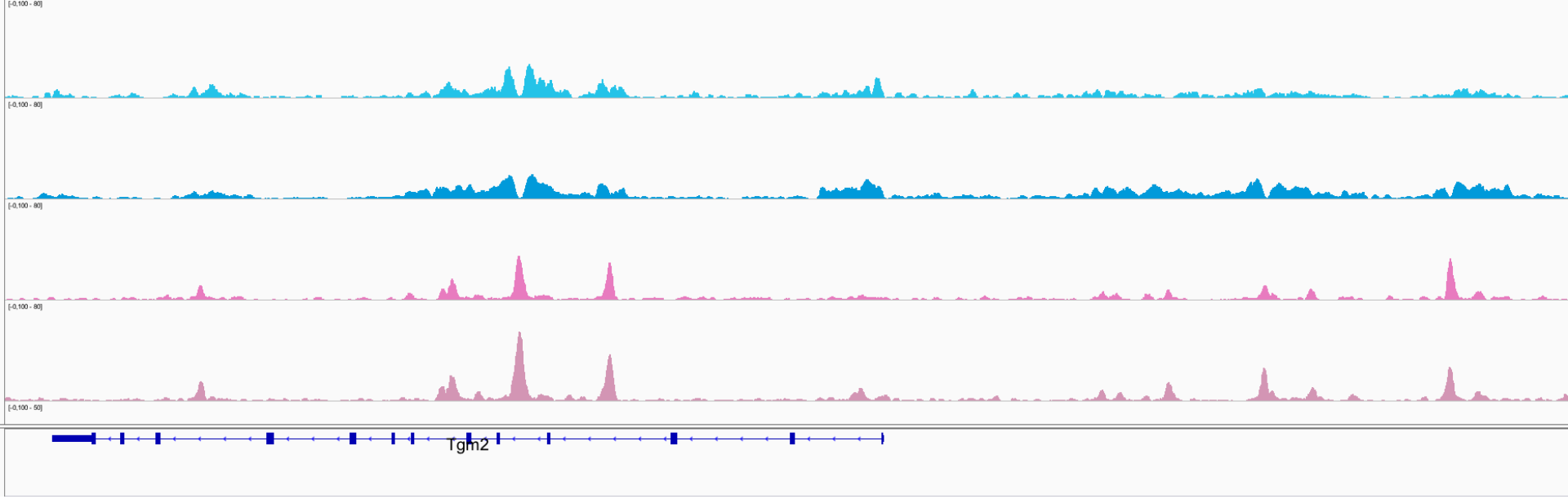


CHIP-SEQ DATA ANALYSIS

Endre Barta, Hungary

University of Debrecen, Center for Clinical Genomics
barta.endre@unideb.hu

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

OUTLINE

- General introduction to the ChIP-seq technology
- Visualisation of genome data
- Command-line ChIP-seq analysis
  - Quality checking of the reads
  - Using SRA
  - Alignment to the genome
  - Peak calling
  - Peak annotation
  - *denovo* motif finding
- Downstream ChIP-seq analysis
  - Comparing different samples
  - Analyzing ChIP region occupancy
- Web pages, program packages helping in ChIP-seq analysis

Functional genomics (transcriptomics) analysis

-A need to understand genome function on a global scale

-Gene regulation shapes cellular function.

To better understand gene-regulation genome-wide, we need to:

- Detect the openness and specific stage of the chromatin (histone modification ChIP-seq)
- Find transcription factor bound sites (TF ChIP-seq)
- Find co-regulators bound to transcription factors (ChIP-seq and other techniques)
- Find active enhancers (ChIP-seq, GRO-seq, RNA-seq)
- Determine transcript specific gene expression levels (GRO-seq, RNA-seq)

Traditional approaches to the problem

- Experimental determination of transcription factors binding sites.
  EMSA, DNAse footprinting

in vitro

- Computational prediction of binding motifs.

Is a motif a real binding site?

- DNA microarray (RNA-seq) analysis of target genes

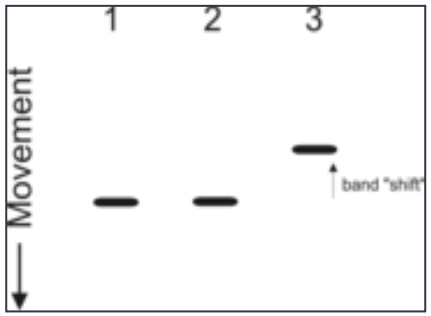
Is the target direct or indirect?

- Reporter gene expression assays of regulatory regions, motifs (deletion, point mutation)

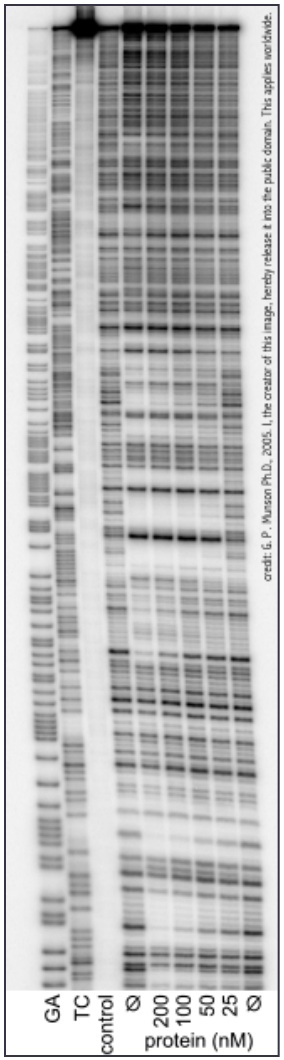
The natural environment of the given sequences can be completely different

The need for a more direct (and genome-wide) in vivo method.

Gel shift (EMSA)

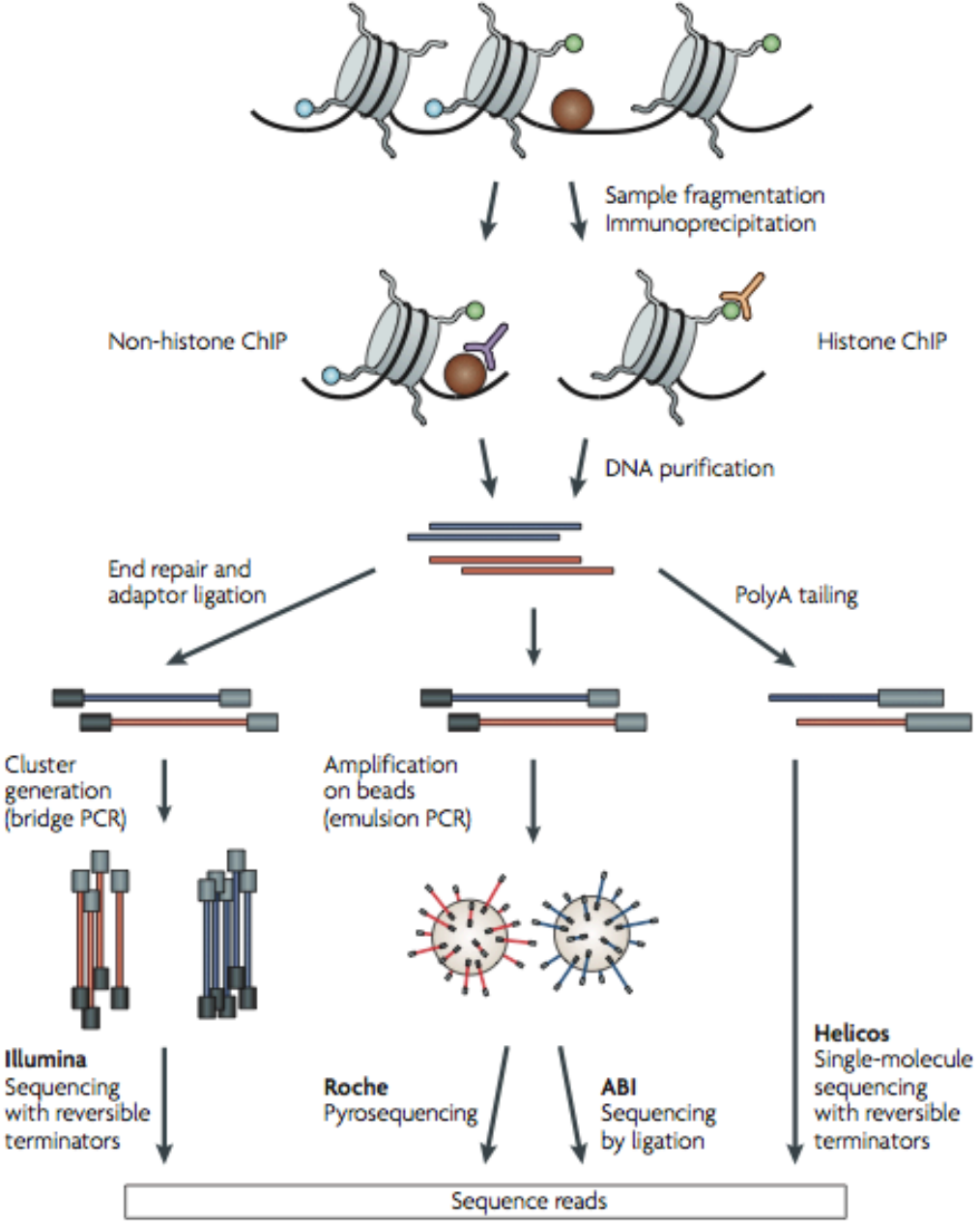


DNAse footprint



ChIP-seq technology

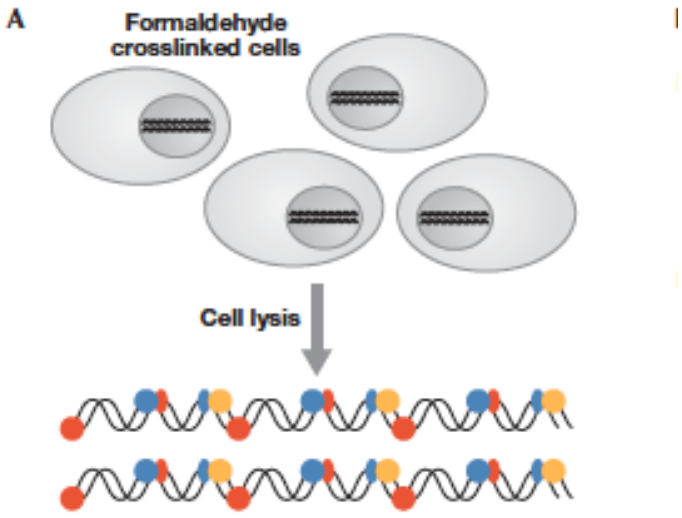
- Chromatine immunoprecipitation followed by next-generation sequencing
- Genome-wide detection of histone modifications or *transcription factor binding sites* (genome positions, where the TFs are bound to the DNS)

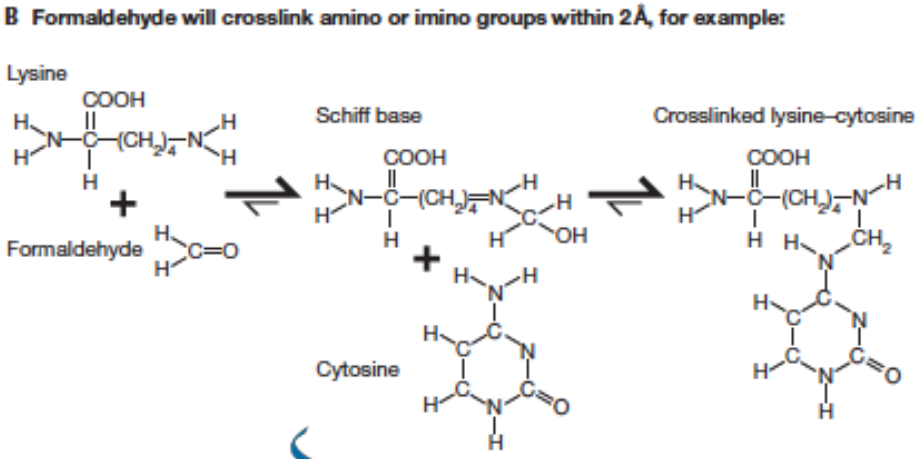


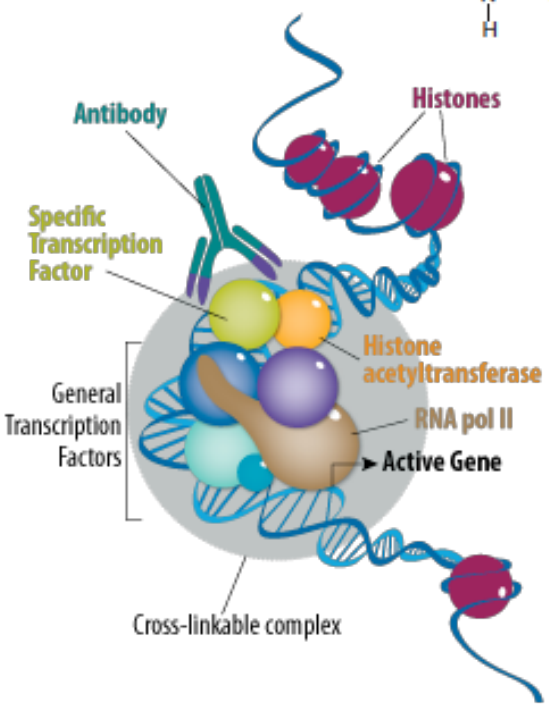
Park P. Nat. Rev. Genetics. 2009; 10:669-680

Chromatin immunoprecipitation (ChIP)

1. Cross linking with FA

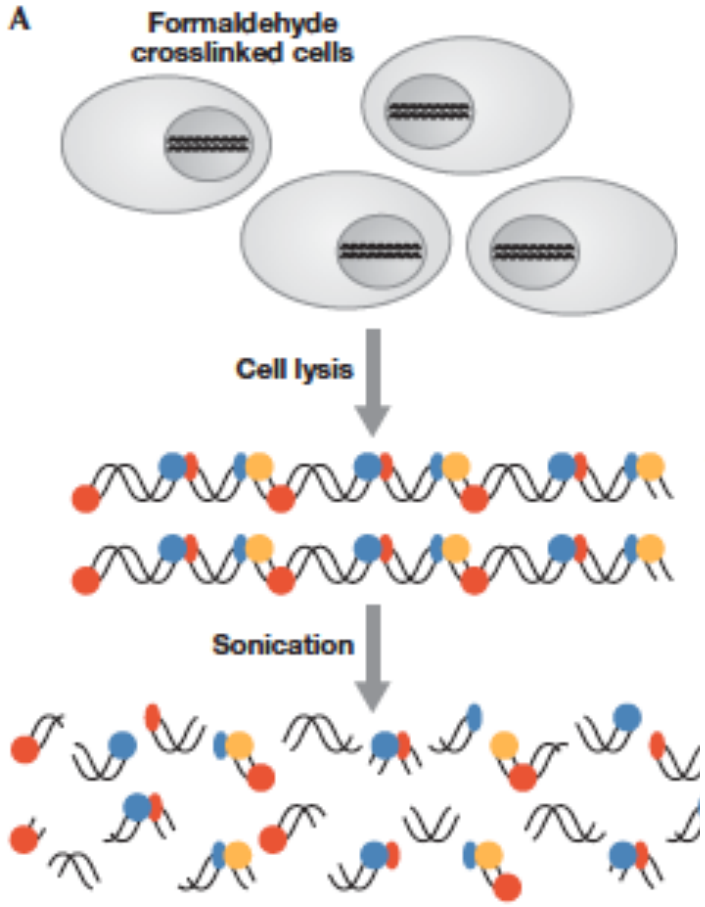


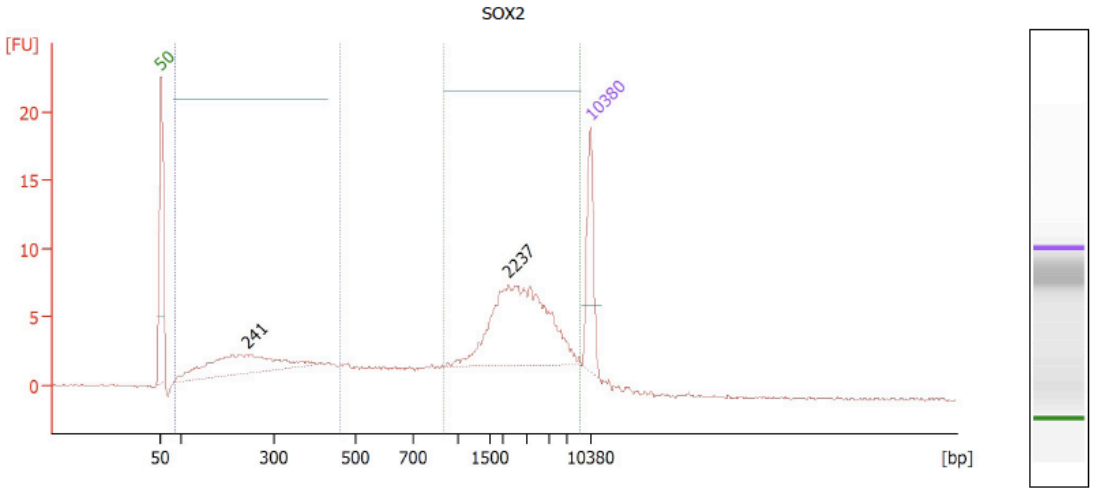




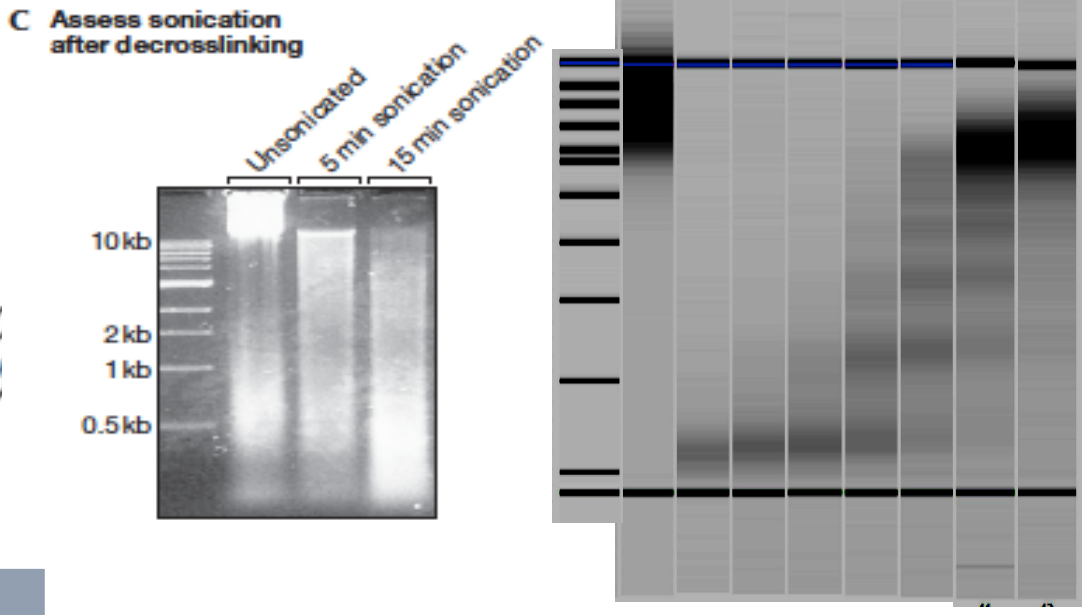
Optimization is crucial.

2. Cell lysis and sonication (or DNase treatment)





Overall Results for sample 1 : SOX2

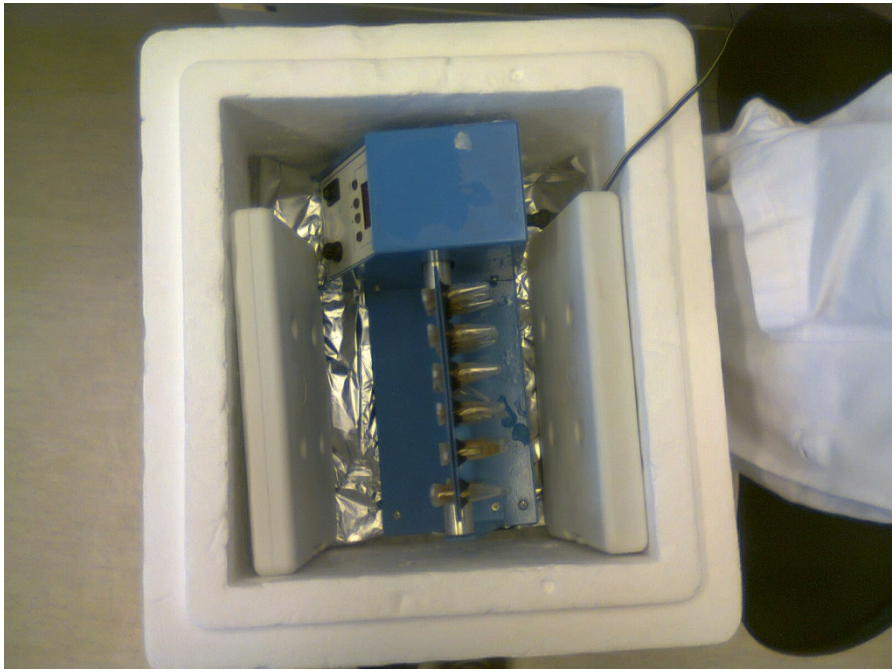


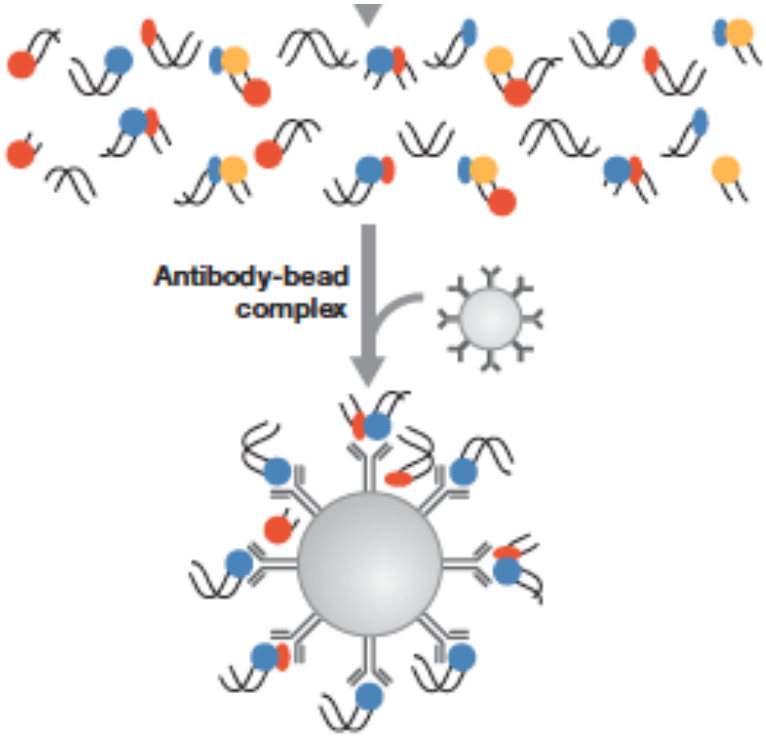
Zsolt Szilágyi, University of Gothenburg

Chromatin immunoprecipitation (ChIP)

3. Immunoprecipitation (IP)

The protein of interest is immunoprecipitated together with the crosslinked DNA

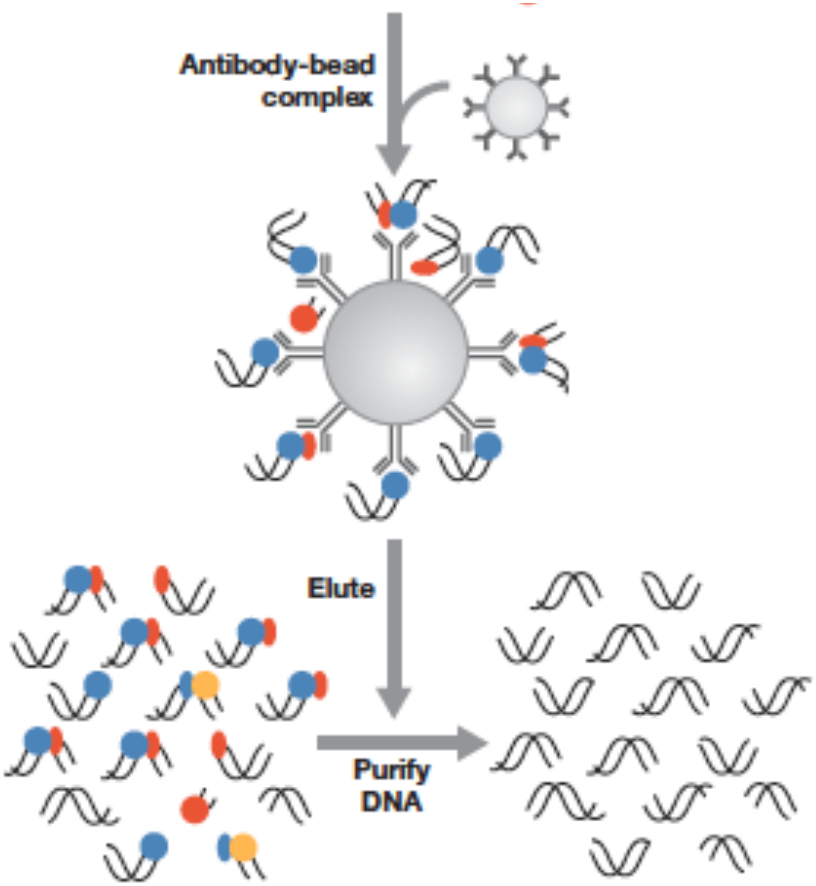


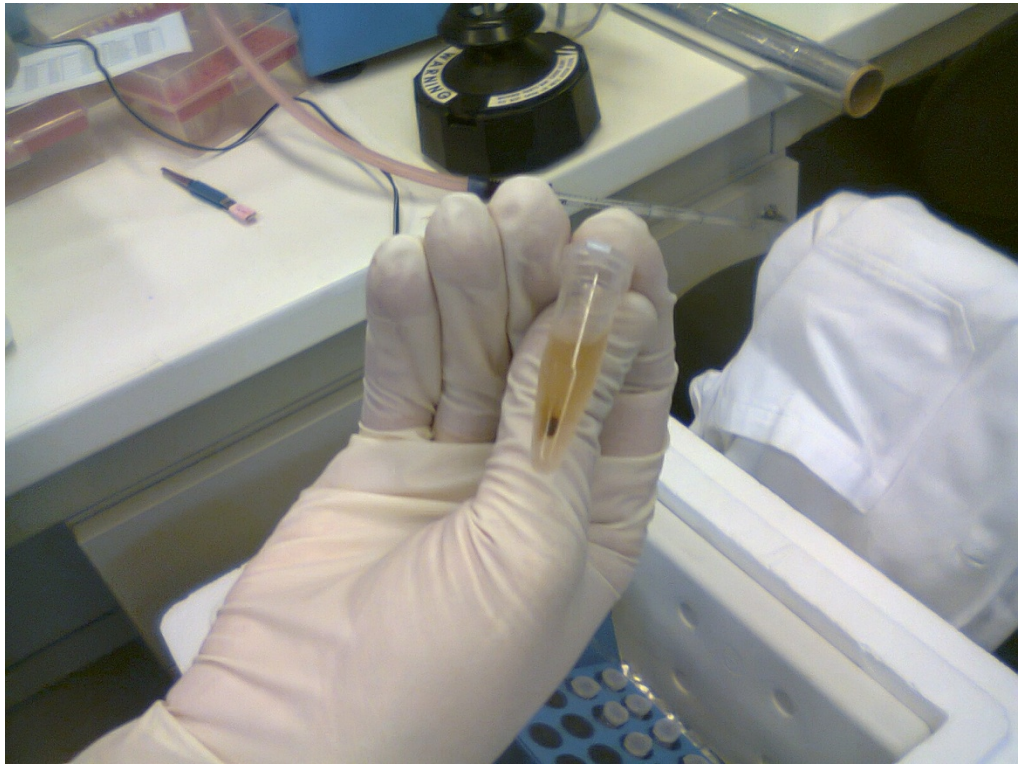


Chromatin immunoprecipitation (ChIP)

Reverse the FA crosslinking

4.Decrosslinking and purification of the DNA





Zsolt Szilágyi, University of Gothenburg

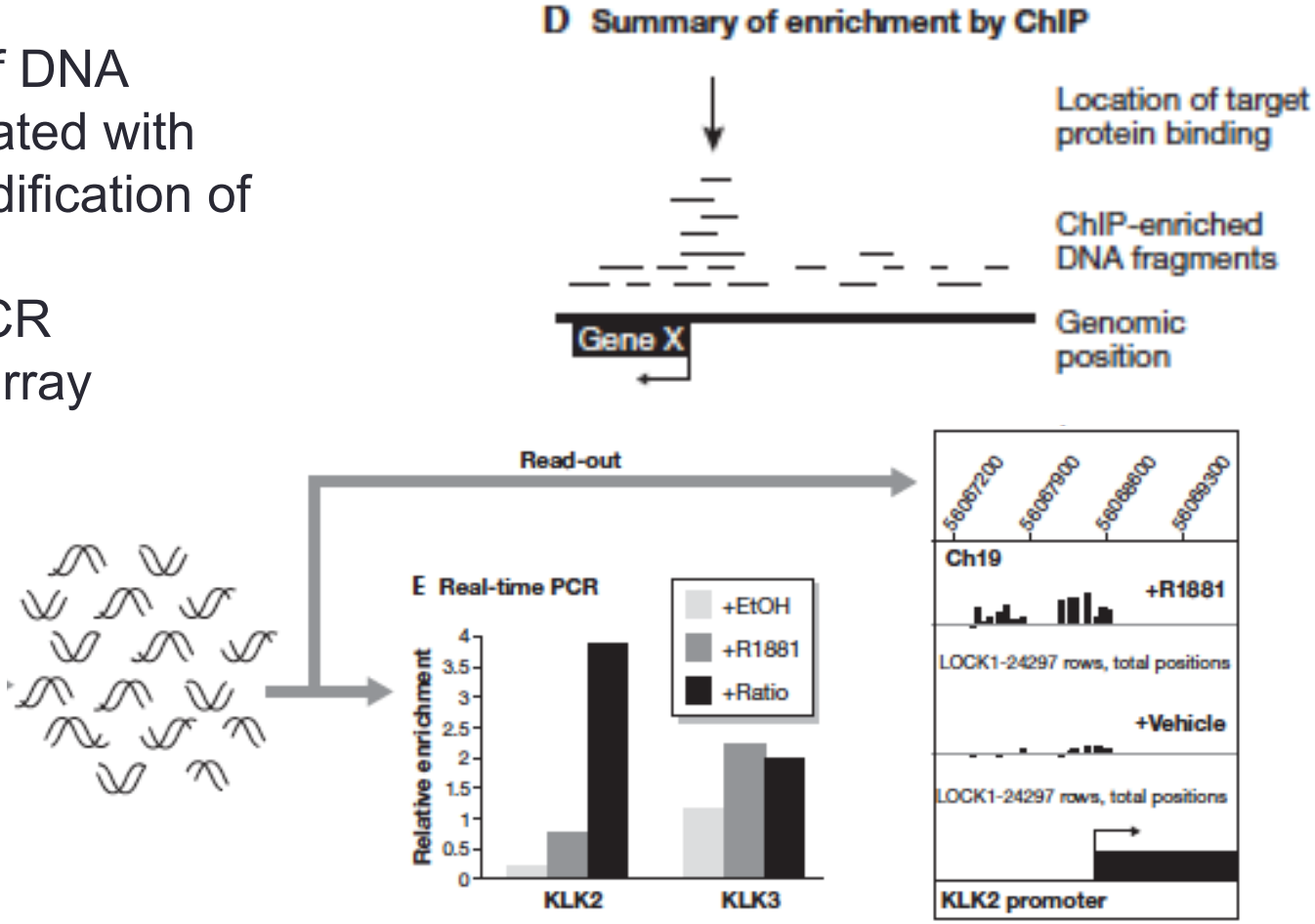
5. Analysis of ChIP DNA

6. Sequencing library preparation

Chromatin immunoprecipitation (ChIP)

Identification of DNA regions associated with the protein/modification of interest

- real-time PCR
- DNA microarray (ChIP-chip)
- Sequencing (ChIP-seq)



ChIP-seq analysis steps in our lab

- Sequencing -> fastq files
- ChIP-seq analyze script (semi-automatic)
  - BAM alignment files
  - Bedgraph files – visualization of peaks
  - Bed files - peak regions
  - Annotation files - (peak positions relative to genes, motif occurrences of two chosen TFBS etc.)
  - GO enrichment analysis
  - denovo motif finding
  - Known motif finding
- Downstream analysis
  - Occupancy analysis
  - Statistics
  - Defining peak subsets, merging, intersecting peaks
  - Comparison of peaksets, subsets
  - Re-doing certain analysis on peaksets, subsets

SEQanswers wiki list of ChIP-seq softwares

ChIP-Seq

SEQanswers
Forums
wiki navigation
Main page
Recent changes
Random page
Help
Software
Toolbox

The bioinformatics applications assigned the Biological domain **ChIP-Seq** (topic_3169) are tabulated below.

Definition:

- Topic concerning the analysis of protein-DNA interactions where chromatin immunoprecipitation (ChIP) is used in combination with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.

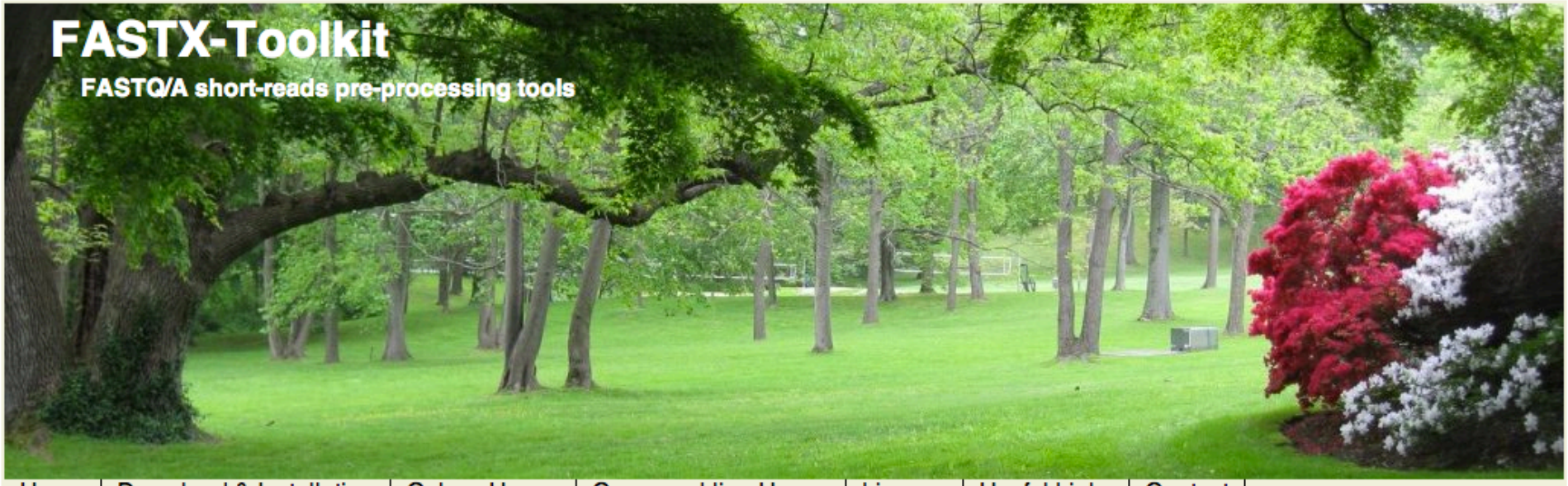
Synonyms:

- Chip-sequencing
- Chip seq
- Chip sequencing

Query returned 52 results.

	Biological domain	Bioinformatics method	Input format	Output format
AREM	**ChIP-Seq**	Peak calling Mapping	Bowtie SAM	BED
Avadis NGS	**ChIP-Seq** DNA-Seq RNA-Seq Small RNA Pathway analysis	Alignment Quality Control Sequence analysis Visualization Biological Contextualization	SAM BAM BED ELAND FASTA FASTQ	
BayesPeak	**ChIP-Seq**	Hidden Markov Model MCMC	BED GFF BAM Or any other format supported by BioConductor	BED GFF BAM Or any other format supported by BioConductor
BEADS	**ChIP-Seq**	Normalization		
CATCH	**ChIP-Seq** ChIP-on-chip	Clustering and alignment	BED WIG	BED CSV
ChIPmeta	Transcription Factor Binding Site identification **ChIP-Seq** ChIP-on-chip	Hidden Markov Model		
ChIPMunk	**ChIP-Seq** Motif analysis Motif discovery	Motif analysis Motif discovery	Multi-fasta Extended multi-fasta	Plain text
CHiPSeq	**ChIP-Seq**	Peak calling		
ChIPseqR	**ChIP-Seq**			
Chipster	**ChIP-Seq** RNA-Seq	QC Filtering Trimming Mapping Peak calling Motif detection	FASTQ SAM BAM	FASTQ SAM BAM

FASTX toolkit



Home | Download & Installation | Galaxy Usage | Command-line Usage | License | Useful Links | Contact

Introduction

The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

Next-Generation sequencing machines usually produce FASTA or FASTQ files, containing multiple short-reads sequences (possibly with quality information).

The main processing of such FASTA/FASTQ files is mapping (aka aligning) the sequences to reference genomes or other databases using specialized programs. Example of such mapping programs are: Blat, SHRiMP, LastZ, MAQ and many many others.

However,
It is sometimes more productive to preprocess the FASTA/FASTQ files before mapping the sequences to the genome - manipulating the sequences to produce better mapping results.

The FASTX-Toolkit tools perform some of these preprocessing tasks.

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

BEDTOOLS

bedtools

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

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.

Burrows-Wheeler Aligner BWA

Home

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 maual page
- Repository

Links:

- SAMtools
- MAQ

SAMtools

SAMTOOLS

SOURCEFORGE.NET®

Home

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

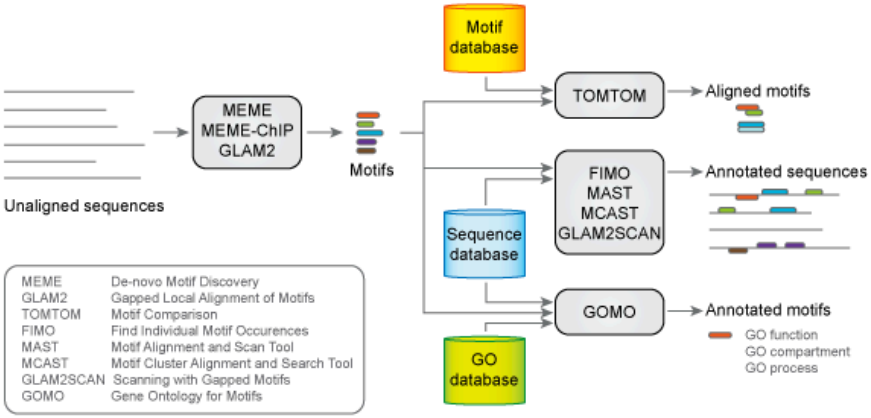
General Introduction
Manual Page (0.1.17)
Variant Calling (mpileup)
Text Alignment Viewer
API Documentation
Example C Program

MEME

The MEME Suite

Motif-based sequence analysis tools

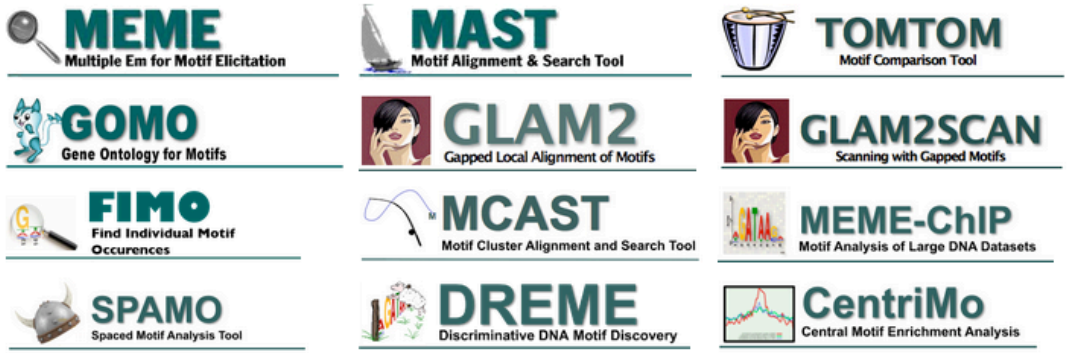
Previous version (4.8.1)



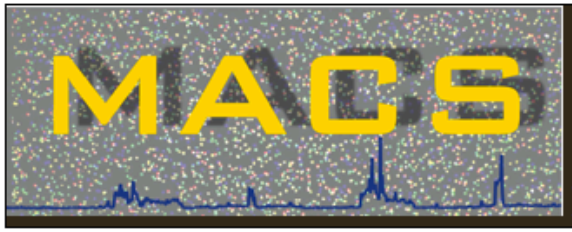
The MEME Suite allows you to:

- discover motifs using MEME, DREME (DNA only) or GLAM2 on groups of related DNA or protein sequences,
- search sequence databases with motifs using MAST, FIMO, MCAST or GLAM2SCAN,
- compare a motif to all motifs in a database of motifs,
- associate motifs with Gene Ontology terms via their putative target genes, and
- analyse motif enrichment using SpaMo or CentriMo.

To submit a query, click on one of the logos below or select "Submit A Job" from the menu at the left.



Maintenance and development of the MEME Suite is funded by the National Center for Research Resources grant NIH/NCRR R01 RR021692. The MEME Suite web server is funded by the National Biomedical Computation Resource.



Model-based Analysis for ChIP-Seq

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".

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 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

NGS software installed on ngsdeb

Programs used for ChIP-seq, GRO-seq and RNA-seq analysis

- HOMER
- SAMTOOLS
- BOWTIE
- ChIPSEEQER
- BEDTOOLS
- BAMTOOLS
- TOPHAT
- TRINITYRNASEQ
- BWA
- MEME
- MACS
- FASTX-TOOLKI
- PICARD

Head Node: 2x6 core, 144GB RAM, 20 Tbyte disk
6x computing nodes: 2x6 core, 48GB RAM 600GB disk

Other programs used for sequence analysis

- VCF-TOOLS
- VCFUTILS
- TABIX
- EMBOSS
- BLAST
- BLAST+
- CBUST
- LASTZ
- MULTIZ
- BLAT
- WEEDER
- DIALIGN
- SRMA
- CLUSTALW
- GLAM2

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)

Naming the samples

- hs_mq_STAT6_1

1. Species: hs or mm
2. Cell type
3. Antibody
4. Replica, other conditions etc.

- Should be consequent (e.g. parallels 1,2,3 not 1,2,a etc) – this will be the name through all the downstream analysis

Parameters that can be set for the run

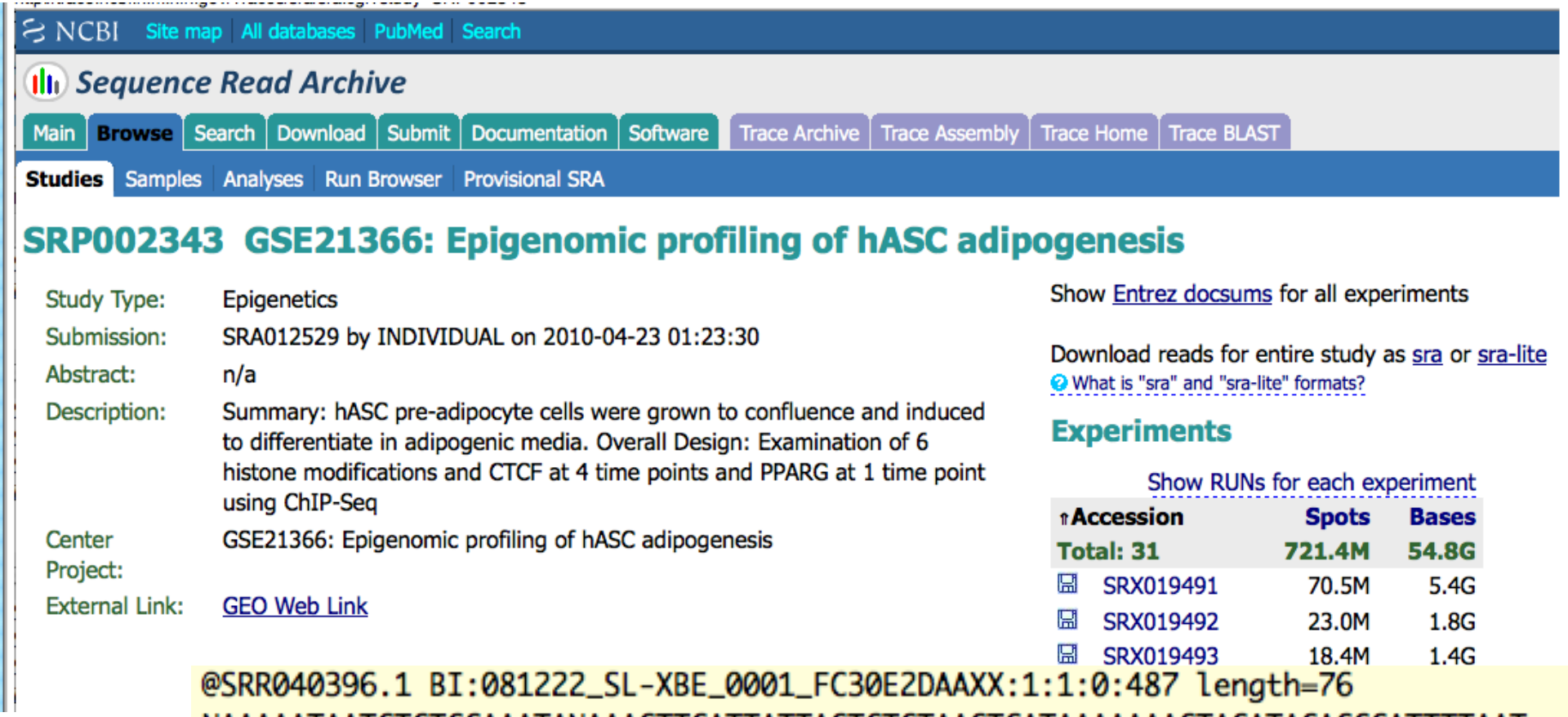
Inside the script

- Procno – for BWA and homer2
- Motif1, motif2 – for explicitly annotating these motifs
- Motiflength – for HOMER (can be 6,10,16 for example)
- Minmotifl, maxmotifl – for MEME
- Nmotif – number of motifs to find for MEME
- Indexdir – for holding indexed genome files for BWA
- Barcode_length – for BWA

As an argumentum

- Location of the listfile
- The analysis directory (where to put or where are already the sample directories
- Optionally a bigdir for storing SRA, fastq and sai files (optimally a local harddisk in a cluster environment)

Raw ChIP-seq sequence data available from the NCBI SRA database (~6000 human and ~5000 mouse)




```
@SRR040396.1 BI:081222_SL-XBE_0001_FC30E2DAAXX:1:1:0:487 length=76
NAAAAATAATGTCTGCAAATANAAACTTCATTATTACTCTGTAACTCATAAAAAAAGTACATAGAGCCATTTTAAT
+SRR040396.1 BI:081222_SL-XBE_0001_FC30E2DAAXX:1:1:0:487 length=76
!:F-:7I5III$5IHIII=I2!I17III@IIIIII1(I0IF.EDIII-E;57)5I-D.<;-I,B.I9608IFI20I
@SRR040396.2 BI:081222_SL-XBE_0001_FC30E2DAAXX:1:1:0:903 length=76
NCAAAGTAGCTGGGAATACAGNTGCCCGCCACCACACCCGGCTAATGATTTGTATTTTTAGTAGAGACGGGGTGTC
+SRR040396.2 BI:081222_SL-XBE_0001_FC30E2DAAXX:1:1:0:903 length=76
!I7?3I&2=6I32--5<,@.H!+&@@=C5I8BF3I1+0+/>+?..0I(+?*I&-668D0%+*)*)F&+6//2'+*&
```

Listfile for the analysis (SRA entries)

Downloads and analyzes the reads from the listed experiments

```
hs_lymphoblastoid_GM15510_normal_NFKB_TNFa      ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX150/SRX150608   factor
hs_lymphoblastoid_GM18505_normal_NFKB_TNFa      ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX150/SRX150362   factor
hs_lymphoblastoid_GM18505_normal_PolII_n        ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX150/SRX150601   factor
hs_MammalEpithel_T47D_cancer_ELF5_doxycycline   ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX093/SRX093408   factor
hs_MammalEpithel_T47D_cancer_ERa_BPA    ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX190/SRX190279   factor
hs_MammalEpithel_T47D_cancer_JunD_n     ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX190/SRX190174   factor
hs_MammalEpithel_T47D_cancer_ProgesteroneReceptor_RU486 ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX185/SRX185709   factor
hs_MammaryEpithel_MammaryEpithel_cancer_HSF1_n  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX155/SRX155763   factor
hs_MammaryGland_MCF10AErSrc_normal_cFos_4OHTAM  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX150/SRX150477   factor
hs_MammaryGland_MCF10AErSrc_normal_E2F4_4OHTAM  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX150/SRX150480   factor
hs_MammaryGland_MCF10AErSrc_normal_STAT3_4OHTAM ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX150/SRX150536   factor
hs_melanoma_A375_cancer_TFAP2C_n        ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX190/SRX190406   factor
hs_MultipleMyeloma_PlasmaCell_cancer_Cdk9_DMSO  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX203/SRX203391   factor
hs_NeuralCrest_hNCC_normal_NR2F1_fromH9hESC     ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX059/SRX059367   factor
hs_NeuralCrest_hNCC_normal_NR2F2_fromH9hESC     ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX059/SRX059368   factor
hs_NeuralCrest_hNCC_normal_TFAP2A_fromH9hESC    ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX131/SRX131914   factor
hs_neuroblastoma_SHSY5Y_cancer_GATA2_n  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX150/SRX150668   factor
hs_neuroblastoma_SHSY5Y_cancer_GATA3_n  ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX150/SRX150646   factor
hs_neuroblastoma_SKNSH_cancer_ELF1_n    ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX190/SRX190207   factor
hs_neuroblastoma_SKNSH_cancer_FOSL2_n   ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX190/SRX190296   factor
hs_neuroblastoma_SKNSH_cancer_FOXM1_n   ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX190/SRX190206   factor
```

SRA toolkit

Using the SRA Toolkit

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.

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...

Listfile with local sequencing samples

Either a fastq file must exist in the sample/fastq directory or a bam file in the bam directory
The script will use it and carry out the analysis

```
hs_hMQ_RSG_STAT6_CS082          factor
hs_hMQ_RSG_RXR_CS083            factor
hs_hMQ_RSG_pu1_CS084            factor
hs_hMQ_RSG_p300_CS085           factor
hs_HepG2_veh_IgG_CS087          control
hs_HepG2_veh_HNF4a_CS086                factor
hs_HepG2_veh_H3K27Ac_CS088              histone
hs_HepG2_veh_H3K4me3_CS089              histone
mm_DC_Sample1_NFKB_CS157                factor
mm_DC_Sample2_NFKB_CS158                factor
mm_DC_Sample3_NFKB_CS159                factor
mm_DC_Sample4_NFKB_CS160                factor
```

Processing local sequencing results

```
hs_hMQ_RSG_STAT6_CS082  AAAA
hs_hMQ_RSG_RXR_CS083    CTGC
hs_hMQ_RSG_pu1_CS084    GCTG
hs_hMQ_RSG_p300_CS085   TGCT
hs_HepG2_veh_HNF4a_CS086        ATTT
hs_HepG2_veh_IgG_CS087  CACG
hs_HepG2_veh_H3K27Ac_CS088      GGAC
hs_HepG2_veh_H3K4me3_CS089      TCGA
```

Demultiplexing:

```
#!/bin/sh
cd /molbio/illumina/fastq/Run11-130129 # lanes to process (lane 2 has TrueSeq barcodes)
for i in 1 3 4 5 6 7 8 ; do
        echo "processing $i at `date`"
        zcat /molbio/illumina/casava/Run11-130129/Project_Run11_ChIP-seq/Sample_Run11_lane${i}/*.fastq.gz |
                fastx_barcode_splitter.pl --bcfile /molbio/projects/illumina/Run11-130129/lists/lane${i}_barcodes.tab --mismatches 1 --bol \
        --suffix ".fastq" --prefix /molbio/illumina/fastq/Run11-130129/ \
        > /molbio/projects/illumina/Run11-130129/logs/barcode_splitter_Run11_lane${i}.log \
        2> /molbio/projects/illumina/Run11-130129/logs/barcode_splitter_Run11_lane${i}.err
        /bin/mv unmatched.fastq lane_${i}_unmatched.fastq
        gzip *.fastq
done > /molbio/projects/illumina/Run11-130129/logs/barcode_splitter.log 2> /molbio/projects/illumina/Run11-130129/logs/barcode_splitter.err
exit
```

```
[barta@ngsdeb Run11-130129]$ pwd
/molbio/projects/illumina/Run11-130129
[barta@ngsdeb Run11-130129]$ for i in 1 3 4 5 6 7 8 ; do for j in `awk '{print $1}' lists/lane${i}_barcodes.tab` ; do mkdir -p analysis/${j}/fastq ; ln -s ../..
/../fastq/${j}.fastq.gz analysis/${j}/fastq/${j}.fastq.gz ; echo -ne "$j\t\tfactor\n" ; done > lists/lane${i}.lst ; done
[barta@ngsdeb Run11-130129]$
```

Creating analysis directories and listfiles

```
#!$/bin/sh

for i in 1 3 4 5 6 7 8 ; do # lanes in the flowcell
        for j in `awk '{print $1}' lists/lane${i}_barcodes.tab` ; do # Sample names
                mkdir -p analysis/${j}/fastq # creating sample directory
                ln -s ../../../fastq/${j}.fastq.gz analysis/${j}/fastq/${j}.fastq.gz # creating symlink from the fastq sample dir to the common fastqdir
                echo -ne "$j\t\tfactor\n" # listfile entry
        done > lists/lane${i}.lst # writing out listfile entries per lanes
done
```

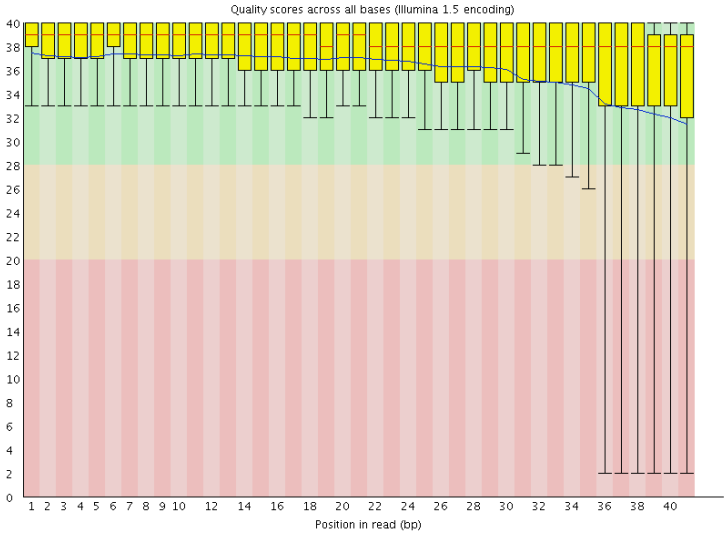
```
[barta@ngsdeb analysis]$ nohup /molbio/bin/ChIP-seq/ChIP-seq_anal-v1_9.sh ../lists/Run11_lane8.lst `pwd` /molbio/localdata/ChIP-seq/bigdir > logs/ChIP-seq_anal-
v1_9_vs_Run11_lane8.log 2> logs/ChIP-seq_anal-v1_9_vs_Run11_lane8.err &
[1] 11126
```

Running the analysis

Processing the raw sequence data

- Download the data in *fastq* format

```
@DARWIN_4090_FC634H2AAXX:5:1:1953:1186#0/1
ATTACCTTGGGTGACTTTGGAGACATTGACTTCTTCNANCC
+DARWIN_4090_FC634H2AAXX:5:1:1953:1186#0/1
ggggggggggcgfgffffcggfgggggcgggggg_B_B_a
```



- Check the quality with fastQC
  - Overall quality of the reads
  - Sequence composition
  - GC bias
  - Adaptor pollution
- Quality and adaptor trimming

Adaptor clipping

- `5' TCGTATGCCGTCTTCTGCTTG 3' - reverse strand NEB RT Primer`
- *fastx_clipper –a* TCGTATGCCGTCTTCTGCTTG

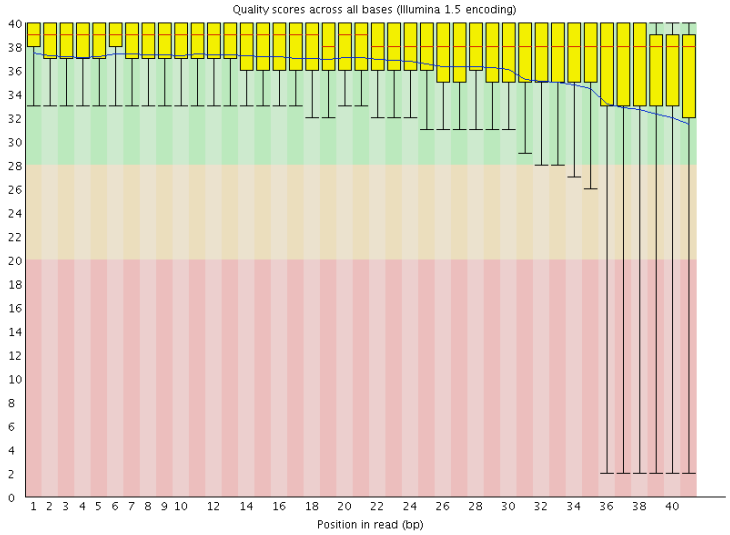
```
@DARWIN_4090_FC634H2AAXX:5:1:18775:1174#0/1
AGTAAGTAGGTGTTGTGTTTTCGTATGCCGNNNTNNNCNTN
                        TCGTATGCCGTCTTCTGCTTG
+DARWIN_4090_FC634H2AAXX:5:1:18775:1174#0/1
hdhhhdfhhdhehedgdhghhhfghdddebBBBBBBBBBBB
```

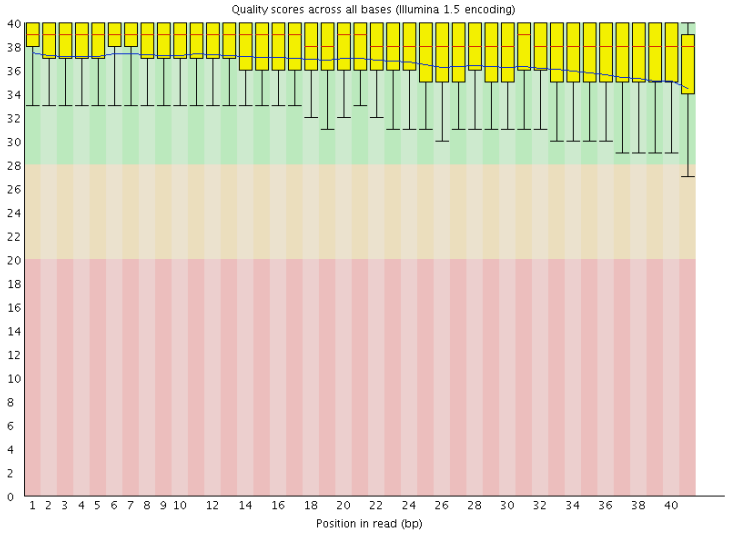
```
@DARWIN_4090_FC634H2AAXX:5:1:18775:1174#0/1
AGTAAGTAGGTGTTGTGTTT
+DARWIN_4090_FC634H2AAXX:5:1:18775:1174#0/1
hdhhhdfhhdhehedgdhgh
```

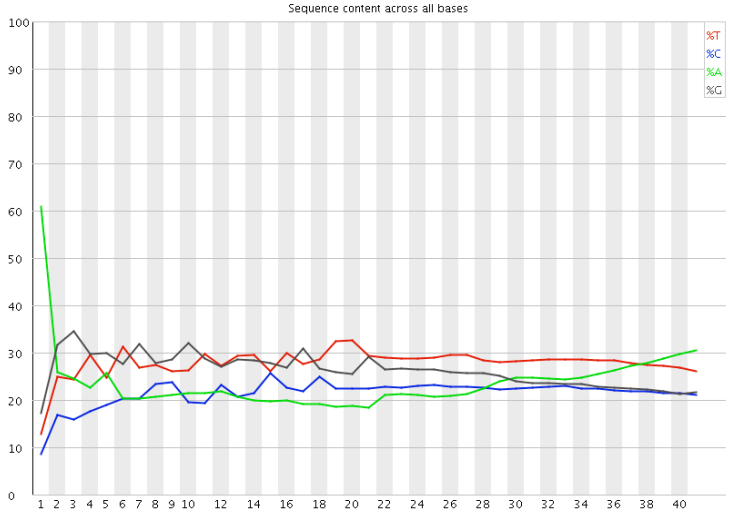
DARWIN_4090_FC634H2AAXX:5:1:18775:1174#0 16 chr19 41311146
37 20M * 0 0
AAACACAACACCTACTTACT hghdgdehehdhhfdhhhdh

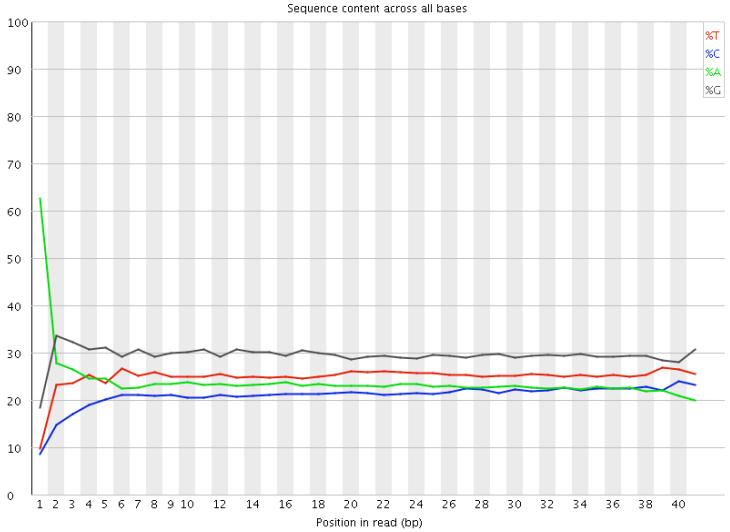
Samples	all reads	mapped reads	Mapped reads %	reads after clipping	mapped reads after clipping	Mapped reads % after clipping	gain in mapped reads	discarded too short read	discarded adapter-only reads
mm_LG268-0-1	32 350 224	12 969 138	40.1	23 529 444	20 364 197	86.54	7 395 059	7 759 743	1 061 037

Quality checking with fastQC before and after clipping









Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAAAAAA	40064	0.12384458296177485	Illumina Single End Adapter 2 (100% over 21bp)

Aligning reads to the reference genome

- BWA program (needs to be downloaded and compiled on a UNIX machine)
  - BWA aln (-> sai internal format)
  - BWA samse (sam multiple alignment format)

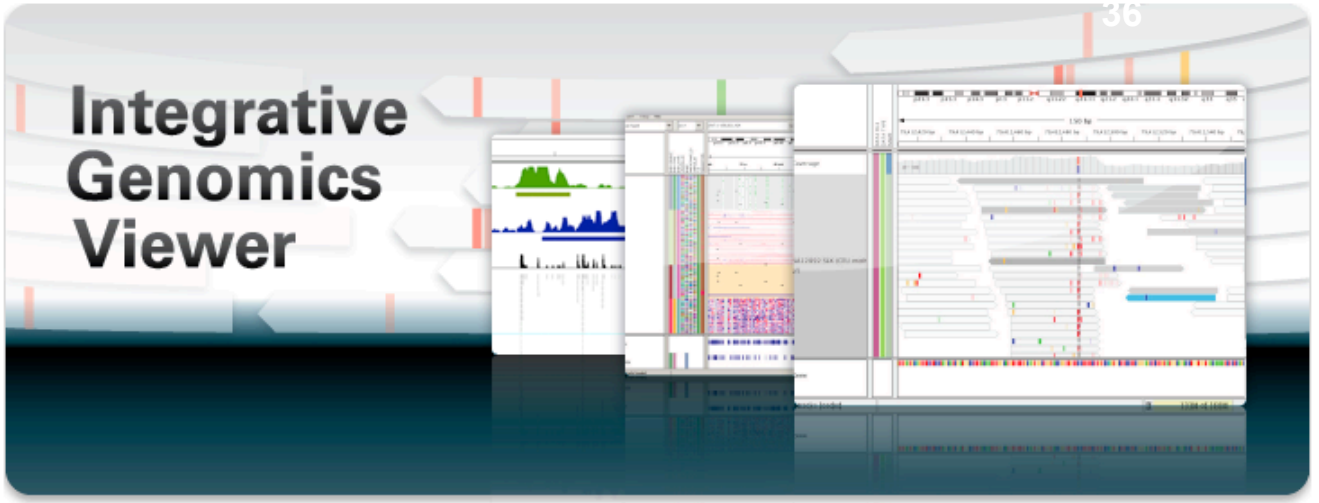

```
if [ ! -f  ${basedir}/${name}/bam/${name}.bam ] ; then
        echo "Indexing to sai for $name at `date`" >> ${logdir}/${name}_ChIP-seq_analyze.log
        cd ${basedir}/${name}/sai
        ( nice -n 20 bwa aln -t $procno -B $barcode_length ${indexdir}/$genome ${basedir}/${name}/fastq/${name}.fastq.gz > ${name}.sai ) >  ${logdir}/${name}-sa
i.log 2> ${logdir}/${name}-sai.err
        echo "finished aligning fastq for $genome for $name at `date`" >> ${logdir}/${name}_ChIP-seq_analyze.log
        echo "------------------------------------------------------------" >> ${logdir}/${name}_ChIP-seq_analyze.log
        # finished fastq alignment starting generating sam files
        echo "Starting converting sai files into sam/bam format" >> ${logdir}/${name}_ChIP-seq_analyze.log
        cd ${basedir}/${name}/bam
        bwa samse ${indexdir}/$genome ${basedir}/${name}/sai/${name}.sai ${basedir}/${name}/fastq/${name}.fastq.gz |
        samtools view -buS -t ${indexdir}/$genome.fai - |
        samtools sort -m 1000000000 - $name
        samtools index ${name}.bam ${name}.bai
        echo "finished converting sai to sam and bam finally on pipeline for $name at `date`" >> ${logdir}/${name}_ChIP-seq_analyze.log
        echo "------------------------------------------------------------" >> ${logdir}/${name}_ChIP-seq_analyze.log
fi
```

SAM file format

SAM= TEXT file, BAM= SAM file compressed and indexed (binary) format

```
@SQ     SN:chr9_random  LN:449403
@SQ     SN:chrM LN:16299
@SQ     SN:chrUn_random LN:5900358
@SQ     SN:chrX LN:166650296
@SQ     SN:chrX_random  LN:1785075
@SQ     SN:chrY LN:15902555
@SQ     SN:chrY_random  LN:58682461
HWI-EAS038:6:1:23:122#0 4       *       0       0       *       *       0       0       TAGCCTTGATGTTTACCTATTGTATCAAAGGGC        OJYMXLTPKOPOXYBBBBBBBBBBBBBBBBBB
B
HWI-EAS038:6:1:25:283#0 0       chr14   27002726        0       33M     *       0       0       AGAGACCCAGGAAATTGAAGTCAGAGCAGTTAG        abaa_Z_`X]PW^BBBBBBBBBB
BBBBBBBB        XT:A:R  NM:i:1  X0:i:3  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:10T22
HWI-EAS038:6:1:26:649#0 0       chr9    27884899        37      33M     *       0       0       CCTTTCTTTTGTCTACTCCTTTCCTCTTGGTAT        abbaabbbbbb```aZ\`a\aa[]
_QWaa`YXS       XT:A:U  NM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:33
HWI-EAS038:6:1:30:918#0 16      chr17   95265601        0       33M     *       0       0       GTGTTTATCAGTCCCAAGGCCACTAGAGGCTTG        BBBBBBBBBBBBBBB[[`\aaZaa
__aaaa`a`       XT:A:R  NM:i:2  X0:i:3  X1:i:0  XM:i:2  XO:i:0  XG:i:0  MD:Z:3G8T20
HWI-EAS038:6:1:32:1507#0        16      chr13   57505480        37      33M     *       0       0       CGGAGCTGGTGGTAGACATTGTGTGCTGCCTAG        \Z]W_[`]ZH]^^Z^T
`bbab_[W\_]bb_W_b       XT:A:U  NM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:33
HWI-EAS038:6:1:32:298#0 4       *       0       0       *       *       0       0       TATAATAAAAATGACATTTTATTAAATACGCCT        `^aaa__\]^[SBBBBBBBBBBBBBBBBBB
B
HWI-EAS038:6:1:32:1938#0        0       chr7    65636851        37      33M     *       0       0       TTTATATTTCTCCCCTTATCATTCCATTTTTTT        ]aa^^X\`YQ\Y[^UY
ZHMHXWZEVFO][BBBB       XT:A:U  NM:i:1  X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:31G1
HWI-EAS038:6:1:32:861#0 4       *       0       0       *       *       0       0       TGCATTCTAAGTTGGTTTAATATAAATCAACAT        ]bUSJGKHWaK_\BBBBBBBBBBBBBBBBBB
B
HWI-EAS038:6:1:32:1814#0        0       chr2    98506740        0       33M     *       0       0       CCACTTGACGACTTCAAAAATGACGAAATCACT        W^R^X`]Z]a]XZ]aZ
W]PYVV\YRW[SUZSST       XT:A:R  NM:i:1  X0:i:12 X1:i:44 XM:i:1  XO:i:0  XG:i:0  MD:Z:14G18
HWI-EAS038:6:1:34:2002#0        0       chr10   97252408        37      33M     *       0       0       CCTAGATTCCTTAGGGTATAAAAGGAGGAGAGC        _a`_ba]_Oa]aV[`a
OHDT^_BBBBBBBBBB        XT:A:U  NM:i:1  X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:29T3
HWI-EAS038:6:1:37:667#0 0       chrX    90652654        37      33M     *       0       0       CAAGTCCAAAAATTCCTTGAAAAATTTCACAAT        Y`_TOMPT^^_[PUWOJQLQQYW]
BBBBBBBB        XT:A:U  NM:i:1  X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:19C13
HWI-EAS038:6:1:37:1236#0        4       *       0       0       *       *       0       0       ATGATTTCTTGTTGTGTATCACTATTCTAGGGG        _Q\LYBBBBBBBBBBBBBBBBBB
BBBBBBBB
HWI-EAS038:6:1:37:262#0 16      chr2    3386587 23      33M     *       0       0       TCTAGTACCCACATGGTGCAAGGAGAGAACCAA        BB]Z[LFTXX]TZYQRXHJUOISU\X]_[UO]
a       XT:A:U  NM:i:1  X0:i:1  X1:i:1  XM:i:1  XO:i:0  XG:i:0  MD:Z:6C26
HWI-EAS038:6:1:38:385#0 0       chr9    35113013        25      33M     *       0       0       AAAAAACGTGAAAAATAAGAAATGCCAACTGAA        [aa``_]PTUUZY[_[R]BBBBBB
BBBBBBBB        XT:A:U  NM:i:2  X0:i:1  X1:i:0  XM:i:2  XO:i:0  XG:i:0  MD:Z:16G9C6
HWI-EAS038:6:1:38:37#0  16      chr16   49998240        37      33M     *       0       0       ATTTGTCTGTGATGATTTTCTGTTCTTTCAATG        B[_XHJJJTMPWWNR__`]__Wa^
`^`R`]a_a       XT:A:U  NM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:33
HWI-EAS038:6:1:40:991#0 16      chr13   75619559        0       33M     *       0       0       TTTAATATTCTATCTTTATTTAGTGCATTTGTT        a_ZQPX`__[`RY[]\]PVT]\\]
WOU[]`V_^       XT:A:R  NM:i:0  X0:i:6619       XM:i:0  XO:i:0  XG:i:0  MD:Z:33
HWI-EAS038:6:1:40:767#0 0       chr11   34713793        25      33M     *       0       0       TAACTTATTCCTTTAGGTCCTGTGTTTTCTATT        aaaO`]aQYBBBBBBBBBBBBBBB
```

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

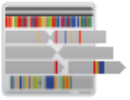


What's New

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

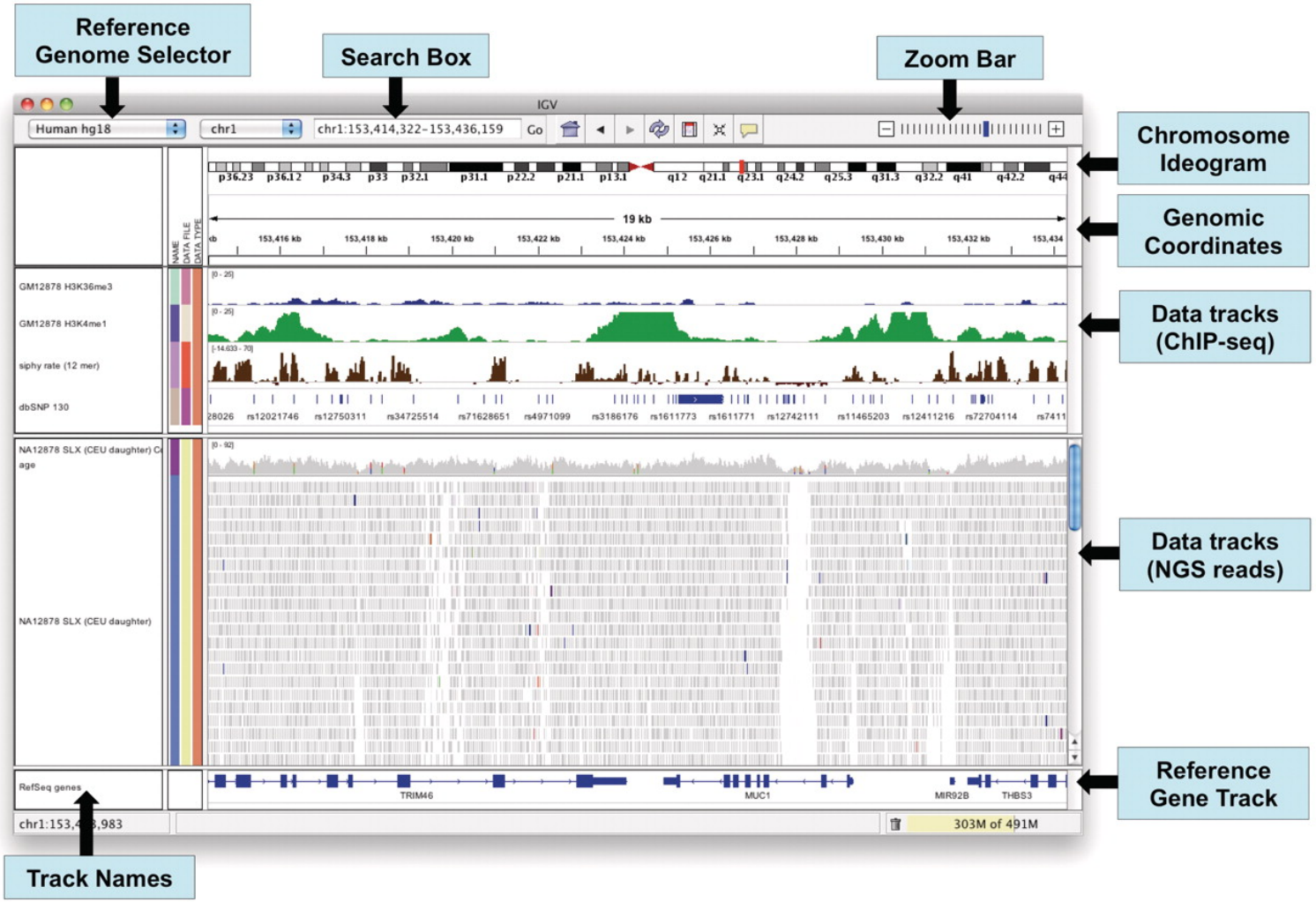
Funding

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.

GENOMESPACE

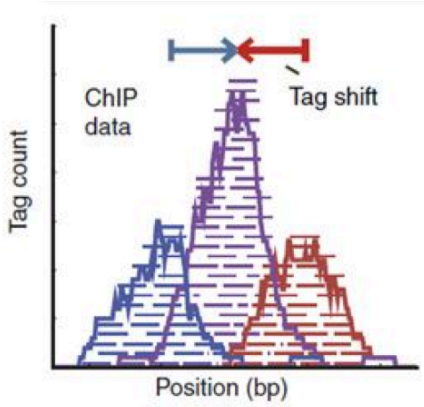
The IGV application window



Thorvaldsdóttir H et al. Brief Bioinform 2012;bib.bbs017

Briefings in Bioinformatics

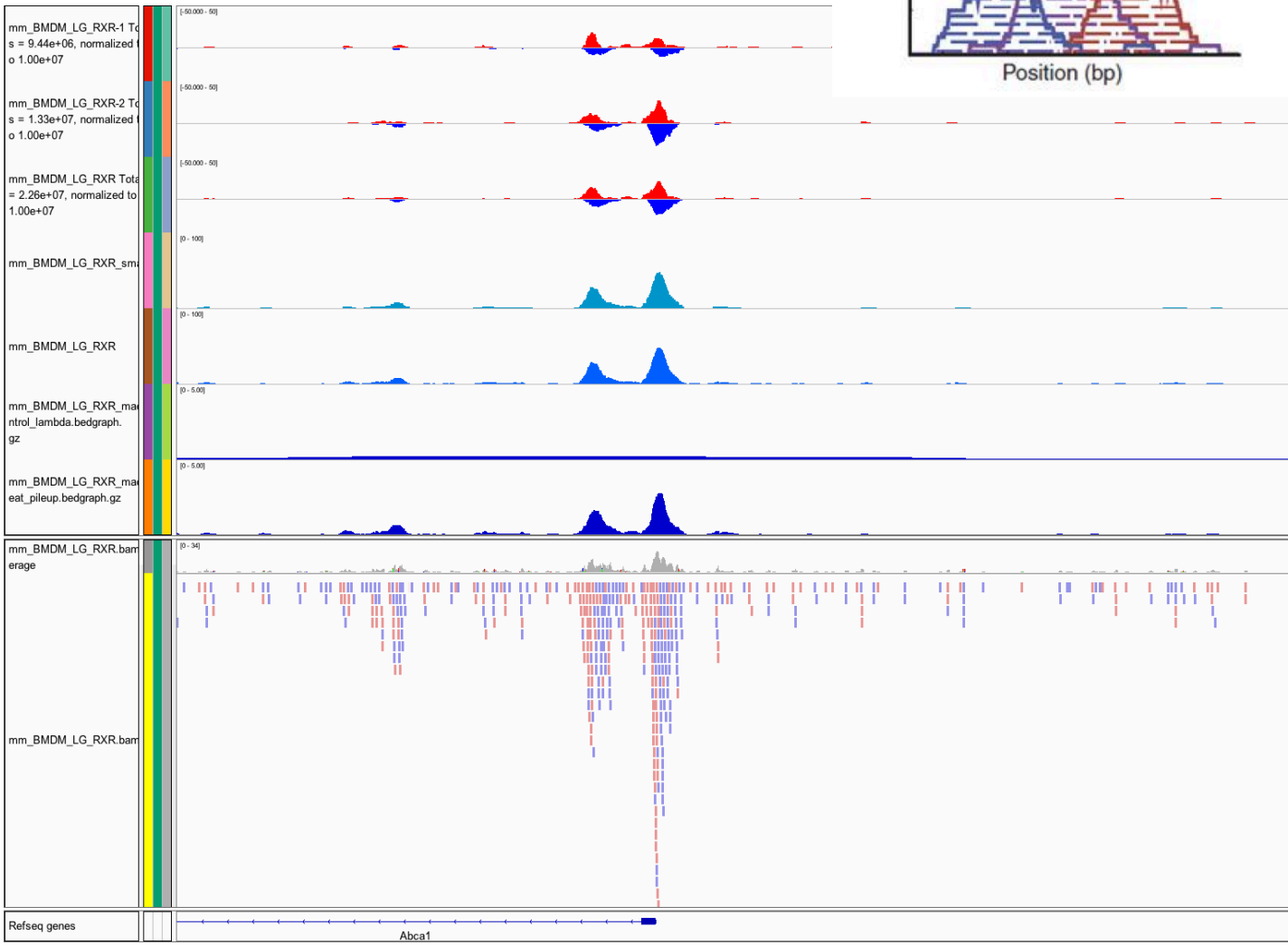
Bedgraph and BAM files



Homer2*
- Coverage-1
- Coverage-2
- Merge
  - Strand specific
  - Small
  - Big

MACS2 (merge)
- Background
- Coverage

Coverage Merge BAM

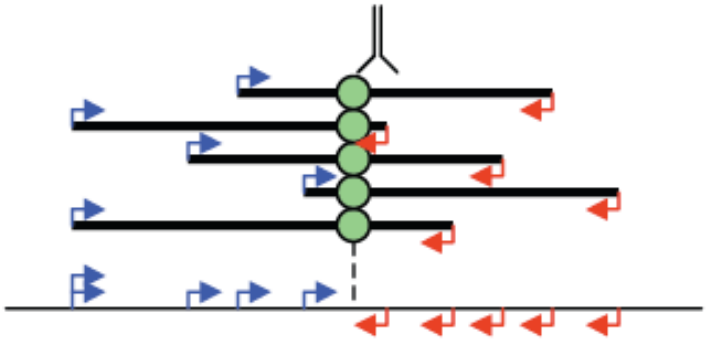


*normalized to 10 million reads
no GC normalization

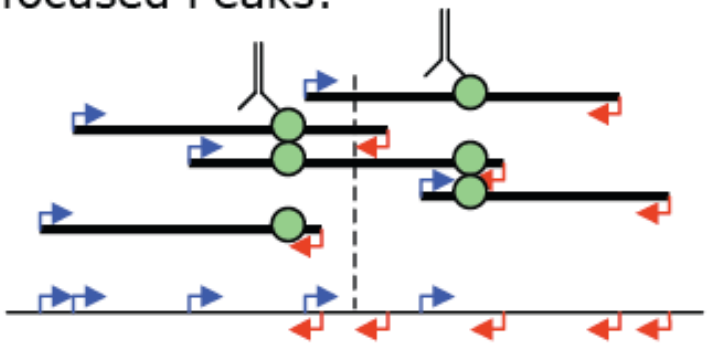
Finding peaks

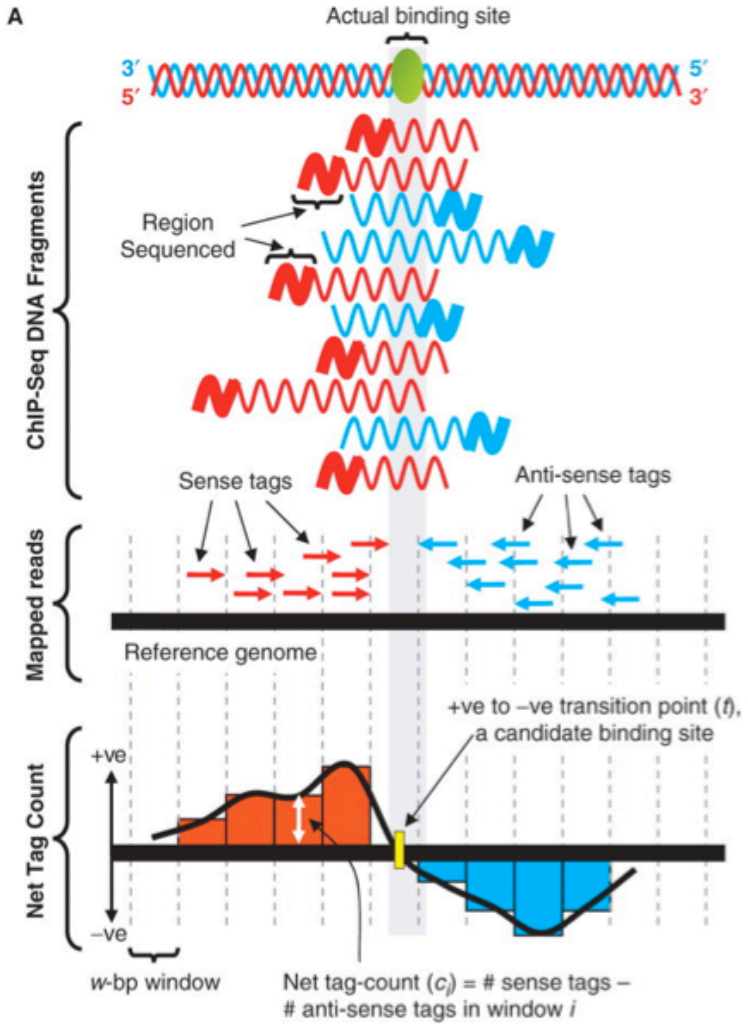
- The number of peaks depends on the methods used and the cutoff values applied.
- More reads doesn't mean necessarily more peaks
- Different methods give only 60-80% similar peaks!

Focused Peaks:



Unfocused Peaks:





Jothi R et al. Nucl. Acids Res. 2008;36:5221-5231

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

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:
```

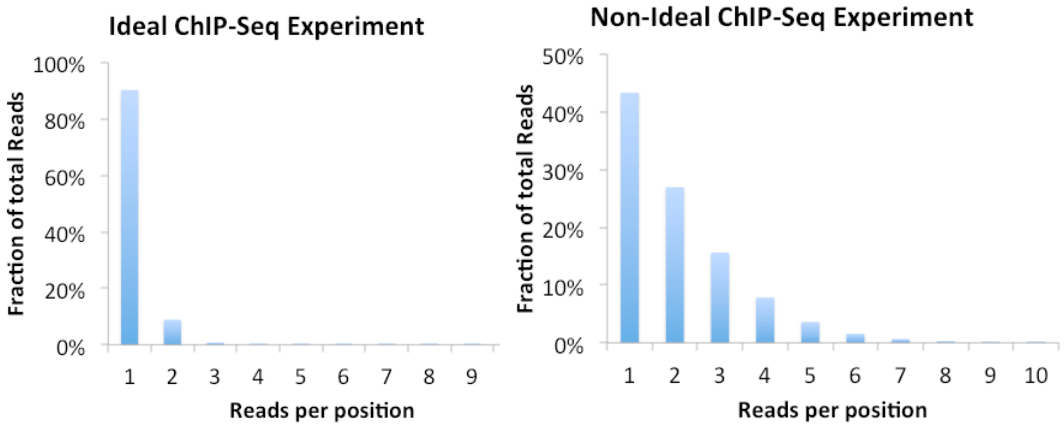
- Column 1: PeakID - a unique name for each peak (very important that peaks have unique names...)
- Column 2: chr - chromosome where peak is located
- Column 3: starting position of peak
- Column 4: ending position of peak
- Column 5: Strand (+/-)
- Column 6: Normalized Tag Counts - number of tags found 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

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

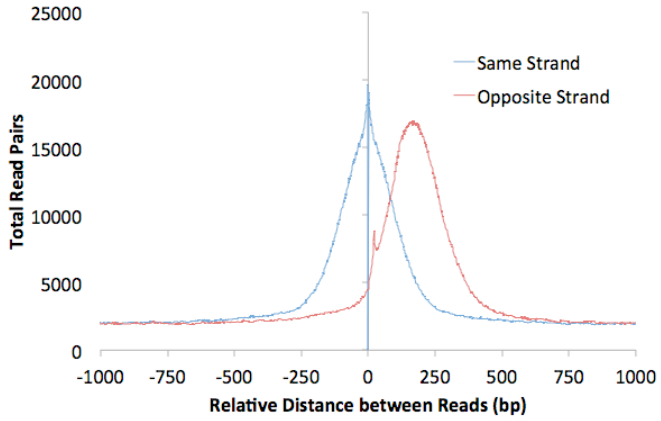
Approximate IP effeciency describes the fraction of tags found in peaks versus. genomic background. This provides an estimate of how well the ChIP worked. Certain antibodies like H3K4me3, ERa, or PU.1 will yield very high IP efficiencies (>20%), while most rand in the 1-20% range. Once this number dips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

Quality control of the ChIP-seq experiments

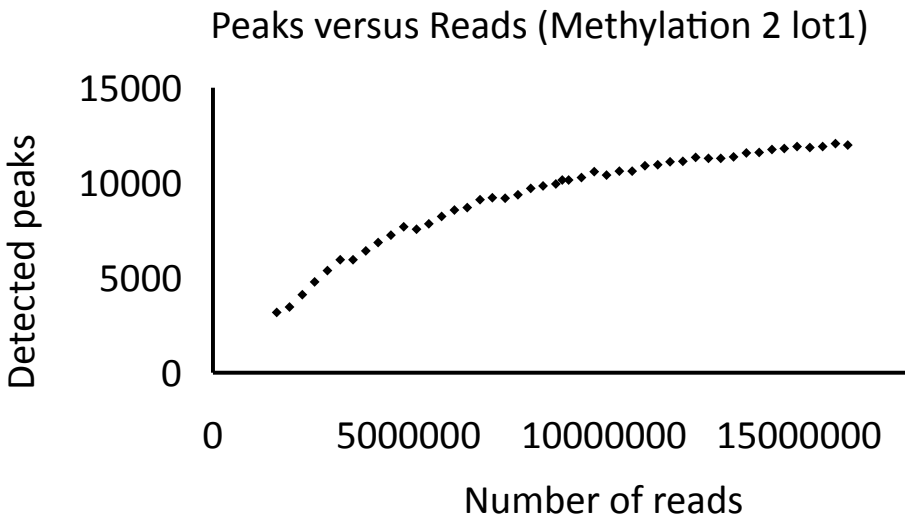
- Clonal tag distribution

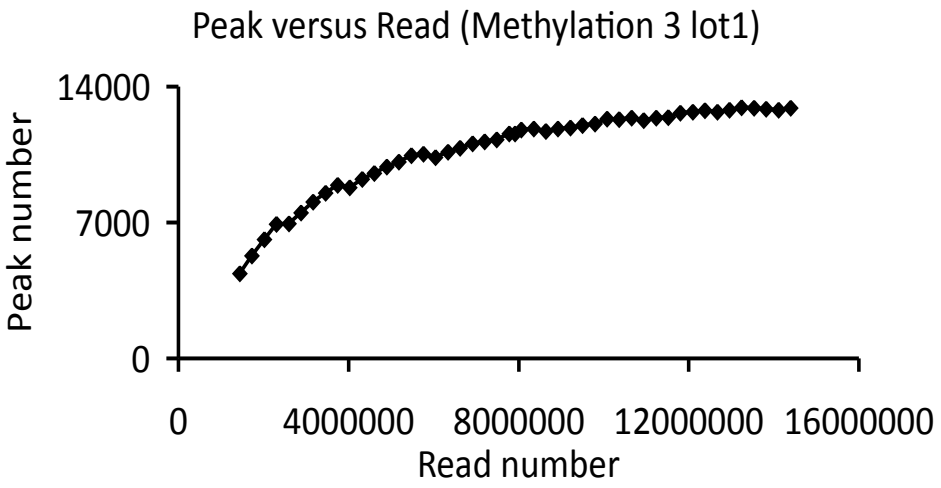


- Autocorrelation analysis



The Peak Read correlation is asymptotic to a horizontal line



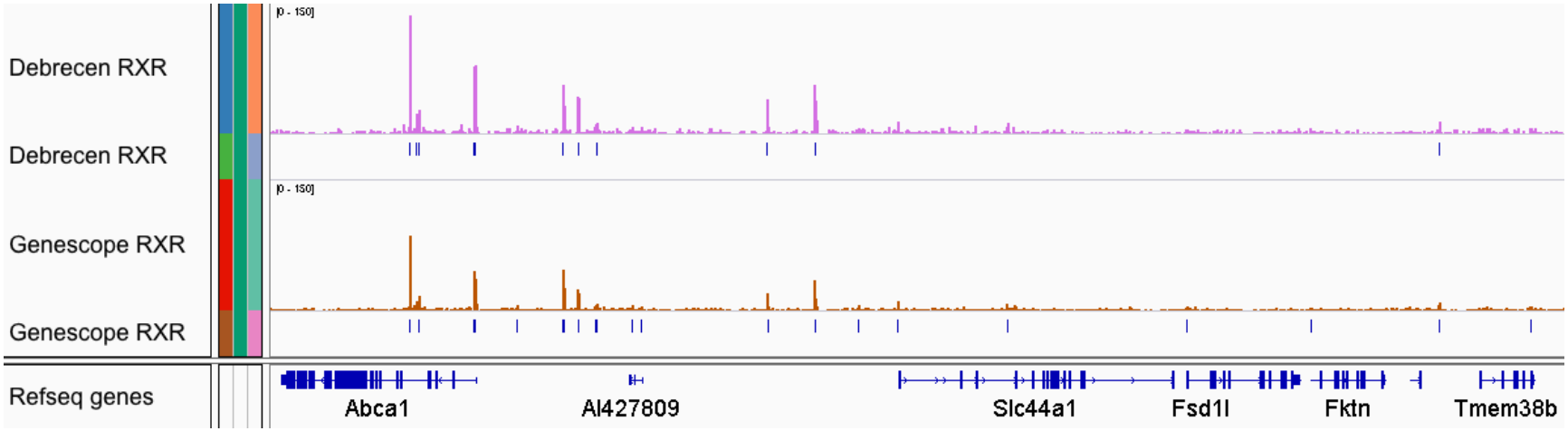


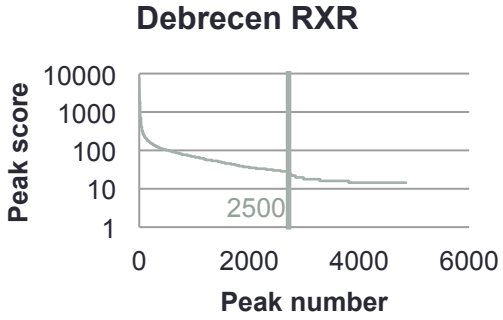
1. Random removal of reads.
2. Repeat seven times.
3. Reanalyse by MACS.
4. Plot.

100% of reads visualized

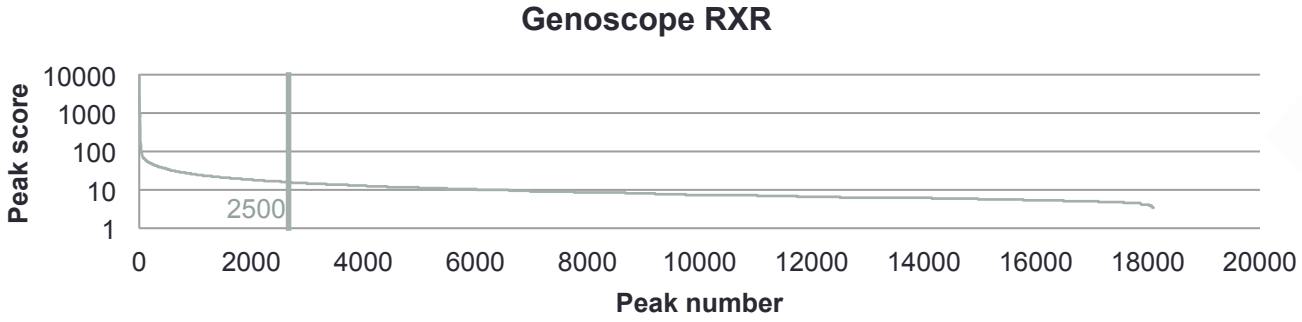
50% of reads visualized

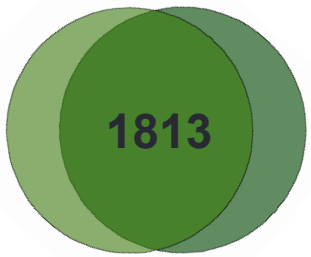
How does the read number affect the peak number





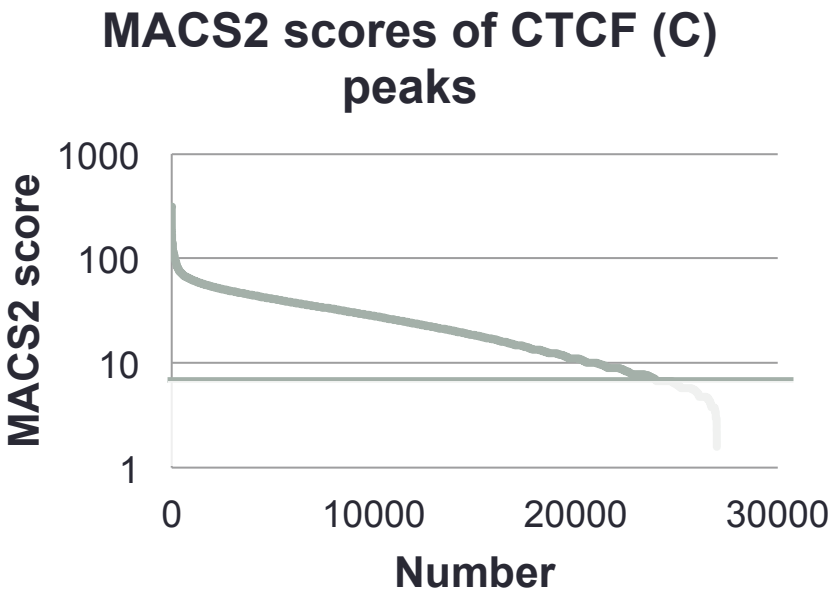
Experiment	No of reads in fastq			Total homer peaks		IP efficiency	
	Genoscope	Debrecen	Fold	Genoscope	Debrecen	Genoscope	Debrecen
mm_BMDM_RXR	60438603	10234482	5.91	18099	4852	2.57	2.60

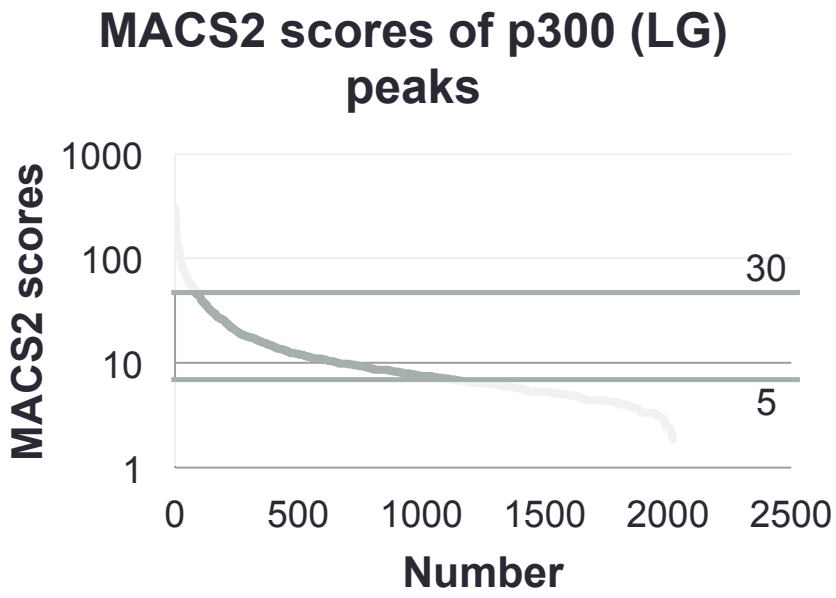




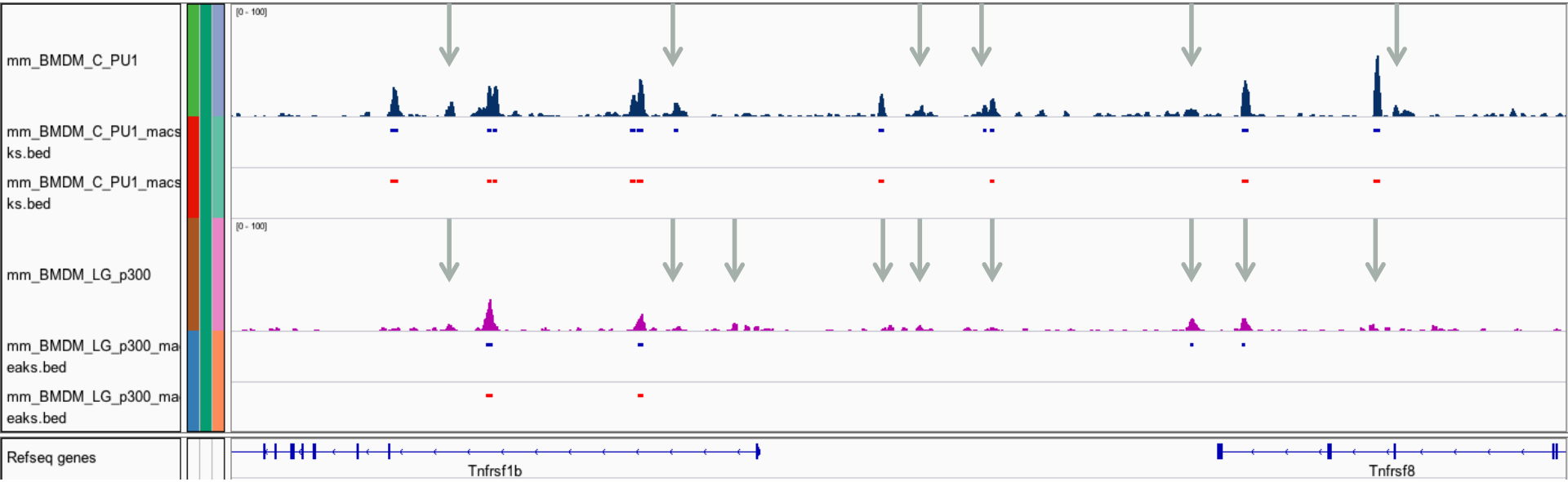
Which are the real peaks?

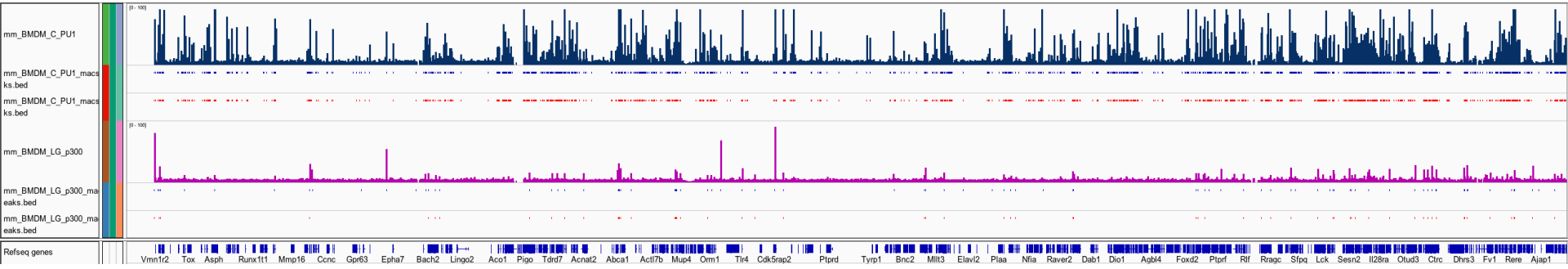
It depends on the type of the ChIP and the number of the reads



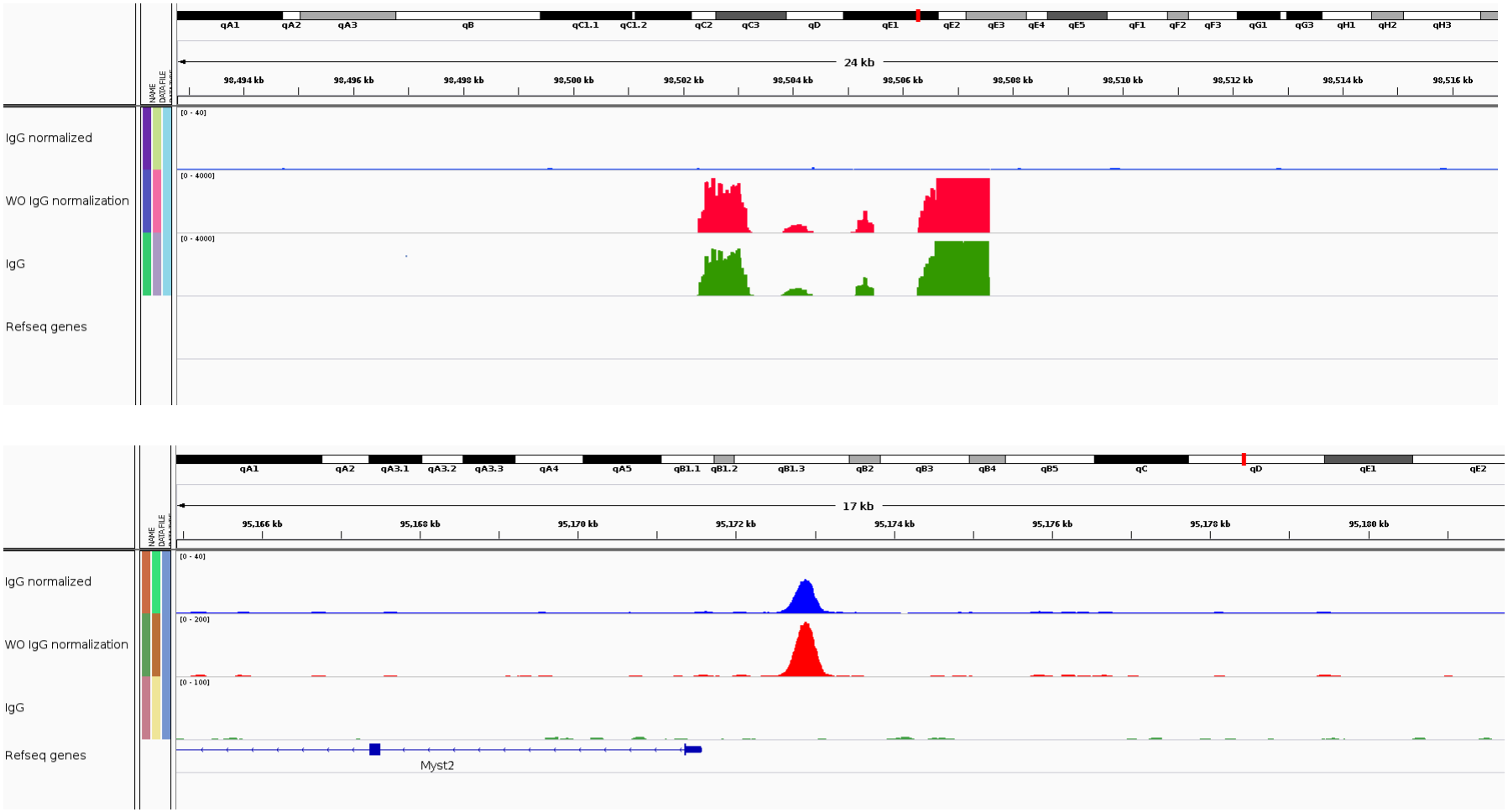


How to define biologically meaningful peaks?

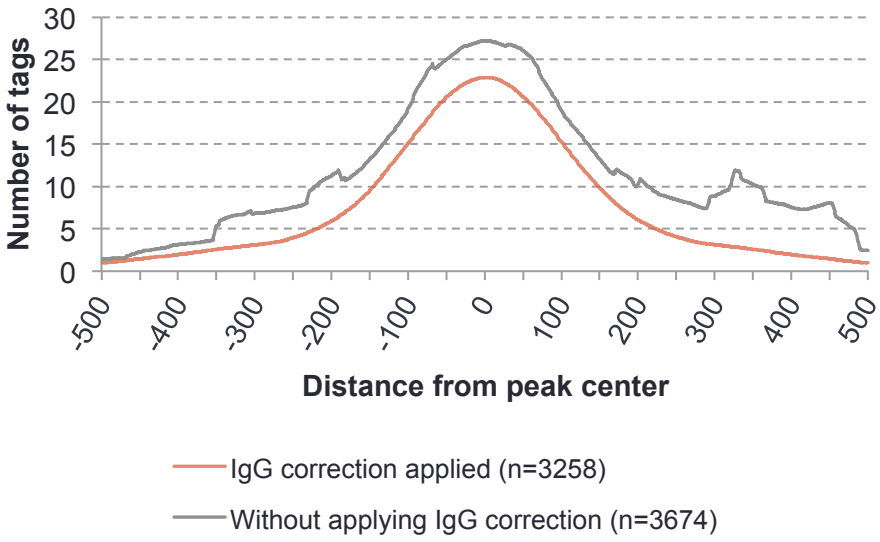




The effect of IgG control on the predicted peaks



The effect of IgG control on the predicted peaks



- Mouse embrionic stem cell
- RXR antibody
- No treatment applied
- 1006 peaks removed

Nr. of reads in fastq	7 825 951	Fragment length (bp)	189
Mean quality	37.15	Nr. of peaks	3 674
Mapped reads (%)	94.97	IP efficiency (%)	5.64

Annotation of the peaks (annotatePeaks.pl)

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

```
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
        5UTR    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
```

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)**

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

Wikipathways enrichment in mouse PPARg adypocyte ChIPs

GO	Term	P-value	LogP	Number in Term	Number in common	Fraction of List	Total List	Total Genes	Common Gene IDs
TNF-alpha-NF-kB_Signaling_Pathway_WP231	TNF-alpha-NF-kB_Signaling_Pathway_WP231	1.59887866489455e-17	-38.6746440317932	211	192	0.0795360397680199	2414	3620	20019,19766,22628,23997,6789 86,56550,56532,18035,56399,1 54,192656,26410,12366,13205, 032,21934,17846,12369,320795
EGFR1_Signaling_Pathway_WP437	EGFR1_Signaling_Pathway_WP437	7.61213846424944e-15	-32.5090322553835	216	192	0.0795360397680199	2414	3620	19766,20416,13710,23938,1712 449,20111,234779,16452,16396 4,215114,11908,18750,18803,2 48,15170,110651,216190,12402
mRNA_processing_WP411	mRNA_processing_WP411	1.26133962094022e-12	-27.3988467685303	126	117	0.0484672742336371	2414	3620	18789,22608,108014,54451,561 9,20637,107701,19134,64340,2 6,108013,20222,102414,70767,
B_Cell_Receptor_Signaling_Pathway_WP23	B_Cell_Receptor_Signaling_Pathway_WP23	2.81043746301287e-12	-26.5976809639124	201	176	0.0729080364540182	2414	3620	19697,12566,20416,76709,2641 68089,20111,18718,101476,234 12507,17973,12929,76108,1924
TGF-beta_Receptor_Signaling_Pathway_WP366	TGF-beta_Receptor_Signaling_Pathway_WP366	7.96448897035415e-11	-23.2534432410339	232	197	0.0816072908036454	2414	3620	218210,227720,12566,12325,17 14282,56458,21814,108058,433 5901,19650,17126,56440,12393
IL-5_Signaling_Pathway_WP127	IL-5_Signaling_Pathway_WP127	1.85062189501834e-09	-20.1077440948757	83	78	0.0323115161557581	2414	3620	11689,22631,11651,20416,6064 5894,20850,20848,16451,26396
MAPK_signaling_pathway_WP382	MAPK_signaling_pathway_WP382	4.38548736704107e-09	-19.2449650731612	162	140	0.0579950289975145	2414	3620	235584,19060,12475,11911,239 19,14103,50932,63953,18211,1 03,240672,22059,16478,19259,
IL-6_Signaling_Pathway_WP364	IL-6_Signaling_Pathway_WP364	3.01227479364653e-08	-17.3179852052258	115	102	0.0422535211267606	2414	3620	73699,11651,20416,51792,1438 4,20851,18099,21939,16452,22
IL-3_Signaling_Pathway_WP286	IL-3_Signaling_Pathway_WP286	4.49775079077704e-08	-16.9171032964016	109	97	0.0401822700911135	2414	3620	22631,11651,20416,14694,1438 479,19303,15461,17179,18708,
IL-1_Signaling_Pathway_WP195	IL-1_Signaling_Pathway_WP195	6.1353992146724e-08	-16.6066055962403	73	68	0.0281690140084507	2414	3620	19697,11651,16177,108960,672
Senescence_and_Autophagy_WP615	Senescence_and_Autophagy_WP615	1.00857209138904e-07	-16.1095600913241	120	105	0.0434962717481359	2414	3620	228361,56717,26416,54673,117 16784,67605,16476,56637,1242

Gene Ontology enrichment analysis

- Peaks can be assigned to genes (not always to the right genes)
- This results in a gene list (like at the microarray analysis)
- The gene lists can be statistically analyzed against other gene list.
- We must usually consider four numbers
  - The number of the whole gene set (usually the gene number)
  - The number of the genes in the GO list
  - The number of genes in the sample
  - The number of the genes in the sample list, which can be found in the GO list
  - HOMER can do GO analysis against different ontologies

HOMER GO analysis (sorted by P-value)

- **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)

P-value	ln(P)	Term	GO Tree
7.024e-43	-97.06	cytoplasm	cellular component
8.437e-40	-89.97	intracellular	cellular component
1.916e-38	-86.85	Transport_of_inorganic_cations-anions_and_amino_acids-oligopeptides_WP1936	wikipathways

# of Genes in Term	# of Target Genes in Term	# of Total Genes	# of Target Genes	Common Genes
7996	1364	19974	2602	Hmbox1,Tomt,Arpc2,Cfh,Raver2,Rybp,Mgea5,Itm2b,Rap1b,Atxn2,Rhob,Mid1ip1,Rbpj,Vp
10855	1723	19974	2602	Hmbox1,Tcfec,Tomt,Wsb1,Arpc2,Cfh,Raver2,Dhx8,Rybp,Mgea5,Itm2b,Med13l,Rap1b,Atx
111	85	4544	914	Slc43a2,Slc38a2,Slc24a6,Slc17a5,Slc6a12,Slc38a1,Slc36a1,Slc9a9,Slc6a6,Ssu72,Dnajc15

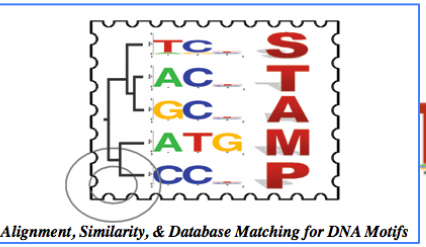
denovo motif finding

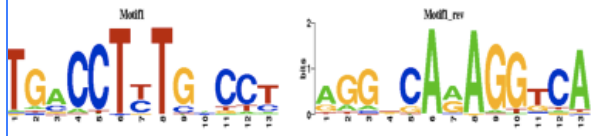
- Aim: Find motifs enriched in ChIP-seq peaks (and different peak subsets)
- Several algorithms exist (MEME, Gibbs sampler etc.)
- Homer works well with ChIP-seq regions
  - It uses genomic intervals (bedfiles) as an input
  - It selects at least the same number of random genomic region with the same sizes
  - It masks out repetitive regions
  - It throws away regions with too many Ns (masked bases)

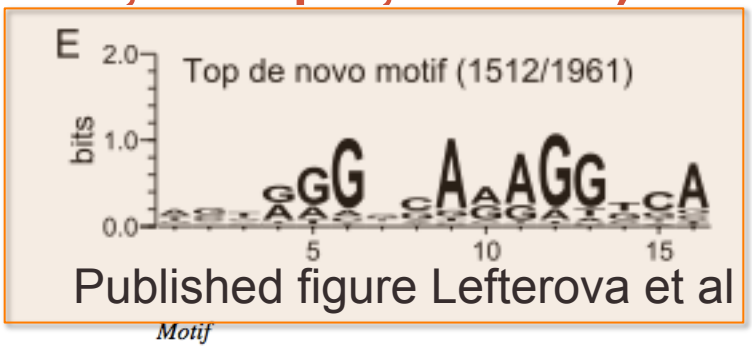
Top peaks give better motif enrichment

	Peaks	w motif	p-value	target %	bg %	fold
PU.1	10883	3266	1E-1274	30.01	4.82	6.23
	5206	1632	1E-643	31.34	5.01	6.26
	1000	442	1E-169	44.24	8.28	5.34
NRhalf	10883	3544	1E-773	32.56	9.52	3.42
	5206	2436	1E-630	46.79	12.77	3.66
	1000	592	1E-258	59.23	10.39	5.70

De novo motif finding (MEME, zoops, w=11)





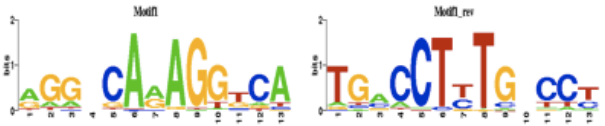


Published figure Lefterova et al

Name	*E value*	*Alignment*	*Motif*
PPARgamma-RXRalpha_M00512	0.0000e+00	---AGGNCARAGGTCA---- NNGRGGTCAAAGGTCANNNN	

Macrophage
521/1961
min: 500,
max: 1500

Motif1



Adypocyte
1622/2634
min:1500
max: 2000

Name	*E value*	*Alignment*	*Motif*
HNF4alpha1_M00411	1.1102e-16	-TGMCCTYTGNCCY NTGACCTTTGNCCY	

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-pvalue	% 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 known motif enrichment analysis

Total Target Sequences = 4219, Total Background Sequences = 45714

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif
1		PU.1(ETS)/ThioMac-PU.1-ChIP-Seq/Homer	1e-489	-1.126e+03	0.0000	2155.0	51.08%	8521.0	18.64%
2		ETS1(ETS)/Jurkat-ETS1-ChIP-Seq/Homer	1e-299	-6.890e+02	0.0000	2703.0	64.07%	16437.6	35.96%
16		CEBP(bZIP)/CEBPb-ChIP-Seq/Homer	1e-61	-1.417e+02	0.0000	1232.0	29.20%	8513.5	18.63%
17		RUNX(Runt)/HPC7-Runx1-ChIP-Seq/Homer	1e-58	-1.348e+02	0.0000	1539.0	36.48%	11517.8	25.20%
18		Esrrb(NR)/mES-Esrrb-ChIP-Seq/Homer	1e-52	-1.198e+02	0.0000	1450.0	34.37%	10935.1	23.92%
19		NF-E2(bZIP)/K562-NFE2-ChIP-Seq/Homer	1e-51	-1.185e+02	0.0000	236.0	5.59%	800.1	1.75%
20		PPARE(NR/DR1)/3T3L1-Pparg-ChIP-Seq/Homer	1e-51	-1.177e+02	0.0000	1931.0	45.77%	15759.0	34.48%
21		RXR(NR/DR1)/3T3L1-RXR-ChIP-Seq/Homer	1e-50	-1.170e+02	0.0000	2149.0	50.94%	18035.2	39.46%

- Enrichment = % of targets sequences with Motif / % of Background sequences with motif

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)

Outputs of the primary analysis III/a.

Experiment	No of reads in fastq	Percent duplicated fastq reads	mean quality	Percent As	Percent Cs	Percent Gs	Percent Ts	Percent Ns	No of total reads in BAM	Percent reads mapped in BAM	HOMER total tags
hs_MQ_minus_H3K4me	3528885	1.41	33.98	24.61	23.03	26.33	25.89	0.148	3528885	68.28%	2165571
hs_MQ_minus_STAT6	3939723	2.15	34.89	28.44	26.52	19.36	25.56	0.117	3939723	65.34%	2196028
hs_MQ_plus_H3K4me	5926449	0.91	34.38	26.82	20.97	23.69	28.37	0.159	5926449	86.19%	4538166
hs_MQ_plus_IgG	3301440	2.00	34.11	27.41	20.73	26.63	25.09	0.14	3301440	37.29%	1047187
hs_MQ_plus_STAT6	2584166	2.03	33.61	27.80	22.46	21.72	27.87	0.148	2584166	46.92%	1029179
hs_MQ_RSGIL4_p300_2	1173753	1.71	34.40	28.08	27.36	18.52	25.89	0.137	1173753	87.25%	881711
hs_MQ_RSGIL4_p300	4226253	0.98	34.88	28.87	26.58	19.13	25.28	0.135	4226253	83.26%	3053184
hs_MQ_RSGIL4_pU1b	2371412	2.08	34.70	28.74	19.18	23.59	28.36	0.123	2371412	90.03%	1848573
hs_MQ_RSGIL4_pU1	1448718	1.93	34.69	26.33	21.43	24.53	27.61	0.098	1448718	68.94%	882943
hs_MQ_RSGIL4_RXRb	8322804	1.28	35.09	28.37	21.81	20.73	28.96	0.132	8322804	95.94%	6934437
hs_MQ_RSGIL4_RXR	5708685	0.94	35.21	27.47	23.00	21.81	27.56	0.159	5708685	71.27%	3589040
hs_MQ_RSGIL4_STAT6	6352861	1.06	34.71	29.61	18.53	24.84	26.92	0.108	6352861	95.31%	5258963
hs_MQ_veh_p300_2	7562405	0.96	34.55	27.84	22.58	18.42	31.03	0.128	7562405	90.20%	5901473
hs_MQ_veh_p300	4616160	1.43	35.05	26.75	23.55	19.22	30.33	0.15	4616160	70.19%	2815536
hs_MQ_veh_pU1b	1558729	1.70	33.92	28.46	22.04	20.98	28.39	0.129	1558729	81.10%	1091559
hs_MQ_veh_pU1	3840460	0.86	34.62	27.60	22.80	21.10	28.37	0.122	3840460	78.24%	2650740
hs_MQ_veh_RXRb	6613994	1.59	35.16	29.32	23.75	18.18	28.64	0.12	6613994	95.97%	5497832
hs_MQ_veh_RXR	8810756	1.24	35.47	27.57	25.24	19.66	27.36	0.167	8810756	80.00%	6228613
hs_MQ_veh_STAT6	6462278	1.40	34.83	29.19	18.61	20.66	31.43	0.117	6462278	95.61%	5352564
mm_BMDM_ATRA1_p300	8728252	8.77	36.37	28.71	26.82	18.38	25.96	0.124	8728252	96.19%	6594828
mm_BMDM_ATRA1_RXR	4050923	6.40	34.77	28.93	19.12	25.11	26.74	0.099	4050923	91.19%	2936993
mm_BMDM_ATRA_p300	9323732	7.73	36.55	28.92	26.68	18.45	25.78	0.162	9323732	96.24%	7144688
mm_BMDM_ATRA_RXR	1347034	5.62	33.84	28.77	19.29	25.90	25.92	0.119	1347034	86.07%	932321
mm_BMDM_GW39651_p300	10540867	7.38	36.37	28.67	18.86	23.26	29.08	0.126	10540867	95.46%	7982633
mm_BMDM_GW39651_RXR	10532149	7.77	36.54	28.56	21.53	21.33	28.43	0.16	10532149	95.14%	8026079
mm_BMDM_GW3965_p300	9297847	7.14	36.36	28.83	18.98	23.02	29.02	0.157	9297847	95.99%	7040892
mm_BMDM_GW3965_RXR	1645654	6.98	34.23	28.47	21.99	20.83	28.58	0.13	1645654	87.46%	1144113

Outputs of the primary analysis III/b.

Percent reads mapped in BAM	HOMER total tags	Average tags per peaks	Peaks width	Fragment length	Total peaks	IP efficiency	% peaks filtered by local signal	% clonal peaks filtered	MACS tagsize	Total tags	% tags filtered out	Fragment length	No of peaks	Experiment
68.28%	2165571	1.06	257	220	14389	9.47	10.60%	0.04%	42	2409631	5.25%	232	15497	hs_MQ_minus_H3K4me
65.34%	2196028	1.18	17	224	277	0.21	38.22%	2.97%	42	2574321	13.82%	317	155	hs_MQ_minus_STAT6
86.19%	4538166	1.08	244	212	14693	6.06	9.51%	0.27%	42	5107860	7.25%	218	16367	hs_MQ_plus_H3K4me
37.29%	1047187	1.19	32	221	389	0.27	21.17%	0.40%	42	1231266	14.20%	335	98	hs_MQ_plus_IgG
46.92%	1029179	1.09	19	230	521	0.33	19.88%	0.46%	42	1212381	7.26%	335	150	hs_MQ_plus_STAT6
87.25%	881711	1.03	13	134	224	0.25	27.88%	0.32%	42	1024134	2.68%	337	57	hs_MQ_RSGIL4_p300_2
83.26%	3053184	1.04	75	213	2608	0.55	11.24%	1.27%	42	3518894	3.86%	310	977	hs_MQ_RSGIL4_p300
90.03%	1848573	1.03	147	125	1037	0.39	11.51%	1.34%	42	2135056	3.14%	211	562	hs_MQ_RSGIL4_pU1b
68.94%	882943	1.04	181	148	6975	2.98	5.11%	0.00%	42	998695	3.96%	250	4067	hs_MQ_RSGIL4_pU1
95.94%	6934437	1.13	167	218	894	0.22	24.96%	7.21%	42	7984897	11.23%	295	565	hs_MQ_RSGIL4_RXRb
71.27%	3589040	1.05	202	169	20360	4.31	6.79%	0.20%	42	4068840	5.03%	180	18775	hs_MQ_RSGIL4_RXR
95.31%	5258963	1.08	58	223	368	0.17	38.77%	9.90%	42	6054658	7.43%	51	849	hs_MQ_RSGIL4_STAT6
90.20%	5901473	1.07	104	211	907	0.22	23.70%	6.75%	42	6821248	6.84%	180	893	hs_MQ_veh_p300_2
70.19%	2815536	1.04	102	194	2363	0.53	9.50%	1.47%	42	3239864	4.39%	302	1026	hs_MQ_veh_p300
81.10%	1091559	1.02	152	133	678	0.37	14.88%	0.38%	42	1264086	2.20%	335	102	hs_MQ_veh_pU1b
78.24%	2650740	1.03	201	160	20385	4.39	7.58%	0.10%	42	3004833	2.85%	180	18123	hs_MQ_veh_pU1
95.97%	5497832	1.06	71	210	566	0.20	31.79%	8.63%	42	6347314	6.06%	54	955	hs_MQ_veh_RXRb
80.00%	6228613	1.07	194	169	31069	5.18	7.84%	0.22%	42	7048891	6.95%	171	31440	hs_MQ_veh_RXR
95.61%	5352564	1.05	131	117	453	0.26	33.33%	10.32%	42	6178432	5.48%	52	854	hs_MQ_veh_STAT6
96.19%	6594828	1.11	31	172	2685	1.39	13.78%	0.19%	42	8395824	16.64%	195	1685	mm_BMDM_ATRA1_p300
91.19%	2936993	1.08	67	171	3823	1.70	11.49%	0.16%	42	3693839	13.98%	185	2076	mm_BMDM_ATRA1_RXR
96.24%	7144688	1.08	29	172	2114	1.15	13.91%	0.32%	42	8973250	13.53%	162	1761	mm_BMDM_ATRA_p300
86.07%	932321	1.05	30	157	388	0.88	32.25%	1.54%	42	1159325	8.94%	336	76	mm_BMDM_ATRA_RXR
95.46%	7982633	1.12	103	171	8435	1.82	9.62%	0.10%	42	10062784	17.07%	186	5650	mm_BMDM_GW39651_p300
95.14%	8026079	1.09	171	172	15377	2.74	11.05%	0.05%	42	10019790	15.00%	178	11641	mm_BMDM_GW39651_RXR

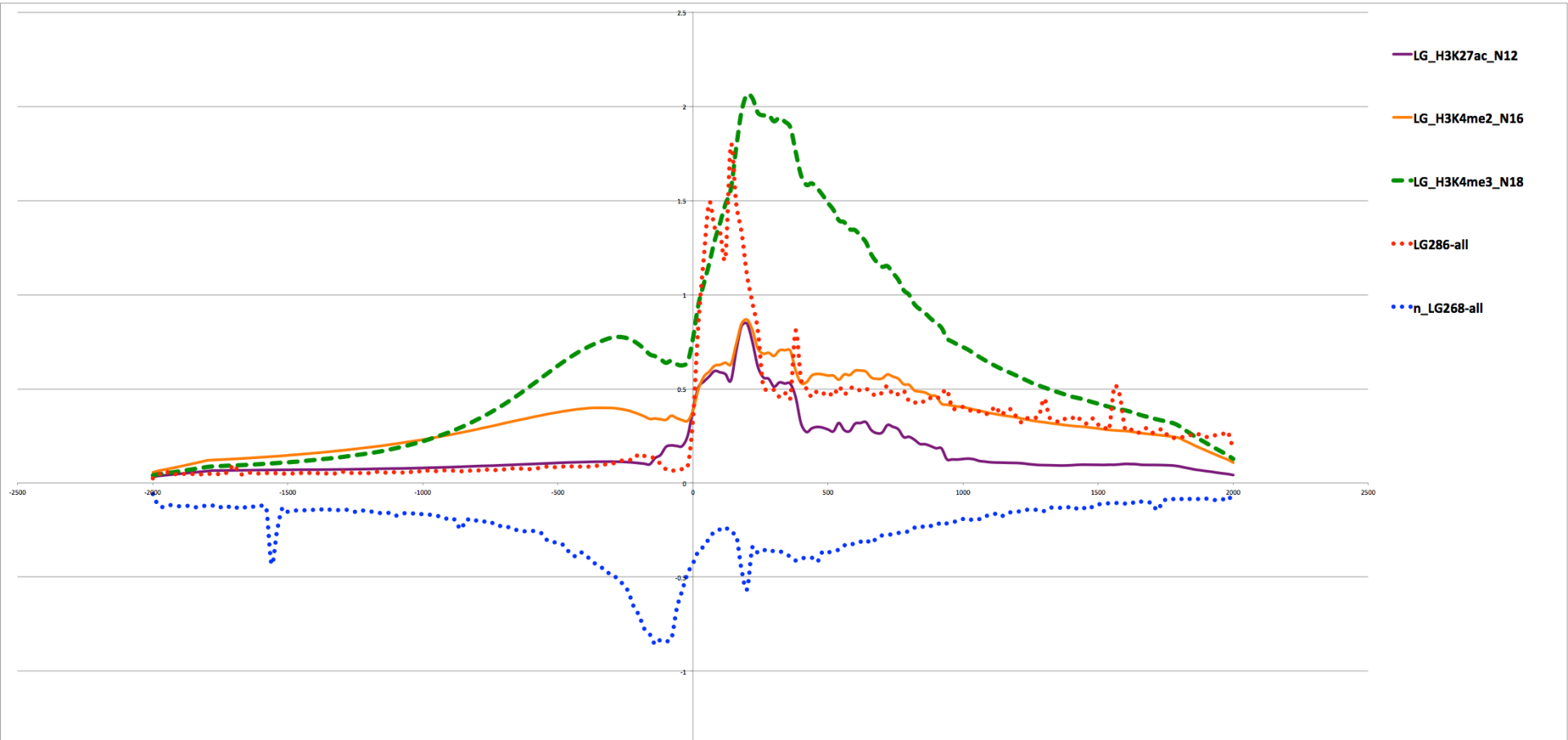
Approximate IP effeciency describes the fraction of tags found in peaks versus. genomic background. This provides an estimate of how well the ChIP worked. Certain antibodies like H3K4me3, ERa, or PU.1 will yield very high IP efficiencies (>20%), while most are in the 1-20% range. Once this number dips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

Downstream analysis

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

Generating profiles or metagenes, or histograms (normalized tag counts in bins)

To compare different ChIP-seq samples, extensive normalization has to be carried out (HOMER's annotatePeak can do this).
The center (0 point) can be either the peak summit, the TSS or the TFBSs



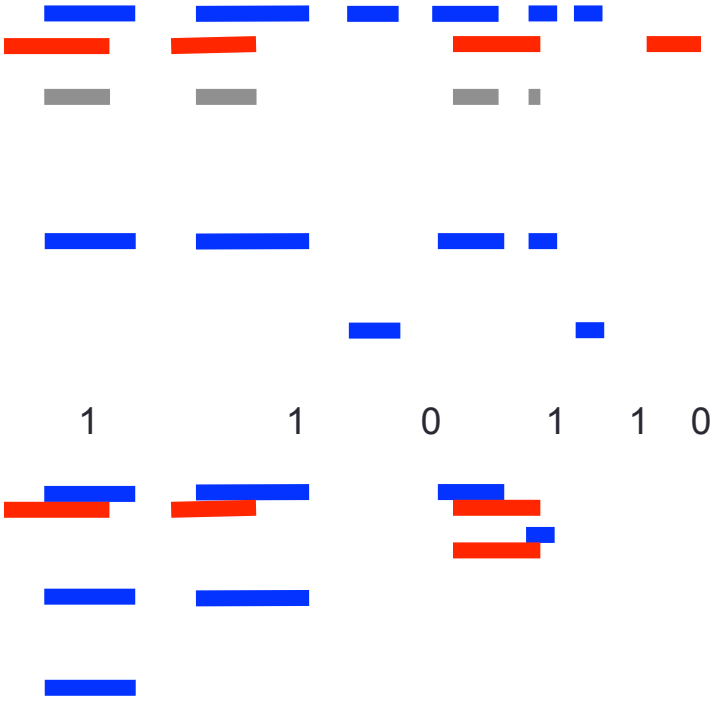
intersectBed

Switches:

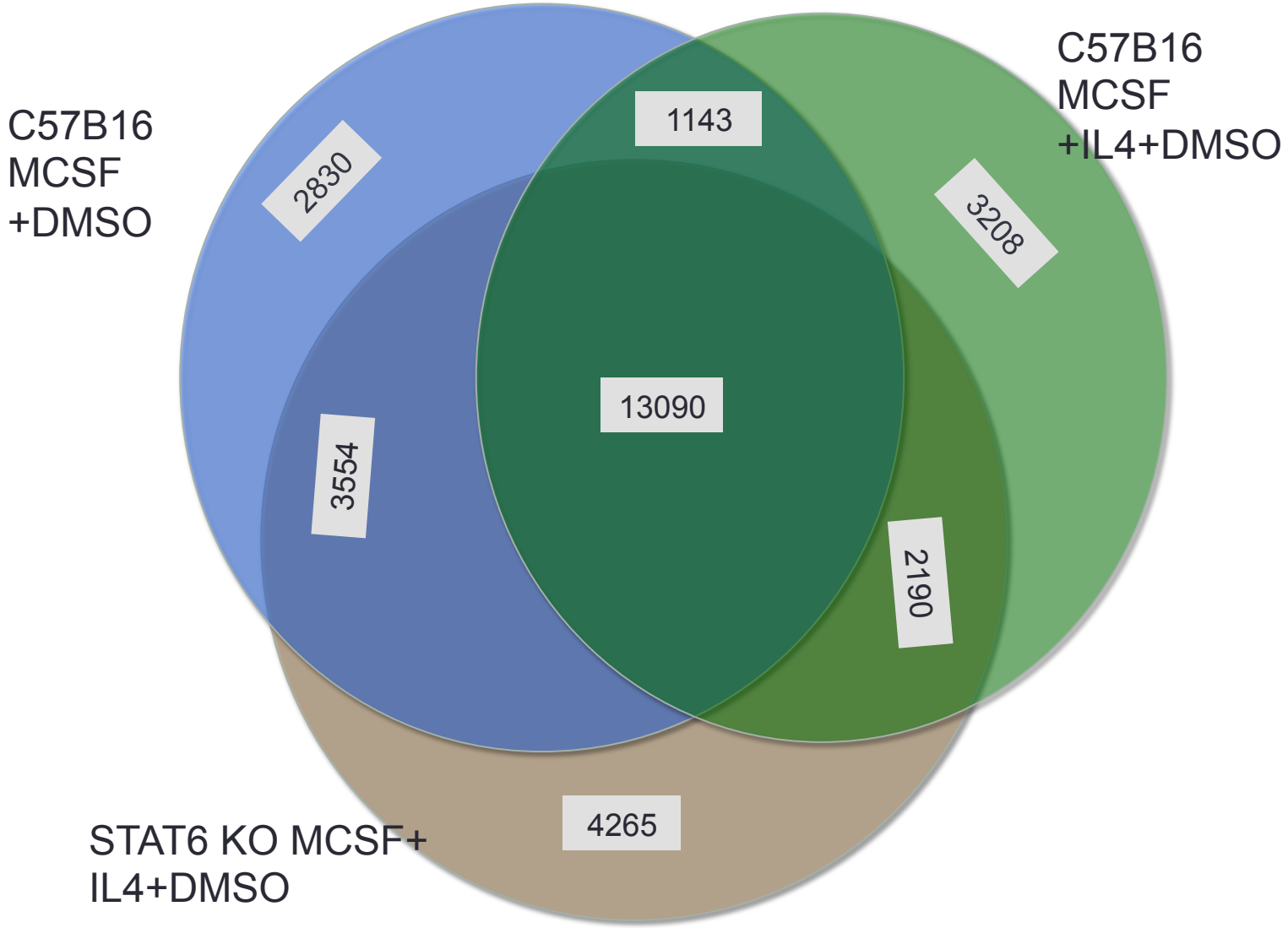
- -a *peakfile1.bed* -b *peakfile2.bed*

  ((-abam => -bed))

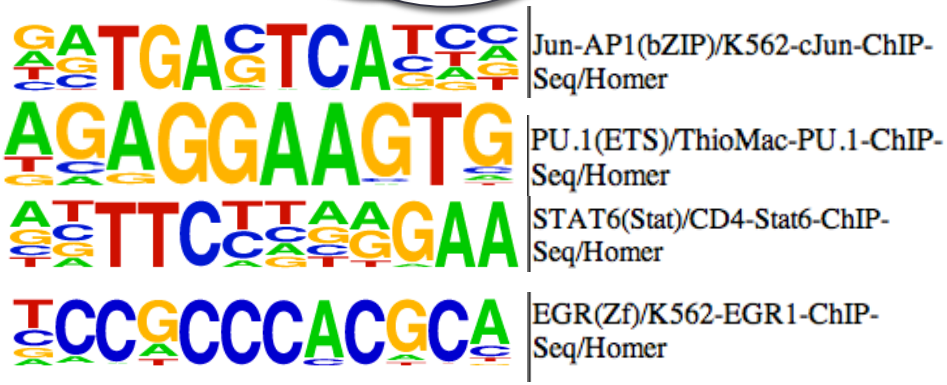
- -u
- -v
- -c (count b on a)
- -wo (fusing beds in a “double bed” table)
- -f (minimum overlap %) – -u -f 0.6
- -r (reciprocal overlap) – -u -f 0.6 -r
- -s (strand specific match)

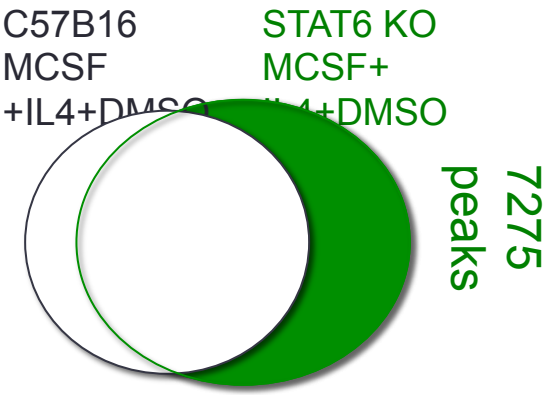


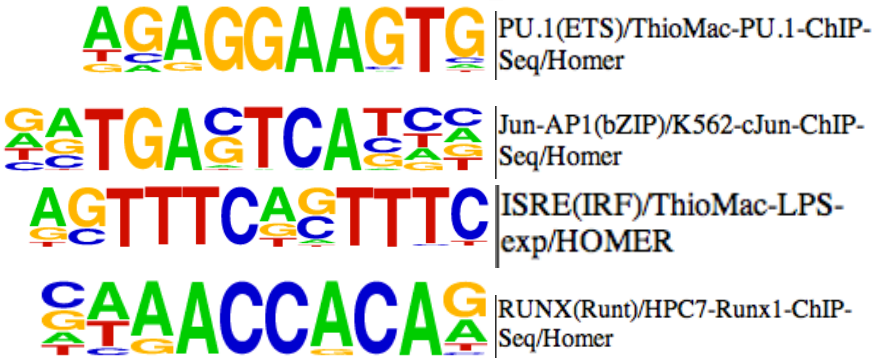
Comparison of ChIP regions from different experiments



Different subsets have different motifs







B_Cell_Receptor_Signaling_Pathway_WP23
Apoptotic_execution_phase_WP1784
BMP_signalling_and_regulation_WP1425
TGF-beta_Receptor_Signaling_Pathway_WP366
Wnt_Signaling_Pathway_NetPath_WP363
Kit_Receptor_Signaling_Pathway_WP304
IL-6_Signaling_Pathway_WP364
Integrin_alphaIIb_beta3_signaling_WP1832
IL-4_signaling_Pathway_WP395

TGF-beta_Receptor_Signaling_Pathway_WP366
EGFR1_Signaling_Pathway_WP437
IL-5_Signaling_Pathway_WP127
B_Cell_Receptor_Signaling_Pathway_WP23
Toll-like_receptor_signaling_pathway_WP75
T_Cell_Receptor_Signaling_Pathway_WP69
IL-3_Signaling_Pathway_WP286
Regulation_of_toll-like_receptor_signaling_pathway_WP1449
Type_II_interferon_signaling_(IFNG)_WP619

DiffBind R package

Home » Bioconductor 2.11 » Software Packages » DiffBind

DiffBind

Differential Binding Analysis of ChIP-Seq peak data

Bioconductor version: Release (2.11)

Compute differentially bound sites from multiple ChIP-seq experiments using affinity (quantitative) data. Also enables occupancy (overlap) analysis and plotting functions.

Author: Rory Stark<rory.stark at cancer.org.uk>, Gordon Brown <gordon.brown at cancer.org.uk>

Maintainer: Rory Stark<rory.stark at cancer.org.uk>

To install this package, start R and enter:

Input files

- 1 input file
  - Comma Separated Values (.csv)
  - 1 header line + 1 line for each sample
- Required fields
  - Sample ID
  - Tissue
  - Factor
  - Condition
  - Replicate ID
  - Read file (bam)
  - Control read file (bam)
  - Peak file

Analysis Pipeline

- Data used in the analysis
  - Read files
  - Peaksets (Homer, MACS)
- 2 pipelines
  - Occupancy Analysis
    - No quality control
    - Less strict
  - Differential Binding Analysis
    - Quality control
    - (Less flexible)

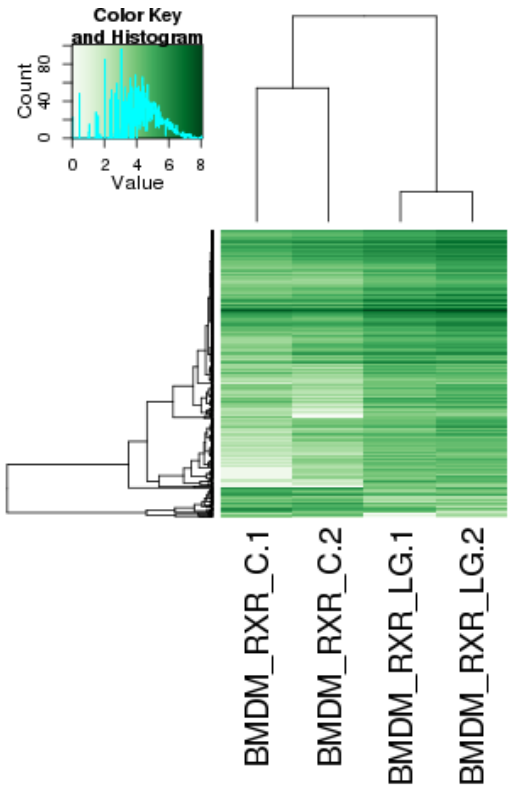
Analysis Pipeline

- Creating a 'dba' object
  - Output based on raw data
- Definition of consensus peakset
  - Read counting
  - Normalization
  - Output based on the consensus peakset
- Contrast, blocking, masking, etc.

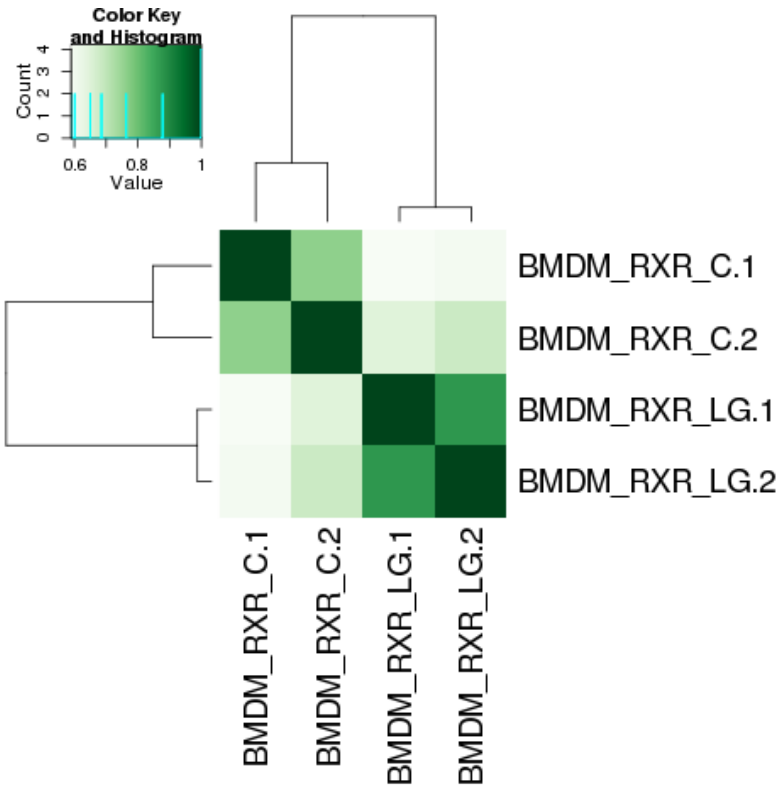
Analysis Pipeline

- Differential Binding Analysis
- Output based on differentially bound sites
  - Heatmaps (correlation, expression)
  - Plots (MA, PCA, boxplot)
  - Venn diagrams
- Save the peaksets for later use

Control vs. LG268

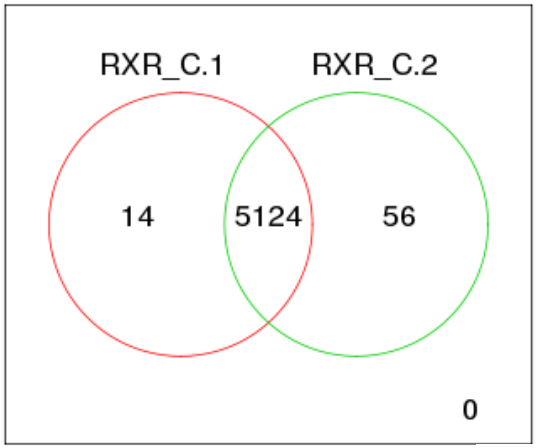


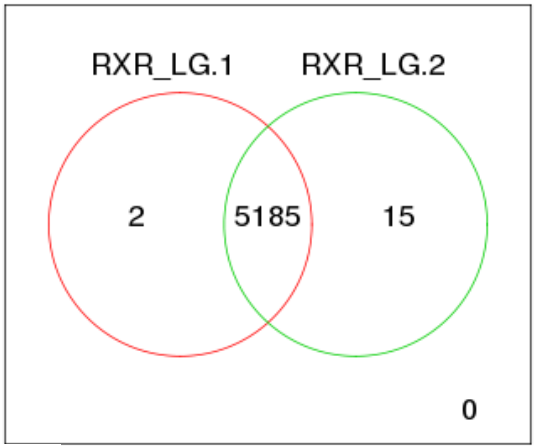
Significant changes

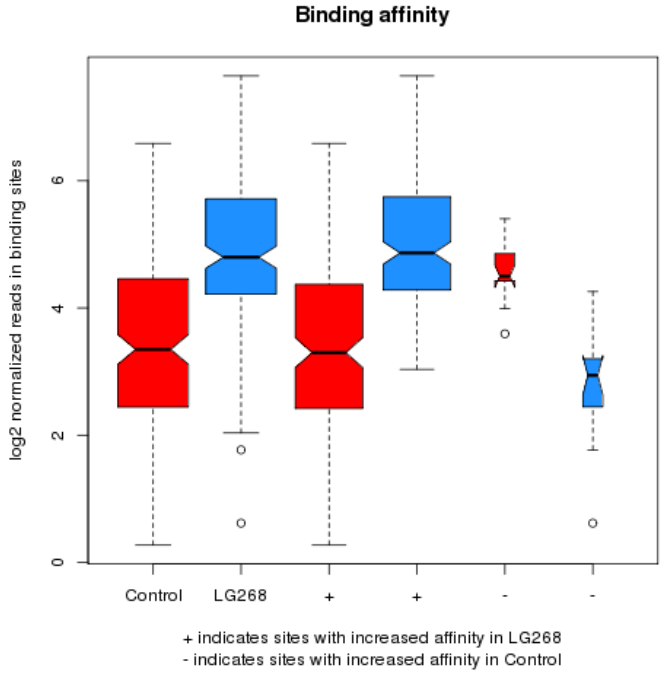


Significant changes

Control vs. LG268



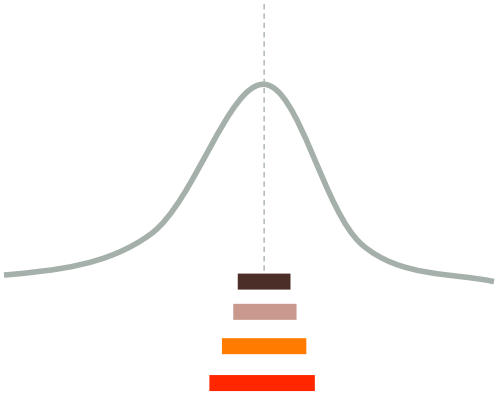




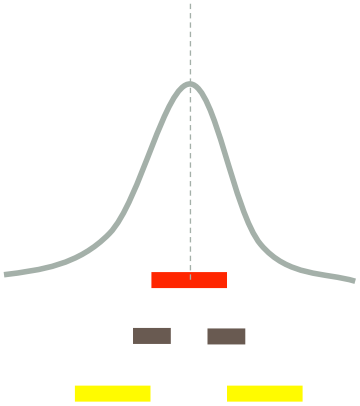
DiffBind: Attila Horváth

Motives are close to peak summits

RXR ChIP peaks

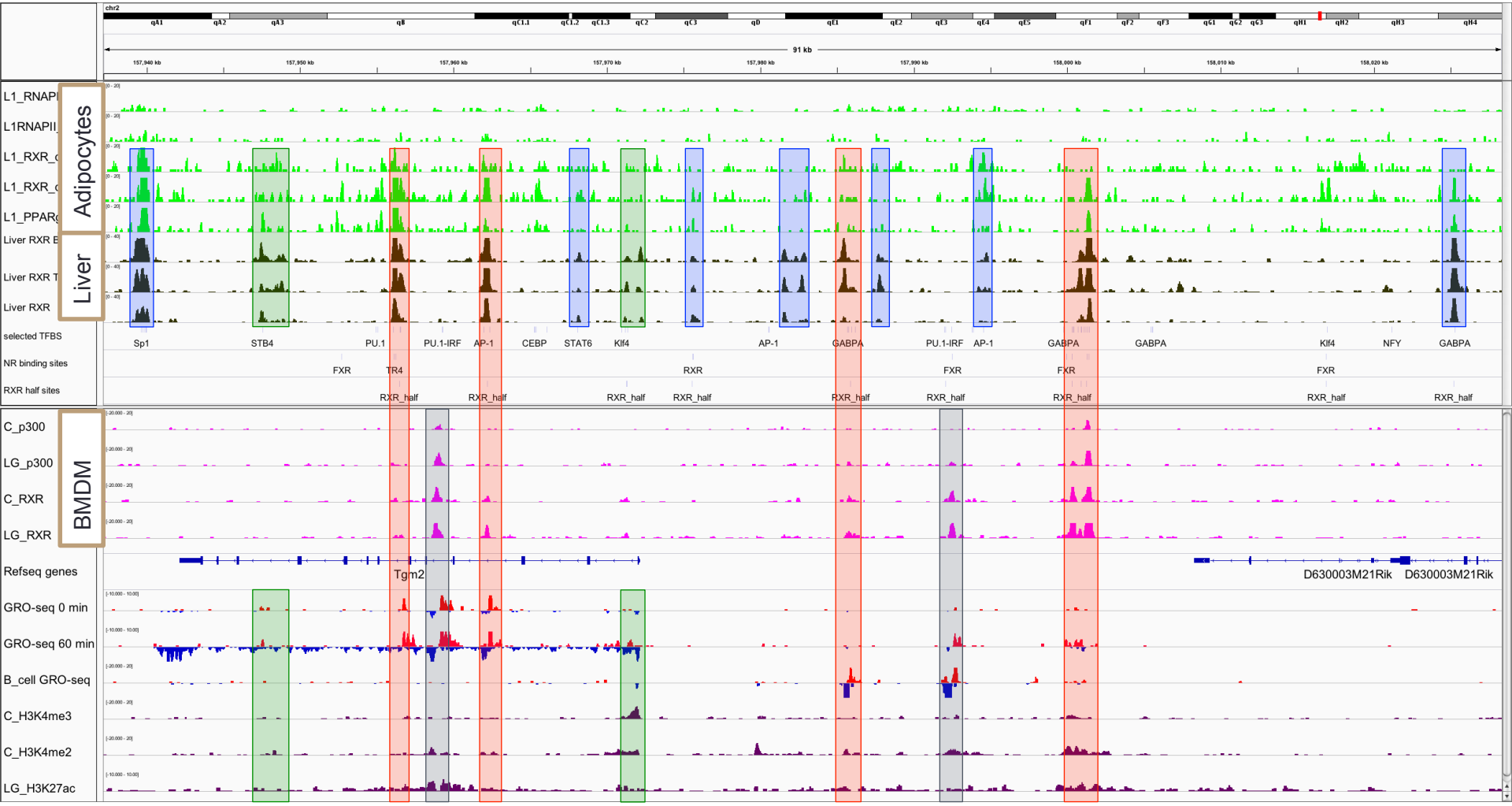


	Width	p-value	target %	bg %	fold
PU.1	50	1E-74	23.17	4.66	4.97
	60	1E-91	28.62	5.94	4.82
	80	1E-142	34.02	5.32	6.39
	100	1E-169	44.24	8.28	5.34
NRhalf	50	1E-151	39.27	6.96	5.64
	60	1E-184	41.05	6.02	6.82
	80	1E-254	52.78	7.63	6.92
	100	1E-258	59.23	10.39	5.70



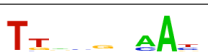
	p-value	target %	bg %	fold
PU.1	1E-86	18.97	5.34	3.55
	1E-135	21.85	4.62	4.73
	1E-169	44.24	8.28	5.34
NRhalf	1E-18	0.64	0.01	64.00
	1E-52	9.52	2.18	4.37
	1E-258	59.23	10.39	5.70

RXR enhancers from different tissues around the Tgm2 gene



Motives on RXR ChIP-seq peaks in different cells (HOMER de novo motif finding)

BMDM RXR

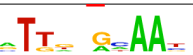
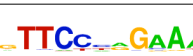
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-765	-1.762e+03	51.85%	16.40%	398.5bp (290.8bp)	
2		1e-201	-4.630e+02	53.37%	33.12%	451.9bp (285.3bp)	
3		1e-141	-3.264e+02	28.42%	14.80%	483.7bp (287.8bp)	
4		1e-125	-2.896e+02	13.79%	5.13%	408.6bp (292.8bp)	
5		1e-92	-2.120e+02	38.27%	25.51%	446.2bp (285.1bp)	

PU.1
NR half
AP-1
PU.1
CEBP

Motives in BMDM:

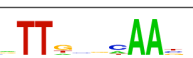
- PU.1
- NR half
- AP-1
- CEPB
- RUNX
- NR DR1
- IRF

L1_4h RXR

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-982	-2.263e+03	53.80%	12.97%	43.3bp (63.2bp)	
2		1e-459	-1.059e+03	27.16%	6.10%	51.1bp (64.5bp)	
3		1e-136	-3.138e+02	16.68%	6.28%	52.4bp (61.3bp)	
4		1e-67	-1.546e+02	25.19%	15.45%	56.0bp (65.2bp)	
5		1e-51	-1.191e+02	5.57%	1.88%	53.2bp (57.0bp)	

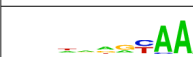
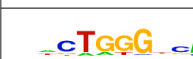
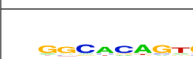
CEBP
AP1
NR half
NF1 half
STAT

L1_D6 RXR

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-1273	-2.933e+03	29.55%	13.25%	43.0bp (59.3bp)	
2		1e-856	-1.972e+03	11.56%	3.51%	49.5bp (61.4bp)	
3		1e-573	-1.321e+03	28.37%	16.89%	50.9bp (58.1bp)	
4		1e-569	-1.312e+03	5.30%	1.15%	36.8bp (57.6bp)	
5		1e-516	-1.189e+03	44.88%	31.91%	50.2bp (62.0bp)	

NR DR1
CEBP
NR half
CTCF
NF1 half

Liver RXR

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-2251	-5.184e+03	62.36%	29.93%	48.9bp (62	
2		1e-865	-1.992e+03	15.63%	4.67%	48.0bp (63	
3		1e-679	-1.564e+03	28.45%	14.23%	51.7bp (64	
4		1e-539	-1.243e+03	5.91%	1.08%	50.6bp (58	
5		1e-479	-1.105e+03	8.79%	2.59%	49.5bp (58	
6		1e-469	-1.081e+03	50.60%	35.66%	56.0bp (64	
7		1e-433	-9.974e+02	8.72%	2.75%	49.5bp (56	
8		1e-249	-5.750e+02	5.18%	1.65%	49.8bp (62	
9		1e-233	-5.370e+02	48.81%	38.24%	53.8bp (59	
10		1e-227	-5.246e+02	10.04%	4.86%	51.2bp (60	

NR half
CEBP
FOXA
FOXA
FOXA
NF1
NR DR1
FOXA
NF1
CEBP

Motif matrix files and logos

RGKKSANRGKKSA ⟶ DRGGTCARAGGTCARN

	>Consensus sequence			Name	Score threshold
	>RGKKSANRGKKSA			DR1	11.0523208315125
R	0.499	0.001	0.499	0.001	
G	0.001	0.001	0.997	0.001	
K	0.001	0.001	0.499	0.499	
K	0.001	0.001	0.499	0.499	
S	0.001	0.499	0.499	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	
K	0.001	0.001	0.499	0.499	
K	0.001	0.001	0.499	0.499	
S	0.001	0.499	0.499	0.001	
A	0.997	0.001	0.001	0.001	

	>Consensus sequence			Name	
	Score threshold				
D	>DRGGTCARAGGTCARN			1-DRGGTCARAGGTCARN	
R	8.661417		*		
G	0.367	0.067	0.311	0.256	
G	0.466	0.001	0.532	0.001	
T	0.189	0.022	0.656	0.133	
C	0.111	0.100	0.667	0.122	
A	0.133	0.256	0.156	0.455	
R	0.166	0.512	0.222	0.100	
A	0.966	0.001	0.011	0.022	
G	0.378	0.067	0.477	0.078	
G	0.821	0.001	0.177	0.001	
T	0.066	0.011	0.922	0.001	
C	0.044	0.022	0.767	0.167	
A	0.111	0.089	0.134	0.666	
R	0.078	0.855	0.045	0.022	
N	0.900	0.001	0.022	0.077	
	0.278	0.211	0.378	0.134	
	0.300	0.211	0.222	0.267	

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

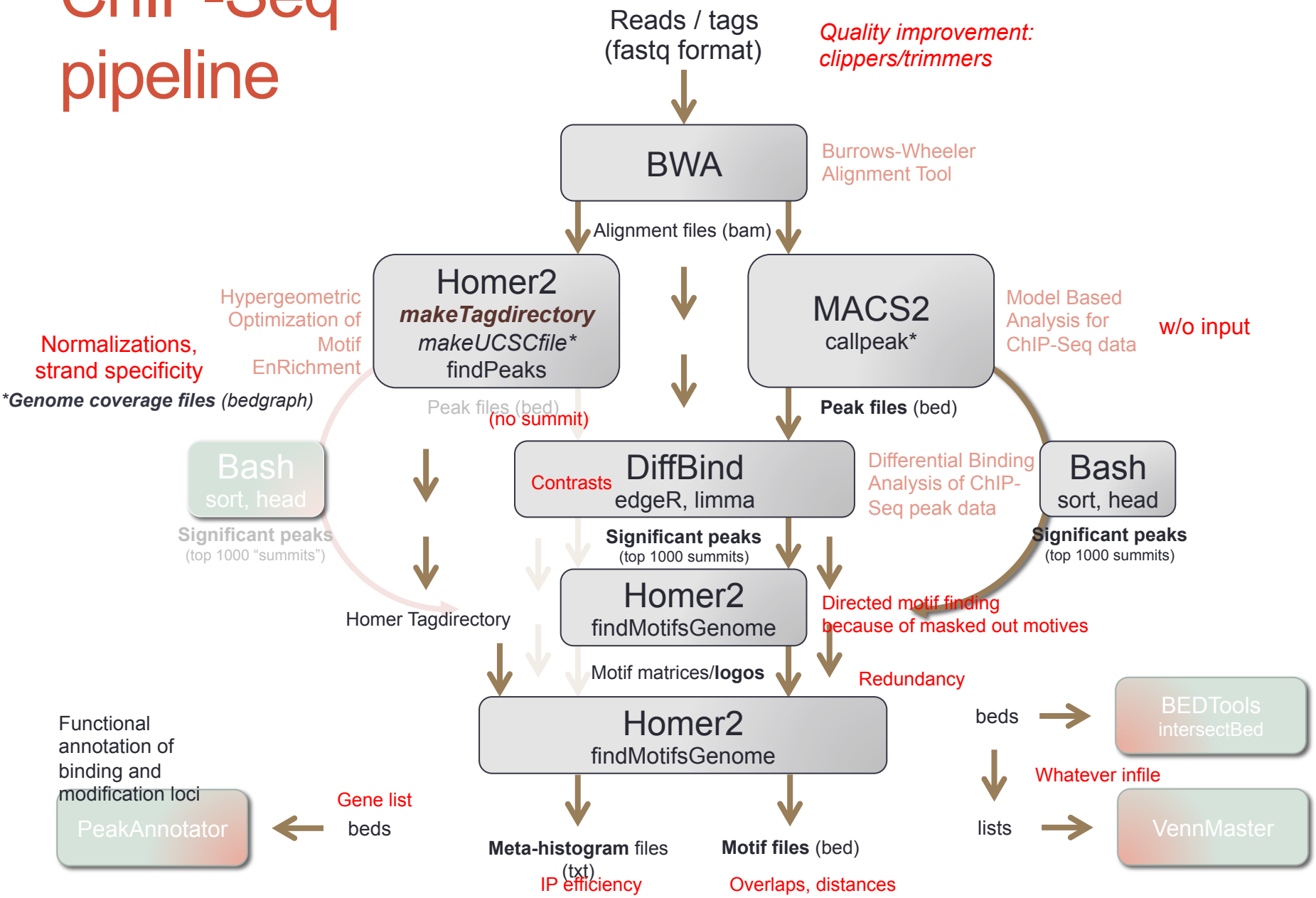
RXR-PU.1 tethering
Motif enrichment

PU.1		1e-650	83.75% 7.20%
AP-1		1e-29	19.81% 7.59%
RUNX		1e-24	8.78% 2.04%
c/EBP		1e-17	3.32% 0.35%
Half		1e-15	4.74% 0.97%
SREBP?		1e-14	3.91% 0.70%

RGKKSA

<<<<<28.46% of peaks

ChIP-Seq pipeline



Chipster (ngs tools from version 2.0)

The Next NGS Challenge
data processing and integration

ChIP-seq chipster tutorial session with Eija Korpelainen COST conference Valencia, May, 2013

CSC

Chipster

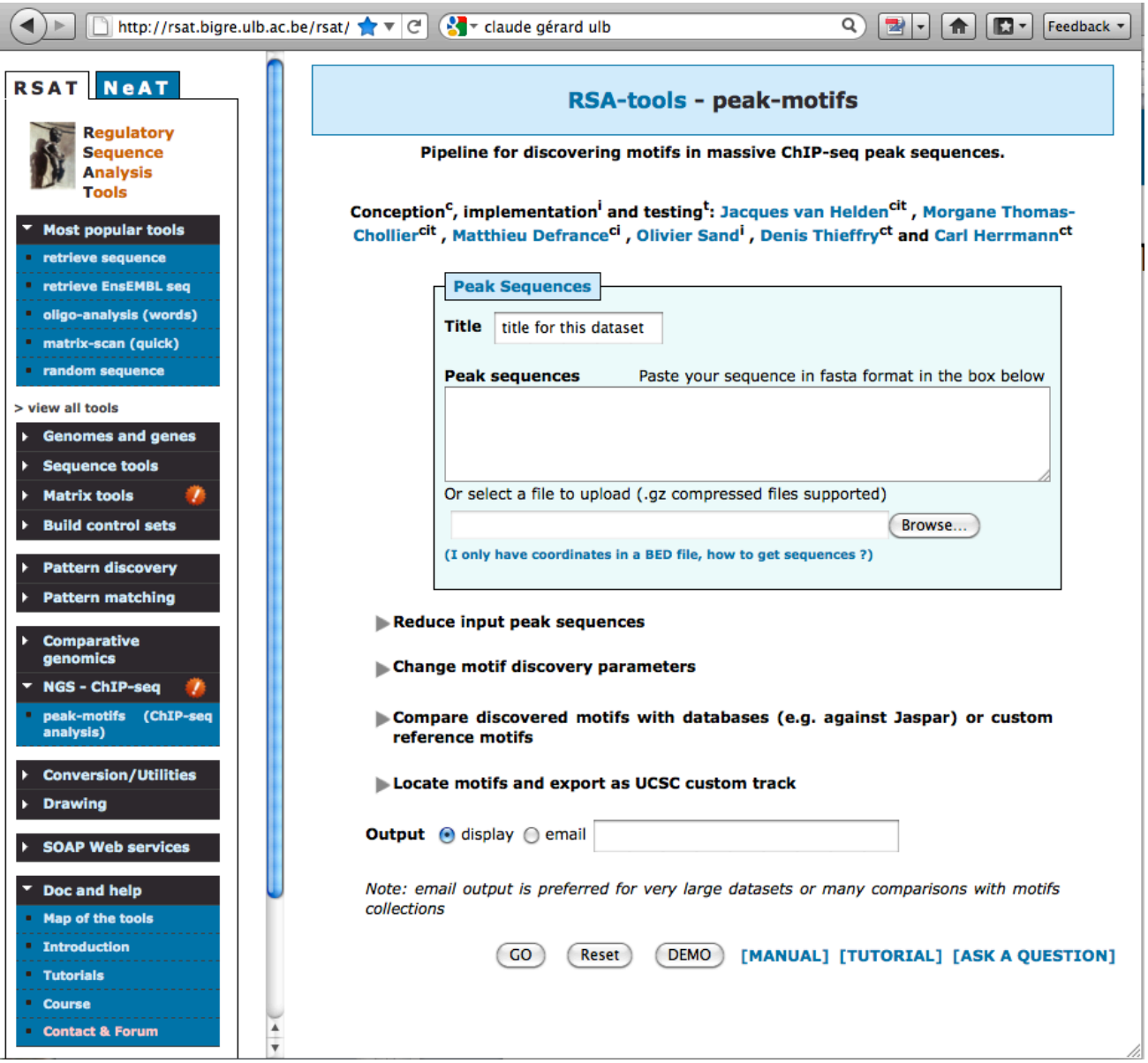
Open source platform for data analysis

- Home
- Getting access
- Analysis tool content
- Supported chip types
- Tutorials
- Manual
- Cite
- FAQ
- Screenshots
- Open source project
- Contact

Welcome to Chipster

Chipster is a user-friendly analysis software for high-throughput data. It contains over 230 analysis tools for next generation sequencing (NGS), microarray and proteomics data. Users can save and share automatic analysis workflows, and visualize data interactively using a built-in genome browser and many other visualizations. Chipster's little brother Embster is available for basic sequence analysis tasks like BLAST.

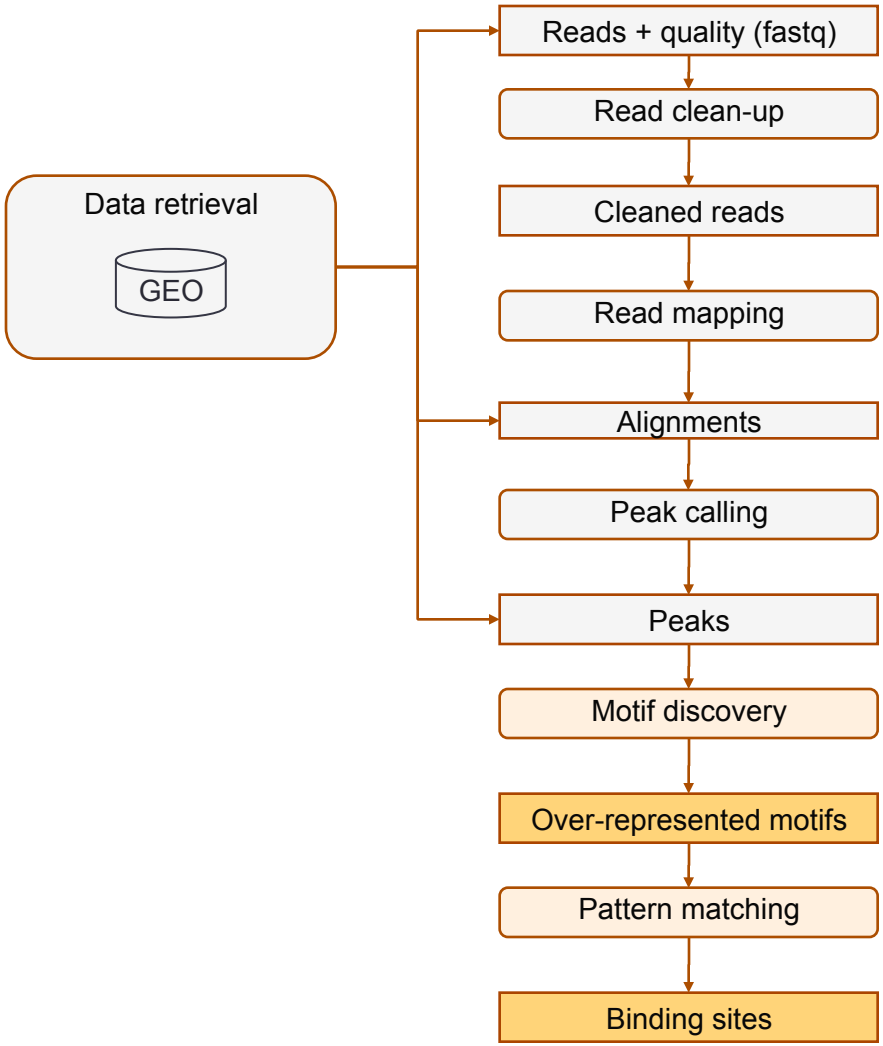
Chipster's client software uses Java Web Start to install itself automatically, and it connects to computing servers for the actual analysis. If you would like to use Chipster running on CSC's server, you need a user account. If you would like to set up your own Chipster server locally, you can download it as a virtual machine.



- Regulatory Sequence Analysis Tools (RSAT)
  - http://rsat.ulb.ac.be/rsat/
- Interfaces
  - Stand-alone apps
  - Web site
  - Web services (SOAP/WSDL API)
- Web interface
  - Simplicity of use (“one click” interface).
  - Advanced options can be accessed optionally.
  - Allows to analyze data set of realistic size (uploaded files).

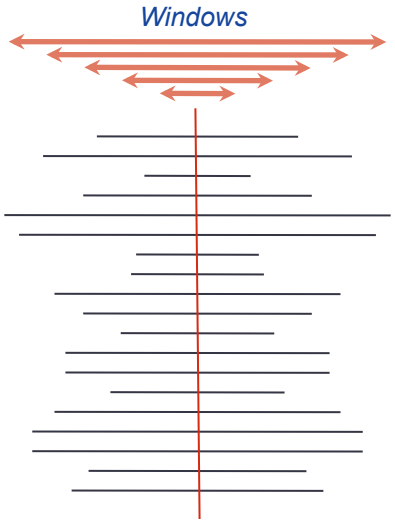
Work flow for chip-seq analysis

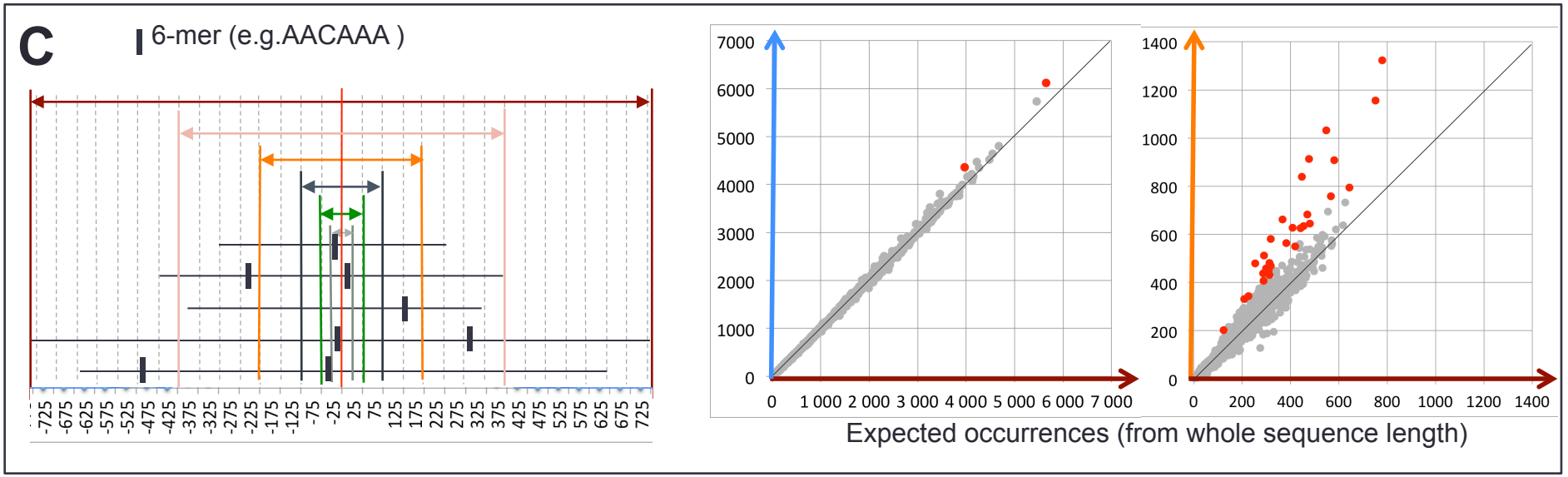
- ChIP-seq data can be retrieved from specialized databases such as Gene Expression Omnibus (GEO).
- The GEO database allows to retrieve sequences at various processing stages.
  - ***Read sequences***: typically, several millions of short sequences (25bp).
  - ***Read locations***: chromosomal coordinates of each read.
  - ***Peak locations***: several thousands of variable size regions (typically between 100bp and 10kb).
- A technological bottleneck lies in the next step: exploitation of full peak collections to discover motifs and predict binding sites.



Local over-representation (program local-words)

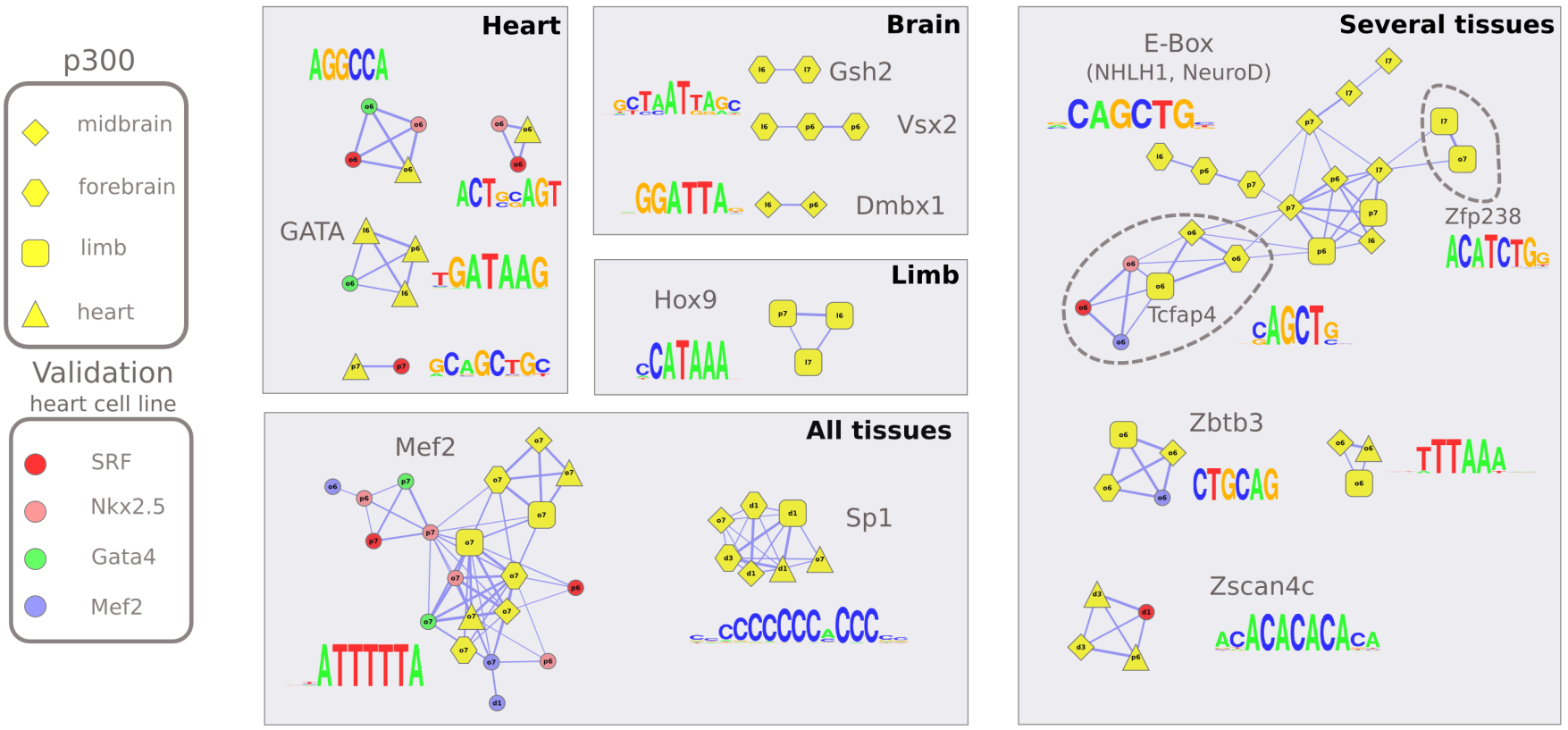
- The program *local-words* detects words that are over-represented in specific position windows.
- The result is thus more informative than for *position-analysis*: in addition to the global positional bias, we detect the precise window where each word is over-represented.





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. Nat Protoc 7(8): 1551-1568.

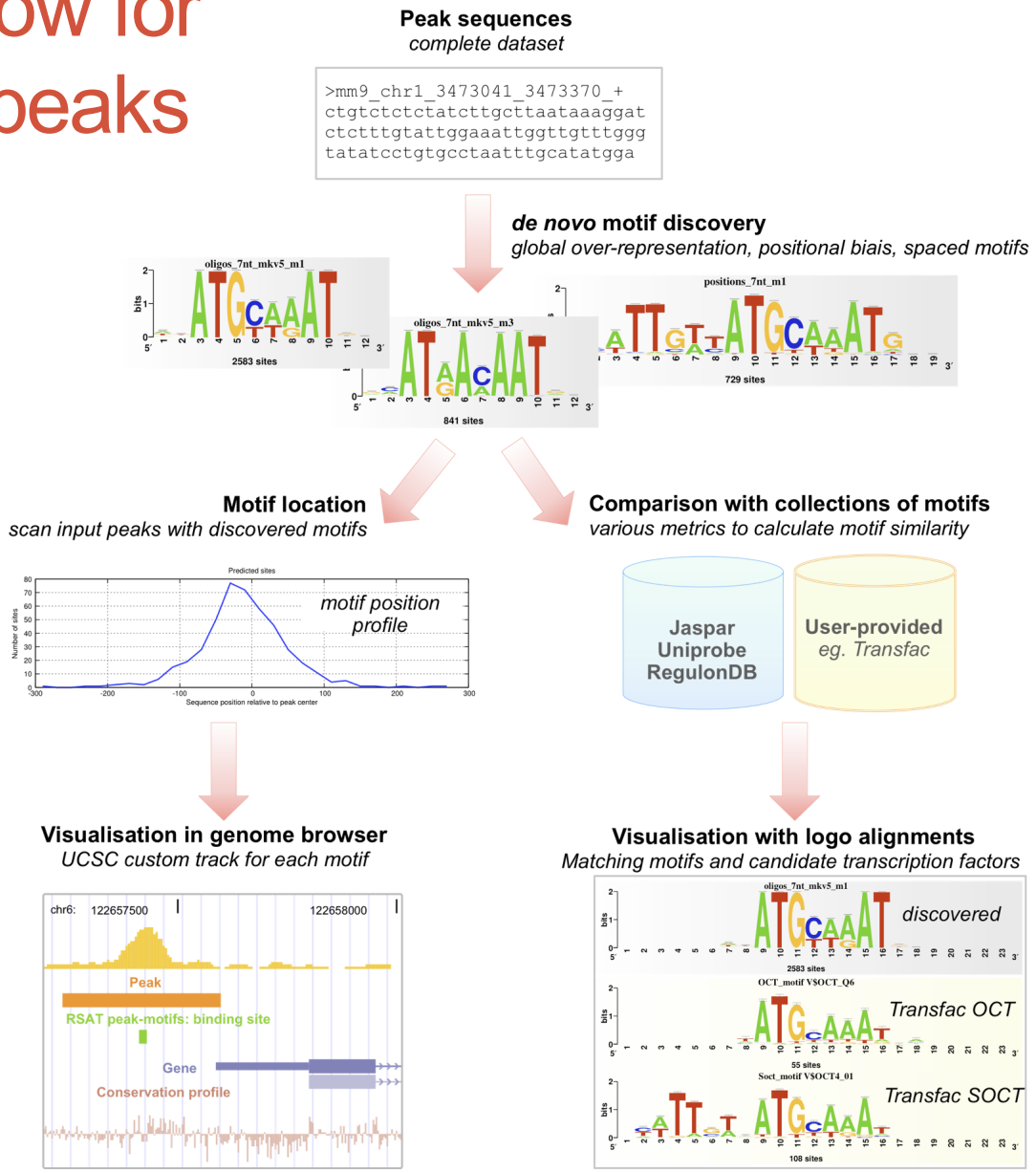
Network of motifs discovered in tissue-specific p300 binding regions



Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. 2012.
RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 40(4): e31.

An integrated work flow for analyzing ChIP-seq peaks

- The program ***peak-motifs*** is a work flow combining a series of RSAT tools optimized for discovered motifs in large sequence sets (tens Mb) resulting from ChIP-seq experiments..
- Multiple pattern discovery algorithms
  - Global over-representation
  - Positional biases
  - Local over-representation
- Discovered motifs are compared with
  - motif databases
  - user-specified reference motifs.
- Prediction of binding sites, which can be uploaded as custom annotation tracks to genon browsers (e.g. UCSC) for visualization.
- Interfaces
  - Stand-alone
  - Web interface
  - Web services (SOAP/WSDL)



Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. 2012.
RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 40(4): e31.

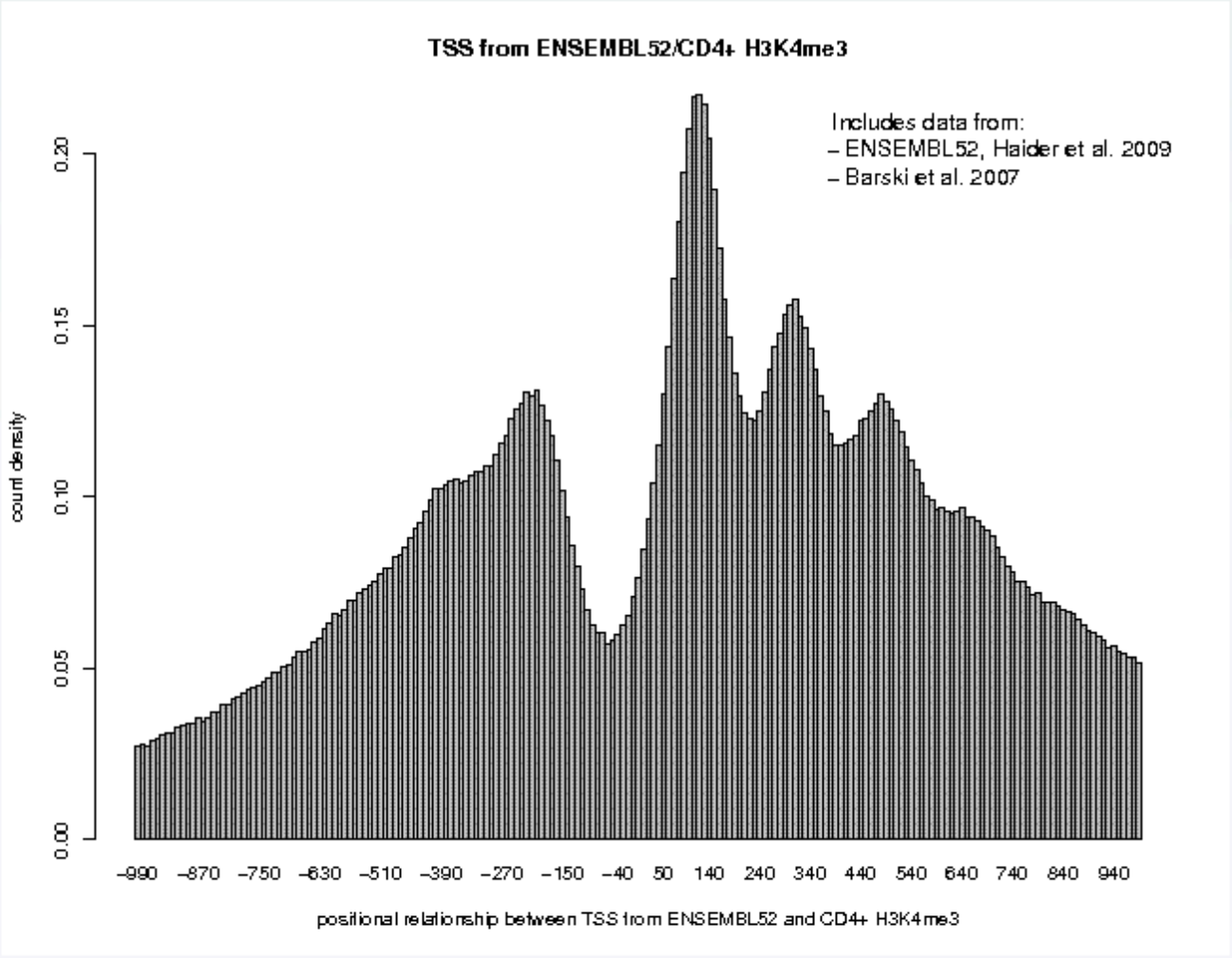
ChIP-seq analysis server @sib

SIB server, correlation tool

ChIP-cor Input Data	
Reference Data Set :	ENSEMBL52, TSS collection
Reference Sample :	TSS from ENSEMBL52 (oriented)
Target Data Set :	Barski 2007, CD4+ cells, histone marks
Target Sample :	CD4+ H3K4me3 (70 bp centered)

Analysis Parameters

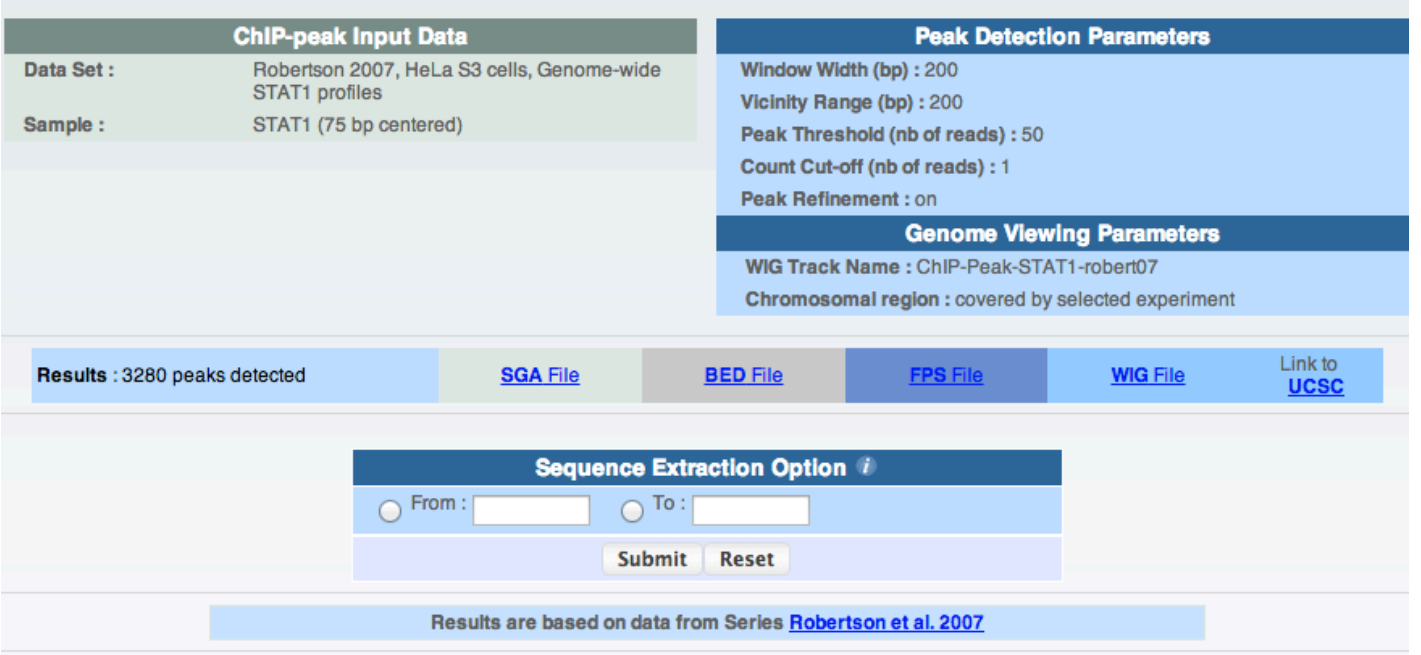
Input Range : -1000 - 1000
Window width: 10
Counts Cut-off value: 1
Normalization: count density



Result Files	Postscript	PDF	TEXT
Statistics and Peak Finding Parameters Estimate		Single Gaussian Fit	Parameters

Results are based on data from Series ENSEMBL52, Haider et al. 2009 and Series Barski et al. 2007

SIB server, ChIP-peak tool



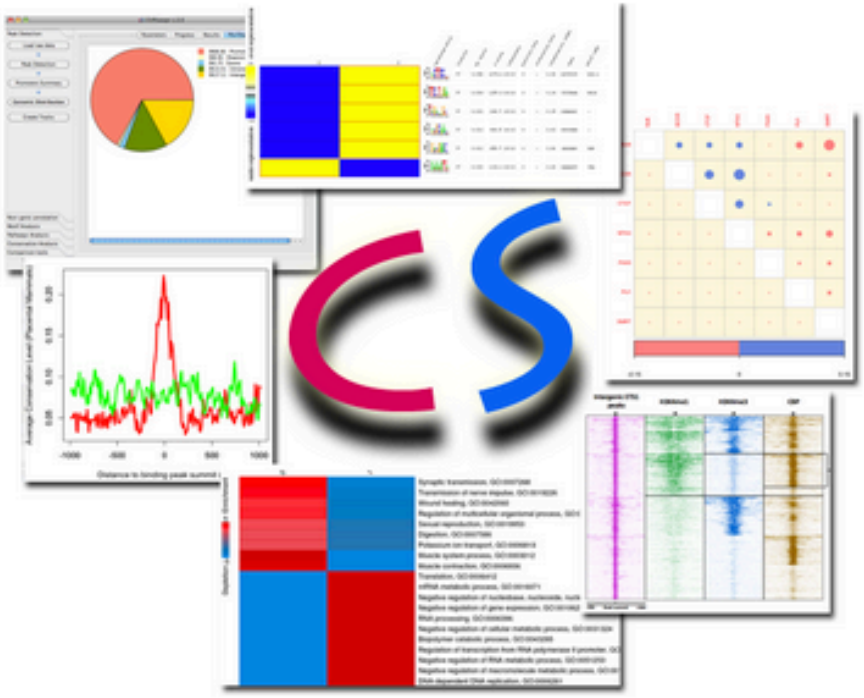
I. Sequence ID (char string)
II. Feature (char string)
III. Sequence Position (integer)
IV. Strand (+/- or 0)
V. Tag Counts (integer)

```
NC_000001.9     STAT1_p 1060813 0       125
NC_000001.9     STAT1_p 1348321 0       116
NC_000001.9     STAT1_p 1358600 0       50
NC_000001.9     STAT1_p 2311841 0       139
NC_000001.9     STAT1_p 2450535 0       80
NC_000001.9     STAT1_p 6144257 0       86
NC_000001.9     STAT1_p 6217254 0       203
NC_000001.9     STAT1_p 6347973 0       66
NC_000001.9     STAT1_p 6387484 0       108
NC_000001.9     STAT1_p 7650437 0       82
NC_000001.9     STAT1_p 8194733 0       98
NC_000001.9     STAT1_p 8509658 0       94
NC_000001.9     STAT1_p 8882745 0       175
NC_000001.9     STAT1_p 8886744 0       52
NC_000001.9     STAT1_p 9093499 0       53
NC_000001.9     STAT1_p 9216177 0       95
```

ChIPseeqer

ChIPseeqer

A comprehensive framework for the analysis of ChIP-seq data

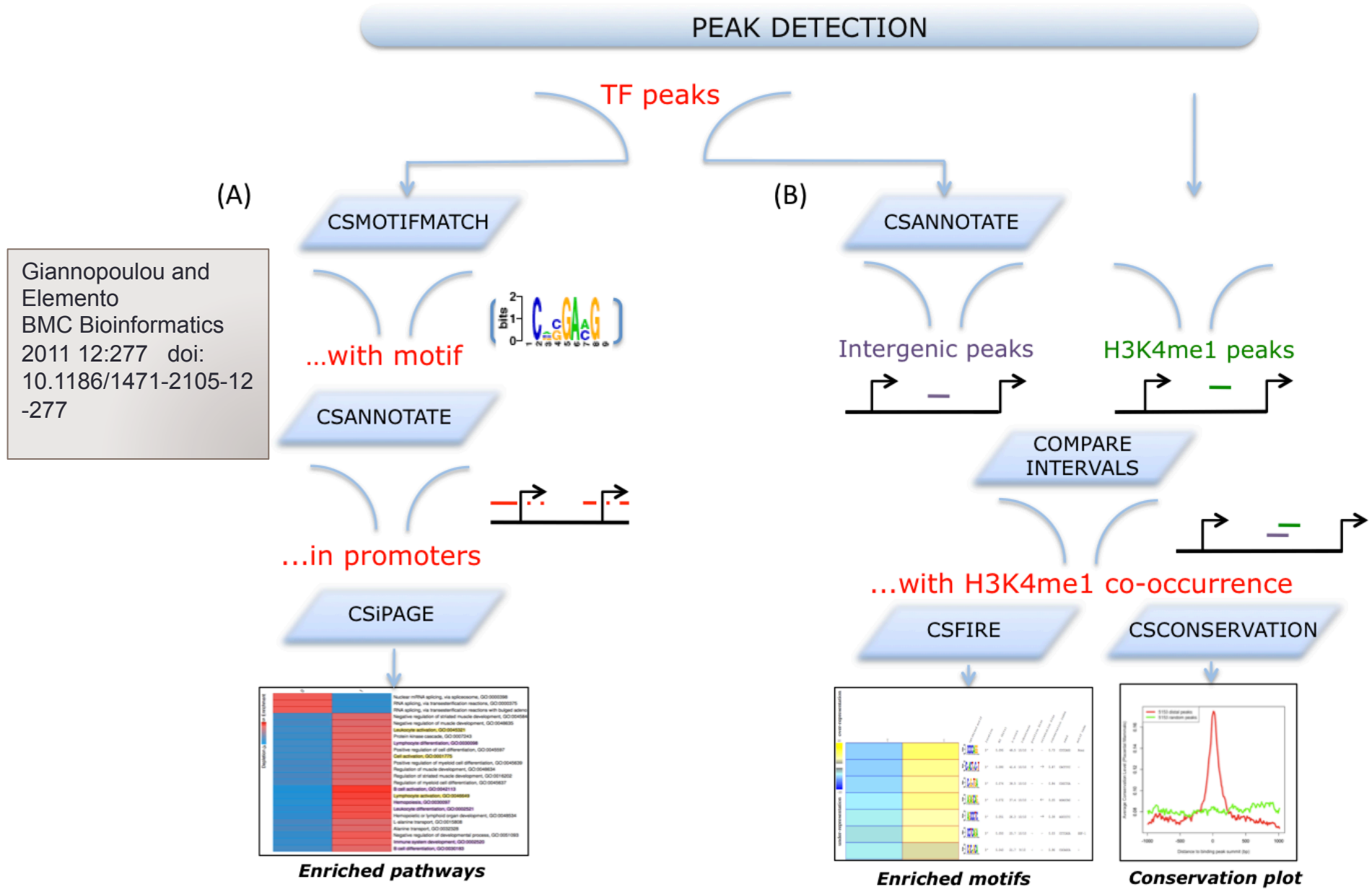


http://physiology.med.cornell.edu/faculty/elemento/lab/chipseq.shtml

ChIPseeqer tasks (command line and GUI)

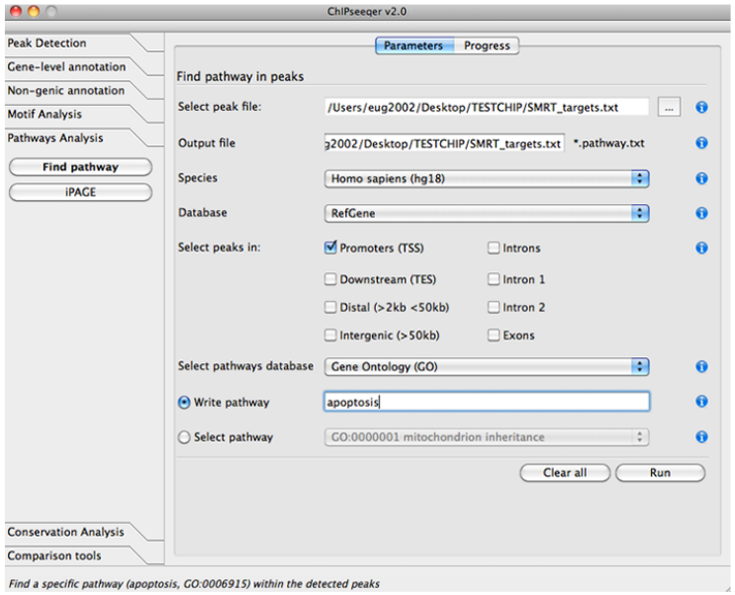
- Peak detection
- Gene-level annotation of peaks
- Pathways enrichment analysis
- Regulatory element analysis, using either a denovo approach, known or user-defined motifs
- Nongenic peak annotation (repeats, CpG island, duplications)
- Conservation analysis
- Clustering analysis
- Visualisation
- Integration and comparison across different ChIP-seq experiments

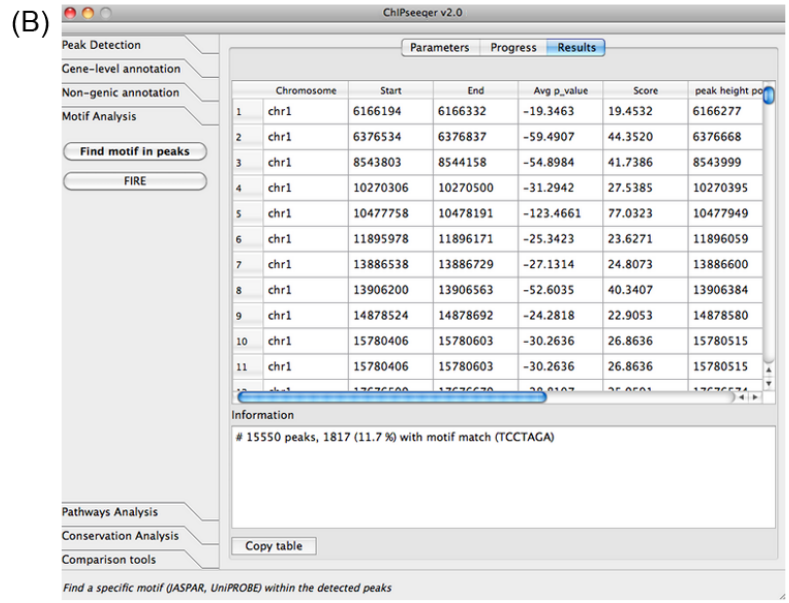
ChIPseeqer workflow

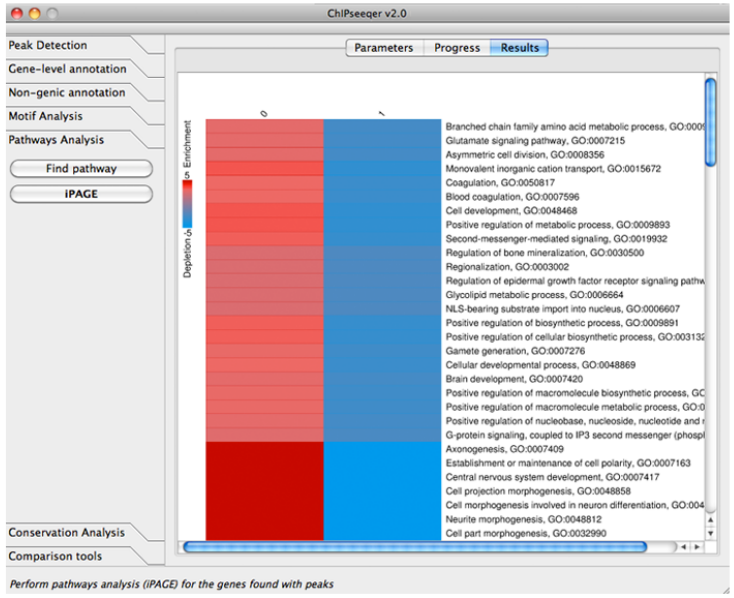


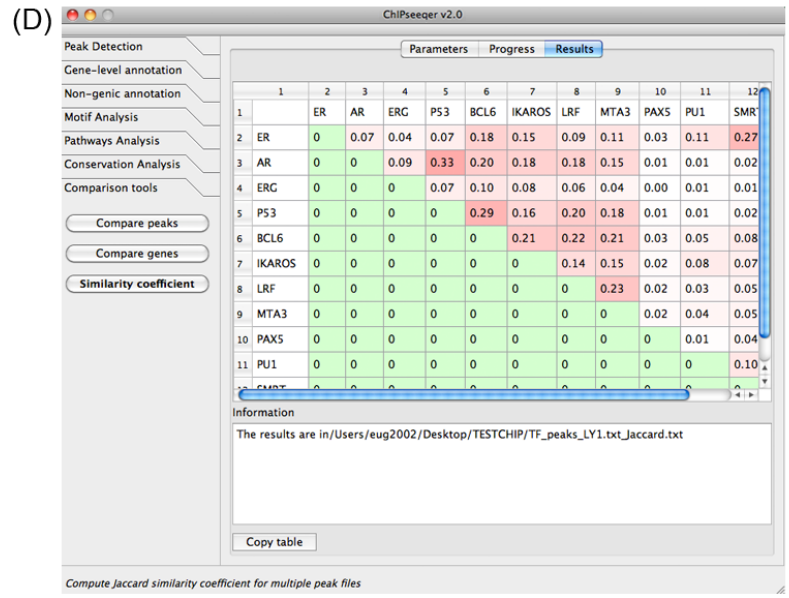
Giannopoulou and Elemento
BMC Bioinformatics 2011 12:277 doi: 10.1186/1471-2105-12-277

ChIPseeqer GUI



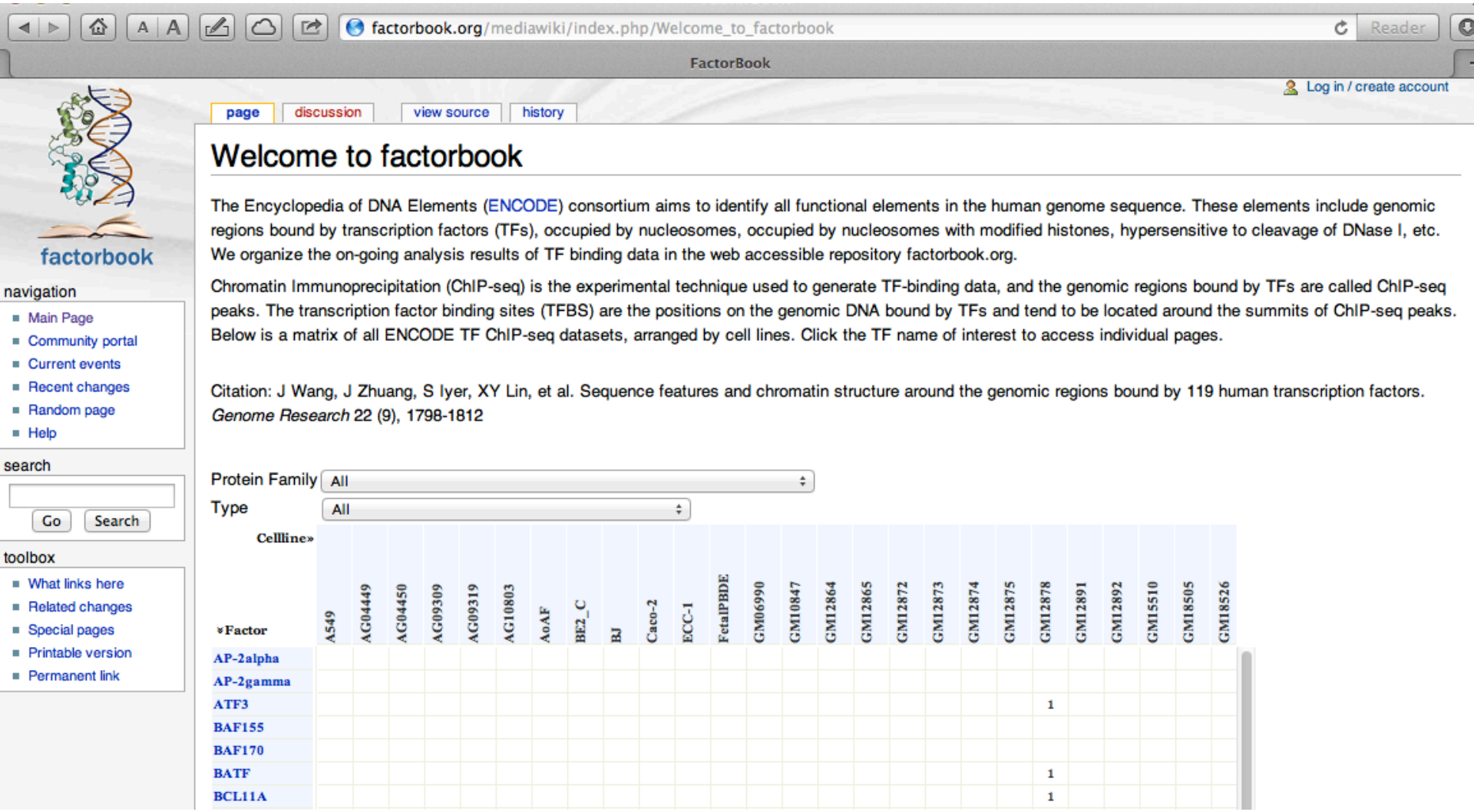






Giannopoulou and Elemento BMC Bioinformatics 2011 12:277 doi: 10.1186/1471-2105-12-277

Factorbook.org (ENCODE ChIP-seq data)



Factorbook (ENCODE) RXR entry

RXRA

retinoid X receptor, alpha

Function

RXRA (retinoid X receptor alpha) is a member of the retinoid X receptor (RXR) and retinoic acid receptor (RAR) steroid and thyroid hormone nuclear receptor superfamily of transcriptional regulators. There are three RXR subtypes: RXRalpha (expressed in the liver, kidneys, epidermis, and intestines); RXRbeta (widely distributed); and RXRgamma (restricted to muscle, the pituitary gland, and certain parts of the brain). RAR/RXR heterodimers bind to the retinoic acid response element (RARE) in the promoter region of target genes, which is composed of tandem 5'-AGGTCA-3' sites. In the absence of ligand, RXR-RAR heterodimers associate with a multiprotein complex containing transcriptional corepressors that induce chromatin condensation. On ligand binding, the corepressors dissociate, and the RXR-RAR heterodimers associate with transcriptional coactivators. Common binding partners for RXR-RAR include PPARA (for transcriptional activity on fatty acid oxidation genes such as ACOX1 and the P450 system genes), liver X receptors, and vitamin D receptors.

ENCODE ChIP-seq Datasets

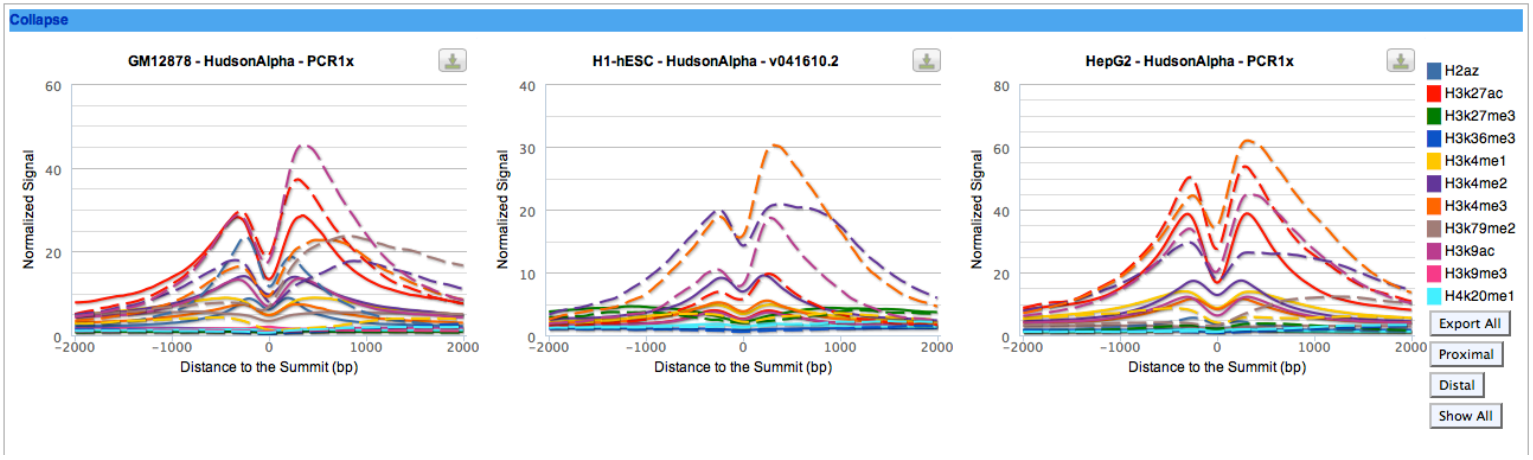
Indicated in the matrix are the numbers of datasets specified by lab and cell line; when the number is greater than 1, multiple ChIP-seq experiments have been performed, some upon treatments. Click the numbers to download the data files on ChIP-seq peaks, alignments, etc.

+/-	HudsonAlpha
GM12878	1
H1-hESC	1
HepG2	1

Average Profiles of Modified Histones around the Summit of ChIP-seq Peaks

Average histone modification profiles are shown for the [-2 kb, +2 kb] window around the summits of TF ChIP-seq peaks, separately for peaks that are proximal ([-1kb, +1kb]) to an annotated transcript (dashed lines) start sites and for peaks that are distal (more than 1kb) to all annotated transcripts (solid lines) start sites. Proximal profiles are arranged such that the transcriptional direction of the nearest transcript is toward the right. Histone modification data were generated by the Broad team, using antibodies to pull down modified histones followed by deep sequencing of the genomic DNA associated with the modified histones. Only histone modification data from the same cell line as the TF ChIP-seq data are shown.

Mouse over a curve to reveal its identity. Mouse over a histone modification in the legend to show its curves and gray out other histone modifications in the figures. Click a histone modification in the legend to toggle on/off its curve in all figures. Click the "Proximal" or "Distal" button in the legend to show only the average histone modification profiles anchored around ChIP-seq peaks that are proximal or distal to annotated transcripts.

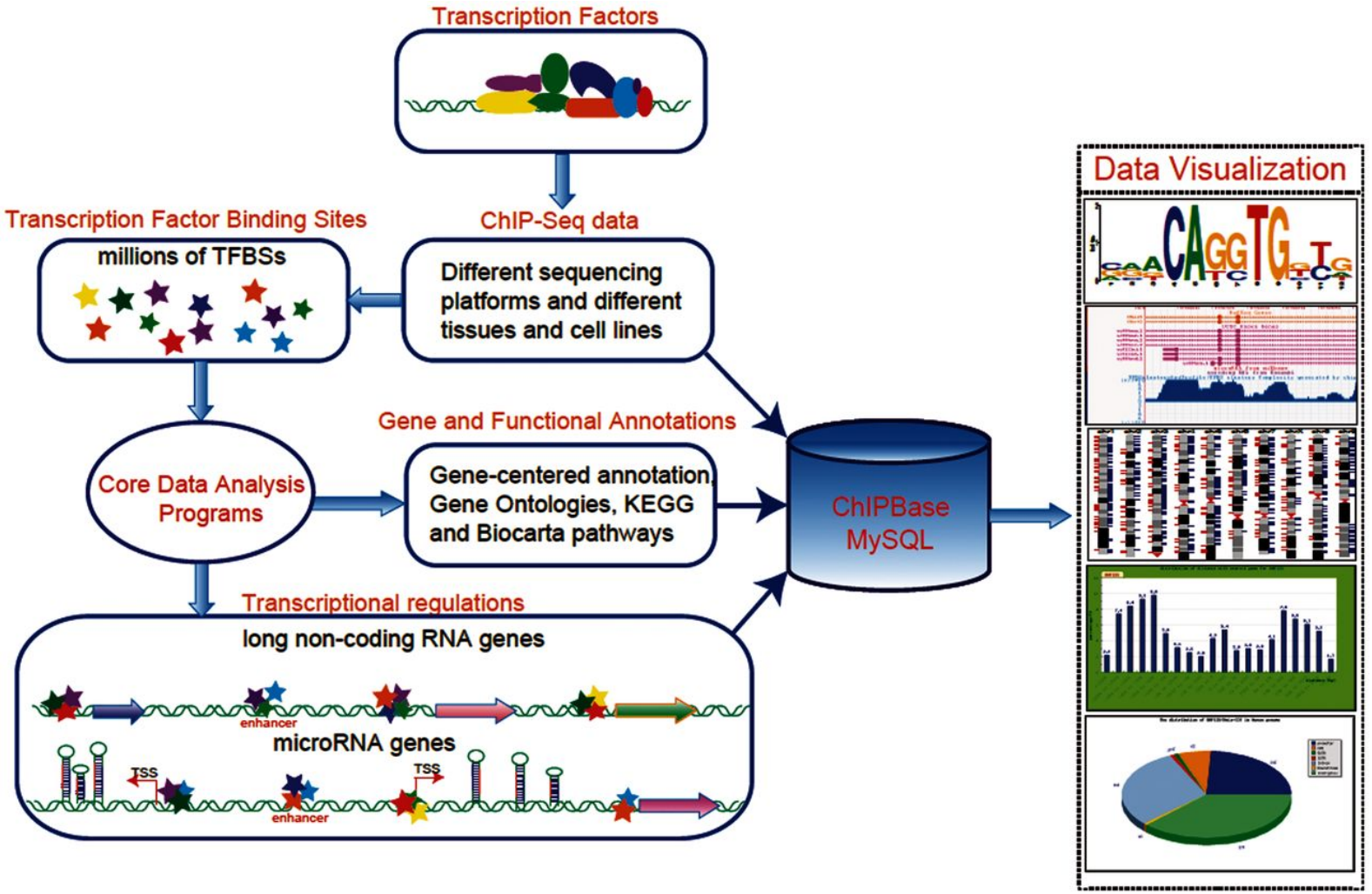


RXRA

PDB ID : 1DSZ

PDB	1DSZ 1RXR 2P1T 2P1V 3E00 3FAL 1FBY 1R0N 3DZY 3H0A 1LBD 1XV9 1YNW 2ACL 3FC6 1BY4 1FM6 2ZY0 3FUG 3OAP 1G1U 1MVC 1XVP 1G5Y 1K74 1XLS 2ZXZ 3DZU 1MZN 2NLL 2P1U 3OZJ 1FM9 1MV9 1RDT 3E94 3KWY
HGNC(alias)	RXRA (NR2B1)
Gene Card	RXRA
Entrez	6256
RefSeq	NM_002957
UniProt	P19793
UCSC	Browser view
Wikipedia	RXRA
Protein Family	Glucocorticoid receptor-like (DNA-binding domain)
Type	Pol II TF
Ensembl Exp.	Human

A system-level overview of the core framework of ChIPBase.



Yang J et al. Nucl. Acids Res. 2012;nar.gks1060

Nucleic Acids Research

Summary of ChIP-seq analysis

- Having a good ChIP is the most important for the analysis
- There are many alternatives for the analysis pipeline
- Analysis speed, percentage of mapped reads etc. are not the most important parameters during the analysis
- Read number influences strongly the peak number
  - Our practice is that for a TF ChIP 6-10 million of reads (depends on the expected number of peaks) is a minimum but it could be sufficient (15-20 million of reads at the histone ChIPs) as well
- Having controls and parallels is good, but not always necessary
- The primary analysis can be run easily from scripts or with workflows
- The most difficult and time-consuming part of the process is the downstream analysis

Center for Clinical GENOMICS and Personalized Medicine

Acknowledgments

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

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

Bioinformatics team:

- Gergely Nagy — PhD student
- Attila Horváth — System administrator, R programmer
- Dávid Jónás — Bioinformatician
- László Steiner — Mathematician
- Erik Czipa — MSc 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