

Chromatin Immunoprecipitation- sequencing (ChIP-seq) Data Analysis

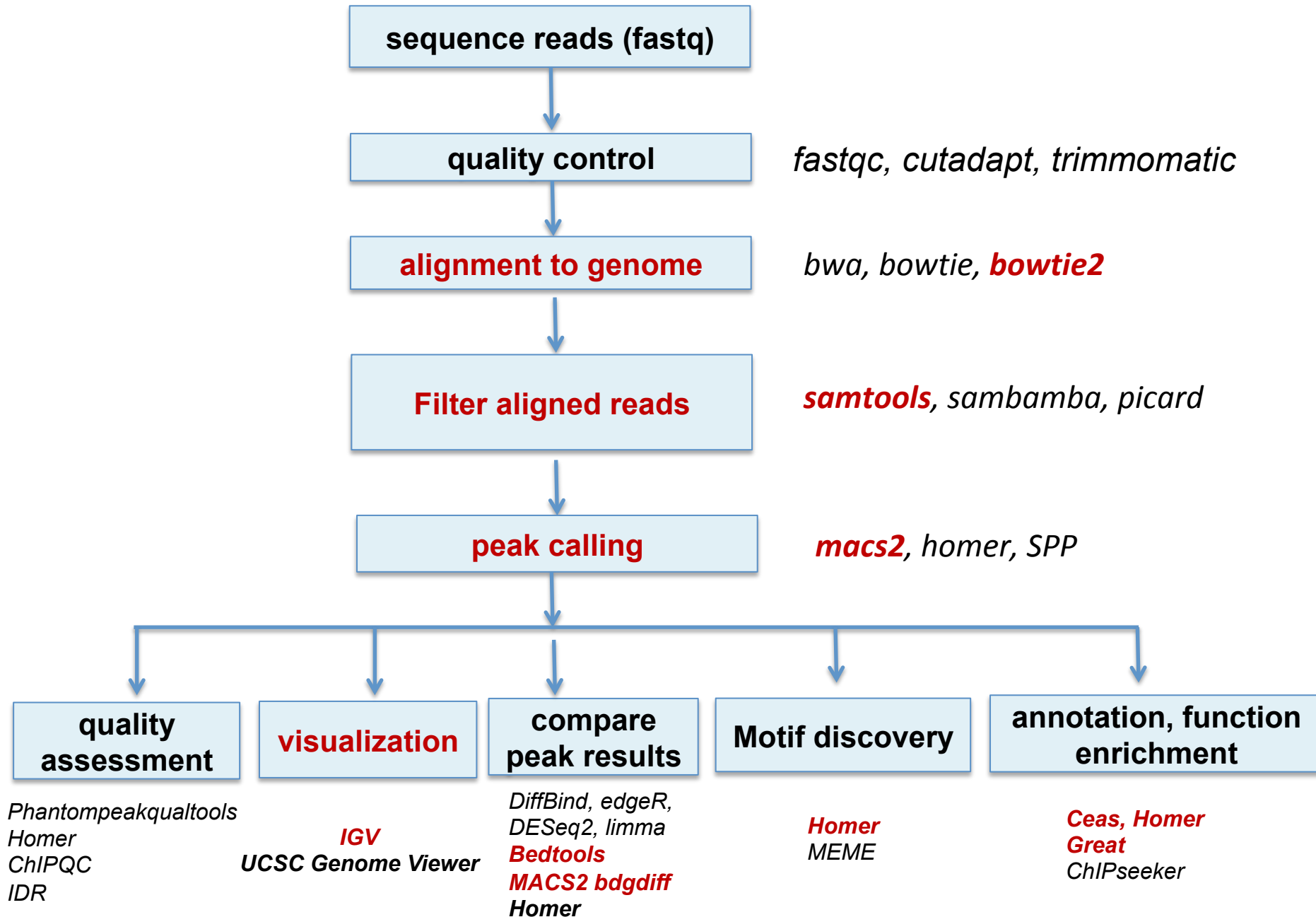
Weihong Yan

wyan@chem.ucla.edu

ChIP-seq Data Analysis

- ❑ Day 1: ChIP-seq Background and Concept
 - ChIP-seq Protocol
 - Quality Control and Guidelines
 - ChIP-seq data from sequence read archive (SRA)
- ❑ Day 2: ChIP-seq data analysis workflow
 - Bowtie2 Alignment
 - MACS2 Peak calling
 - IGV genome visualization
- ❑ Day 3: Peak Annotation and Functional Analysis
 - CEAS: Cis-regulatory element annotation system
 - BEDTools for genome arithmetic
 - HOMER peak annotation, functional and motif analysis
 - GREAT

ChIP-seq Data Analysis Workflow



http://en.wikipedia.org/wiki/FASTQ_format

- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a [FASTA](#) title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT
+
!''*(((((***+)))))(((((.1***-+'''))**55CCF>>>>>CCCCCCCC65
```

! " # \$ % & ' () * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J K L M N O P Q R S T U V W X Y Z [\] ^ _ ` a b c d e f g h i j k l m n o p q r s t u v w x y z { | } ~

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36  
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC  
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36  
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```


Align Reads to Reference Genome

BWA performs gapped alignments and can be used to detect indels and SNPs. BWA is generally used for DNA projects

Bowtie doesn't perform gapped alignments. It runs faster and requires smaller memory footprint.

Bowtie2 is fast and can perform local and gapped alignment. It performs better for reads longer than 50bp.

Bowtie 2

<https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml>



Bowtie 2

Fast and sensitive read alignment



JOHNS HOPKINS
UNIVERSITY

The bowtie2 aligner

bowtie2 takes a Bowtie 2 index and a set of sequencing read files and outputs a set of alignments in SAM format.

"Alignment" is the process by which we discover how and where the read sequences are similar to the reference sequence. An "alignment" is a result from this process, specifically: an alignment is a way of "lining up" some or all of the characters in the read with some characters from the reference in a way that reveals how they're similar. For example:

```
Read:      GACTGGGCGATCTCGACTTCG
           ||||  ||| ||||| |||
Reference: GACTG--CGATCTCGACATCG
```

Local alignment example

The following is a "local" alignment because some of the characters at the ends of the read do not participate. In this case, 4 characters are omitted (or "soft trimmed" or "soft clipped") from the beginning and 3 characters are omitted from the end. This sort of alignment can be produced by Bowtie 2 only in local mode.

```
Read:      ACGGTTGCGTTAATCCGCCACG
Reference: TAACTTGCGTTAAATCCGCCTGG
```

Alignment:

```
Read:      ACGGTTGCGTTAA-TCCGCCACG
           ||||| |||
Reference: TAACTTGCGTTAAATCCGCCTGG
```

Bowtie 2 Alignment Program

```
[wyman@n6431 data]$ bowtie2 --help
Bowtie 2 version 2.4.2 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r> | --interleaved <i> | -b <bam>} [-S <sam>]

<bt2-idx>  Index filename prefix (minus trailing .x.bt2).
          NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
<m1>       Files with #1 mates, paired with files in <m2>.
          Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
<m2>       Files with #2 mates, paired with files in <m1>.
          Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
<r>        Files with unpaired reads.
          Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
<i>        Files with interleaved paired-end FASTQ/FASTA reads.
          Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
<bam>      Files are unaligned BAM sorted by read name.
<sam>      File for SAM output (default: stdout)

<m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.

Options (defaults in parentheses):

Input:
  -q          query input files are FASTQ .fq/.fastq (default)
  --tab5      query input files are TAB5 .tab5
  --tab6      query input files are TAB6 .tab6
  --qseq      query input files are in Illumina's qseq format
  -f          query input files are (multi-)FASTA .fa/.mfa
  -r          query input files are raw one-sequence-per-line
  -F k:<int>,i:<int> query input files are continuous FASTA where reads
          are substrings (k-mers) extracted from a FASTA file <s>
          and aligned at offsets 1, 1+i, 1+2i, ... end of reference
          <m1>, <m2>, <r> are sequences themselves, not files
  -c          skip the first <int> reads/pairs in the input (none)
  -s/--skip <int> stop after first <int> reads/pairs (no limit)
  -u/--upto <int> trim <int> bases from 5'/left end of reads (0)
  -5/--trim5 <int> trim <int> bases from 5'/left end of reads (0)
  -3/--trim3 <int> trim <int> bases from 3'/right end of reads (0)
  --trim-to [3;5]<int> trim reads exceeding <int> bases from either 3' or 5' end
          If the read end is not specified then it defaults to 3 (0)
  --phred33    qualities are Phred+33 (default)
  --phred64    qualities are Phred+64
  --int-quals   qualities encoded as space-delimited integers

Presets:
  Same as:
  For --end-to-end:
  --very-fast    -D 5 -R 1 -N 0 -L 22 -i S,0,2.50
  --fast         -D 10 -R 2 -N 0 -L 22 -i S,0,2.50
  --sensitive    -D 15 -R 2 -N 0 -L 22 -i S,1,1.15 (default)
  --very-sensitive -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

For --local:
  --very-fast-local -D 5 -R 1 -N 0 -L 25 -i S,1,2.00
  --fast-local     -D 10 -R 2 -N 0 -L 22 -i S,1,1.75
  --sensitive-local -D 15 -R 2 -N 0 -L 20 -i S,1,0.75 (default)
  --very-sensitive-local -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

Alignment:
  -N <int>    max # mismatches in seed alignment; can be 0 or 1 (0)
  -L <int>    length of seed substrings; must be >3, <32 (22)
  -i <int>    interval between seed substrings w/r/t read len (S,1,1.15)
  --n-unc    func for max # non-A/C/G/Ts permitted in aln (L,0,0.15)
  --dpad <int> include <int> extra ref chars on sides of DP table (15)
  --gbar <int> disallow gaps within <int> nucs of read extremes (4)
  --ignore-quals treat all quality values as 30 on Phred scale (off)
  --nowf       do not align forward (original) version of read (off)
  --norc       do not align reverse-complement version of read (off)
  --no-1mm-upfront do not allow 1 mismatch alignments before attempting to
          scan for the optimal seeded alignments
  --end-to-end  entire read must align; no clipping (on)
  OR
  --local       local alignment; ends might be soft clipped (off)
```

```
Scoring:
  --ma <int>    match bonus (0 for --end-to-end, 2 for --local)
  --mp <int>    max penalty for mismatch; lower qual = lower penalty (6)
  --np <int>    penalty for non-A/C/G/Ts in read/ref (1)
  --rdg <int>,<int> read gap open, extend penalties (5,3)
  --rfg <int>,<int> reference gap open, extend penalties (5,3)
  --score-min <func> min acceptable alignment score w/r/t read length
          (G,20,8 for local, L,-0.6,-0.6 for end-to-end)

Reporting:
  (default)      look for multiple alignments, report best, with MAPQ
  OR
  -k <int>       report up to <int> alns per read; MAPQ not meaningful
  OR
  -a/--all       report all alignments; very slow, MAPQ not meaningful

Effort:
  -D <int>       give up extending after <int> failed extends in a row (15)
  -R <int>       for reads w/ repetitive seeds, try <int> sets of seeds (2)

Paired-end:
  -I/--minins <int> minimum fragment length (0)
  -X/--maxins <int> maximum fragment length (500)
  --fr/--rf/--ff    -1, -2 mates align fw/rev, rev/fw, fw/fw (--fr)
  --no-mixed        suppress unpaired alignments for paired reads
  --no-discordant   suppress discordant alignments for paired reads
  --dovetail        concordant when mates extend past each other
  --no-contain      not concordant when one mate alignment contains other
  --no-overlap      not concordant when mates overlap at all

BAM:
  --align-paired-reads
          Bowtie2 will, by default, attempt to align unpaired BAM reads.
          Use this option to align paired-end reads instead.
  --preserve-tags   Preserve tags from the original BAM record by
          appending them to the end of the corresponding SAM output.

Output:
  -t/--time        print wall-clock time taken by search phases
  --un <path>      write unpaired reads that didn't align to <path>
  --al <path>      write unpaired reads that aligned at least once to <path>
  --un-conc <path> write pairs that didn't align concordantly to <path>
  --al-conc <path> write pairs that aligned concordantly at least once to <path>
          (Note: for --un, --al, --un-conc, or --al-conc, add '-gz' to the option name, e.g.
          --un-gz <path>, to gzip compress output, or add '-bz2' to bzip2 compress output.)
  --quiet          print nothing to stderr except serious errors
  --met-file <path> send metrics to file at <path> (off)
  --met-stderr     send metrics to stderr (off)
  --met <int>      report internal counters & metrics every <int> secs (1)
  --no-unal        suppress SAM records for unaligned reads
  --no-head        suppress header lines, i.e. lines starting with @
  --no-sq          suppress @SQ header lines
  --rg-id <text>   set read group id, reflected in @RG line and RG:Z: opt field
  --rg <text>      add <text> ("lab:val") to @RG line of SAM header.
          Note: @RG line only printed when --rg-id is set.
  --omit-sec-seq   put '*' in SEQ and QUAL fields for secondary alignments.
  --sam-no-qname-trunc
          Suppress standard behavior of truncating readname at first whitespace
          at the expense of generating non-standard SAM.
  --xq            Use '='/'X', instead of 'M', to specify matches/mismatches in SAM record.
  --soft-clipped-unmapped-tlen
          Exclude soft-clipped bases when reporting TLEN
  --sam-append-comment
          Append FASTA/FASTQ comment to SAM record

Performance:
  --p/--threads <int> number of alignment threads to launch (1)
  --reorder        force SAM output order to match order of input reads
  --mm            use memory-mapped I/O for index; many 'bowtie's can share

Other:
  --qc-filter      filter out reads that are bad according to QSEQ filter
```

Bowtie2 Sam Output

    <https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#sam-output>

Click to go back, hold to see history

- Inferred fragment length. Size is negative if the mate's alignment occurs upstream of this alignment. Size is 0 if the mates did not align concordantly. However, size is non-0 if the mates aligned discordantly to the same chromosome.
- Read sequence (reverse-complemented if aligned to the reverse strand)
- ASCII-encoded read qualities (reverse-complemented if the read aligned to the reverse strand). The encoded quality values are on the [Phred quality scale](#) and the encoding is ASCII-offset by 33 (ASCII char !), similarly to a [FASTQ](#) file.
- Optional fields. Fields are tab-separated. `bowtie2` outputs zero or more of these optional fields for each alignment, depending on the type of the alignment:

AS:i:<N>

Alignment score. Can be negative. Can be greater than 0 in `--local` mode (but not in `--end-to-end` mode). Only present if SAM record is for an aligned read.



AS:i:<N>

Alignment score for the best-scoring alignment found other than the alignment reported. Can be negative. Can be greater than 0 in `--local` mode (but not in `--end-to-end` mode). Only present if the SAM record is for an aligned read and more than one alignment was found for the read. Note that, when the read is part of a concordantly-aligned pair, this score could be greater than `AS:i`.

YS:i:<N>

Alignment score for opposite mate in the paired-end alignment. Only present if the SAM record is for a read that aligned as part of a paired-end alignment.

XN:i:<N>

The number of ambiguous bases in the reference covering this alignment. Only present if SAM record is for an aligned read.

XM:i:<N>

The number of mismatches in the alignment. Only present if SAM record is for an aligned read.

XO:i:<N>

The number of gap opens, for both read and reference gaps, in the alignment. Only present if SAM record is for an aligned read.

XG:i:<N>

The number of gap extensions, for both read and reference gaps, in the alignment. Only present if SAM record is for an aligned read.

NM:i:<N>

The edit distance; that is, the minimal number of one-nucleotide edits (substitutions, insertions and deletions) needed to transform the read string into the reference string. Only present if SAM record is for an aligned read.

YF:Z:<S>

String indicating reason why the read was filtered out. See also: [Filtering](#). Only appears for reads that were filtered out.

YT:Z:<S>

Value of `UU` indicates the read was not part of a pair. Value of `CP` indicates the read was part of a pair and the pair aligned concordantly. Value of `DP` indicates the read was part of a pair and the pair aligned discordantly. Value of `UP` indicates the read was part of a pair but the pair failed to align either concordantly or discordantly.

MD:Z:<S>

A string representation of the mismatched reference bases in the alignment. See [SAM Tags format specification](#) for details. Only present if SAM record is for an aligned read.



Bowtie 2

Fast and sensitive read alignment



JOHNS HOPKINS
UNIVERSITY

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.



Indexes		
H. sapiens, GRCh38 no-alt analysis set	NCBI	
H. sapiens, GRCh38 no-alt +decoy set	NCBI	
H. sapiens, GRCh38 + major SNVs	NCBI+1KG	
How we built this, FASTA		
H. sapiens, GRCh37	NCBI	
H. sapiens, Ash1v1.7	JHU	
H. sapiens, Ash1v2.0	JHU	
H. sapiens, CHM13 v2	T2T	
H. sapiens, CHM13 v1, plus Y	T2T	
H. sapiens, ...		
H. sapiens, ...		
M. Musculus		
M. Musculus		
M. Musculus		
P. troglodytes		
P. troglodytes		
M. mulatta		
B. taurus		
S. scrofa		
C. familiaris		
C. familiaris		
R. norvegicus		
R. norvegicus		
G. gallus, GRCg6a	NCBI	
G. gallus, Galgal4	Ensembl	

Download hg19 bowtie2 index to hoffman2

```
% cd ~/myscratch/workshop
```

```
% wget https://genome-index.s3.amazonaws.com/bt/hg19.zip
```

```
% unzip hg19.zip
```

```
[wyann7430 workshop]$ cd ~/myscratch/workshop
[wyann7430 workshop]$ wget https://genome-index.s3.amazonaws.com/bt/hg19.zip
--2023-10-26 00:06:35-- https://genome-index.s3.amazonaws.com/bt/hg19.zip
Resolving genome-index.s3.amazonaws.com (genome-index.s3.amazonaws.com)... 52.217.114.185, 52.216.208.49, 52.217.85.68, ...
Connecting to genome-index.s3.amazonaws.com (genome-index.s3.amazonaws.com)|52.217.114.185|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3694403333 (3.4G) [application/zip]
Saving to: 'hg19.zip'

100%[=====]

2023-10-26 00:12:32 (9.95 MB/s) - 'hg19.zip' saved [3694403333/3694403333]

[wyann7430 workshop]$ unzip hg19.zip
Archive: hg19.zip
  inflating: hg19.1.bt2
  inflating: hg19.2.bt2
  inflating: hg19.3.bt2
  inflating: hg19.4.bt2
  inflating: hg19.rev.1.bt2
  inflating: hg19.rev.2.bt2
  inflating: make_hg19.sh
```

ChIP-seq Read Alignment and Filtering

- Run bowtie2 in hoffman2 cluster
- Filter reads aligned to chrM, unlocalized (random) and unplaced (chrUn_) chromosomes and reads aligned to genome more than once (reads contain “XS” tag in the sam output)
- Use samtools program to:
 - convert and sort bowtie2 output to bam files

Bowtie2 Alignment

Run bowtie2 in hoffman2 cluster

```
% qcrsh -l h_rt=4:00:00,h_data=4G -pe shared 4
% cd ~/myscratch/workshop
% module load bowtie2
% bowtie2 -p 4 -x hg19 -q INP_30m_IFNa.fastq --local --no-unal -k 1 -S INP_30m_IFNa.sam &>INP_30m_IFNa.log
% bowtie2 -p 4 -x hg19 -q INP_6h_IFNa.fastq --local --no-unal -k 1 -S INP_6h_IFNa.sam &>INP_6h_IFNa.log
% bowtie2 -p 4 -x hg19 -q STAT1_30m_IFNa.fastq --local --no-unal -k 1 -S STAT1_30m_IFNa.sam &>STAT1_30m_IFNa.log
% bowtie2 -p 4 -x hg19 -q STAT1_6h_IFNa.fastq --local --no-unal -k 1 -S STAT1_6h_IFNa.sam &>STAT1_6h_IFNa.log
```

```
[wyang@n7430 workshop]$ bowtie2 -p 4 -x hg19 -q INP_30m_IFNa.fastq --local --no-unal -k 1 -S INP_30m_IFNa.sam &>INP_30m_IFNa.log
[wyang@n7430 workshop]$ bowtie2 -p 4 -x hg19 -q INP_6h_IFNa.fastq --local --no-unal -k 1 -S INP_6h_IFNa.sam &>INP_6h_IFNa.log
[wyang@n7430 workshop]$ bowtie2 -p 4 -x hg19 -q STAT1_30m_IFNa.fastq --local --no-unal -k 1 -S STAT1_30m_IFNa.sam &>STAT1_30m_IFNa.log
[wyang@n7430 workshop]$ bowtie2 -p 4 -x hg19 -q STAT1_6h_IFNa.fastq --local --no-unal -k 1 -S STAT1_6h_IFNa.sam &>STAT1_6h_IFNa.log
```

The above commands can also be run in a shell script:

```
% cp /u/scratch/w/wyang/workshop/bowtie2Align.sh ~/myscratch/workshop
% cd ~/myscratch/workshop
% ./bowtie2Align.sh INP_30m_IFNa
% ./bowtie2Align.sh INP_6h_IFNa
% ./bowtie2Align.sh STAT1_30m_IFNa
% ./bowtie2Align.sh STAT1_6h_IFNa
```

```
[wyang@n7430 workshop]$ more bowtie2Align.sh
#!/bin/bash

module load bowtie2

bowtie2 -p 4 -x hg19 -q $1.fastq --local --no-unal -k 1 -S $1.sam &>$1.log
```

Bowtie2 Alignment Output

```
% bowtie2 -p 4 -x hg19 -q INP_30m_IFNa.fastq --local --no-unal -k 1 -S INP_30m_IFNa.sam &>INP_30m_IFNa.log
```

INP_30m_IFNa.sam: alignment result in sam format

INP_30m_IFNa.log: summary of alignment result (mapping rate)

SRR502228.213	0	chr10	32190285	255	27M	*	0	0	CCCCCTCTCTACTAAAAACAAAAA	B7=>B<ABABBCABA<ACBB@B@?	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.214	16	chr16	67375980	255	27M	*	0	0	GCCTTAAGTAAAGTTATATATTTCTATG	?.:@BB@ABBBABABBBB@;@BBB	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.215	16	chr10	42596894	255	27M	*	0	0	GAATGGAATCATCTTCTAATGGAAAGG	;AB@?AB<?;5?;8=BB@BBBCBBC	AS:i:48	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:16A10	YT:Z:UU
SRR502228.216	16	chr3	30312454	255	27M	*	0	0	TTCCACAGAATGACITTTTCTTCTTGAA	9.+6?A@BB@<9B?=@<<:1?>A@<BC	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.217	0	chr4	92106344	255	27M	*	0	0	TGTTTATTTTGTGCTTCCAAATGCTCCT	=B74>AB757678BBAAABBB<?>A;	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.219	16	chr5	79094001	255	27M	*	0	0	TCCCTAAATACTTGCACCTTGAOCITC	=?/5-?BB>A?>5B<-2?74?30?BB>	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.220	16	chr8	59539612	255	27M	*	0	0	TTATTGTGTTTTTGTGCTGCTGATTTGT	?0?AA@0B0(AAA)AA0B0BB7B	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.222	0	chr7	139949826	255	27M	*	0	0	GTTGGCCAGGCTGCTCTCAACCTCTG	CB07?<@<<?BB?>=@BBB?A6?B5	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.223	0	chr1	186928049	255	27M	*	0	0	TTTAGGATGTACACTAAATAATTTATTT	BCBC0>@BBB?BAB>9@BBA?ABA?	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.193	16	chr1	211803534	255	27M	*	0	0	TTCAGAAATCATGGAGGAGTCGTGTGG	@AABCCC;?CCCCCCCC@BACCCC?	AS:i:50	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:9C17	YT:Z:UU
SRR502228.194	16	chr6	110908704	255	27M	*	0	0	TAAAAAAGGAGAAAGGGGATATCACCA	@A>CB@;/?020CA7>BBB>>BCBB@	AS:i:49	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:7T19	YT:Z:UU
SRR502228.195	16	chr2	37597468	255	27M	*	0	0	TATATTTTGAATGGAACATAAAAAA	@5*/95B@B<C7BAB>6+++B0?CAB@	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.196	16	chr22	30095708	255	27M	*	0	0	GTCCGGCAAGAATATGAAAAAGTTTAA	>0..3?*>B?;@>B@;B<A0>>C>B;	AS:i:54	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:27	YT:Z:UU
SRR502228.198	16	chr7	50440532	255	27M	*	0	0	TTTCGAATAGGAGTTTGAACATCCA	AB:03??;B@<09A@:2?BB;BBCCB	AS:i:47	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:6C20	YT:Z:UU

<https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#sam-output>

```
[wyan@n7047 workshop]$ more INP_30m_IFNa.log
26699669 reads; of these:
  26699669 (100.00%) were unpaired; of these:
    6910835 (25.88%) aligned 0 times
    19788834 (74.12%) aligned exactly 1 time
    0 (0.00%) aligned >1 times
74.12% overall alignment rate
```


Sequence Alignment/Map (SAM) Tools

Not Secure http://www.htslib.org



Samtools

Home

Download ▾

Workflows ▾

Documentation ▾

Support ▾

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

Samtools	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
BCFtools	Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
HTSlib	A C library for reading/writing high-throughput sequencing data

```
[lwyang@n6038 data]$ samtools
Program: samtools (Tools for alignments in the SAM format)
Version: 1.15 (using htslib 1.15)

Usage:  samtools <command> [options]

Commands:
  -- Indexing
    dict      create a sequence dictionary file
    faidx     index/extract FASTA
    fqidx     index/extract FASTQ
    index     index alignment

  -- Editing
    calmd     recalculate MD/NM tags and '=' bases
    fixmate   fix mate information
    reheader  replace BAM header
    targetcut cut fosmid regions (for fosmid pool only)
    addreplacerg adds or replaces RG tags
    markdup   mark duplicates
    ampliconclip clip oligos from the end of reads

  -- File operations
    collate   shuffle and group alignments by name
    cat       concatenate BAMs
    consensus produce a consensus Pileup/FASTA/FASTQ
    merge     merge sorted alignments
    mpileup   multi-way pileup
    sort      sort alignment file
    split     splits a file by read group
    quickcheck quickly check if SAM/BAM/CRAM file appears intact
    fastq     converts a BAM to a FASTQ
    fasta     converts a BAM to a FASTA
```

```
fastq      converts a BAM to a FASTQ
fasta      converts a BAM to a FASTA
import     Converts FASTA or FASTQ files to SAM/BAM/CRAM

-- Statistics
  bedcov    read depth per BED region
  coverage  alignment depth and percent coverage
  depth     compute the depth
  flagstat  simple stats
  idxstats  BAM index stats
  phase     phase heterozygotes
  stats     generate stats (former bamcheck)
  ampliconstats generate amplicon specific stats

-- Viewing
  flags     explain BAM flags
  head      header viewer
  tview     text alignment viewer
  view      SAM<->BAM<->CRAM conversion
  depad     convert padded BAM to unpadded BAM
  samples   list the samples in a set of SAM/BAM/CRAM files

-- Misc
  help [cmd] display this help message or help for [cmd]
  version   detailed version information
```

Filter Aligned Reads and Convert Alignment Format

Filter non unique mapped reads (read with XS tag in the sam output), reads aligned to chrM, unlocalized (random) and unplaced (chrUn_) regions

```
% module load samtools
```

```
% cat INP_30m_IFNa.sam |grep -v 'chrM\|chrUn\|random\|XS' >INP_30m_IFNa_filter.sam
```

Use samtools program to perform “sam to bam” conversion and then sort bam files

```
% samtools view -@ 4 -bS INP_30m_IFNa_filter.sam |samtools sort -@ 4 - >INP_30m_IFNa.bam
```

```
[wyann@n1976 workshop]$ module load samtools
[wyann@n1976 workshop]$ cat INP_30m_IFNa.sam |grep -v 'chrM\|chrUn\|random\|XS' >INP_30m_IFNa_filter.sam
[wyann@n1976 workshop]$ samtools view -@ 4 -bS INP_30m_IFNa_filter.sam |samtools sort -@ 4 - >INP_30m_IFNa.bam
[bam_sort_core] merging from 4 files and 4 in-memory blocks...
```

Or Combine above commands into a shell script and run for each sample

```
% cp /u/scratch/w/wyann/workshop/processSam.sh ~/myscratch/workshop
```

```
% ./processSam.sh INP_30m_IFNa
```

```
% ./processSam.sh INP_6h_IFNa
```

```
% ./processSam.sh STAT1_30m_IFNa
```

```
% ./processSam.sh STAT1_6h_IFNa
```

```
[wyann@n7047 workshop]$ more processSam.sh
#!/bin/bash

module load samtools

cat $1.sam |grep -v 'chrM\|Un\|random\|XS' >$1_filter.sam

samtools view -@ 4 -bS $1_filter.sam |samtools sort -@ 4 - >$1.bam
```

```
[wyann@n7047 workshop]$ ./processSam.sh INP_30m_IFNa
[bam_sort_core] merging from 4 files and 4 in-memory blocks...
[wyann@n7047 workshop]$ ./processSam.sh INP_6h_IFNa
[bam_sort_core] merging from 4 files and 4 in-memory blocks...
[wyann@n7047 workshop]$ ./processSam.sh STAT1_30m_IFNa
[bam_sort_core] merging from 0 files and 4 in-memory blocks...
[wyann@n7047 workshop]$ ./processSam.sh STAT1_6h_IFNa
[bam_sort_core] merging from 0 files and 4 in-memory blocks...
```

Normalize aligned reads for equal coverage

1. Count number of reads for each sample

```
% module load samtools
```

```
% samtools view -c INP_30m_IFNa.bam
```

```
% samtools view -c INP_6h_IFNa.bam
```

```
% samtools view -c STAT1_30m_IFNa.bam
```

```
% samtools view -c STAT1_6h_IFNa.bam
```

```
[[wyan@n7047 workshop]$ samtools view -c INP_30m_IFNa.bam
19041856
[[wyan@n7047 workshop]$ samtools view -c INP_6h_IFNa.bam
19660692
[[wyan@n7047 workshop]$ samtools view -c STAT1_30m_IFNa.bam
10778033
[[wyan@n7047 workshop]$ samtools view -c STAT1_6h_IFNa.bam
11019174
```

2. Subsample input controls to 11million reads

```
% samtools view -s 0.58 INP_30m_IFNa.bam -b >INP_30m_IFNa_11e6.bam
```

```
% samtools view -s 0.58 INP_6h_IFNa.bam -b >INP_6h_IFNa_11e6.bam
```

```
% samtools view -c INP_30m_IFNa_11e6.bam
```

```
% samtools view -c INP_6h_IFNa_11e6.bam
```

```
[[wyan@n7047 workshop]$ samtools view -s 0.58 INP_30m_IFNa.bam -b >INP_30m_IFNa_11e6.bam
[[wyan@n7047 workshop]$ samtools view -s 0.58 INP_6h_IFNa.bam -b >INP_6h_IFNa_11e6.bam
[[wyan@n7047 workshop]$ samtools view -c INP_30m_IFNa_11e6.bam
11044013
[[wyan@n7047 workshop]$ samtools view -c INP_6h_IFNa_11e6.bam
11405138
```

Identifying ChIP-seq enrichment using MACS

Jianxing Feng^{1,3}, Tao Liu^{2,3}, Bo Qin¹, Yong Zhang¹ & Xiaole Shirley Liu²

¹Department of Bioinformatics, School of Life Sciences and Technology, Tongji University, Shanghai, China. ²Department of Biostatistics and Computational Biology, Harvard School of Public Health, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. ³These authors contributed equally to this work. Correspondence should be addressed to Y.Z. (yzhang@tongji.edu.cn) and X.S.L. (xsliu@jimmy.harvard.edu).

Published online 30 August 2012; doi:10.1038/nprot.2012.101

Model-based analysis of ChIP-seq (MACS) is a computational algorithm that identifies genome-wide locations of transcription/ chromatin factor binding or histone modification from ChIP-seq data. MACS consists of four steps: removing redundant reads, adjusting read position, calculating peak enrichment and estimating the empirical false discovery rate (FDR). In this protocol, we provide a detailed demonstration of how to install MACS and how to use it to analyze three common types of ChIP-seq data sets with different characteristics: the sequence-specific transcription factor FoxA1, the histone modification mark H3K4me3 with sharp enrichment and the H3K36me3 mark with broad enrichment. We also explain how to interpret and visualize the results of MACS analyses. The algorithm requires ~3 GB of RAM and 1.5 h of computing time to analyze a ChIP-seq data set containing 30 million reads, an estimate that increases with sequence coverage. MACS is open source and is available from <http://liulab.dfci.harvard.edu/MACS/>.

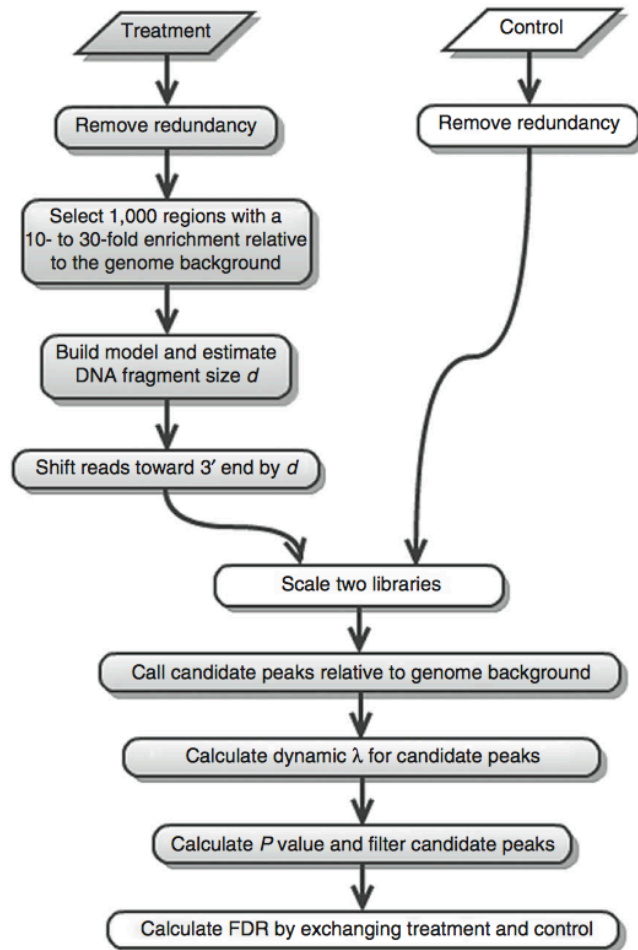
INTRODUCTION

Researchers have widely used the process of ChIP-seq¹ to map transcription factor binding sites and histone modification status on a genome-wide scale². ChIP comprises a few basic steps: cross-linking

version of MACS (1.4.2) to publicly available ChIP-seq data²⁰ on a local computer. MACS is also available at the web-based ChIP-seq analysis portal Cistrome²¹, which provides a complete workflow

<http://www.nature.com/nprot/journal/v7/n9/pdf/nprot.2012.101.pdf>

Model-based Analysis of ChIP-seq (MACS)



- ✓ Removing redundant reads: MACS retains no more than one read per genomic location
- ✓ Building the peak model: Reads that map to the positive and negative strands often appear to the left and right of the protein-DNA interaction location.
- ✓ Calculate peak enrichment using local background normalization

Figure 1 | Workflow of MACS 1.4.2. If the control sample is missing, then the steps shown in white boxes will be skipped (remove redundancy of the control sample, scale two libraries and calculate FDR by exchanging treatment and control).

MACS2 2.2.9.1

`pip install MACS2`



Released: Jul 21, 2023

Model Based Analysis for ChIP-Seq data

Navigation

☰ Project description

Project description

MACS: Model-based Analysis for ChIP-Seq

Install MACS2

% `pip install MACS2`

% `ls $HOME/.local/bin`

macs2 should be in `$HOME/.local/bin`

```
[wyman@login2 bin]$ pip install MACS2
Defaulting to user installation because normal site-packages is not writeable
Collecting MACS2
  Downloading MACS2-2.2.9.1.tar.gz (2.0 MB)
    |████████████████████| 2.0 MB 31.2 MB/s
Installing build dependencies ... done
Getting requirements to build wheel ... done
Preparing wheel metadata ... done
Requirement already satisfied: numpy>=1.19 in /u/local/apps/python/3.9.6/gcc-4.8.5/lib/python3.9/site-packages (from MACS2) (1.22.2)
Requirement already satisfied: Cython==0.29 in /u/local/apps/python/3.9.6/gcc-4.8.5/lib/python3.9/site-packages (from MACS2) (0.29.24)
Building wheels for collected packages: MACS2
  Building wheel for MACS2 (PEP 517) ... done
  Created wheel for MACS2: filename=MACS2-2.2.9.1-cp39-cp39-linux_x86_64.whl size=1612660 sha256=1031b07fa30dc8c87118ec644a2bd21571ffa6fe7a71684029ad4fbd7aea65d2
  Stored in directory: /u/home/w/wyman/.cache/pip/wheels/99/ad/d1/103e02a7131c6f7d4702d85ff58830e621b45ef2f77f3977b5
Successfully built MACS2
Installing collected packages: MACS2
Successfully installed MACS2-2.2.9.1
WARNING: You are using pip version 21.2.4; however, version 23.3.1 is available.
You should consider upgrading via the '/u/local/apps/python/3.9.6/gcc-4.8.5/bin/python3.9 -m pip install --upgrade pip' command.
```


MACS2 Peak Calling

```
[wyman@n1874 data]$ $HOME/.local/bin/macs2 -h
usage: macs2 [-h] [--version] {callpeak,bdgpeakcall,bdgbroadcall,bdgcmp,bdgopt,cmbreps,bdgdifff,filterdup,predictd,pileup,randsample,refinepeak} ...

macs2 -- Model-based Analysis for ChIP-Sequencing

positional arguments:
  {callpeak,bdgpeakcall,bdgbroadcall,bdgcmp,bdgopt,cmbreps,bdgdifff,filterdup,predictd,pileup,randsample,refinepeak}
  callpeak              Main MACS2 Function: Call peaks from alignment results.
  bdgpeakcall           Call peaks from bedGraph output. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS2 are acceptable.
  bdgbroadcall          Call broad peaks from bedGraph output. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS2 are acceptable.
  bdgcmp               Deduct noise by comparing two signal tracks in bedGraph. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS2 are acceptable.
  bdgopt               Operations on score column of bedGraph file. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS2 are acceptable.
  cmbreps              Combine BEDGraphs of scores from replicates. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS2 are acceptable.
  bdgdifff             Differential peak detection based on paired four bedgraph files. Note: All regions on the same chromosome in the bedGraph file should be continuous so only bedGraph files from MACS2 are acceptable.
  filterdup            Remove duplicate reads at the same position, then save the rest alignments to BED or BEDPE file. If you use '--keep-dup all option', this script can be utilized to convert any acceptable format into BED or BEDPE format.
  predictd             Predict d or fragment size from alignment results. *Will NOT filter duplicates*
  pileup              Pileup aligned reads with a given extension size (fragment size or d in MACS language). Note there will be no step for duplicate reads filtering or sequencing depth scaling, so you may need to do certain pre/post-processing.
  randsample           Randomly sample number/percentage of total reads.
  refinepeak          (Experimental) Take raw reads alignment, refine peak summits and give scores measuring balance of waston/crick tags. Inspired by SPP.

optional arguments:
  -h, --help            show this help message and exit
  --version             show program's version number and exit
```

```
[wyman@n7047 workshop]$ $HOME/.local/bin/macs2 callpeak -h
usage: macs2 callpeak [-h] -t TFILE [TFILE ...] [-c [CFILE ...]] [-f {AUTO,BAM,SAM,BED,ELAND,ELANDMULTI,ELANDEXPORT,BOWTIE,BAMPE,BEDPE}] [-g GSIZE] [-s TSIZE] [--keep-dup KEEPDUPLICATES] [--outdir OUTDIR]
                        [-n NAME] [-B] [--verbose VERBOSE] [--trackline] [--SPMR] [--nomodel] [--shift SHIFT] [--extsize EXTSIZE] [--bw BW] [--d-min D_MIN] [-m MFOLD MFOLD] [--fix-bimodal] [-q QVALUE] [-p PVALUE]
                        [--scale-to {large,small}] [--down-sample] [--seed SEED] [--tempdir TEMPDIR] [--nolambda] [--slocal SMALLLOCAL] [--llocal LARGELOCAL] [--max-gap MAXGAP] [--min-length MINLEN] [--broad]
                        [--broad-cutoff BROADCUTOFF] [--cutoff-analysis] [--call-summits] [--fe-cutoff FECUTOFF] [--buffer-size BUFFER_SIZE] [--to-large] [--ratio RATIO]

optional arguments:
  -h, --help            show this help message and exit
```

- t ChIP-seq treatment file
- c ChIP-seq control file
- f input file format
- g genome size
- n output file name
- q minimum FDR (q-value) cutoff for peak detection
- B (--bdg) save pileup pileup result into a bedGraph file

MACS2 Peak Calling

Call peaks with macs2 and make model plots

```
% module load python
% module load R
% cd ~/myscratch/workshop
% $HOME/.local/bin/macs2 callpeak -t STAT1_30m_IFNa.bam -c INP_30m_IFNa_11e6.bam -g hs -n STAT1_30m_IFNa -f BAM -q 0.05 --bdg
% $HOME/.local/bin/macs2 callpeak -t STAT1_6h_IFNa.bam -c INP_6h_IFNa_11e6.bam -g hs -n STAT1_6h_IFNa -f BAM -q 0.05 --bdg
% Rscript STAT1_30m_IFNa_model.r
% Rscript STAT1_6h_IFNa_model.r
```

```
[wyman@login2 workshop]$ qsh -l h_data=4G,h_rt=4:00:00 -pe shared 4
Last login: Tue Oct 24 00:21:27 2023 from 172.16.96.79
[wyman@n7431 ~]$ module load python
[wyman@n7431 ~]$ module load R
Modulefile for intel is already loaded
[wyman@n7431 ~]$ cd ~/myscratch/workshop
[wyman@n7431 workshop]$ $HOME/.local/bin/macs2 callpeak -t STAT1_30m_IFNa.bam -c INP_30m_IFNa_11e6.bam -g hs -n STAT1_30m_IFNa -f BAM -q 0.05 --bdg
INFO @ Thu, 26 Oct 2023 15:07:04:
# Command line: callpeak -t STAT1_30m_IFNa.bam -c INP_30m_IFNa_11e6.bam -g hs -n STAT1_30m_IFNa -f BAM -q 0.05 --bdg
# ARGUMENTS LIST:
# name = STAT1_30m_IFNa
# format = BAM
# ChIP-seq file = ['STAT1_30m_IFNa.bam']
# control file = ['INP_30m_IFNa_11e6.bam']
# effective genome size = 2.70e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# The maximum gap between significant sites is assigned as the read length/tag size.
# The minimum length of peaks is assigned as the predicted fragment length "d".
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
# Broad region calling is off
# Paired-End mode is off
```


Output files

1. `NAME_peaks.xls` is a tabular file which contains information about called peaks. You can open it in excel and sort/filter using excel functions. Information include:

- chromosome name
- start position of peak
- end position of peak
- length of peak region
- absolute peak summit position
- pileup height at peak summit
- $-\log_{10}(\text{pvalue})$ for the peak summit (e.g. $\text{pvalue} = 1\text{e-}10$, then this value should be 10)
- fold enrichment for this peak summit against random Poisson distribution with local lambda,
- $-\log_{10}(\text{qvalue})$ at peak summit

Coordinates in XLS is 1-based which is different from BED format. When `--broad` is enabled for broad peak calling, the pileup, p-value, q-value, and fold change in the XLS file will be the mean value across the entire peak region, since peak summit won't be called in broad peak calling mode.

2. `NAME_peaks.narrowPeak` is BED6+4 format file which contains the peak locations together with peak summit, p-value, and q-value. You can load it to the UCSC genome browser. Definition of some specific columns are:

- 5th: integer score for display. It's calculated as `int(-10*log10pvalue)` or `int(-10*log10qvalue)` depending on whether `-p` (pvalue) or `-q` (qvalue) is used as score cutoff. Please note that currently this value might be out of the [0-1000] range defined in [UCSC ENCODE narrowPeak format](#). You can let the value saturated at 1000 (i.e. $\text{p/q-value} = 10^{-100}$) by using the following 1-liner
`awk: awk -v OFS="\t" '{\$5=\$5>1000?1000:\$5} {print}' NAME_peaks.narrowPeak`
- 7th: fold-change at peak summit
- 8th: $-\log_{10}\text{pvalue}$ at peak summit
- 9th: $-\log_{10}\text{qvalue}$ at peak summit
- 10th: relative summit position to peak start

The file can be loaded directly to the UCSC genome browser. Remove the beginning track line if you want to analyze it by other tools.

<https://github.com/mac3-project/MACS/blob/master/docs/callpeak.md>

3. `NAME_summits.bed` is in BED format, which contains the peak summits locations for every peak. The 5th column in this file is the same as what is in the `narrowPeak` file. If you want to find the motifs at the binding sites, this file is recommended. The file can be loaded directly to the UCSC genome browser. Remove the beginning track line if you want to analyze it by other tools.
4. `NAME_peaks.broadPeak` is in BED6+3 format which is similar to `narrowPeak` file, except for missing the 10th column for annotating peak summits. This file and the `gappedPeak` file will only be available when `--broad` is enabled. Since in the broad peak calling mode, the peak summit won't be called, the values in the 5th, and 7-9th columns are the mean value across all positions in the peak region. Refer to `narrowPeak` if you want to fix the value issue in the 5th column.
5. `NAME_peaks.gappedPeak` is in BED12+3 format which contains both the broad region and narrow peaks. The 5th column is the score for showing grey levels on the UCSC browser as in `narrowPeak`. The 7th is the start of the first narrow peak in the region, and the 8th column is the end. The 9th column should be RGB color key, however, we keep 0 here to use the default color, so change it if you want. The 10th column tells how many blocks including the starting 1bp and ending 1bp of broad regions. The 11th column shows the length of each block and 12th for the start of each block. 13th: fold-change, 14th: $-\log_{10}pvalue$, 15th: $-\log_{10}qvalue$. The file can be loaded directly to the UCSC genome browser. Refer to `narrowPeak` if you want to fix the value issue in the 5th column.
6. `NAME_model.r` is an R script which you can use to produce a PDF image of the model based on your data. Load it to R by:

```
$ Rscript NAME_model.r
```

Then a pdf file `NAME_model.pdf` will be generated in your current directory. Note, R is required to draw this figure.

7. The `NAME_treat_pileup.bdg` and `NAME_control_lambda.bdg` files are in bedGraph format which can be imported to the UCSC genome browser or be converted into even smaller bigWig files. The `NAME_treat_pielup.bdg` contains the pileup signals (normalized according to `--scale-to` option) from ChIP/treatment sample. The `NAME_control_lambda.bdg` contains local biases estimated for each genomic location from the control sample, or from treatment sample when the control sample is absent. The subcommand `bdgcmp` can be used to compare these two files and make a bedGraph file of scores such as p-value, q-value, log-likelihood, and log fold changes.

MACS2 Outputs

```
INFO @ Thu, 26 Oct 2023 15:07:44: #3 Call peaks...
INFO @ Thu, 26 Oct 2023 15:07:44: #3 Pre-compute pvalue-qvalue table...
INFO @ Thu, 26 Oct 2023 15:08:36: #3 In the peak calling step, the following will be performed simultaneously:
INFO @ Thu, 26 Oct 2023 15:08:36: #3 Write bedGraph files for treatment pileup (after scaling if necessary)... STAT1_30m_IFNa_treat_pileup.bdg
INFO @ Thu, 26 Oct 2023 15:08:36: #3 Write bedGraph files for control lambda (after scaling if necessary)... STAT1_30m_IFNa_control_lambda.bdg
INFO @ Thu, 26 Oct 2023 15:08:36: #3 Pileup will be based on sequencing depth in treatment.
INFO @ Thu, 26 Oct 2023 15:08:36: #3 Call peaks for each chromosome...
INFO @ Thu, 26 Oct 2023 15:09:31: #4 Write output xls file... STAT1_30m_IFNa_peaks.xls
INFO @ Thu, 26 Oct 2023 15:09:31: #4 Write peak in narrowPeak format file... STAT1_30m_IFNa_peaks.narrowPeak
INFO @ Thu, 26 Oct 2023 15:09:31: #4 Write summits bed file... STAT1_30m_IFNa_summits.bed
INFO @ Thu, 26 Oct 2023 15:09:31: Done!
```

```
[wyan@n7431 workshop]$ head STAT1_30m_IFNa_peaks.narrowPeak
chr1 948587 948834 STAT1_30m_IFNa_peak_1 263 . 15.2101 33.1972 26.3043 185
chr1 4363313 4363350 STAT1_30m_IFNa_peak_2 22 . 5.07005 7.06389 2.21643 17
chr1 8812693 8812751 STAT1_30m_IFNa_peak_3 68 . 7.92055 12.4714 6.85696 25
chr1 25682452 25682699 STAT1_30m_IFNa_peak_4 193 . 12.6751 25.9006 19.3333 143
chr1 27998706 27998749 STAT1_30m_IFNa_peak_5 27 . 4.75425 7.68122 2.79245 23
chr1 27998787 27998968 STAT1_30m_IFNa_peak_6 192 . 10.5273 25.7778 19.2251 38
chr1 44514033 44514101 STAT1_30m_IFNa_peak_7 31 . 5.57705 8.12087 3.18899 49
chr1 54842906 54842981 STAT1_30m_IFNa_peak_8 14 . 4.07507 5.98914 1.44108 14
chr1 110453381 110453447 STAT1_30m_IFNa_peak_9 55 . 7.15826 10.9664 5.57929 6
chr1 154580619 154580751 STAT1_30m_IFNa_peak_10 70 . 7.60507 12.6942 7.06535 28
```

1. chromosome
2. start coordinate
3. end coordinate
4. name
5. score
6. strand

Standard BED file fields

7. signalValue - Measurement of overall enrichment for the region
8. pValue - Statistical significance (-log10)
9. qValue - Statistical significance using false discovery rate (-log10)
10. peak - Point-source called for this peak; 0-based offset from chromStart

narrowPeak specific fields

Reformat MACS2 output files for Downstream Analysis and Visualization

Get peaks in 4 column bed format for downstream analysis, such as CEAS and GREAT analysis

```
% cd ~/myscratch/workshop
% cut -f 1,2,3,4 STAT1_30m_IFNa_peaks.narrowPeak >STAT1_30m_peaks.bed
% cut -f 1,2,3,4 STAT1_6h_IFNa_peaks.narrowPeak >STAT1_6h_peaks.bed
```

Convert bedGraph (bdg) to BigWig (bw) file

```
% cd ~/myscratch/workshop
% wget http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86\_64/bedGraphToBigWig
% chmod u+x bedGraphToBigWig
% wget https://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/hg19.chrom.sizes
% ./bedGraphToBigWig STAT1_6h_IFNa_treat_pileup.bdg hg19.chrom.sizes STAT1_6h_IFNa_treat_pileup.bw
% ./bedGraphToBigWig STAT1_30m_IFNa_treat_pileup.bdg hg19.chrom.sizes STAT1_30m_IFNa_treat_pileup.bw
% ./bedGraphToBigWig STAT1_6h_IFNa_control_lambda.bdg hg19.chrom.sizes STAT1_6h_IFNa_control_lambda.bw
% ./bedGraphToBigWig STAT1_30m_IFNa_control_lambda.bdg hg19.chrom.sizes STAT1_30m_IFNa_control_lambda.bw
```

```
[wyann@n7431 workshop]$ ./bedGraphToBigWig STAT1_6h_IFNa_treat_pileup.bdg hg19.chrom.sizes STAT1_6h_IFNa_treat_pileup.bw
[wyann@n7431 workshop]$ ./bedGraphToBigWig STAT1_30m_IFNa_treat_pileup.bdg hg19.chrom.sizes STAT1_30m_IFNa_treat_pileup.bw
[wyann@n7431 workshop]$ ./bedGraphToBigWig STAT1_6h_IFNa_control_lambda.bdg hg19.chrom.sizes STAT1_6h_IFNa_control_lambda.bw
[wyann@n7431 workshop]$ ./bedGraphToBigWig STAT1_30m_IFNa_control_lambda.bdg hg19.chrom.sizes STAT1_30m_IFNa_control_lambda.bw
```

Copy results to local computer

Open a new terminal in local computer

```
% cd $HOME/Desktop
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_30m_IFNa_model.pdf ./
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_6h_IFNa_model.pdf ./
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_30m_IFNa_peaks.narrowPeak ./
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_6h_IFNa_peaks.narrowPeak ./
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_30m_IFNa_control_lambda.bw ./
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_30m_IFNa_treat_pileup.bw ./
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_6h_IFNa_control_lambda.bw ./
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_6h_IFNa_treat_pileup.bw ./
```

MACS2 Outputs

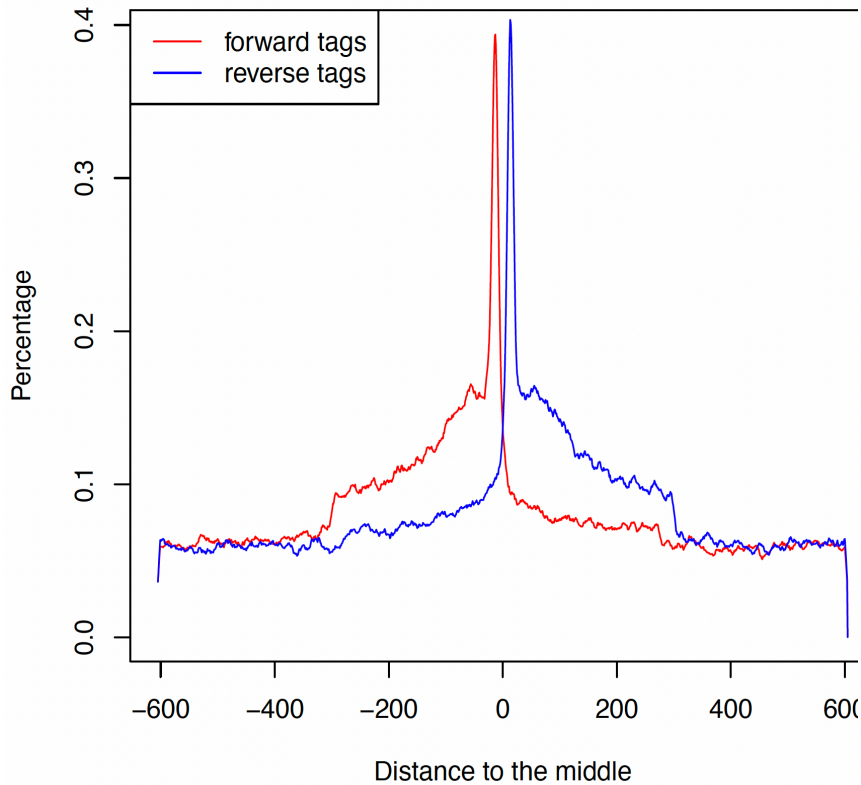
```
# This file is generated by MACS version 2.2.9.1
# Command line: callpeak -t STAT1_30m_IFNa.bam -c INP_30m_IFNa_11e6.bam -g hs -n STAT1_30m_IFNa -f BAM -q 0.05 --bdg
# ARGUMENTS LIST:
# name = STAT1_30m_IFNa
# format = BAM
# ChIP-seq file = ['STAT1_30m_IFNa.bam']
# control file = ['INP_30m_IFNa_11e6.bam']
# effective genome size = 2.70e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# The maximum gap between significant sites is assigned as the read length/tag size.
# The minimum length of peaks is assigned as the predicted fragment length "d".
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
# Broad region calling is off
# Paired-End mode is off

# tag size is determined as 27 bps
# total tags in treatment: 10778033
# tags after filtering in treatment: 10513428
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.02
# total tags in control: 11044013
# tags after filtering in control: 10812191
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.02
# d = 27
# alternative fragment length(s) may be 27,78,96,513,587 bps
chr  start  end  length  abs_summit  pileup  -log10(pvalue)  fold_enrichment  -log10(qvalue)  name
chr1  948588  948834  247  948773  29  33.1972  15.2101  26.3043  STAT1_30m_IFNa_peak_1
chr1  4363314  4363350  37  4363331  9  7.06389  5.07005  2.21643  STAT1_30m_IFNa_peak_2
chr1  8812694  8812751  58  8812719  9  12.4714  7.92055  6.85696  STAT1_30m_IFNa_peak_3
chr1  25682453  25682699  247  25682596  24  25.9006  12.6751  19.3333  STAT1_30m_IFNa_peak_4
chr1  27998707  27998749  43  27998730  13  7.68122  4.75425  2.79245  STAT1_30m_IFNa_peak_5
chr1  27998788  27998968  181  27998826  30  25.7778  10.5273  19.2251  STAT1_30m_IFNa_peak_6
chr1  44514034  44514101  68  44514083  10  8.12087  5.57705  3.18899  STAT1_30m_IFNa_peak_7
chr1  54842907  54842981  75  54842921  11  5.98914  4.07507  1.44108  STAT1_30m_IFNa_peak_8
chr1  110453382  110453447  66  110453388  8  10.9664  7.15826  5.57929  STAT1_30m_IFNa_peak_9
chr1  154580620  154580751  132  154580648  14  12.6942  7.60507  7.06535  STAT1_30m_IFNa_peak_10
chr1  155163211  155163237  27  155163221  11  9.2151  6.08406  4.07734  STAT1_30m_IFNa_peak_11
```

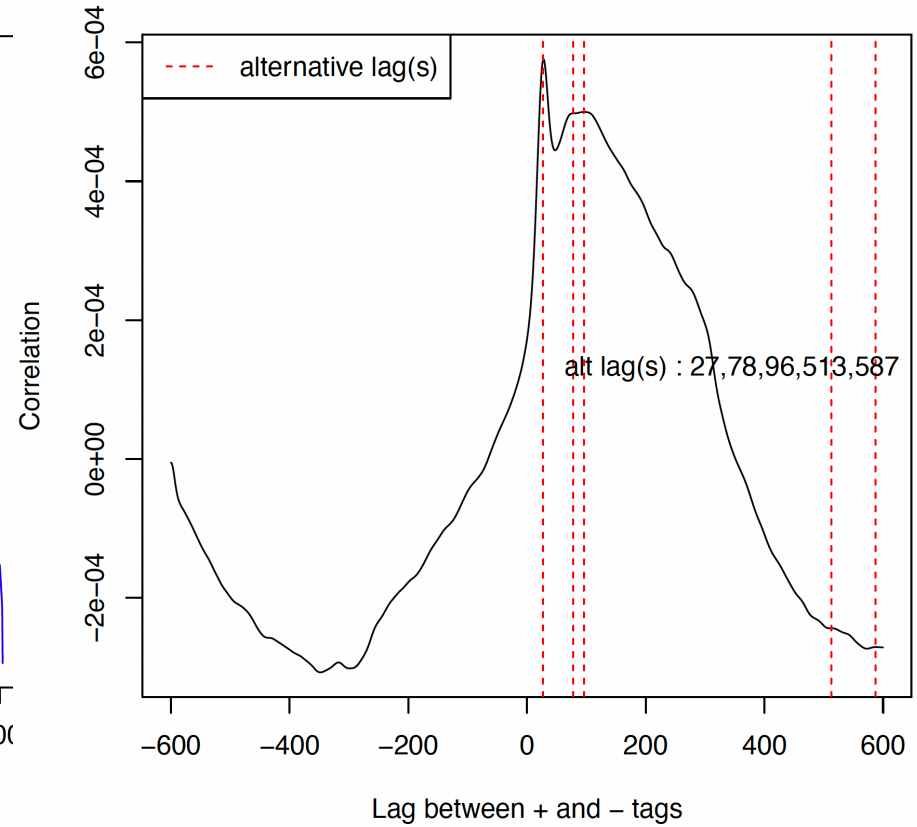
NRF	Complexity
NRF < 0.5	concerning
0.5	acceptable
$0.5 \leq \text{NRF} \leq 0.9$	compliant
NRF > 0.9	Ideal

MACS2 Output

Peak Model



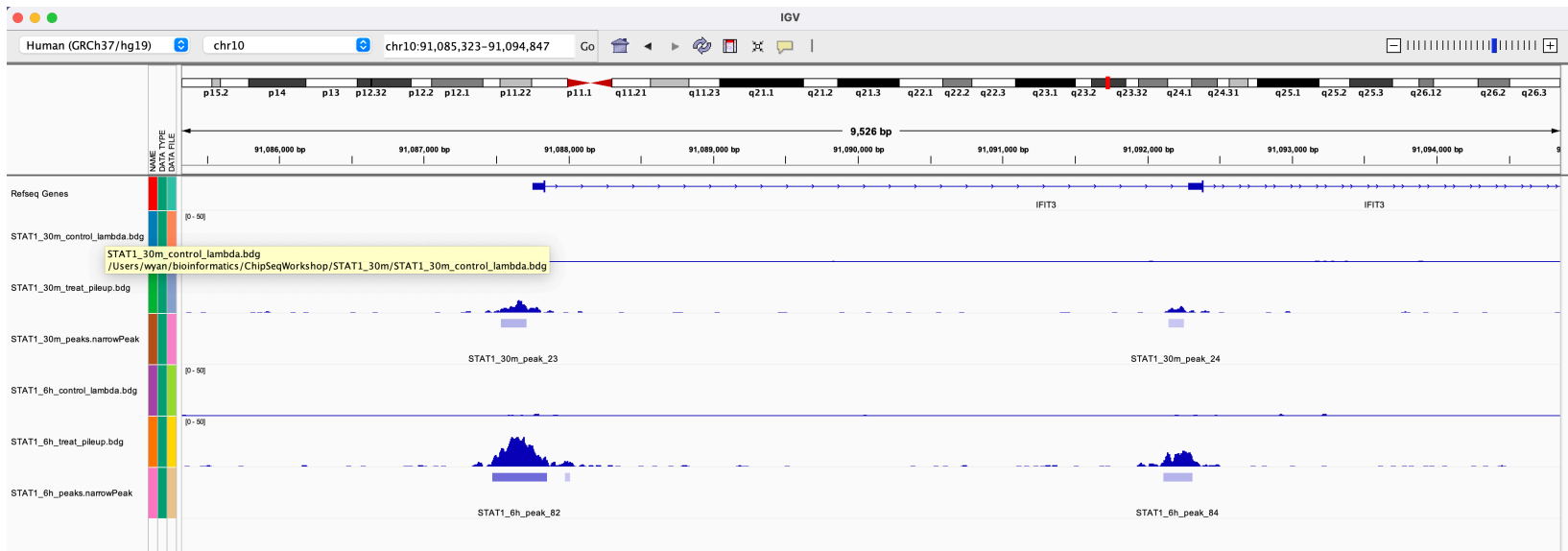
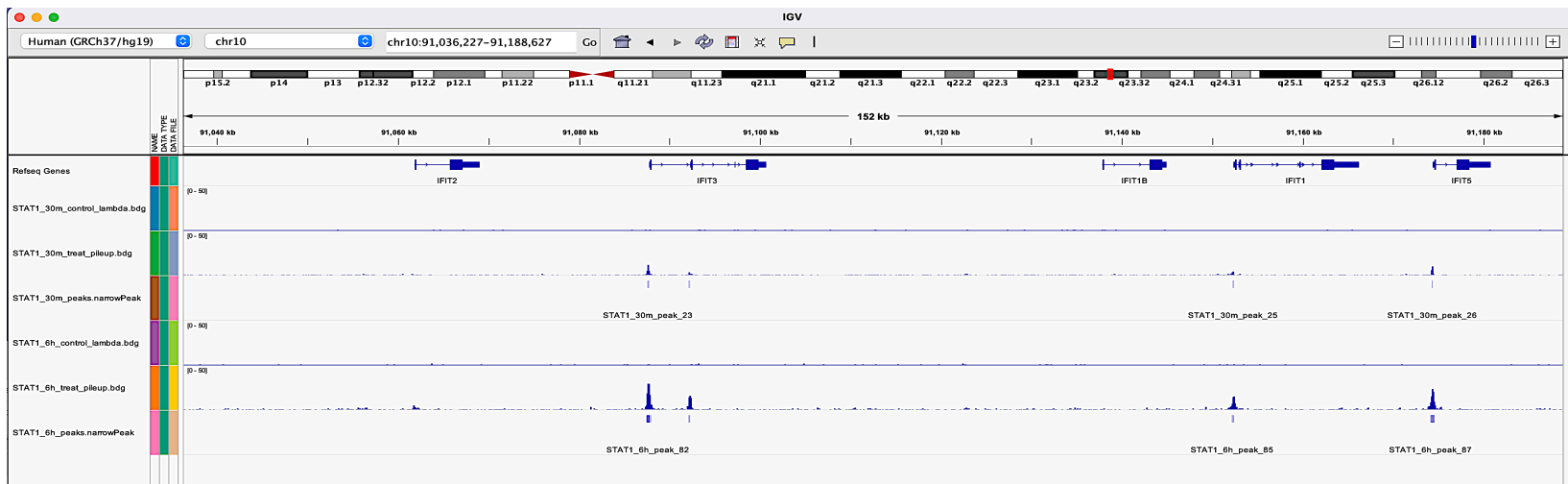
Cross-Correlation



$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

Integrative Genomics Viewer

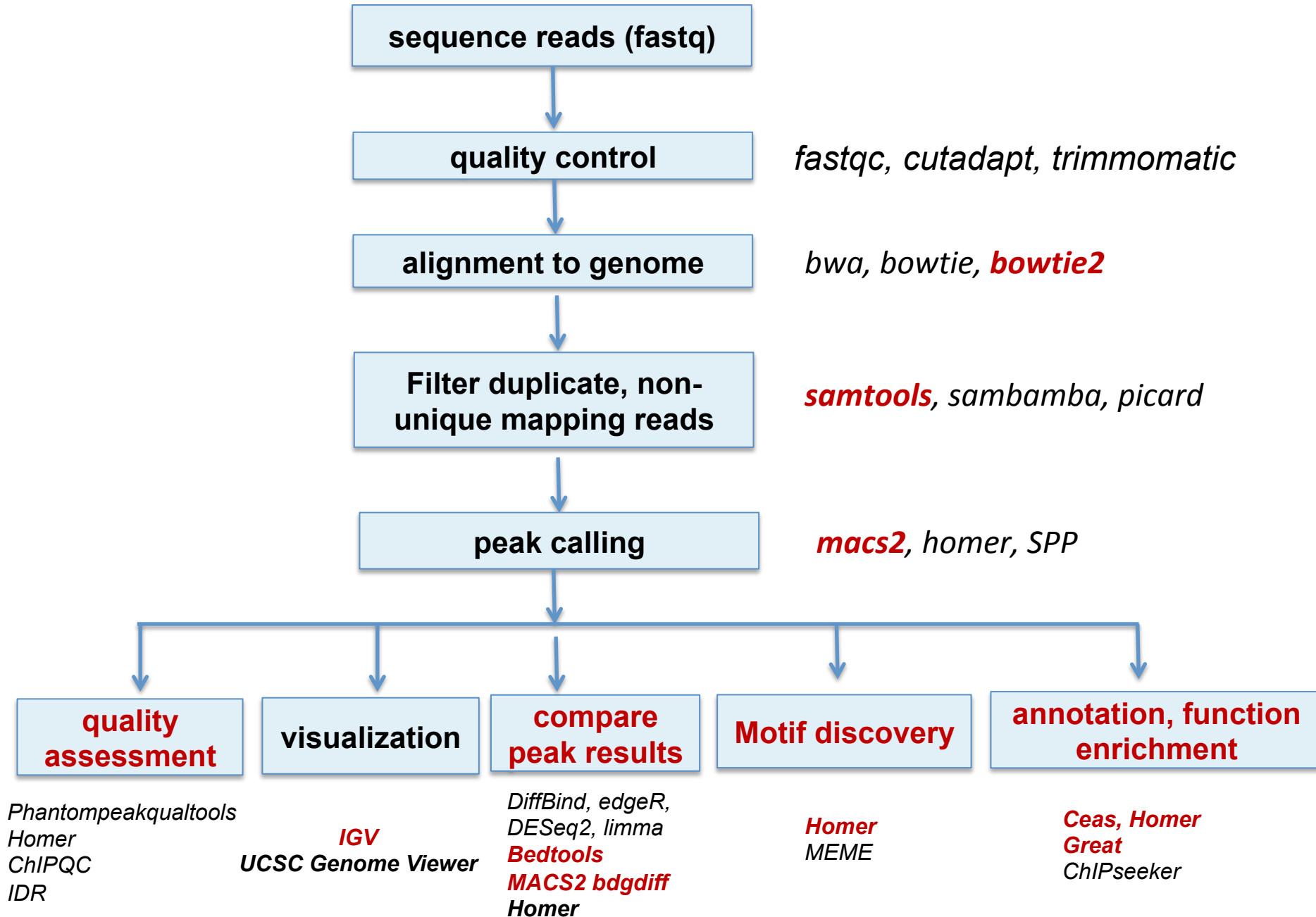


Browse IFIT genes in genome browser
IFIT: interferon induced proteins with tetratricopeptide repeats

ChIP-seq Data Analysis

- ❑ Day 1: ChIP-seq Background and Concept
 - ChIP-seq Protocol
 - Quality Control and Guidelines
 - ChIP-seq data from sequence read archive (SRA)
- ❑ Day 2: ChIP-seq data analysis workflow
 - Bowtie2 Alignment
 - MACS2 Peak calling
 - IGV genome visualization
- ❑ Day 3: Peak Annotation and Functional Analysis
 - CEAS: Cis-regulatory element annotation system
 - BEDTools for genome arithmetic
 - HOMER peak annotation, functional and motif analysis
 - GREAT

ChIP-seq Data Analysis Workflow



Call Differential Binding Events

```
[wyang@n7026 STAT1_30m_6h_bdgdiff]$ ~/.local/bin/macs2 bdgdiff -h
usage: macs2 bdgdiff [-h] --t1 T1BDG --t2 T2BDG --c1 C1BDG --c2 C2BDG [-C CUTOFF] [--d1 DEPTH1] [--d2 DEPTH2] [--outdir OUTDIR] [--o-prefix OPREFIX] [-o OFILE]
OFILE OFILE)

optional arguments:
  -h, --help            show this help message and exit
  --t1 T1BDG            MACS pileup bedGraph for condition 1. Incompatible with callpeak --SPMR output. REQUIRED
  --t2 T2BDG            MACS pileup bedGraph for condition 2. Incompatible with callpeak --SPMR output. REQUIRED
  --c1 C1BDG            MACS control lambda bedGraph for condition 1. Incompatible with callpeak --SPMR output. REQUIRED
  --c2 C2BDG            MACS control lambda bedGraph for condition 2. Incompatible with callpeak --SPMR output. REQUIRED
  -C CUTOFF, --cutoff CUTOFF
                        logLR cutoff. DEFAULT: 3 (likelihood ratio=1000)
  -l MINLEN, --min-len MINLEN
                        Minimum length of differential region. Try bigger value to remove small regions. DEFAULT: 200
  -g MAXGAP, --max-gap MAXGAP
                        Maximum gap to merge nearby differential regions. Consider a wider gap for broad marks. Maximum gap should be smaller than minimum length (-g). DEFAULT: 100
  --d1 DEPTH1, --depth1 DEPTH1
                        Sequencing depth (# of non-redundant reads in million) for condition 1. It will be used together with --d2. See description for --d2 below for how to assign the
                        em. Default: 1
  --d2 DEPTH2, --depth2 DEPTH2
                        Sequencing depth (# of non-redundant reads in million) for condition 2. It will be used together with --d1. DEPTH1 and DEPTH2 will be used to calculate scaling
                        factor for each sample, to down-scale larger sample
                        to the level of smaller one. For example, while comparing 10 million condition 1 and 20 million condition 2, use --d1 10 --d2 20, then pileup value in bedGraph
                        for condition 2 will be divided by 2. Default: 1
  --outdir OUTDIR       If specified all output files will be written to that directory. Default: the current working directory
  --o-prefix OPREFIX    Output file prefix. Actual files will be named as PREFIX_cond1.bed, PREFIX_cond2.bed and PREFIX_common.bed. Mutually exclusive with -o/--ofile.
  -o OFILE OFILE OFILE, --ofile OFILE OFILE OFILE
                        Output filenames. Must give three arguments in order: 1. file for unique regions in condition 1; 2. file for unique regions in condition 2; 3. file for common
                        regions in both conditions. Note: mutually exclusive
                        with --o-prefix.
```

-o file1 file2 file3

File1: peaks highly enriched in condition 1 comparing to condition 2.

File2: peaks highly enriched in condition 2 comparing to condition 1.

File3: peaks highly enriched in both condition 1 and condition 2. The difference between condition 1 and condition 2 is not significant.

The last column in the file represent the difference between two conditions in log10 likelihood ratios (LR).

Call Differential Binding Events

Find peaks that are differentially enriched in STAT1_30m and STAT1_6h

Get sequencing depth from MACS2 callpeak output

```
% cd ~/myscratch/workshop  
% grep "tags after filtering" STAT1_30m_IFNa_peaks.xls  
% grep "tags after filtering" STAT1_6h_IFNa_peaks.xls
```

```
[wyan@n7026 workshop]$ grep "tags after filtering" STAT1_30m_IFNa_peaks.xls  
# tags after filtering in treatment: 10513428  
# tags after filtering in control: 10812191  
[wyan@n7026 workshop]$ grep "tags after filtering" STAT1_6h_IFNa_peaks.xls  
# tags after filtering in treatment: 10608588  
# tags after filtering in control: 11170440
```

Run bdgdiff from MACS2

```
% module load python  
% ~/.local/bin/macs2 bdgdiff --t1 STAT1_30m_IFNa_treat_pileup.bdg --t2  
STAT1_6h_IFNa_treat_pileup.bdg --c1 STAT1_30m_IFNa_control_lambda.bdg --c2  
STAT1_6h_IFNa_control_lambda.bdg -l 30 -g 20 --d1 10513428 --d2 10608588 --outdir  
STAT1_30m_6h_bdgdiff -o STAT1_30m_uniq.bed STAT1_6h_uniq.bed STAT1_30m_6h_common.bed
```

```
[wyan@n7026 workshop]$ ~/.local/bin/macs2 bdgdiff --t1 STAT1_30m_IFNa_treat_pileup.bdg --t2 STAT1_6h_IFNa_treat_pileup.bdg --c1 STAT1_30m_IFNa_control_lambda.bdg --c2  
STAT1_6h_IFNa_control_lambda.bdg -l 30 -g 20 --d1 10513428 --d2 10608588 --outdir STAT1_30m_6h_bdgdiff -o STAT1_30m_uniq.bed STAT1_6h_uniq.bed STAT1_30m_6h_common.bed  
INFO @ Thu, 02 Nov 2023 16:13:59: Read and build treatment 1 bedGraph...  
INFO @ Thu, 02 Nov 2023 16:14:22: Read and build control 1 bedGraph...  
INFO @ Thu, 02 Nov 2023 16:15:03: Read and build treatment 2 bedGraph...  
INFO @ Thu, 02 Nov 2023 16:15:27: Read and build control 2 bedGraph...  
INFO @ Thu, 02 Nov 2023 16:19:21: Write peaks...  
INFO @ Thu, 02 Nov 2023 16:19:21: Done
```

Go to STAT1_30m_6h_bdgdiff directory to examine differential binding regions

```
% cd STAT1_30m_6h_bdgdiff  
% ls
```

```
[wyan@n7026 workshop]$ cd ~/myscratch/workshop/STAT1_30m_6h_bdgdiff  
[wyan@n7026 STAT1_30m_6h_bdgdiff]$ ls  
STAT1_30m_6h_common.bed STAT1_30m_uniq.bed STAT1_6h_uniq.bed
```

Cis-regulatory Element Annotation System

Summary

We present a tool designed to characterize genome-wide protein-DNA interaction patterns from ChIP-chip and ChIP-Seq of both sharp and broad binding factors. As a stand-alone extension of our web application CEAS (Cis-regulatory Element Annotation System), it provides statistics on ChIP enrichment at important genome features such as specific chromosome, promoters, gene bodies, or exons, and infers genes most likely to be regulated by a binding factor. CEAS also enables biologists to visualize the average ChIP enrichment signals over specific genomic features, allowing continuous and broad ChIP enrichment to be perceived which might be too subtle to detect from ChIP peaks alone.

CEAS Overview

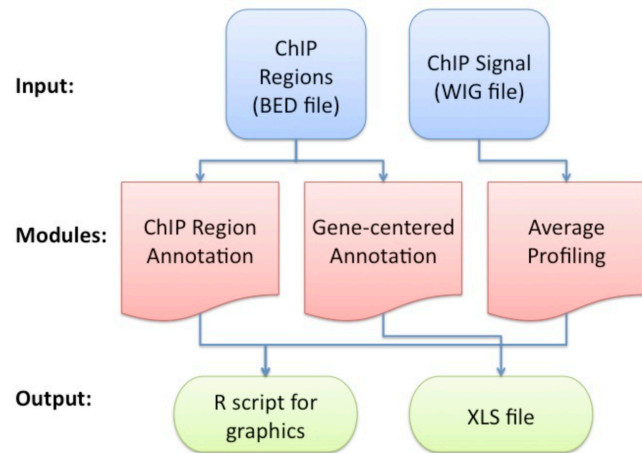


Figure 1 The work-flow of CEAS. A gene annotation table file, a BED file with ChIP regions, and a WIG file with ChIP enrichment signal are required. CEAS consists of three modules: ChIP region annotation, gene-centered annotation, and average profiling within/near important genomic features. As output, CEAS produces an R script of graphical results (or a PDF file if R can be directly called in the same environment as Python) and a tab-delimited with XLS extension of gene-centered annotation.

<https://liulab-dfci.github.io/resources/software/CEAS.zip>

Cis-regulatory Element Annotation System

Summary

We present a tool designed to characterize genome-wide protein-DNA interaction patterns from ChIP-chip and ChIP-Seq of both sharp and broad binding factors. As a stand-alone extension of our web application CEAS (Cis-regulatory Element Annotation System), it provides statistics on ChIP enrichment at important genome features such as specific chromosome, promoters, gene bodies, or exons, and infers genes most likely to be regulated by a binding factor. CEAS also enables biologists to visualize the average ChIP enrichment signals over specific genomic features, allowing continuous and broad ChIP enrichment to be perceived which might be too subtle to detect from ChIP peaks alone.

Modules

ChIP region annotation

CEAS estimates the relative enrichment level of ChIP regions in each gene feature with respect to the whole genome. For this, it first calculates the percentages of the ChIP regions that reside in the following four categories: (a) promoters, (b) bidirectional promoters, (c) downstreams of a gene, and (d) gene bodies (3'UTRs, 5'UTRs, coding exons, and introns). In addition to these categories, the user can add another extra category of interest such as non-coding regions as an optional input parameter. 'Promoters' correspond to the upstream regions of the transcription start site (TSS) of genes. Three promoter sizes can be specified by the user (1kb, 2kb, and 3kb by default). For instance, if the user set promoter to be the 1kb, 3kb, and 10kb upstream of TSS, CEAS computes the cumulative percentages of ChIP regions that fall in $\leq 1\text{kb}$, $\leq 3\text{kb}$, and $\leq 10\text{kb}$ upstream of the TSS of genes, respectively. 'Bidirectional promoters' are promoter regions between divergently transcribed genes whose TSS are closer in proximity than user-defined distances (two options, 2.5kb and 5kb by default). 'Downstreams' refer to the regions immediately downstream of genes, spanning up to the same search range as in 'promoters' from the transcription termination site (TTS). 'Gene bodies' are further categorized into UTR regions (3' and 5' UTRs), coding exons and introns. After the percentages of ChIP regions respective categories, P-values for the significance of the relative enrichment with respect to the background are calculated using one-sided binomial test.

As a final summary of ChIP region annotation, CEAS draws a pie chart of how ChIP regions spread over the categories. If ChIP regions do not fall into any of the categories, they are considered to be 'distal intergenic'.

Gene-centered annotation

Identifying genes associated with ChIP regions is important to infer the direct regulatory gene targets of the binding factor of interest. CEAS provides the distances to the centers of the nearest ChIP regions upstream and downstream of every RefSeq gene's TSS, allowing biologists to determine the potential target genes of the binding factor based on the distances. Moreover, in case that a broad ChIP peak covers the whole or part of a gene body, it is useful to know how much of the gene, including its promoter or downstream, is occupied by the ChIP region in addition to how far the TSS of the gene is from the ChIP region center. To this end, CEAS divides every gene into three equal fractions and, for each fraction, calculates the percentage of the area covered by ChIP regions. It also estimates the percentages of the promoter and downstream of the gene (3kb upstream of TSS and downstream of TTS by default) that are covered by ChIP regions. The results are saved as a tab-delimited text file with XLS extension for easy Excel visualization, which contains a row of annotations for every RefSeq gene.

Average signal profiling within/near important genomic features

Since ChIP region and gene-centered annotation operate on discrete ChIP regions identified by a peak-calling algorithm, some subtle binding patterns may fail to be captured, depending on the cut-off of peak calling. Therefore, CEAS displays continuous ChIP enrichment signal within/around important gene features to help biologists visualize the average binding patterns. It draws the average signals around TSS and TTS in a user-defined range ($\pm 3\text{kb}$ from TSS and TTS by default). In addition, CEAS gives average signals on meta-gene, meta-concatenated-exon, meta-concatenated-intron, meta-exons and meta-intron, where the prefix 'meta' indicates that every element (e.g. every gene) is normalized residing within the above categories are obtained, they are compared against the genome background percentages for those to have the same length (e.g., 3kb for meta-gene). The difference between meta-concatenated-exon and meta-exons is that the first concatenates all exons of a gene before calculating the average gene (like a meta-cDNA) profile, whereas the latter calculates the average exon profile of all exons. These plots allow biologists to gain insight on how ChIP enrichment varies over the gene body (or exons and introns). CEAS provides an additional function of drawing the average ChIP signals of multiple sub-groups of genes (e.g. up vs. down genes), allowing the user to compare between the gene groups. In addition, we provide a separate script, named 'sitepro', in our CEAS package, which draws the average signal (from a WIG) in a given list of sites of interest (from a BED). This script enables biologists to visualize the average signals in any arbitrary regions (e.g. transcription factor binding sites) in addition to the pre-defined genomic regions.

Cis-regulatory Element Annotation System

Install ceas in hoffman2

```
% cd ~/myscratch/workshop/  
% wget https://liulab-dfci.github.io/resources/software/CEAS.zip  
% unzip CEAS.zip  
% cd CEAS/src  
% tar xvf CEAS-Package-0.9.9.tar.gz  
% cd CEAS-Package-0.9.9  
% python setup.py install --prefix=~/.local  
After installation, ceas program is in $HOME/.local/bin directory
```

Copy hg19 gene annotation table to the workshop directory

```
% cd ~/myscratch/workshop/CEAS/src  
% tar xvf hg19.tar  
% cp hg19/refGene ~/myscratch/workshop/hg19.refGene
```

ChIP region and Gene centered Annotation

Run ceas to annotate peaks identified in STAT1_30m and STAT1_6h

```
% module load R  
% cd ~/myscratch/workshop  
% $HOME/.local/bin/ceas --name=STAT1_30m_ceas -b STAT1_30m_peaks.bed -g hg19.refGene  
% $HOME/.local/bin/ceas --name=STAT1_6h_ceas -b STAT1_6h_peaks.bed -g hg19.refGene
```

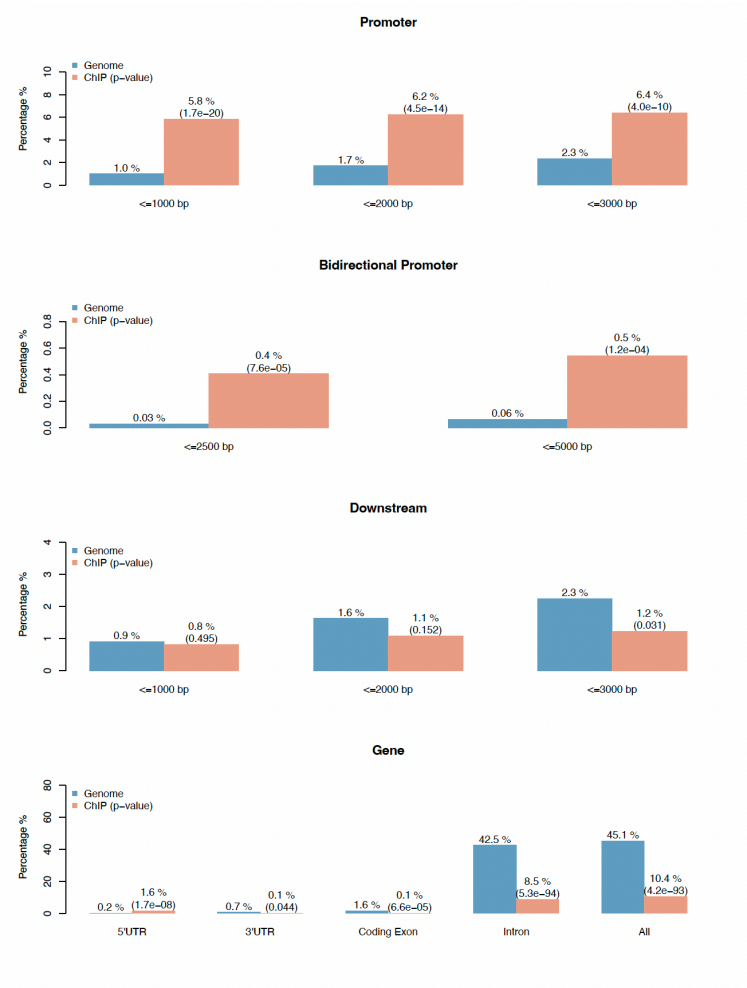
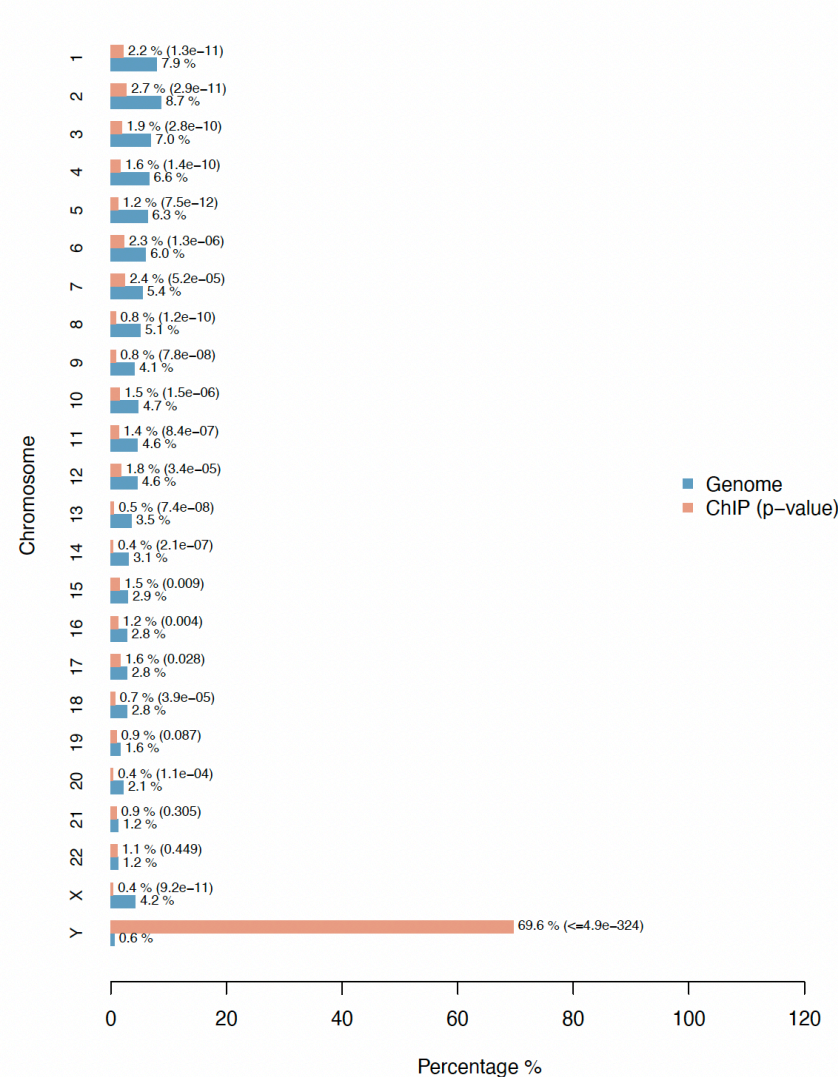
Copy CEAS outputs to local computer

```
% cd $HOME/Desktop  
% Desktop %scp hoffman2.idre.ucla.edu:~/myscratch/workshop/"*ceas*" ./
```

```
wyan@Ws-iMac Desktop % scp hoffman2.idre.ucla.edu:~/myscratch/workshop/"*ceas*" ./  
wyan@hoffman2.idre.ucla.edu's password:
```

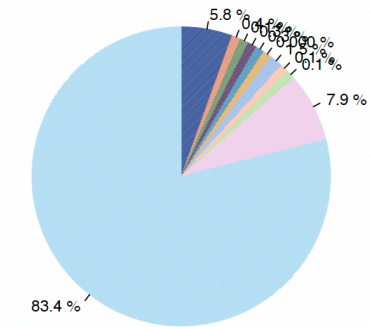
STAT1_30m_ceas.R	100%	22KB	733.5KB/s
STAT1_30m_ceas.pdf	100%	20KB	611.4KB/s
STAT1_30m_ceas.xls	100%	2573KB	3.3MB/s
STAT1_6h_ceas.R	100%	28KB	1.4MB/s
STAT1_6h_ceas.pdf	100%	23KB	1.1MB/s
STAT1_6h_ceas.xls	100%	2575KB	6.3MB/s

Chromosomal Distribution of ChIP Regions



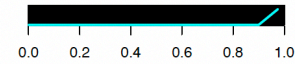
STAT1_30m_peaks annotation

Distribution of ChIP Regions

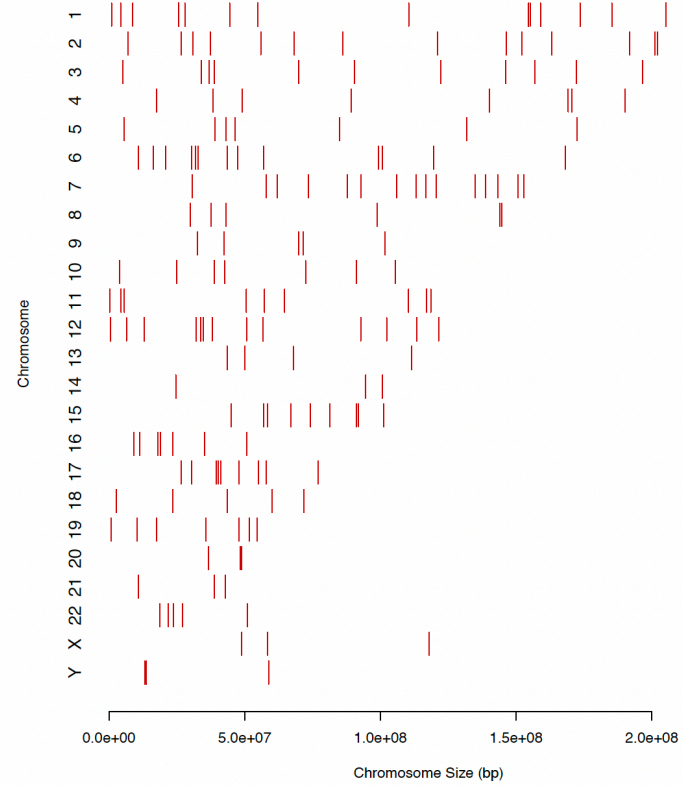


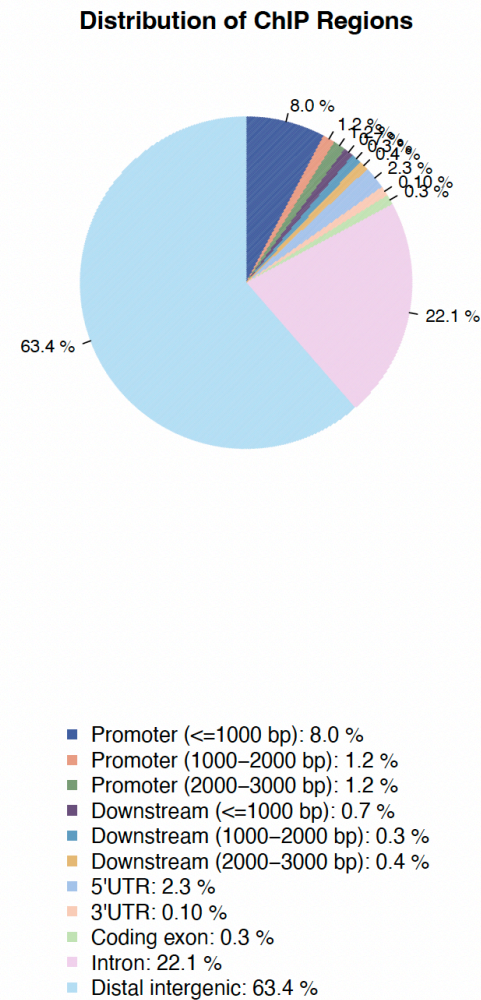
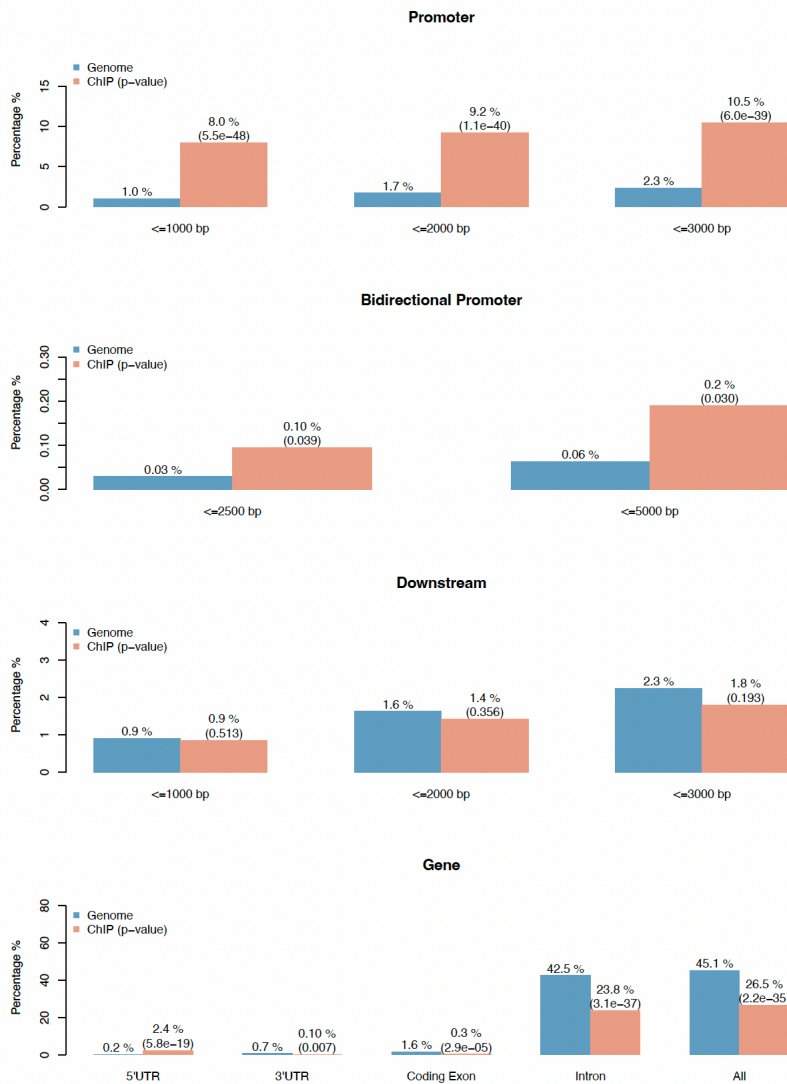
- Promoter (<=1000 bp): 5.8 %
- Promoter (1000–2000 bp): 0.4 %
- Promoter (2000–3000 bp): 0.1 %
- Downstream (<=1000 bp): 0.3 %
- Downstream (1000–2000 bp): 0.3 %
- Downstream (2000–3000 bp): 0.000 %
- 5'UTR: 1.5 %
- 3'UTR: 0.1 %
- Coding exon: 0.1 %
- Intron: 7.9 %
- Distal intergenic: 83.4 %

Distribution of Peak Heights



ChIP Regions (Peaks) over Chromosomes





STAT1_6h_peaks annotation

RefSeq

Distances from TSS/TTS to nearest ChIP region

Occupancy rates of ChIP regions at genes, including promoters, downstreams, and exons

#name	chr	txStart	txEnd	strand	dist u TSS	dist d TSS	dist u TTS	dist d TTS	3000bp u TSS	3000bp d TSS	1/3 gene	2/3 gene	3/3 gene	3000bp d TT: exons
NM_152876	chr10	90750287	90775541	+	10031059	337374	10056313	312120	0	0	0	0	0	0
NM_003956	chr10	90965695	90967071	-	120590	10247843	121966	10246467	0	0	0	0	0	0
NM_001127605	chr10	90973327	91011660	-	76001	10292432	114334	10254099	0	0	0	0	0	0
NM_000235	chr10	90973329	91011660	-	76001	10292432	114332	10254101	0	0	0	0	0	0
NM_001547	chr10	91061705	91069032	+	10342477	25956	10349804	18629	0	0	0	0	0	0
NM_001549	chr10	91087601	91100724	+	10368373	60	8511	51469	0.04	0.1	0.07	0.05	0	0.09
NM_001031683	chr10	91092238	91100724	+	25	59955	8511	51469	0.04	0.03	0.03	0	0	0.03
NM_001010987	chr10	91137812	91144961	+	45599	14381	52748	7232	0	0	0	0	0	0
NM_001548	chr10	91152321	91163742	+	128	9	11252	10349	0.04	0.05	0.04	0	0	0.06
NM_012420	chr10	91174324	91180758	+	233	32	6402	2790401	0.08	0.08	0.11	0	0	0.06
NM_213606	chr10	91190055	91295313	-	2675846	120957	2781104	15699	0	0	0	0	0	0
NM_148978	chr10	91342748	91403647	-	2567512	229291	2628411	168392	0	0	0	0	0	0
NM_148977	chr10	91342748	91405215	-	2565944	230859	2628411	168392	0	0	0	0	0	0
NM_138316	chr10	91342748	91403647	-	2567512	229291	2628411	168392	0	0	0	0	0	0
NR_026835	chr10	91451056	91457685	-	2513474	283329	2520103	276700	0	0	0	0	0	0
NM_016195	chr10	91461366	91534700	+	287010	2509793	360344	2436459	0	0	0	0	0	0
NM_000872	chr10	92500577	92617671	-	1353488	1443315	1470582	1326221	0	0	0	0	0	0
NM_019860	chr10	92500577	92617671	-	1353488	1443315	1470582	1326221	0	0	0	0	0	0
NM_019859	chr10	92500577	92617671	-	1353488	1443315	1470582	1326221	0	0	0	0	0	0
NM_006413	chr10	92631708	92661946	+	1457352	1339451	1487590	1309213	0	0	0	0	0	0
NM_001104546	chr10	92631708	92668312	+	1457352	1339451	1493956	1302847	0	0	0	0	0	0

STAT1_6h_ceas.xls

Field	Description
chr	Chromosome of a RefSeq gene
txStart	Transcription starting site (TSS) of a RefSeq gene
txEnd	Transcription terminating site (TTS) of a RefSeq gene
strand	Strand of a RefSeq Gene
dist u txStart	Distance to the nearest ChIP region (center) upstream of txStart (bp)
dist d txStart	Distance to the nearest ChIP region (center) downstream of txStart (bp)
dist u txEnd	Distance to the nearest ChIP region (center) upstream of txEnd (bp)
dist d txEnd	Distance to the nearest ChIP region (center) downstream of txEnd (bp)
3kb u txStart	Occupancy rate of ChIP regions in 3kb upstream of txStart (0.0 - 1.0)
3kb d txStart	Occupancy rate of ChIP regions in 3kb downstream of txStart (0.0 - 1.0)
1/3 gene	Occupancy rate of ChIP regions in the 1st third of a gene (0.0 - 1.0)
2/3 gene	Occupancy rate of ChIP regions in the 2nd third of a gene (0.0 - 1.0)
3/3 gene	Occupancy rate of ChIP regions in the 3rd third of a gene (0.0 - 1.0)
3kb d txEnd	Occupancy rate of ChIP regions in 3kb downstream of txEnd (0.0 - 1.0)
exons	Occupancy rate of ChIP regions in the exons (0.0-1.0)

CEAS – Average signal profiling within/near important genomic features

Signal profile of STAT1 in STAT2_6h binding regions

Convert bedGraph (bdg) to Wiggle (.wig) file

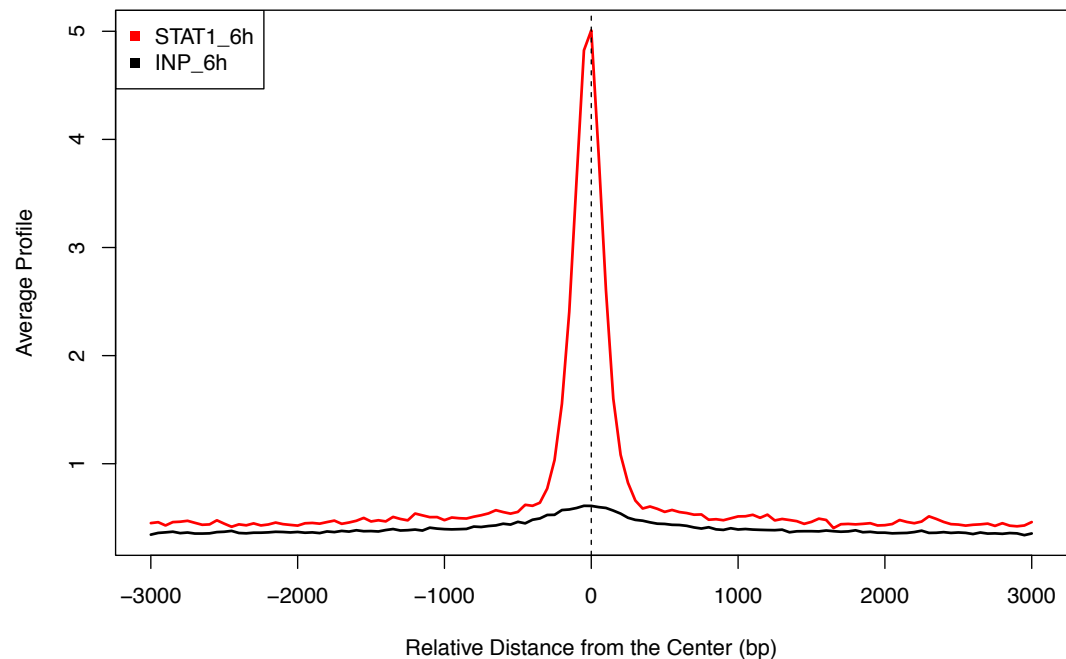
```
% cd ~/myscratch/workshop
% cp /u/scratch/w/wyan/workshop/bedgraph_to_wig.pl ./
% cp /u/scratch/w/wyan/workshop/STAT2_6h_peaks.bed ./
% perl bedgraph_to_wig.pl --bedgraph STAT1_6h_IFNa_treat_pileup.bdg --wig STAT1_6h_IFNa_treat_pileup.wig --step 50
% perl bedgraph_to_wig.pl --bedgraph STAT1_6h_IFNa_control_lambda.bdg --wig STAT1_6h_IFNa_control_lambda.wig --step 50
% $HOME/.local/bin/sitepro --name=stat2peaks --pf-res=50 --span=3000 -b STAT2_6h_peaks.bed -w STAT1_6h_IFNa_treat_pileup.wig -w
STAT1_6h_IFNa_control_lambda.wig
```

```
[wyman@login2 workshop]$ perl bedgraph_to_wig.pl --bedgraph STAT1_6h_IFNa_treat_pileup.bdg --wig STAT1_6h_IFNa_treat_pileup.wig --step 50
[wyman@login2 workshop]$ perl bedgraph_to_wig.pl --bedgraph STAT1_6h_IFNa_control_lambda.bdg --wig STAT1_6h_IFNa_control_lambda.wig --step 50
[wyman@login2 workshop]$ $HOME/.local/bin/sitepro --name=stat2peaks --pf-res=50 --span=3000 -b STAT2_6h_peaks.bed -w STAT1_6h_IFNa_treat_pileup.wig -w STAT1_6h_IFNa_control_lambda.wig -l STAT1_6h -l INP_6h
```

Copy stat2peaks.pdf to local computer

```
% cd Desktop
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/stat2peaks.pdf ./
```

Average Profiles around the Center of Sites





bedtools: a powerful toolset for genome arithmetic

Collectively, the **bedtools** utilities are a swiss-army knife of tools for a wide-range of genomics analysis tasks. The most widely-used tools enable *genome arithmetic*: that is, set theory on the genome. For example, **bedtools** allows one to *intersect*, *merge*, *count*, *complement*, and *shuffle* genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF. While each individual tool is designed to do a relatively simple task (e.g., *intersect* two interval files), quite sophisticated analyses can be conducted by combining multiple bedtools operations on the UNIX command line.

bedtools is developed in the [Quinlan laboratory](#) at the [University of Utah](#) and benefits from fantastic contributions made by scientists worldwide.

```
[wyman@n7431 ~]$ module load bedtools
[wyman@n7431 ~]$ bedtools
bedtools is a powerful toolset for genome arithmetic.

Version:      v2.30.0
About:        developed in the quinlanlab.org and by many contributors worldwide.
Docs:         http://bedtools.readthedocs.io/
Code:         https://github.com/arq5x/bedtools2
Mail:         https://groups.google.com/forum/#!forum/bedtools-discuss

Usage:        bedtools <subcommand> [options]

The bedtools sub-commands include:

[ Genome arithmetic ]
  intersect    Find overlapping intervals in various ways.
  window       Find overlapping intervals within a window around an interval.
  closest      Find the closest, potentially non-overlapping interval.
  coverage     Compute the coverage over defined intervals.
  map          Apply a function to a column for each overlapping interval.
  genomecov    Compute the coverage over an entire genome.
  merge        Combine overlapping/nearby intervals into a single interval.
  cluster      Cluster (but don't merge) overlapping/nearby intervals.
  complement   Extract intervals _not_ represented by an interval file.
  shift        Adjust the position of intervals.
  subtract     Remove intervals based on overlaps b/w two files.
  slop         Adjust the size of intervals.
  flank        Create new intervals from the flanks of existing intervals.
  sort         Order the intervals in a file.
  random       Generate random intervals in a genome.
  shuffle      Randomly redistribute intervals in a genome.
  sample       Sample random records from file using reservoir sampling.
  spacing      Report the gap lengths between intervals in a file.
  annotate     Annotate coverage of features from multiple files.
```

Bedtools: a toolset for genomic arithmetic



```
Usage: bedtools intersect [OPTIONS] -a <bed/gff/vcf/bam> -b <bed/gff/vcf/bam>
```

```
Note: -b may be followed with multiple databases and/or  
wildcard (*) character(s).
```

```
Options:
```

- ```
-wa
```

 Write the original entry in A for each overlap.
- ```
-wb
```

 Write the original entry in B for each overlap.
- Useful for knowing `_what_` A overlaps. Restricted by `-f` and `-r`.
- ```
-loj
```

 Perform a "left outer join". That is, for each feature in A  
report each overlap with B. If no overlaps are found,  
report a NULL feature for B.
- ```
-wo
```

 Write the original A and B entries plus the number of base
pairs of overlap between the two features.
- Overlaps restricted by `-f` and `-r`.
Only A features with overlap are reported.
- ```
-wao
```

 Write the original A and B entries plus the number of base  
pairs of overlap between the two features.  
- Overlapping features restricted by `-f` and `-r`.  
However, A features w/o overlap are also reported  
with a NULL B feature and `overlap = 0`.
- ```
-u
```

 Write the original A entry `_once_` if `_any_` overlaps found in B.
- In other words, just report the fact `>=1` hit was found.
- Overlaps restricted by `-f` and `-r`.
- ```
-c
```

 For each entry in A, report the number of overlaps with B.  
- Reports 0 for A entries that have no overlap with B.  
- Overlaps restricted by `-f`, `-F`, `-r`, and `-s`.
- ```
-C
```

 For each entry in A, separately report the number of
- overlaps with each B file on a distinct line.
- Reports 0 for A entries that have no overlap with B.
- Overlaps restricted by `-f`, `-F`, `-r`, and `-s`.
- ```
-v
```

 Only report those entries in A that have `_no overlaps_` with B.  
- Similar to "grep -v" (an homage).



## Find the peaks present in STAT1\_30m and STAT1\_6h

```
% module load bedtools
% cd ~/myscratch/workshop
% bedtools intersect -wa -u -a STAT1_6h_peaks.bed -b STAT1_30m_peaks.bed >STAT1_6h_overlapping_30m.bed
```

## Count number of overlapped peaks

```
% wc -l STAT1_6h_overlapping_30m.bed
```

## Find the peaks unique to STAT1\_6h

```
% bedtools intersect -v -a STAT1_6h_peaks.bed -b STAT1_30m_peaks.bed >STAT1_6h_notOverlapping_30m.bed
```

## Count number of unique peaks found in STAT1\_6h

```
% wc -l STAT1_6h_notOverlapping_30m.bed
```

```
[wyman@n7431 workshop]$ bedtools intersect -wa -u -a STAT1_6h_peaks.bed -b STAT1_30m_peaks.bed >STAT1_6h_overlapping_30m.bed
[wyman@n7431 workshop]$ wc -l STAT1_6h_overlapping_30m.bed
426 STAT1_6h_overlapping_30m.bed
[wyman@n7431 workshop]$ bedtools intersect -v -a STAT1_6h_peaks.bed -b STAT1_30m_peaks.bed >STAT1_6h_notOverlapping_30m.bed
[wyman@n7431 workshop]$ wc -l STAT1_6h_notOverlapping_30m.bed
625 STAT1_6h_notOverlapping_30m.bed
[wyman@n7431 workshop]$ wc -l STAT1_6h_peaks.bed
1051 STAT1_6h_peaks.bed
```



## HOMER (v4.11, 10-24-2019)

Software for motif discovery and next generation sequencing analysis

HOMER (Hypergeometric Optimization of Motif EnRichment) is a suite of tools for Motif Discovery and next-gen sequencing analysis. It is a collection of command line programs for UNIX-style operating systems written in Perl and C++. HOMER was primarily written as a *de novo* motif discovery algorithm and is well suited for finding 8-20 bp motifs in large scale genomics data. HOMER contains many useful tools for analyzing ChIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C and numerous other types of functional genomics sequencing data sets.

## HOMER Analysis Tools:

- [makeTagDirectory](#)  
creation of tag directories and quality control
- [findPeaks](#)  
perform peak calling process
- [findMotifsGenome.pl](#)  
motif analysis
- [annotatePeaks.pl](#)  
perform peak annotation, associate peaks with nearby genes,  
and gene ontology analysis
- [makeBigWig.pl](#)  
create BigWig file from bedGraph. BigWig file can be used for  
UCSC genome browser visualization
- [getDifferentialPeaks](#)  
Identify peaks with significantly different tag densities





# HOMER

Software for motif discovery and ChIP-Seq analysis

## Next-gen Sequencing Analysis: Creating a "Tag Directory" with makeTagDirectory

### What does makeTagDirectory do?

**makeTagDirectory** basically parses through the alignment file and splits the tags into separate files based on their chromosome. As a result, several \*.tags.tsv files are created in the output directory. These are made to very efficiently return to the data during downstream analysis. This also helps speed up the analysis of very large data sets without running out of memory.

### Basic Quality Control Analysis:

The following 4 basic quality control results are produced by default and are relatively inexpensive in terms of processing power to create while paring alignment files into Tag Directories, and do not require any extra information to produce. These files are meant to be opened with a text editor or graphed using EXCEL or similar program. More detailed descriptions of these files and how to interpret them are found in the sequencing technique-specific tutorials. Also, if you want to rerun QC analysis on an existing Tag Directory, run the command using the "-update" option:

**makeTagDirectory ExistingTagDirectory/ -update [additional options...]**

#### Basic Tag information

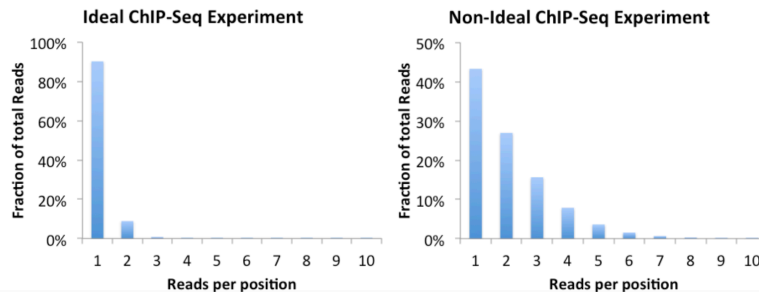
**tagInfo.txt** - Contains basic configuration information, such as the total number of reads, the total number of unique positions with aligned reads (by genome and chromosome), and various other statistics. One of the more important parameters is "fragmentLengthEstimate=##", which provides an estimate of the length of fragments used for sequencing.

#### Read Length Distribution

**tagLengthDistribution.txt** - File contains a histogram of read lengths used for alignment.

#### Clonal Tag Distribution

**tagCountDistribution.txt** - File contains a histogram of clonal read depth, showing the number of reads per unique position. If an experiment is "over-sequenced", you start seeing the same reads over and over instead of unique reads. Sometimes this is a sign there was not enough starting material for sequencing library preparation. Below are examples of ideal and non-ideal results - in the case of the non-ideal experiment, you probably don't want to sequence that library anymore.



# Create tag directories for STAT1\_30m\_IFNa and STAT1\_6h\_IFNa

```
% module load homer
% cd ~/myscratch/workshop/
% makeTagDirectory STAT1_30m_IFNa_tagDir STAT1_30m_IFNa.bam
% makeTagDirectory STAT1_6h_IFNa_tagDir STAT1_6h_IFNa.bam
% ls STAT1_30m_IFNa_tagDir/
% head -n 8 STAT1_30m_IFNa_tagDir/tagCountDistribution.txt
```

```
[wyman@n7431 workshop]$ module load homer
[wyman@n7431 workshop]$ cd ~/myscratch/workshop/
[wyman@n7431 workshop]$ makeTagDirectory STAT1_30m_IFNa_tagDir STAT1_30m_IFNa.bam
Will parse file: STAT1_30m_IFNa.bam

Creating directory: STAT1_30m_IFNa_tagDir and removing existing *.tags.tsv

Treating STAT1_30m_IFNa.bam as a bam file
Reading alignment file STAT1_30m_IFNa.bam

Optimizing tag files...
Estimated genome size = 3094793862
Estimated average read density = 0.003483 per bp
Total Tags = 10778033.0
Total Positions = 10531118
Average tag length = 26.9
Median tags per position = 1 (ideal: 1)
Average tags per position = 1.023
Fragment Length Estimate: 68
Peak Width Estimate: 89
Autocorrelation quality control metrics:
 Same strand fold enrichment: 1.3
 Diff strand fold enrichment: 1.3
 Same / Diff fold enrichment: 1.1

Guessing sample is ChIP-Seq - may have low enrichment with lots of background
```

```
[wyman@n7431 workshop]$ ls STAT1_30m_IFNa_tagDir/
chr10.tags.tsv chr13.tags.tsv chr16.tags.tsv chr19.tags.tsv chr21.tags.tsv chr3.tags.tsv chr6.tags.tsv chr9.tags.tsv tagAutocorrelation.txt
chr11.tags.tsv chr14.tags.tsv chr17.tags.tsv chr1.tags.tsv chr22.tags.tsv chr4.tags.tsv chr7.tags.tsv chrX.tags.tsv tagCountDistribution.txt
chr12.tags.tsv chr15.tags.tsv chr18.tags.tsv chr20.tags.tsv chr2.tags.tsv chr5.tags.tsv chr8.tags.tsv chrY.tags.tsv tagInfo.txt
[wyman@n7431 workshop]$ head -n 8 STAT1_30m_IFNa_tagDir/tagCountDistribution.txt
Tags per tag position (Median = 1, tags per genomic bp = 0.003) Fraction of Positions
0 0.000000
1 0.983212
2 0.014388
3 0.001215
4 0.000476
5 0.000252
6 0.000138
```



# HOMER

Software for motif discovery and next-gen sequencing analysis

## Finding Enriched Motifs in Genomic Regions (*findMotifsGenome.pl*)

HOMER was initially developed to automate the process of finding enriched motifs in ChIP-Seq peaks. More generally, HOMER analyzes genomic positions, not limited to only ChIP-Seq peaks, for enriched motifs. The main idea is that all the user really needs is a file containing genomic coordinates (i.e. a HOMER peak file or BED file), and HOMER will generally take care of the rest. To analyze a peak file for motifs, run the following command:

```
findMotifsGenome.pl <peak/BED file> <genome> <output directory> -size # [options]
```

```
[wyran@n1887 data]$ findMotifsGenome.pl
```

Program will find de novo and known motifs in regions in the genome

Usage: findMotifsGenome.pl <pos file> <genome> <output directory> [additional options]

Example: findMotifsGenome.pl peaks.txt mm8r peakAnalysis -size 200 -len 8

Possible Genomes:

```
hg19 human
hg38 human
mm10 mouse
-- or --
```

Custom: provide the path to genome FASTA files (directory or single file)

Heads up: will create the directory "prepared/" in same location.

Basic options:

```
-mask (mask repeats/lower case sequence, can also add 'r' to genome, i.e. mm9r)
-bg <background position file> (genomic positions to be used as background, default=automatic)
 removes background positions overlapping with target positions unless -keepOverlappingBg is used
 -chopify (chop up large background regions to the avg size of target regions)
-len <#>[,<#>,<#>...] (motif length, default=8,10,12) [NOTE: values greater 12 may cause the program
 to run out of memory - in these cases decrease the number of sequences analyzed (-N),
 or try analyzing shorter sequence regions (i.e. -size 100)]
-size <#> (fragment size to use for motif finding, default=200)
 -size <#,>#> (i.e. -size -100,50 will get sequences from -100 to +50 relative from center)
 -size given (uses the exact regions you give it)
-S <#> (Number of motifs to optimize, default: 25)
-mis <#> (global optimization: searches for strings with # mismatches, default: 2)
-norevopp (don't search reverse strand for motifs)
-nomotif (don't search for de novo motif enrichment)
-rna (output RNA motif logos and compare to RNA motif database, automatically sets -norevopp)
```

Scanning sequence for motifs

```
-find <motif file> (This will cause the program to only scan for motifs)
```

Known Motif Options/Visualization

```
-mset <vertebrates|insects|worms|plants|yeast|all> (check against motif collects, default: auto)
-basic (just visualize de novo motifs, don't check similarity with known motifs)
-bits (scale sequence logos by information content, default: doesn't scale)
-nocheck (don't search for de novo vs. known motif similarity)
-mcheck <motif file> (known motifs to check against de novo motifs,
-float (allow adjustment of the degeneracy threshold for known motifs to improve p-value[dangerous])
-noknown (don't search for known motif enrichment, default: -known)
-mknown <motif file> (known motifs to check for enrichment,
-nofacts (omit humor)
-seqlogo (use weblogo/seqlogo/ghostscript to generate logos, default uses SVG now)
```

Sequence normalization options:

```
-gc (use GC% for sequence content normalization, now the default)
-cpg (use CpG% instead of GC% for sequence content normalization)
-noweight (no CG correction)
Also -nlen <#>, -olen <#>, see homer2 section below.
```

Advanced options:

```
-h (use hypergeometric for p-values, binomial is default)
-N <#> (Number of sequences to use for motif finding, default=max(50k, 2x input)
-local <#> (use local background, # of equal size regions around peaks to use i.e. 2)
-redundant <#> (Remove redundant sequences matching greater than # percent, i.e. -redundant 0.5)
-maxN <#> (maximum percentage of N's in sequence to consider for motif finding, default: 0.7)
-maskMotif <motif file1> [motif file 2]... (motifs to mask before motif finding)
-opt <motif file1> [motif file 2]... (motifs to optimize or change length of)
-rand (randomize target and background sequences labels)
-ref <peak file> (use file for target and background - first argument is list of peak ids for targets)
-oligo (perform analysis of individual oligo enrichment)
-dumpFasta (Dump fasta files for target and background sequences for use with other programs)
-preparse (force new background files to be created)
-preparedDir <directory> (location to search for prepared file and/or place new files)
-keepFfiles (keep temporary files)
-fdr <#> (Calculate empirical FDR for de novo discovery #=number of randomizations)
```

homer2 specific options:

```
-homer2 (use homer2 instead of original homer, default)
-nlen <#> (length of lower-order oligos to normalize in background, default: -nlen 3)
 -nmax <#> (Max normalization iterations, default: 160)
 -neutral (weight sequences to neutral frequencies, i.e. 25%, 6.25%, etc.)
-olen <#> (lower-order oligo normalization for oligo table, use if -nlen isn't working well)
-p <#> (Number of processors to use, default: 1)
-e <#> (Maximum expected motif instance per bp in random sequence, default: 0.01)
-cache <#> (size in MB for statistics cache, default: 500)
-quickMask (skip full masking after finding motifs, similar to original homer)
-minlp <#> (stop looking for motifs when seed logp score gets above #, default: -100)
```

Original homer specific options:

```
-homer1 (to force the use of the original homer)
-depth [low|med|high|allnight] (time spent on local optimization default: med)
```



# HOMER

Software for motif discovery and next-gen sequencing analysis

---

## Finding Enriched Motifs in Genomic Regions (*findMotifsGenome.pl*)

HOMER was initially developed to automate the process of finding enriched motifs in ChIP-Seq peaks. More generally, HOMER analyzes genomic positions, not limited to only ChIP-Seq peaks, for enriched motifs. The main idea is that all the user really needs is a file containing genomic coordinates (i.e. a HOMER peak file or BED file), and HOMER will generally take care of the rest. To analyze a peak file for motifs, run the following command:

```
findMotifsGenome.pl <peak/BED file> <genome> <output directory> -size # [options]
```

## Run findMotifsGenome.pl to find de novo and known transcription factor motifs for STAT1\_30m\_IFNa and STAT1\_6h\_IFNa

```
% cd ~/myscratch/workshop
```

```
% findMotifsGenome.pl STAT1_6h_IFNa_summits.bed hg19 STAT1_6h_motifs -preparedDir STAT1_6h_preparedDir
```

## Transfer motif result to local computer

Open a new terminal

```
% cd ~/Desktop
```

```
% scp -r hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_6h_motifs ./
```

| STAT1_6h_motifs            |                  |          |              | Search |  |
|----------------------------|------------------|----------|--------------|--------|--|
| Name                       | Date Modified    | Size     | Kind         |        |  |
| homerMotifs.all.motifs     | Today at 5:42 PM | 30 KB    | Document     |        |  |
| homerMotifs.motifs8        | Today at 5:42 PM | 9 KB     | Document     |        |  |
| homerMotifs.motifs10       | Today at 5:42 PM | 10 KB    | Document     |        |  |
| homerMotifs.motifs12       | Today at 5:42 PM | 11 KB    | Document     |        |  |
| homerResults               | Today at 5:42 PM | --       | Folder       |        |  |
| homerResults.html          | Today at 5:42 PM | 105 KB   | HTML text    |        |  |
| knownResults               | Today at 5:42 PM | --       | Folder       |        |  |
| knownResults.html          | Today at 5:42 PM | 376 KB   | HTML text    |        |  |
| knownResults.txt           | Today at 5:42 PM | 48 KB    | Plain Text   |        |  |
| motifFindingParameters.txt | Today at 5:42 PM | 93 bytes | Plain Text   |        |  |
| seq.autonorm.tsv           | Today at 5:42 PM | 2 KB     | TSV Document |        |  |

File: ///Users/ryan/Desktop/STAT1\_6h\_motifs/knownResults.html

## Homer Known Motif Enrichment Results (STAT1\_6h\_motifs)

[Homer de novo Motif Results](#)

[Gene Ontology Enrichment Results](#)

[Known Motif Enrichment Results \(txt file\)](#)

Total Target Sequences = 1052, Total Background Sequences = 37753

| Rank | Motif                                                                               | Name                                                  | P-value | log P-value | q-value (Benjamini) | # Target Sequences with Motif | % of Targets Sequences with Motif | # Background Sequences with Motif | % of Background Sequences with Motif | Motif File                          | SVC                 |
|------|-------------------------------------------------------------------------------------|-------------------------------------------------------|---------|-------------|---------------------|-------------------------------|-----------------------------------|-----------------------------------|--------------------------------------|-------------------------------------|---------------------|
| 1    |    | ISRE(IRF)/ThioMac-LPS-Expression(GSE23622)/Homer      | 1e-336  | -7.738e+02  | 0.0000              | 279.0                         | 26.52%                            | 275.6                             | 0.73%                                | <a href="#">motif file (matrix)</a> | <a href="#">SVG</a> |
| 2    |    | IRF3(IRF)/BMDM-Irf3-ChIP-Seq(GSE67343)/Homer          | 1e-317  | -7.315e+02  | 0.0000              | 446.0                         | 42.40%                            | 1566.5                            | 4.15%                                | <a href="#">motif file (matrix)</a> | <a href="#">SVG</a> |
| 3    |    | IRF2(IRF)/Erythroblas-IRF2-ChIP-Seq(GSE36985)/Homer   | 1e-306  | -7.055e+02  | 0.0000              | 302.0                         | 28.71%                            | 472.6                             | 1.25%                                | <a href="#">motif file (matrix)</a> | <a href="#">SVG</a> |
| 4    |    | IRF8(IRF)/BMDM-IRF8-ChIP-Seq(GSE77884)/Homer          | 1e-289  | -6.659e+02  | 0.0000              | 422.0                         | 40.11%                            | 1562.0                            | 4.14%                                | <a href="#">motif file (matrix)</a> | <a href="#">SVG</a> |
| 5    |    | IRF1(IRF)/PBM-IRF1-ChIP-Seq(GSE43036)/Homer           | 1e-228  | -5.269e+02  | 0.0000              | 283.0                         | 26.90%                            | 718.1                             | 1.90%                                | <a href="#">motif file (matrix)</a> | <a href="#">SVG</a> |
| 6    |   | ZNF189(Zf)/HEK293-ZNF189.GFP-ChIP-Seq(GSE58341)/Homer | 1e-222  | -5.118e+02  | 0.0000              | 488.0                         | 46.39%                            | 3347.0                            | 8.87%                                | <a href="#">motif file (matrix)</a> | <a href="#">SVG</a> |
| 7    |  | PU.1:IRF8(ETS:IRF)/pDC-Irf8-ChIP-Seq(GSE66899)/Homer  | 1e-115  | -2.665e+02  | 0.0000              | 200.0                         | 19.01%                            | 865.0                             | 2.29%                                | <a href="#">motif file (matrix)</a> | <a href="#">SVG</a> |
| 8    |  | TEAD(TEA)/Fibroblast-PU.1-ChIP-Seq(Unpublished)/Homer | 1e-86   | -1.991e+02  | 0.0000              | 358.0                         | 34.03%                            | 4162.7                            | 11.03%                               | <a href="#">motif file (matrix)</a> | <a href="#">SVG</a> |

ISRE: Interferon stimulated response element

IRF: Interferon regulatory factor



# HOMER

Software for motif discovery and next-gen sequencing analysis

## Annotating Regions in the Genome (*annotatePeaks.pl*)

Homer contains a useful, all-in-one program for performing peak annotation called **annotatePeaks.pl**. In addition to associating peaks with nearby genes, **annotatePeaks.pl** can perform Gene Ontology Analysis, genomic feature association analysis (Genome Ontology), associate peaks with gene expression data, calculate ChIP-Seq Tag densities from different experiments, and find motif occurrences in peaks. **annotatePeaks.pl** can also be used to create histograms and heatmaps. Description of the annotation functions are covered below, while quantification of tags, motifs, histograms, etc. are covered [here](#).

## Use *annotatePeaks.pl* to associate peaks with genes and perform GO analysis

```
% cd ~/myscratch/workshop
```

```
% annotatePeaks.pl STAT1_6h_IFNa_peaks.narrowPeak hg19 -go STAT1_6h_GO >STAT1_6h_annotatePeaks.txt
```

## Transfer *STAT1\_6h\_annotatePeaks.txt* and *biological\_process.txt* to local computer

```
% cd ~/Desktop
```

```
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_6h_annotatePeaks.txt ./
```

```
% scp hoffman2.idre.ucla.edu:~/myscratch/workshop/STAT1_6h_GO/biological_process.txt
```

```
[lwyann@n6671 data]$ annotatePeaks.pl STAT1_6h/STAT1_6h_peaks.narrowPeak hg19 -go STAT1_6h_GO >STAT1_6h_annotatePeaks.txt

Peak file = STAT1_6h/STAT1_6h_peaks.narrowPeak
Genome = hg19
Organism = human
Will perform Gene Ontology analysis - output to directory = STAT1_6h_GO
Peak/BED file conversion summary:
BED/Header formatted lines: 1052
peakfile formatted lines: 0
Duplicated Peak IDs: 0

Peak File Statistics:
Total Peaks: 1052
Redundant Peak IDs: 0
Peaks lacking information: 0 (need at least 5 columns per peak)
Peaks with misformatted coordinates: 0 (should be integer)
Peaks with misformatted strand: 0 (should be either +/- or 0/1)

Peak file looks good!

Reading Positions...
Finding Closest TSS...
Annotating:.....
Annotation Number of peaks Total size (bp) Log2 Ratio (obs/exp) LogP enrichment (+values depleted)
3UTR 5.0 26762632 -0.864 2.219
miRNA 0.0 96670 -0.047 0.033
ncRNA 1.0 6991458 -1.249 1.161
TSS 13.0 32286633 0.244 -1.177
pseudo 1.0 2089637 0.493 -0.676
Exon 5.0 37906581 -1.331 4.294
Intron 229.0 1258916630 -0.902 87.639
Intergenic 643.0 1691354378 0.161 -11.206
Promoter 150.0 35809872 3.623 -251.691
5UTR 5.0 2586029 2.507 -6.162
snRNA 0.0 357 -0.000 0.000
scRNA 0.0 97 -0.000 0.000
NOTE: If this part takes more than 2 minutes, there is a good chance
your machine ran out of memory: consider hitting ctrl+C and rerunning
the command with "-noann"
To capture annotation stats in a file, use "-annStats <filename>" next time
Annotating:.....
Annotation Number of peaks Total size (bp) Log2 Ratio (obs/exp) LogP enrichment (+values depleted)
3UTR 5.0 26762632 -0.863 2.217
Other 0.0 3935577 -1.226 1.338
```





# HOMER

Software for motif discovery and next-gen sequencing analysis

## Annotating Regions in the Genome (*annotatePeaks.pl*)

Homer contains a useful, all-in-one program for performing peak annotation called **annotatePeaks.pl**. In addition to associating peaks with nearby genes, **annotatePeaks.pl** can perform Gene Ontology Analysis, genomic feature association analysis (Genome Ontology), associate peaks with gene expression data, calculate ChIP-Seq Tag densities from different experiments, and find motif occurrences in peaks. **annotatePeaks.pl** can also be used to create histograms and heatmaps. Description of the annotation functions are covered below, while quantification of tags, motifs, histograms, etc. are covered [here](#).

### STAT1\_6h\_annotatePeaks.txt

|    | A                             | B     | C         | D         | E      | F          | G            | H                                   | I                          | J              | K            | L         | M            | N            | O            | P           | Q            | R              | S              | T |
|----|-------------------------------|-------|-----------|-----------|--------|------------|--------------|-------------------------------------|----------------------------|----------------|--------------|-----------|--------------|--------------|--------------|-------------|--------------|----------------|----------------|---|
| 1  | PeakID (cmd=annotatePeaks.pl) | Chr   | Start     | End       | Strand | Peak Score | Focus Ratio/ | Annotation                          | Detailed Annotation        | Distance to TS | Nearest Prot | Entrez ID | Nearest Unig | Nearest Refs | Nearest Ensc | Gene Name   | Gene Alias   | Gene Descrip   | Gene Type      |   |
| 2  | STAT1_6h_peak_710             | chrY  | 58827077  | 58827650  | +      | 1064       | NA           | Intergenic                          | (CATTG)n Satellite Satelli | -273040        | NM_005840    | 10251     | Hs.381912    | NM_005840    | ENSG000001   | SPRY3       | spry-3       | sprouty RTK    | protein-coding |   |
| 3  | STAT1_6h_peak_850             | chrY  | 58862663  | 58862816  | +      | 920        | NA           | Intergenic                          | (CATTG)n Satellite Satelli | -237664        | NM_005840    | 10251     | Hs.381912    | NM_005840    | ENSG000001   | SPRY3       | spry-3       | sprouty RTK    | protein-coding |   |
| 4  | STAT1_6h_peak_180             | chr14 | 24630167  | 24630582  | +      | 778        | NA           | promoter-TSS (NM_006084)            | promoter-TSS (NM_006084)   | -100           | NM_006084    | 10379     | Hs.1706      | NM_006084    | ENSG000002   | IRF9        | IRF-9 ISGF3  | interferon re  | protein-coding |   |
| 5  | STAT1_6h_peak_451             | chr3  | 146262252 | 146262698 | +      | 671        | NA           | promoter-TSS (NM_00136)             | promoter-TSS (NM_00136)    | -81            | NM_021105    | 5359      | Hs.130759    | NM_021105    | ENSG000001   | PLSCR1      | MMTRA18      | phospholipid   | protein-coding |   |
| 6  | STAT1_6h_peak_701             | chrY  | 58823367  | 58823793  | +      | 649        | NA           | Intergenic                          | (CATTG)n Satellite Satelli | -276823        | NM_005840    | 10251     | Hs.381912    | NM_005840    | ENSG000001   | SPRY3       | spry-3       | sprouty RTK    | protein-coding |   |
| 7  | STAT1_6h_peak_323             | chr2  | 55920882  | 55921265  | +      | 645        | NA           | promoter-TSS (NM_03310)             | promoter-TSS (NM_03310)    | -94            | NM_033109    | 87178     | Hs.388733    | NM_033109    | ENSG000001   | PNPT1       | COPXD13 DF   | polyribonucle  | protein-coding |   |
| 8  | STAT1_6h_peak_356             | chr2  | 191884933 | 191885295 | +      | 631        | NA           | Intergenic                          | CpG                        | -1138          | NR_136318    | 105373805 | Hs.570280    | NR_136318    | ENSG000002   | LOC1053738  | -            | uncharacteri   | ncRNA          |   |
| 9  | STAT1_6h_peak_50              | chr1  | 173640051 | 173640508 | +      | 631        | NA           | Intergenic                          | LTR71A LTR ERV1            | -1289          | NR_158771    | 339416    | Hs.130054    | NM_198493    | ENSG000001   | ANKRD45     | CT117        | ankyrin repei  | protein-coding |   |
| 10 | STAT1_6h_peak_432             | chr3  | 5191105   | 5191565   | +      | 586        | NA           | intron (NM_018184, LTR71A LTR ERV1) | 27358                      | NM_018184      | 55207        | Hs.250009 | NM_018184    | ENSG000001   | ARL8B        | ARL10C Gie  | ADP ribosyla | protein-coding |                |   |
| 11 | STAT1_6h_peak_265             | chr17 | 39647074  | 39647130  | +      | 565        | NA           | promoter-TSS (NM_00377)             | promoter-TSS (NM_00377)    | -986           | NM_003771    | 8689      | Hs.248189    | NM_003771    | ENSG000001   | KRT36       | HA6 KRTHA6   | keratin 36     | protein-coding |   |
| 12 | STAT1_6h_peak_445             | chr3  | 122282939 | 122283295 | +      | 560        | NA           | promoter-TSS (NM_00114)             | promoter-TSS (NM_00114)    | 30             | NM_001146    | 83666     | Hs.518200    | NM_031458    | ENSG000001   | PARP9       | ARTD9 BAL    | poly(ADP-rib   | protein-coding |   |
| 13 | STAT1_6h_peak_389             | chr21 | 11186592  | 11186642  | +      | 549        | NA           | Intergenic                          | HSAT1 Satellite Satellite  | -87680         | NM_001187    | 574       | Hs.545789    | NM_001187    | BAGE         | BAGE1 CT2.8 | melanoma     | protein-coding |                |   |
| 14 | STAT1_6h_peak_163             | chr12 | 113375904 | 113376372 | +      | 535        | NA           | promoter-TSS (NM_00618)             | promoter-TSS (NM_00618)    | -141           | NM_006187    | 4940      | Hs.528634    | NM_006187    | ENSG000001   | OAS3        | p100 p1000   | 2'-5'-oligoad  | protein-coding |   |

### biological\_process.txt

|    | A          | B                                                                  | C          | D          | E             | F           | G             | H            | I           | J           | K                                                                   | L | M | N | O | P |
|----|------------|--------------------------------------------------------------------|------------|------------|---------------|-------------|---------------|--------------|-------------|-------------|---------------------------------------------------------------------|---|---|---|---|---|
| 1  | TermID     | Term                                                               | Enrichment | logP       | Genes in Term | Target Gene | Fraction of T | Total Target | Total Genes | Entrez Gene | Gene Symbols                                                        |   |   |   |   |   |
| 2  | GO:0051607 | defense response to virus                                          | 7.88E-38   | -85.434427 | 192           | 47          | 0.13055556    | 360          | 18054       | 23586,81622 | DDX58,UNC93B1,RTP4,TRIM5,TRIM22,NT5C3A,IL23A,OAS1,MX2,IFITM1,IF     |   |   |   |   |   |
| 3  | GO:0034340 | response to type I interferon                                      | 5.69E-34   | -76.549684 | 72            | 31          | 0.08611111    | 360          | 18054       | #####       | MX1,IFITM3,HLA-E,IFI35,OAS1,IFIT3,OAS1,MX2,XAF1,STAT2,IFITM1,ISG15, |   |   |   |   |   |
| 4  | GO:0009615 | response to virus                                                  | 5.87E-32   | -71.912939 | 281           | 49          | 0.13611111    | 360          | 18054       | 23586,81622 | DDX58,UNC93B1,RTP4,TRIM5,TRIM22,NT5C3A,IL23A,OAS1,MX2,IFITM1,IF     |   |   |   |   |   |
| 5  | GO:0071357 | cellular response to type I interferon                             | 6.31E-32   | -71.841146 | 67            | 29          | 0.08055556    | 360          | 18054       | #####       | MX1,IFITM3,HLA-E,IFI35,OAS1,IFIT3,OAS1,MX2,XAF1,STAT2,IFITM1,ISG15, |   |   |   |   |   |
| 6  | GO:0060337 | type I interferon signaling pathway                                | 6.31E-32   | -71.841146 | 67            | 29          | 0.08055556    | 360          | 18054       | #####       | MX1,IFITM3,HLA-E,IFI35,OAS1,IFIT3,OAS1,MX2,XAF1,STAT2,IFITM1,ISG15, |   |   |   |   |   |
| 7  | GO:0098542 | defense response to other organism                                 | 2.38E-26   | -59.001489 | 842           | 72          | 0.2           | 360          | 18054       | 81622,64108 | UNC93B1,RTP4,TRIM5,NT5C3A,IL23A,BATF2,EIF2AK2,SP100,TRIM25,IRF2,    |   |   |   |   |   |
| 8  | GO:0045087 | innate immune response                                             | 1.10E-24   | -55.165698 | 664           | 62          | 0.17222222    | 360          | 18054       | 23586,81622 | DDX58,UNC93B1,PARP14,TRIM5,HLA-E,IFI35,TRIM22,IL23A,OAS1,MX2,IFIT   |   |   |   |   |   |
| 9  | GO:0006952 | defense response                                                   | 1.78E-23   | -52.384192 | 1230          | 83          | 0.23055556    | 360          | 18054       | 81622,64108 | UNC93B1,RTP4,TRIM5,TAP1,NT5C3A,HMOX1,BATF2,TACR1,EIF2AK2,S          |   |   |   |   |   |
| 10 | GO:0009607 | response to biotic stimulus                                        | 6.56E-23   | -51.077923 | 1227          | 82          | 0.22777778    | 360          | 18054       | 1634,81622, | DCN,UNC93B1,RTP4,KLHL6,TRIM5,CHMP5,NT5C3A,IL23A,BATF2,EIF2AK2,S     |   |   |   |   |   |
| 11 | GO:0051707 | response to other organism                                         | 7.13E-23   | -50.995011 | 1201          | 81          | 0.225         | 360          | 18054       | 1634,81622, | DCN,UNC93B1,RTP4,KLHL6,TRIM5,CHMP5,NT5C3A,IL23A,BATF2,EIF2AK2,S     |   |   |   |   |   |
| 12 | GO:0043207 | response to external biotic stimulus                               | 7.93E-23   | -50.888365 | 1203          | 81          | 0.225         | 360          | 18054       | 1634,81622, | DCN,UNC93B1,RTP4,KLHL6,TRIM5,CHMP5,NT5C3A,IL23A,BATF2,EIF2AK2,S     |   |   |   |   |   |
| 13 | GO:0048525 | negative regulation of viral process                               | 1.10E-20   | -45.953445 | 101           | 25          | 0.06944444    | 360          | 18054       | #####       | MX1,PLSCR1,IFITM3,TRIM5,OAS1,CHMP3,OAS1,IFITM1,ISG15,ISG20,EIF2A    |   |   |   |   |   |
| 14 | GO:1903901 | negative regulation of viral life cycle                            | 4.37E-20   | -44.576829 | 85            | 23          | 0.06388889    | 360          | 18054       | #####       | MX1,PLSCR1,IFITM3,TRIM5,CHMP3,OAS1,IFITM1,ISG15,ISG20,EIF2A         |   |   |   |   |   |
| 15 | GO:1903900 | regulation of viral life cycle                                     | 1.83E-18   | -40.842783 | 149           | 27          | 0.075         | 360          | 18054       | #####       | MX1,PLSCR1,IFITM3,TRIM5,OAS1,CHMP3,OAS1,IFITM1,ISG15,ISG20,DDX3:    |   |   |   |   |   |
| 16 | GO:0050792 | regulation of viral process                                        | 1.87E-18   | -40.819656 | 208           | 31          | 0.08611111    | 360          | 18054       | #####       | MX1,PLSCR1,IFITM3,TRIM5,OAS1,CHMP3,OAS1,IFITM1,ISG15,ISG20,DDX3:    |   |   |   |   |   |
| 17 | GO:0043901 | negative regulation of multi-organism process                      | 6.20E-18   | -39.621468 | 233           | 32          | 0.08888889    | 360          | 18054       | #####       | PARP14,TRIM5,HLA-E,OAS1,IFITM1,EIF2AK2,DHX58,LGALS9,TRIM25,IFIT5,   |   |   |   |   |   |
| 18 | GO:0045071 | negative regulation of viral genome replication                    | 5.01E-17   | -37.532974 | 59            | 18          | 0.05          | 360          | 18054       | #####       | MX1,PLSCR1,IFITM3,TRIM5,OAS1,IFITM1,ISG15,ISG20,EIF2AK2,RSAD2,SHF1  |   |   |   |   |   |
| 19 | GO:0043903 | regulation of symbiosis, encompassing mutualism through parasitism | 6.73E-17   | -37.237805 | 235           | 31          | 0.08611111    | 360          | 18054       | #####       | MX1,PLSCR1,IFITM3,TRIM5,OAS1,CHMP3,OAS1,IFITM1,ISG15,ISG20,DDX3:    |   |   |   |   |   |
| 20 | GO:0006955 | immune response                                                    | 9.50E-17   | -36.892753 | 1544          | 82          | 0.22777778    | 360          | 18054       | 81622,89857 | UNC93B1,KLHL6,TRIM5,TAP1,IL19,IL23A,NBN,CD164,EIF2AK2,SP100,TRIM    |   |   |   |   |   |
| 21 | GO:0060333 | interferon-gamma-mediated signaling pathway                        | 1.39E-16   | -36.512646 | 72            | 19          | 0.05277778    | 360          | 18054       | #####       | MT2A,TRIM5,HLA-E,TRIM22,OAS1,OAS2,SP100,TRIM34,TRIM25,NM            |   |   |   |   |   |



GREAT version 4.0.4 current (08/19/2019 to now) ▾

## GREAT predicts functions of cis-regulatory regions.

Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions. Cis-regulatory regions can be identified via both experimental methods (e.g. [ChIP-seq](#)) and by computational methods (e.g. [comparative genomics](#)). For more see our [Nature Biotech Paper](#).

### News

-  Aug. 30, 2022: Our [WhichTF publication](#) that describes GREAT v4 update is now out. Please check our updated [citation page](#).
- Aug. 19, 2019: GREAT version 4 [adds support for human hg38 assembly and updates ontology datasets for all supported assemblies](#).
- Sep. 8, 2018: GREAT has served over 1 million job submissions.
- Oct. 23, 2017: GREAT is moved to a VM to eliminate proxy errors.
- June 22, 2017: GREAT hardware upgrade to meet increasing submission volume.
- Nov. 16, 2015: The [GREAT user help forums](#) are frozen.
- Feb. 15, 2015: GREAT version 3 [switches to Ensembl genes, adds support for zebrafish danRer7 and mouse mm10 assemblies, and adds new ontologies](#).
- Apr. 3, 2012: GREAT version 2 [adds new annotations to human and mouse ontologies and visualization tools for data exploration](#).
- Feb. 18, 2012: The [GREAT user help forums](#) are opened.
- May 2, 2010: GREAT version 1 is launched, concurrent to [Nature Biotechnology publication](#) (reprint, Faculty of 1000 "Must Read"). [How to Cite GREAT?](#)

[More news items...](#)

#### Species Assembly

- ☐ Human: GRCh38 ([UCSC hg38, Dec. 2013](#))
- ☒ Human: GRCh37 ([UCSC hg19, Feb. 2009](#)) ←
- ☐ Mouse: GRCm38 ([UCSC mm10, Dec. 2011](#))
- ☐ Mouse: NCBI build 37 ([UCSC mm9, Jul. 2007](#))

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

#### Test regions

- ☒ BED file:  STAT1\_6...eaks.bed ←
- ☐ BED data:

[What should my test regions file contain?](#)  
[How can I create a test set from a UCSC Genome Browser annotation track?](#)

#### Background

- ☒ Whole genome

GREAT version 4.0.4 current (08/19/2019 to now)

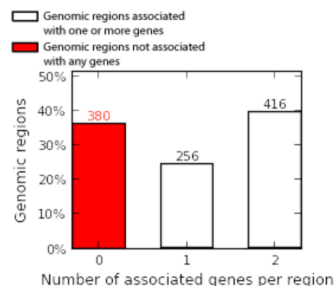
## + Job Description

## + Region-Gene Association Graphs

What do these graphs illustrate?

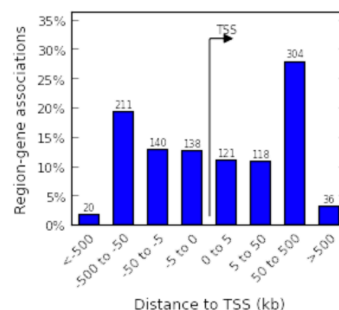
Number of associated genes per region

[Download as PDF.](#)



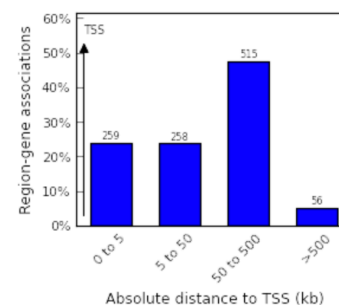
Binned by orientation and distance to TSS

[Download as PDF.](#)



Binned by absolute distance to TSS

[Download as PDF.](#)



## GO Biological Process (20+ terms)

Global controls

Table controls: [Export](#)

Shown top rows in this table: [20](#)

[Set](#)

Term annotation count: Min: [1](#)

Max: [Inf](#)

[Set](#)

Visualize this table: [\[select one\]](#)

| <a href="#">Term Name</a>                                       | Binom Rank | Binom Raw P-Value | Binom FDR Q-Val | Binom Fold Enrichment | Binom Observed Region Hits | Binom Region Set Coverage | Hyper Rank | Hyper FDR Q-Val | Hyper Fold Enrichment | Hyper Observed Gene Hits | Hyper Total Genes | Hyper Gene Set Coverage |
|-----------------------------------------------------------------|------------|-------------------|-----------------|-----------------------|----------------------------|---------------------------|------------|-----------------|-----------------------|--------------------------|-------------------|-------------------------|
| <a href="#">response to type I interferon</a>                   | 1          | 9.4057e-46        | 1.2364e-41      | 18.1886               | 51                         | 4.85%                     | 3          | 1.1470e-21      | 11.0564               | 31                       | 66                | 3.93%                   |
| <a href="#">type I interferon signaling pathway</a>             | 2          | 1.0073e-43        | 6.6204e-40      | 18.7363               | 48                         | 4.56%                     | 4          | 1.9708e-20      | 11.1908               | 29                       | 61                | 3.68%                   |
| <a href="#">defense response to virus</a>                       | 3          | 1.3461e-37        | 5.8981e-34      | 7.1997                | 72                         | 6.84%                     | 1          | 1.7569e-23      | 6.6495                | 50                       | 177               | 6.35%                   |
| <a href="#">response to virus</a>                               | 4          | 9.1412e-28        | 3.0040e-24      | 4.4579                | 80                         | 7.60%                     | 2          | 4.6761e-22      | 5.2310                | 58                       | 261               | 7.36%                   |
| <a href="#">defense response to other organism</a>              | 5          | 3.4492e-26        | 9.0679e-23      | 3.7164                | 92                         | 8.75%                     | 5          | 1.1582e-16      | 3.6869                | 65                       | 415               | 8.25%                   |
| <a href="#">negative regulation of viral process</a>            | 6          | 6.8843e-24        | 1.5082e-20      | 7.6286                | 43                         | 4.09%                     | 8          | 3.8392e-13      | 7.1411                | 27                       | 89                | 3.43%                   |
| <a href="#">negative regulation of viral life cycle</a>         | 7          | 4.0547e-22        | 7.6142e-19      | 7.7960                | 39                         | 3.71%                     | 12         | 1.3855e-12      | 7.0347                | 26                       | 87                | 3.30%                   |
| <a href="#">negative regulation of viral genome replication</a> | 8          | 1.0356e-21        | 1.7017e-18      | 10.7125               | 31                         | 2.95%                     | 13         | 6.5379e-12      | 9.2311                | 20                       | 51                | 2.54%                   |
| <a href="#">innate immune response</a>                          | 9          | 4.9210e-20        | 7.1875e-17      | 2.7343                | 106                        | 10.08%                    | 11         | 1.3411e-12      | 2.7826                | 74                       | 626               | 9.39%                   |
| <a href="#">negative regulation of multi-organism process</a>   | 10         | 7.4257e-17        | 9.7611e-14      | 4.8847                | 43                         | 4.09%                     | 18         | 2.7531e-9       | 4.9653                | 27                       | 128               | 3.43%                   |
| <a href="#">response to interferon-gamma</a>                    | 11         | 6.9336e-15        | 8.2856e-12      | 4.1876                | 44                         | 4.18%                     | 16         | 1.1675e-9       | 4.8073                | 29                       | 142               | 3.68%                   |
| <a href="#">interferon-gamma-mediated signaling pathway</a>     | 12         | 4.5185e-14        | 4.9497e-11      | 6.0569                | 29                         | 2.76%                     | 25         | 5.0423e-8       | 6.6204                | 18                       | 64                | 2.28%                   |
| <a href="#">regulation of viral process</a>                     | 13         | 1.6771e-13        | 1.6958e-10      | 3.5478                | 48                         | 4.56%                     | 19         | 3.2368e-9       | 4.0884                | 33                       | 190               | 4.19%                   |