	
G 0.044	0.767	0.044	0.022	0.167	
T 0.111	0.134	0.111	0.089	0.666	
C 0.078	0.045	0.078	0.855	0.022	
A 0.900	0.022	0.900	0.001	0.077	
R 0.278	0.378	0.278	0.211	0.134	
N 0.300	0.267	0.300	0.211	0.222	

Log P value Unused Match in Target and Background, P value
 * -226.958685 0 T:159.0(19.06%),B:970.0(2.08%),P:1e-98

BEDtools



bedtools: a flexible suite of utilities for comparing genomic features.

[Project Home](#) [Downloads](#) [Wiki](#) [Issues](#) [Source](#)

[Summary](#) [People](#)

Project Information

+31 Recommend this on Google

Starred by 107 users
[Project feeds](#)

Code license
[GNU GPL v2](#)

Labels

bioinformatics, genomics,
bed, sam, bam, overlap,
features, sequencing,
intersect, coverage, gff, vcf,
bedgraph, intervals,
genomearithmetric

Members
aaronqui...@gmail.com

Featured

Downloads
[BEDTools.v2.17.0.tar.gz](#)
[Show all »](#)

Wiki pages
[Contributors](#)
[FAQ](#)
[Help](#)

Documentation

The documentation for bedtools has moved!

The existing PDF manual will be phased out in the next few months.

The most up to date documentation is at bedtools.readthedocs.org.

BEDTools Summary

The BEDTools utilities allow one to address common genomics tasks such as finding feature overlaps and computing coverage. The utilities are largely based on four widely-used file formats: [BED](#), [GFF/GTF](#), [VCF](#), and [SAM/BAM](#). Using BEDTools, one can develop sophisticated pipelines that answer complicated research questions by "streaming" several BEDTools together. The following are examples of common questions that one can address with BEDTools.

1. Intersecting two BED files in search of overlapping features.
2. Culling/refining/computing coverage for BAM alignments based on genome features.
3. Merging overlapping features.
4. Screening for *paired-end* (PE) overlaps between PE sequences and existing genomic features.
5. Calculating the depth and breadth of sequence coverage across defined "windows" in a genome.
6. Screening for overlaps between "split" alignments and genomic features.

Citation

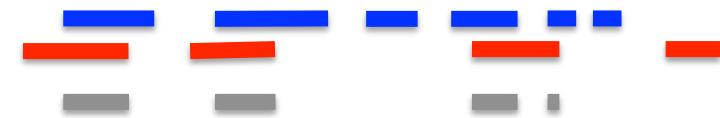
Please cite the following article if you use BEDTools in your research:

- Quinlan AR and Hall IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.

BEDtools: intersectBed

Switches:

- -a *peakfile1.bed* -b *peakfile2.bed*



((-abam => -bed))

- -u



- -v



- -c (count b on a)



1	1	0	1	1	0
---	---	---	---	---	---

- -wo (fusing beds in a “double bed” table)



- -f (minimum overlap %) – -u -f 0.6 →



- -r (reciprocal overlap) – -u -f 0.6 -r →



- -s (strand specific match)

VennMaster

Peak overlaps

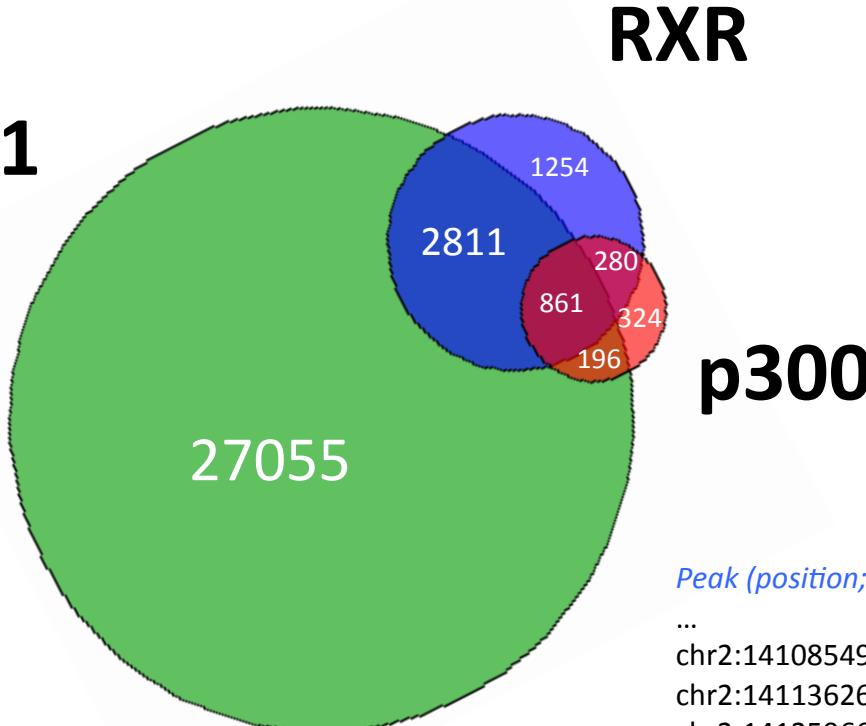
A listfile

Unit (peak) *Group*

...
A-30917 PU.1
A-30918 PU.1
A-30919 PU.1
A-30920 PU.1
A-30921 PU.1
A-30922 PU.1
A-30923 PU.1
A-1 RXR
A-2 RXR
A-3 RXR
A-4 RXR
A-5 RXR
A-6 RXR
A-7 RXR
. .
. .

PU.1

→ VennMaster



New list file format
automatically from beds:

filename ↓

<i>Peak (position; merge)</i>	<i>Group</i>
...	...
chr2:14108549-14108910	PU1peaks
chr2:14113626-14114054	PU1peaks
chr2:14125966-14126214	PU1peaks
chr2:14128322-14128869	PU1peaks
chr2:14150196-14150481	PU1peaks
chr2:14150692-14151155	PU1peaks
chr2:14155587-14155954	PU1peaks
...	...

Annotation of the peaks (annotatePeaks.pl)

- Genomic localization
- Closest TSS
- Motif occurrences
- Enrichment in different ontologies

Gene Ontology Enrichment Results

[Homer *de novo* Motif Enrichment Results](#)

[Known Motif Enrichment Results](#)

Text file version of complete results (i.e. open with Excel)

- **Biological Process:** Functional groupings of proteins ([Gene Ontology](#))
- **Molecular Function:** Mechanistic actions of proteins ([Gene Ontology](#))
- **Cellular Component:** Protein localization ([Gene Ontology](#))
- **KEGG Pathways:** Groups of proteins in the same pathways (From [KEGG](#)) (last update > year ago)
- **Interactions:** Groups of proteins interacting with the same protein (From [NCBI Gene](#))
- **Interpro:** Proteins with similar domains and features ([Interpro](#))
- **Pfam:** Proteins with similar domains and features ([Pfam](#))
- **SMART:** Proteins with similar domains and features ([SMART](#))
- **Gene3D:** Proteins with similar domains and features ([Gene3D Database](#))
- **Prosite:** Proteins with similar domains and features ([Prosite Database](#))
- **PRINTS:** Proteins with similar domains and features ([PRINTS Database](#))
- **Chromosome Location:** Genes with similar chromosome localization
- **miRNA Targets:** Genes targeted by similar miRNAs ([miRBase](#)) (last update > year ago)
- **MSigDB:** Genes sets for pathways, factor/miRNA target predictions, expression patterns, etc. ([GSEA/MSigDB](#))
- **wikipathways:** Genes sets for pathways ([Wikipathways](#))

Annotating:....	
3UTR	42.0
Other	9.0
TTS	78.0
LINE	126.0
srpRNA	2.0
SINE	106.0
DNA	23.0
Exon	69.0
Intron	2358.0
Intergenic	1946.0
Promoter	358.0
SUTR	7.0
scRNA	4.0
CpG-Island	12.0
Low_complexity	2.0
LTR	185.0
Simple_repeat	142.0
Satellite	49.0
rRNA	22.0

Method: Generate a list of genes and compare the list statistically with the list of genes present in a given ontology

Outputs of the primary analysis I.

1. BAM format alignment files for visualization and for occupancy analysis *name.bam*
2. Bedgraph files for visualization
 1. *name.bedgraph.gz* : normalized (10 million) extended reads from both strand
 2. *name_small.bedgraph.gz*: normalized (10 million), extended reads shown in the positive strand (summit shows the binding site)
 3. *name_big.bedgraph.gz* : same as above but with the highest resolution (and sometimes in a bigger size)
3. Bed files for visualization and for further analysis
 1. ChIP regions from HOMER analysis (*name_macs_peaks.bed*).
Summit +- 100 bp
 2. ChIP regions from MACS analysis (*name-homerpeaks.bed*)
 3. MACS peak summits (*name_macs_summits.bed*)

Outputs of the primary analysis II.

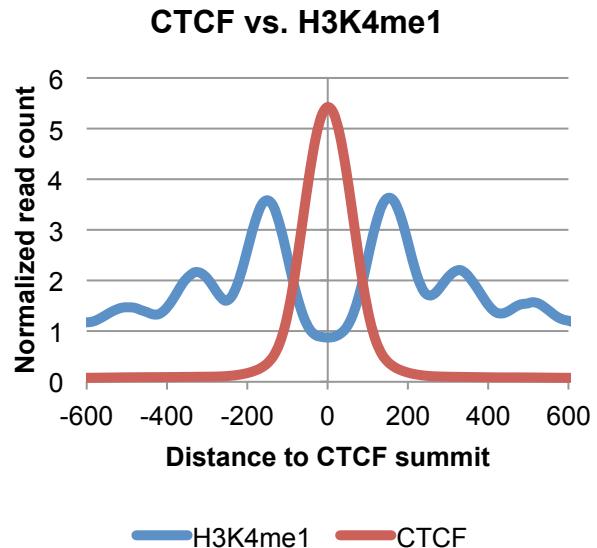
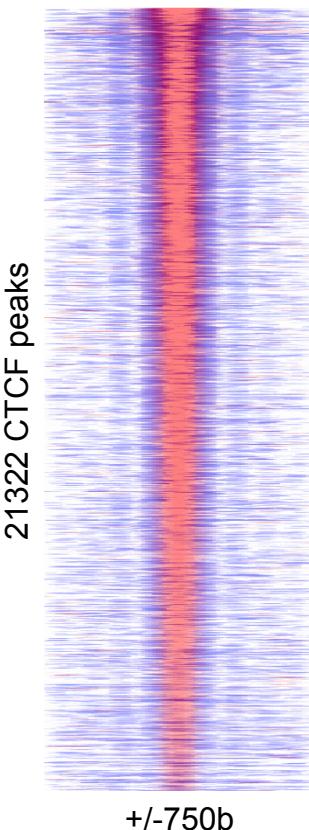
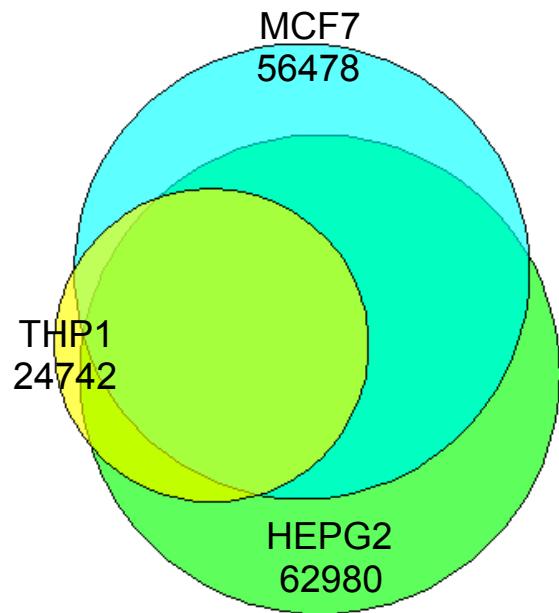
4. Annotation file (*name_homermotifsannot.txt*), tab delimited, can be directly imported into the excel or other programs). There is an other file (*name_macs-homermotifsannot.txt*) for MACS peak annotation
5. *denovo* motif finding, known motif enrichment and GO annotation enrichment for the best 1000 peaks from both the HOMER and MACS peak predictions. HTML format, which can be opened directly from any internet browser (*homerResults.html*)
6. Overall statistics of the experiments (generated with a separate script)

Downstream analysis

Comparing different samples

- Overlapping regions (intersectBed)
- Occupancy analysis (diffBind)
- Generating profiles
- Re-analyze peak subsets for motif occurrences

Goals of the module



Acknowledgments

<http://genomics.med.unideb.hu/>

- László Nagy director
- Bálint L Bálint head of the lab

Bioinformatics team:

- Gergely Nagy predoctor
- Attila Horváth system administrator, R programmer
- Dávid Jónás bioinformatician
- László Steiner mathematician
- Erik Czipa PhD student

<http://nlab.med.unideb.hu/>

László Nagy	lab leader
Bálint L Bálint	senior researcher
Zsuzsanna Nagy	senior researcher
Bence Dániel	PhD student
Zoltán Simándi	PhD student
Péter Brázda	PhD student
Ixchelt Cuaranta	PhD student

Supported by NKTH/ANR Regulomix
OMFB-00313/2010

