ChIP-seq analysis: reads alignment

Table 1 | **A subset of software tools available for three key steps in the analysis of sequence data**

Software tool	Web address	Notes
Short-read aligners		
BWA	http://bio-bwa.sourceforge.net	Fast and efficient; based on the Burrows–Wheeler transform
Bowtie	http://bowtie-bio.sourceforge.net	Similar to BWA, part of suite of tools that includes TopHat and CuffLinks for RNA-seq processing
GSNAP	http://research-pub.gene.com/gmap	Considers a set of variant allele inputs to better align to heterozygous sites
Wikipedia list of aligners	http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment	A comprehensive list of available short-read aligners, with descriptions and links to download the software
Peak callers		
MACS	http://liulab.dfci.harvard.edu/MACS	Fits data to a dynamic Poisson distribution; works with and without control data
PeakSeq	http://info.gersteinlab.org/PeakSeq	Takes into account differences in mappability of genomic regions; enrichment based on FDR calculation
ZINBA	http://code.google.com/p/zinba	Can incorporate multiple genomic factors, such as mappability and GC content; can work with point-source and broad-source peak data
Differential peak calling		
edgeR	http://www.bioconductor.org/packages/2.9/bioc/html/edgeR.html	Uses negative binomial distribution to model differences in tag counts; uses replicates to better estimate significant differences
DESeq	http://www-huber.embl.de/users/anders/DESeq	Also uses negative binomial distribution modelling, but differs in the calculation of the mean and variance of the distribution
baySeq	http://www.bioconductor.org/packages/release/bioc/html/baySeq.html	Uses empirical Bayes approach to identify significant differences; assumes negative binomial distribution of data
SAMSeq	http://www.stanford.edu/~junli07/research.html#SAM	Based on the popular SAM software; a non-parametric method that uses resampling to normalize for differences in sequencing depth

BWA, Burrows–Wheeler Aligner; DESeq, analysis of high-throughput sequencing to detect differential expression; FDR, false-discovery rate; GSNAP, Genomic Short-read Nucleotide Alignment Program; MACS, Model-based Analysis for ChIP–seq; RNA-seq, high-throughput RNA sequencing; SAM, significance analysis of microarrays; ZINBA, Zero-Inflated Negative Binomial Algorithm.

Mappability
The uniqueness of a stretch of DNA sequence compared with a whole-genome sequence. Short sequence reads can be confidently mapped to unique sequence, but less confidently mapped to sequence that occurs multiple times in a genome.

ChIP-seq analysis: reads alignment

Bowtie

An ultrafast memory-efficient short read aligner

JOHNS HOPKINS UNIVERSITY

Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

OSI certified

Recent news

Hiring Postdocs

- The Langmead and Salzberg labs have open positions for postdoctoral researchers. See the posting and please apply if you're interested in working with either or both of us.

0.12.9 release - 12/16/12

- Fixed a bug whereby read names would not be truncated at first whitespace character in unmapped or maxed-out SAM records.
- Fixed errors and warnings when compiling with `clang++`.
- Fixed most errors and warnings when compiling with recent versions of `g++`, though you may need to add `EXTRA_FLAGS=-Wno-enum-compare` to avoid all warnings.

0.12.8 release - 5/6/12

Site Map

Home
News archive
Getting started
Manual
Tools that use Bowtie

Latest Release

Bowtie 0.12.9 12/16/12

Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

For release updates, subscribe to the mailing list.

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.

OSI certified

ChIP-seq analysis: reads alignment

Options

(typically processors number, mismatches, redundancy etc…)

Genome index

(Different organisms, build)
(downloadable as prebuilt or create your own with bowtie-build as long as a fasta file of the genome is available)

Input file

Fasta, fastq, raw, color space
(http://www.activemotif.com/catalog/819/chip-sequencing-service)

ChIP-seq analysis: reads alignment

```
roberto@BCTHUMPER:~/Roberto_hg19_MS/sese/giorgio/SxaQSEQsWA048L6$ /home/kurdistani/bowtie-0.12.7/bowtie
No index, query, or output file specified!
Usage:
  bowtie [options]* <ebwt> {-1 <m1> -2 <m2> | --12 <r> | <s>} [<hit>]

  <m1>    Comma-separated list of files containing upstream mates (or the
          sequences themselves, if -c is set) paired with mates in <m2>
  <m2>    Comma-separated list of files containing downstream mates (or the
          sequences themselves if -c is set) paired with mates in <m1>
  <r>     Comma-separated list of files containing Crossbow-style reads.  Can be
          a mixture of paired and unpaired.  Specify "-" for stdin.
  <s>     Comma-separated list of files containing unpaired reads, or the
          sequences themselves, if -c is set.  Specify "-" for stdin.
  <hit>   File to write hits to (default: stdout)
Input:
  -q                 query input files are FASTQ .fq/.fastq (default)
  -f                 query input files are (multi-)FASTA .fa/.mfa
  -r                 query input files are raw one-sequence-per-line
  -c                 query sequences given on cmd line (as <mates>, <singles>)
  -C                 reads and index are in colorspace
  -Q/--quals <file>  QV file(s) corresponding to CSFASTA inputs; use with -f -C
  --Q1/--Q2 <file>   same as -Q, but for mate files 1 and 2 respectively
  -s/--skip <int>    skip the first <int> reads/pairs in the input
  -u/--qupto <int>   stop after first <int> reads/pairs (excl. skipped reads)
  -5/--trim5 <int>   trim <int> bases from 5' (left) end of reads
  -3/--trim3 <int>   trim <int> bases from 3' (right) end of reads
  --phred33-quals    input quals are Phred+33 (default)
  --phred64-quals    input quals are Phred+64 (same as --solexa1.3-quals)
  --solexa-quals     input quals are from GA Pipeline ver. < 1.3
  --solexa1.3-quals  input quals are from GA Pipeline ver. >= 1.3
  --integer-quals    qualities are given as space-separated integers (not ASCII)
Alignment:
  -v <int>           report end-to-end hits w/ <=v mismatches; ignore qualities
    or
  -n/--seedmms <int> max mismatches in seed (can be 0-3, default: -n 2)
  -e/--maqerr <int>  max sum of mismatch quals across alignment for -n (def: 70)
  -l/--seedlen <int> seed length for -n (default: 28)
  --nomaqround       disable Maq-like quality rounding for -n (nearest 10 <= 30)
  -I/--minins <int>  minimum insert size for paired-end alignment (default: 0)
  -X/--maxins <int>  maximum insert size for paired-end alignment (default: 250)
  --fr/--rf/--ff     -1, -2 mates align fw/rev, rev/fw, fw/fw (default: --fr)
  --nofw/--norc      do not align to forward/reverse-complement reference strand
  --maxbts <int>     max # backtracks for -n 2/3 (default: 125, 800 for --best)
```

ChIP-seq analysis: reads alignment

```
roberto@BCTHUMPER:~/Roberto_hg19_MS/sese/giorgio/SxaQSEQsWA048L6$ /home/kurdistani/bowtie-0.12.7/bowtie
No index, query, or output file specified!
Usage:
  bowtie [options]* <ebwt> {-1 <m1> -2 <m2> | --12 <r> | <s>} [<hit>]

  <m1>    Comma-separated list of files containing upstream mates (or the
          sequences themselves, if -c is set) paired with mates in <m2>
  <m2>    Comma-separated list of files containing downstream mates (or the
          sequences themselves if -c is set) paired with mates in <m1>
  <r>     Comma-separated list of files containing Crossbow-style reads.  Can be
          a mixture of paired and unpaired.  Specify "-" for stdin.
  <s>     Comma-separated list of files containing unpaired reads, or the
          sequences themselves, if -c is set.  Specify "-" for stdin.
  <hit>   File to write hits to (default: stdout)
Input:
  -q                 query input files are FASTQ .fq/.fastq (default)
  -f                 query input files are (multi-)FASTA .fa/.mfa
  -r                 query input files are raw one-sequence-per-line
  -c                 query sequences given on cmd line (as <mates>, <singles>)
  -C                 reads and index are in colorspace
  -Q/--quals <file>  QV file(s) corresponding to CSFASTA inputs; use with -f -C
  --Q1/--Q2 <file>   same as -Q, but for mate files 1 and 2 respectively
  -s/--skip <int>    skip the first <int> reads/pairs in the input
  -u/--qupto <int>   stop after first <int> reads/pairs (excl. skipped reads)
  -5/--trim5 <int>   trim <int> bases from 5' (left) end of reads
  -3/--trim3 <int>   trim <int> bases from 3' (right) end of reads
  --phred33-quals    input quals are Phred+33 (default)
  --phred64-quals    input quals are Phred+64 (same as --solexa1.3-quals)
  --solexa-quals     input quals are from GA Pipeline ver. < 1.3
  --solexa1.3-quals  input quals are from GA Pipeline ver. >= 1.3
  --integer-quals    qualities are given as space-separated integers (not ASCII)
Alignment:
  -v <int>           report end-to-end hits w/ <=v mismatches; ignore qualities
    or
  -n/--seedmms <int> max mismatches in seed (can be 0-3, default: -n 2)
  -e/--maqerr <int>  max sum of mismatch quals across alignment for -n (def: 70)
  -l/--seedlen <int> seed length for -n (default: 28)
  --nomaqround       disable Maq-like quality rounding for -n (nearest 10 <= 30)
  -I/--minins <int>  minimum insert size for paired-end alignment (default: 0)
  -X/--maxins <int>  maximum insert size for paired-end alignment (default: 250)
  --fr/--rf/--ff     -1, -2 mates align fw/rev, rev/fw, fw/fw (default: --fr)
  --nofw/--norc      do not align to forward/reverse-complement reference strand
  --maxbts <int>     max # backtracks for -n 2/3 (default: 125, 800 for --best)
```

ChIP-seq analysis: reads alignment

```
roberto@BCTHUMPER:~/Roberto_hg19_MS/sese/giorgio/SxaQSEQsWA048L6$ /home/kurdistani/bowtie-0.12.7/bowtie -p (n) -m 1 --strata --be
st -t -v 2 /home/kurdistani/bowtie-0.12.7/indexes/yeast/scer.
scer.1.ebwt      scer.2.ebwt      scer.3.ebwt      scer.4.ebwt      scer.rev.1.ebwt  scer.rev.2.ebwt
roberto@BCTHUMPER:~/Roberto_hg19_MS/sese/giorgio/SxaQSEQsWA048L6$ /home/kurdistani/bowtie-0.12.7/bowtie -p (n) -m 1 --strata --be
st -t -v 2 /home/kurdistani/bowtie-0.12.7/indexes/yeast/scer.
```

/path/to/bowtie/bowtie -p (n) -m 1 --strata --best -t -v 2 /path/to/bowtie/indexes/yeast/scer reb1_20_ip.fastq > reb1_20_ip.bowtie

/path/to/bowtie/bowtie -p (n) -m 1 --strata --best -t -v 2 /path/to/bowtie/indexes/yeast/scer reb1_20_ip.fastq > reb1_20_ip.bowtie

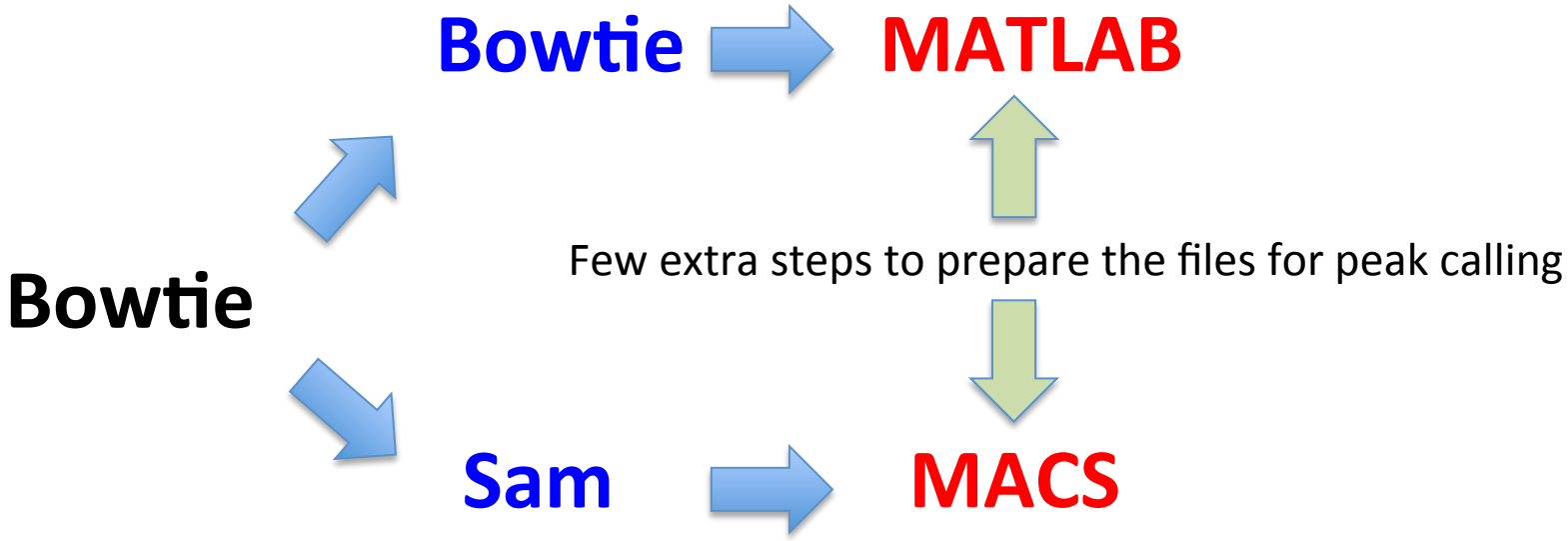
/home/kurdistani/bowtie-0.12.7/bowtie -p 6 -m 1 --strata --best -t -v 2 /home/kurdistani/bowtie-0.12.7/indexes/yeast/scer reb1_20_ip.fastq > reb1_20_ip.bowtie
Time loading forward index: 00:00:00
Time loading mirror index: 00:00:00
End-to-end 2/3-mismatch full-index search: 00:01:04
reads processed: 5000000
reads with at least one reported alignment: 1626706 (32.53%)
reads that failed to align: 2221985 (44.44%)
reads with alignments suppressed due to -m: 1151309 (23.03%)
Reported 1626706 alignments to 1 output stream(s)
Time searching: 00:01:04
Overall time: 00:01:04

ChIP-seq analysis: reads alignment

/home/kurdistani/bowtie-0.12.7/bowtie -p 6 -m 1 --strata --best -t -v 2 /home/kurdistani/bowtie-0.12.7/indexes/yeast/scer reb1_20_ip.fastq > reb1_20_ip.bowtie
Time loading forward index: 00:00:00
Time loading mirror index: 00:00:00
End-to-end 2/3-mismatch full-index search: 00:01:04
reads processed: 5000000
reads with at least one reported alignment: 1626706 (32.53%)
reads that failed to align: 2221985 (44.44%)
reads with alignments suppressed due to -m: 1151309 (23.03%)
Reported 1626706 alignments to 1 output stream(s)
Time searching: 00:01:04
Overall time: 00:01:04

/home/kurdistani/bowtie-0.12.7/bowtie -p 6 -m 1 -S --strata --best -t -v 2 /home/kurdistani/bowtie-0.12.7/indexes/yeast/scer reb1_20_ip.fastq > reb1_20_ip.sam
Time loading forward index: 00:00:00
Time loading mirror index: 00:00:00
End-to-end 2/3-mismatch full-index search: 00:01:13
reads processed: 5000000
reads with at least one reported alignment: 1626706 (32.53%)
reads that failed to align: 2221985 (44.44%)
reads with alignments suppressed due to -m: 1151309 (23.03%)
Reported 1626706 alignments to 1 output stream(s)
Time searching: 00:01:13
Overall time: 00:01:13

ChIP-seq analysis: reads alignment



ChIP-seq analysis: reads alignment

Ultimate goal of peak calling?

Determine whether there is enrichment compare to a control

ChIP-seq analysis: reads alignment

ChIP-seq analysis: peak calling

Most important files for subsequent analysis:

BED files:

Genome browser files:

UCSC Genome Bioinformatics

Home - Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - Help

Frequently Asked Questions: Data File Formats

General formats:

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- bigBed format
- bigWig format
- Chain format
- GenePred table format
- GFF format
- GTF format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

ChIP-seq analysis: peak calling

MACS

https://github.com/taoliu/MACS/blob/macs_v1/README.rst

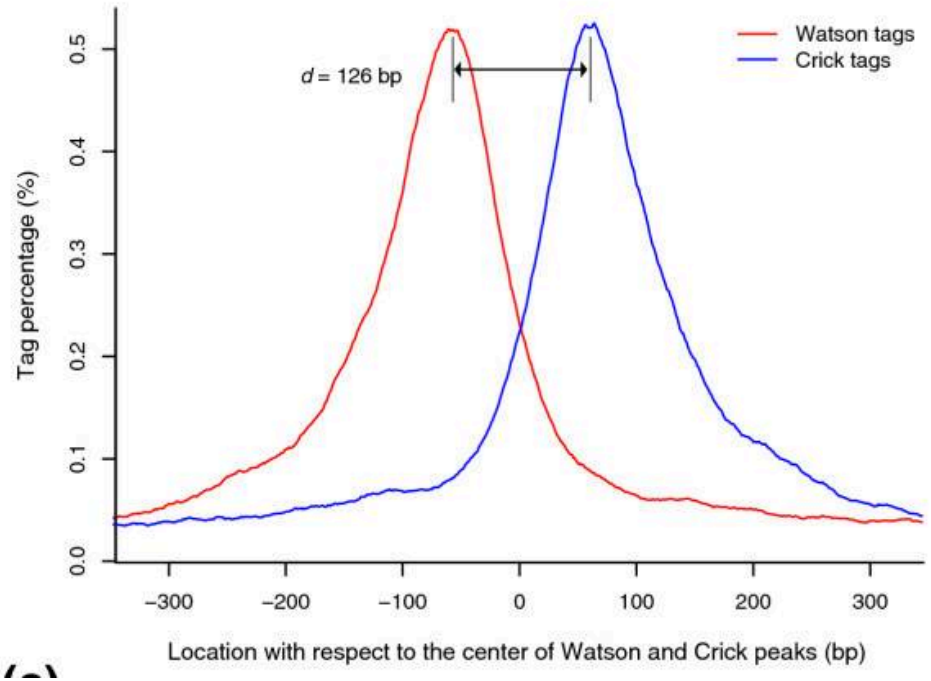
ChIP-seq analysis: peak calling

Open Access

Method

Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang[¤*], Tao Liu[¤*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nusbaum[¶], Richard M Myers[¥], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]



MACS requirements

BAM files

BAM is the compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments.

For custom track display, the main advantage of indexed BAM over PSL and other human-readable alignment formats is that only the portions of the files needed to display a particular region are transferred to UCSC. This makes it possible to display alignments from files that are so large that the connection to UCSC would time out when attempting to upload the whole file to UCSC.

ChIP-seq analysis: peak calling

SAMtools

SOURCEFORGE.NET®

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

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

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

ChIP-seq analysis: peak calling

What Information is in the SAM/BAM Header

The SAM/BAM header is not required, but if it is there, it contains generic information for the SAM/BAM file.

The header may contain the version information for the SAM/BAM file and information regarding whether or not and how the file is sorted.

It also contains supplemental information for alignment records like information about the reference sequences, the processing that was used to generate the various reads in the file, and the programs that have been used to process the different reads. The alignment records may then point to this supplemental information identifying which ones the specific alignment is associated with.

For example, a group of reads in the SAM/BAM file may all be assigned to the same reference sequence. Rather than every alignment containing information about the reference sequence, this information is put in the header, and the alignment "points" to the appropriate reference sequence in the header via the RNAME field. The header contains generic information about this reference like its length.

The SAM/BAM Header also may contain comments which are free-form text lines that can contain any information.

Header lines start with an '@'.
Example SAM
Example Header Lines

ChIP-seq analysis: peak calling

Using macs required few passages:

- samtools view –Sb SAM > BAM
- samtools sort in.bam > bam.sorted

ChIP-seq analysis: peak calling

Using macs for peak calling in unix:

- macs14 –t test.bam –c control.bam –f BAM –n name –g genome –w --bdg

- macs2 callpeak -t test.bam -c control.bam -f BAM -g hs –n name -B -q 0.01

ChIP-seq analysis: peak calling

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, peak summit position related to the start position of peak region, number of tags in peak region, -10*log10(pvalue) for the peak region (e.g. pvalue =1e-10, then this value should be 100), fold enrichment for this region against random Poisson distribution with local lambda, FDR in percentage. Coordinates in XLS is 1-based which is different with BED format.

2. NAME_peaks.bed is BED format file which contains the peak locations. You can load it to UCSC genome browser or Affymetrix IGB software. The 5th column in this file is the -10*log10pvalue of peak region.

3. NAME_summits.bed is in BED format, which contains the peak summits locations for every peaks. The 5th column in this file is the summit height of fragment pileup. If you want to find the motifs at the binding sites, this file is recommended.

4. NAME_negative_peaks.xls is a tabular file which contains information about negative peaks. Negative peaks are called by swapping the ChIP-seq and control channel.

5. NAME_model.r is an R script which you can use to produce a PDF image about the model based on your data. Load it to R by:

   ```
   R --vanilla < 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.

6. NAME_treat/control_afterfiting.wig.gz files in NAME_MACS_wiggle directory are wiggle format files which can be imported to UCSC genome browser/GMOD/Affy IGB. The .bdg.gz files are in bedGraph format which can also be imported to UCSC genome browser or be converted into even smaller bigWig files.

ChIP-seq analysis: peak calling

MATLAB

bowtieout files

-	277382		0
-	268828		0
-	83708	0	
-	55911	0	
-	404570		0
-	78840	0	
-	518096		0
-	535814		0
-	468305		0
-	165708		0

Chr length file

chrI	230218
chrII	813184
chrIII	316620
chrIV	1531933
chrV	576874
chrVI	270161
chrVII	1090940
chrVIII	562643
chrIX	439888
chrX	745751
chrXI	666816
chrXII	1078177
chrXIII	924431
chrXIV	784333
chrXV	1091291
chrXVI	948066
chrM	85779

Genes file

I	42176	42719	+	RNA	reference	YAL055W
I	42880	45022	-	RNA	reference	YAL054C
I	45898	48250	+	RNA	reference	YAL053W
I	48563	51707	+	RNA	reference	YAL051W
I	51854	52595	-	RNA	reference	YAL049C
I	52800	54789	-	RNA	reference	YAL048C
I	54583	54913	+	RNA	reference	YAL047W-A
I	54988	56857	-	RNA	reference	YAL047C
I	57028	57385	-	RNA	reference	YAL046C
I	57487	57796	-	RNA	reference	YAL045C
I	57517	57850	+	RNA	reference	YAL044W-A

ChIP-seq analysis: peak calling

Using MATLAB for peak calling:

```
function [numpeaks] = chip_seq_ratio(ipname,inputname,grbed,tilingfile,Plim,neigh,windowsize)

%%ipname = filename of IP file which will be read as
%%ipname-chr#-bowtieout.txt (e.g. h3k36chr1bowtieout.txt)
%%inputname = filename of input file which will be read as
%%inputname-chr#-bowtieout.txt (e.g. h3cchr1bowtieout.txt)
%%grbed = 0 or 1.  0 doesn't output any gr or bed files, while 1 does
%%tilingfile = 0 or 1.  0 doesn't output any Excel TSS or TTS tiling files, while 1 does
%%Plim = value of log10(Pvalue) threshold (e.g. -4)
%%negh = 0 or 1.  1 requires that significant windows also have significant neighbors
%%windowsize = size of window used to tile genome (must be multiple of ten)
%%% Sample command:  chip_seq_ratio('k36chr','h3cchr',0,0,4,0,100)

[chr gstart gend gpm RNA reference gs] = textread('scer_genes.txt','%s %d %d %c %s %s %s');
%TEST%[chr gstart gend gpm RNA reference gs] = textread('../mm8_kg4mp.txt','%s %d %d %c %s %s %s');
%%[iechr iegstart iegend iegpm] = textread('../../seq_gene_cds.txt','%s %d %d %c');
[chrlen chrlength] = textread('scer_chr_length.txt','%s %d');
%TEST%[chrlen chrlength] = textread('../mm8_chrlens.txt','%s %d');

uchr = {'I','II','III','IV','V','VI','VII','VIII','IX','X','XI','XII','XIII','XIV','XV','XVI'};


%%%%compute sum of counts of both samples for normalization

sumchrcounts1 = 0;
sumchrcounts2 = 0;

fprintf('computing total counts for two libraries\n');

for j=1:length(uchr)
```