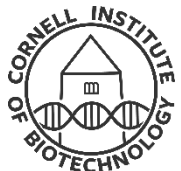


Epigenomic data analysis

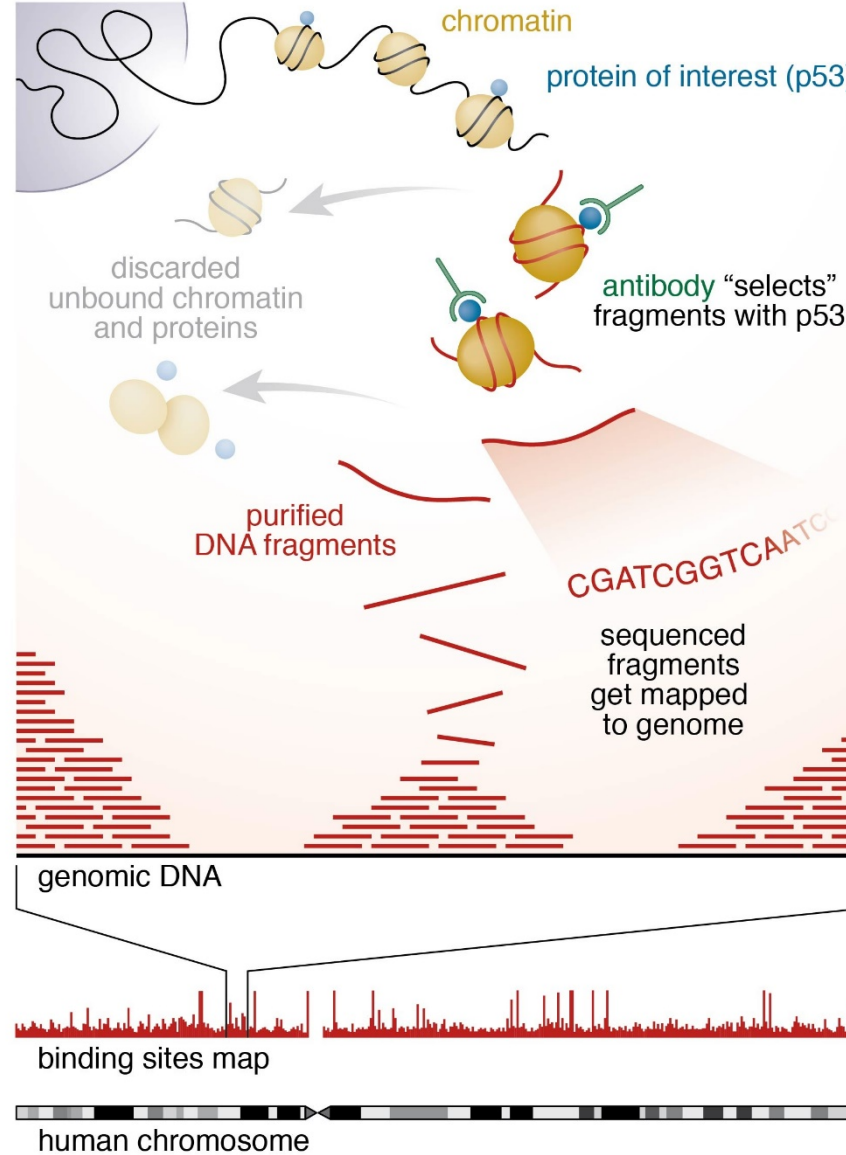
Week 3

Qi Sun, William Lai and Jeff Glaubitz

Bioinformatics Facility & Epigenomics Facility

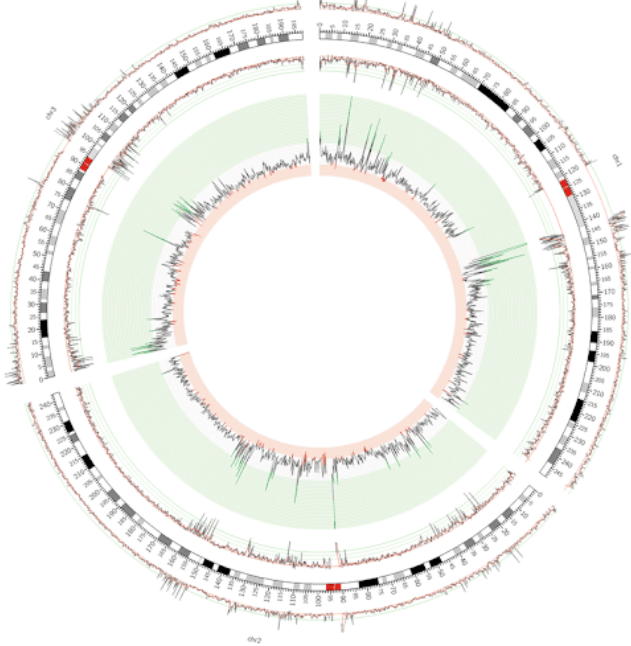


Peak identification and downstream analysis

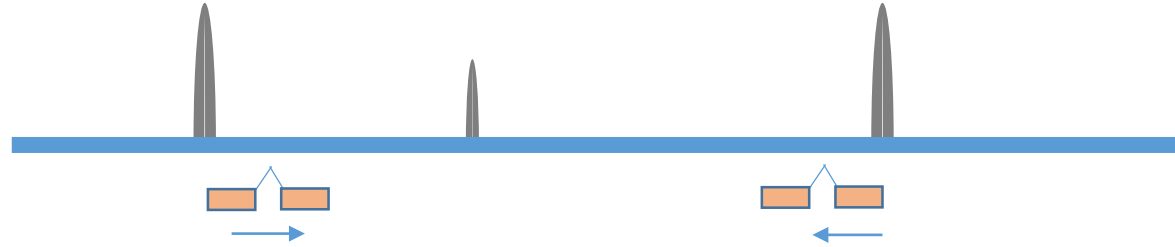


Examples of downstream analysis

1. Density of peaks across the genome

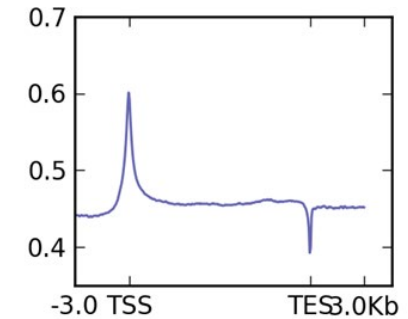


2. Identification of genes near the peaks

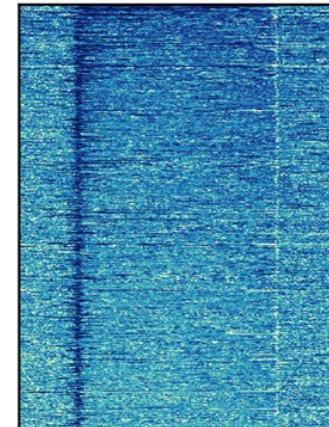
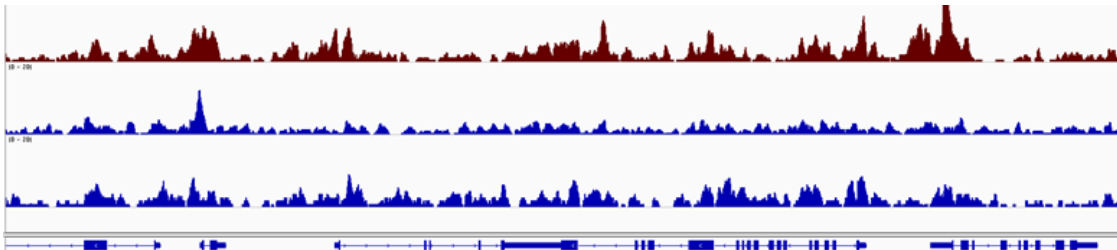


4. Scale regions.

(e.g. transcription start to transcription end)



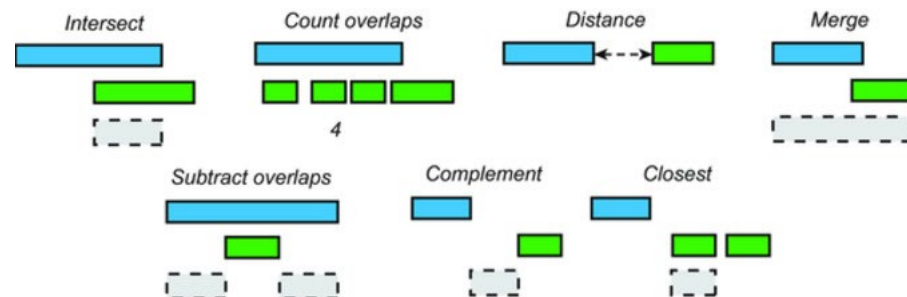
3. Count the number of reads in each feature





BEDTools: Overview

- Fast, flexible tools for comparing large sets of genomic features
- Perform “genome arithmetic” (e.g., How many? How far?)
- Written by Aaron Quinlan (University of Utah)
- *Curr Protoc Bioinform* 2014 <https://doi.org/10.1002/0471250953.bi1112s47>
- BEDTools documentation: <http://bedtools.readthedocs.org/>
- bedtools-discuss Google group: <https://groups.google.com/g/bedtools-discuss>





BEDTools: Full list of 43 tools

<https://bedtools.readthedocs.io/en/latest/content/overview.html#summary-of-available-tools>

annotate	Annotate coverage of features from multiple files.	map	Apply a function to a column for each overlapping interval.
bamtobed	Convert BAM alignments to BED (& other) formats.	maskfasta	Use intervals to mask sequences from a FASTA file.
bamtofastq	Convert BAM records to FASTQ records.	merge	Combine overlapping/nearby intervals into a single interval.
bed12tobed6	Breaks BED12 intervals into discrete BED6 intervals.	multicov	Counts coverage from multiple BAMs at specific intervals.
bedpetobam	Convert BEDPE intervals to BAM records.	multiinter	Identifies common intervals among multiple interval files.
bedtobam	Convert intervals to BAM records.	nuc	Profile the nucleotide content of intervals in a FASTA file.
closest	Find the closest, potentially non-overlapping interval.	overlap	Computes the amount of overlap from two intervals.
cluster	Cluster (but don't merge) overlapping/nearby intervals.	pairtobed	Find pairs that overlap intervals in various ways.
complement	Extract intervals _not_ represented by an interval file.	pairtopair	Find pairs that overlap other pairs in various ways.
coverage	Compute the coverage over defined intervals.	random	Generate random intervals in a genome.
expand	Replicate lines based on lists of values in columns.	reldist	Calculate the distribution of relative distances b/w two files.
fisher	Calculate Fisher statistic b/w two feature files.	sample	Sample random records from file using reservoir sampling.
flank	Create new intervals from the flanks of existing intervals.	shift	Adjust the position of intervals.
genomecov	Compute the coverage over an entire genome.	shuffle	Randomly redistribute intervals in a genome.
getfasta	Use intervals to extract sequences from a FASTA file.	slop	Adjust the size of intervals.
groupby	Group by common cols. & summarize oth. cols.	sort	Order the intervals in a file.
igv	Create an IGV snapshot batch script.	spacing	Report the gap lengths between intervals in a file.
intersect	Find overlapping intervals in various ways.	split	Split a file into multiple files with equal records or base pairs.
jaccard	Calculate the Jaccard statistic b/w two sets of intervals.	subtract	Remove intervals based on overlaps b/w two files.
links	Create a HTML page of links to UCSC locations.	tag	Tag BAM alignments based on overlaps with interval files.
makewindows	Make interval "windows" across a genome.	unionbedg	Combines coverage intervals from multiple BEDGRAPH files.
		window	Find overlapping intervals within a window around an interval.



BEDTools: 27 most useful, classified

Modify single input interval file

merge	Combine overlapping/nearby intervals into a single interval.
slop	Adjust the size of intervals.
flank	Create new intervals from the flanks of existing intervals.
complement	Extract intervals _not_ represented by an interval file.

Coverage & bam utilities

coverage	Compute the coverage over defined intervals.
genomecov	Compute the coverage over an entire genome.
multicov	Counts coverage from multiple BAMs at specific intervals.
bamtoBED	Convert BAM alignments to BED (& other) formats.
bamtoFASTQ	Convert BAM records to FASTQ records.

Compare 2 or more interval files

intersect	Find overlapping intervals in various ways.
closest	Find the closest, potentially non-overlapping interval.
subtract	Remove intervals based on overlaps b/w two files.
window	Find overlapping intervals within a window around an interval.
overlap	Computes the amount of overlap from two intervals.

Database-style summaries

map	Apply a function to a column for each overlapping interval.
groupby	Group by common cols. & summarize oth. cols. (~ SQL "groupBy")
expand	Replicate lines based on lists of values in columns.

Genome & FASTA utilities

makewindows	Make interval "windows" across a genome.
getfasta	Use intervals to extract sequences from a FASTA file.
nuc	Profile the nucleotide content of intervals in a FASTA file.
maskfasta	Use intervals to mask sequences from a FASTA file.

Statistics & hypothesis testing

jaccard	Calculate the Jaccard statistic b/w two sets of intervals.
fisher	Calculate Fisher statistic b/w two feature files.
reldist	Calculate the distribution of relative distances b/w two files.
shuffle	Randomly redistribute intervals in a genome.
random	Generate random intervals in a genome.
sample	Sample random records from file using reservoir sampling.



BEDTools: General considerations

- Most (but not all!) of the tools are documented here:
<https://bedtools.readthedocs.io/en/latest/content/overview.html#summary-of-available-tools>
- Command line help is available too:

```
bedtools -h 2>&1 | less          # see all of the available BEDTools commands
bedtools makewindows -h 2>&1 | less # get help for the makewindows tool
```

- Multiple input formats accepted (bed, gff, vcf, *bam*, *fasta*)
- Some tools require sorted input, those that don't often work faster
- So you might as well pre-sort always
 - by chromosome name, then position:

```
myUnsorted.bed sort -k1,1 -k2,2n > mySorted.bed
```



BEDTools: Single input interval file

Modify single input interval file

[merge](#)

Combine overlapping/nearby intervals into a single interval.

[slop](#)

Adjust the size of intervals.

[flank](#)

Create new intervals from the flanks of existing intervals.

[complement](#)

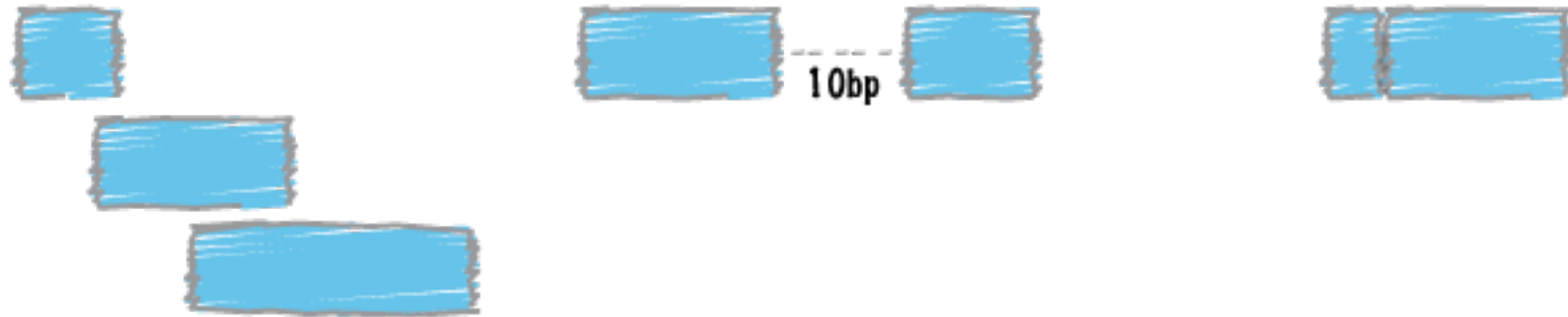
Extract intervals not represented by an interval file.



BEDTools: merge

```
bedtools merge [OPTIONS] -i <BED/GFF/VCF/BAM>
```

Input (I)



merge |



merge |
(-d 10)



merge |
(-n)

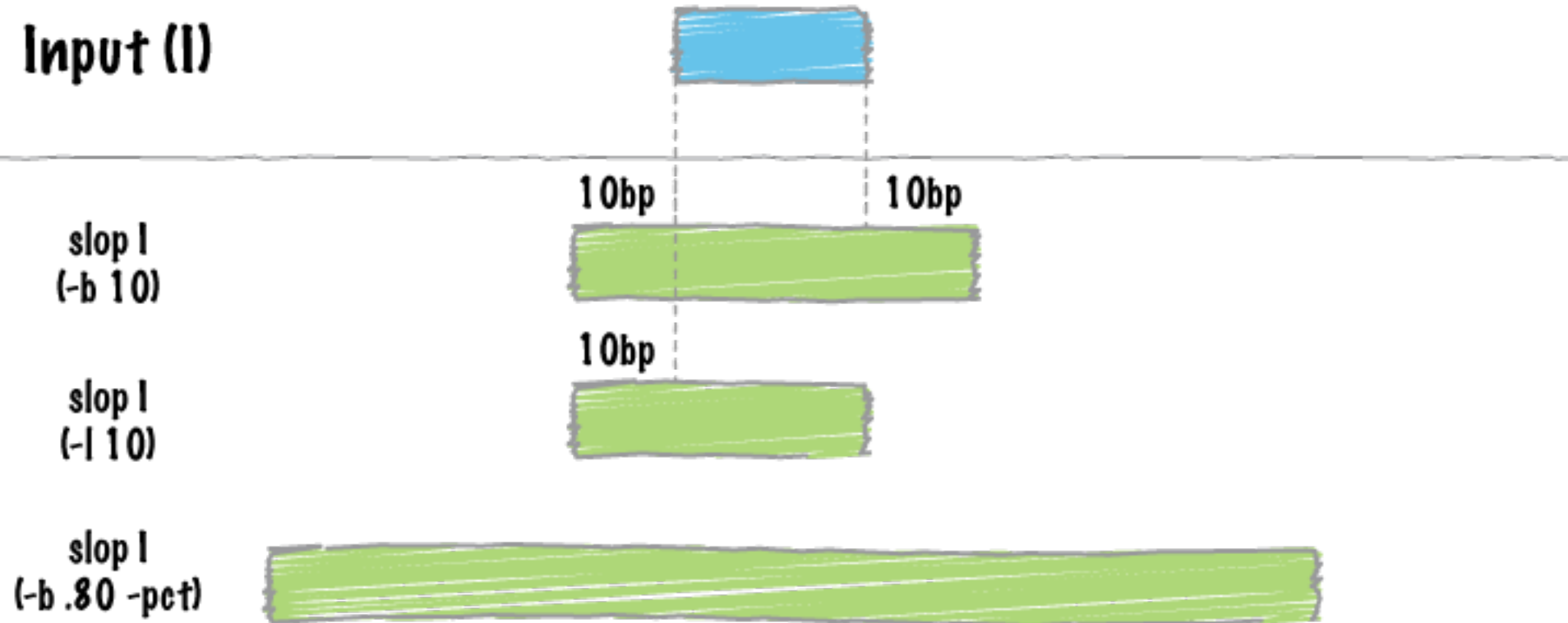




BEDTools: slop

```
bedtools slop [OPTIONS] -i <BED/GFF/VCF> -g <GENOME> [-b or (-l and -r)]
```

Input (I)

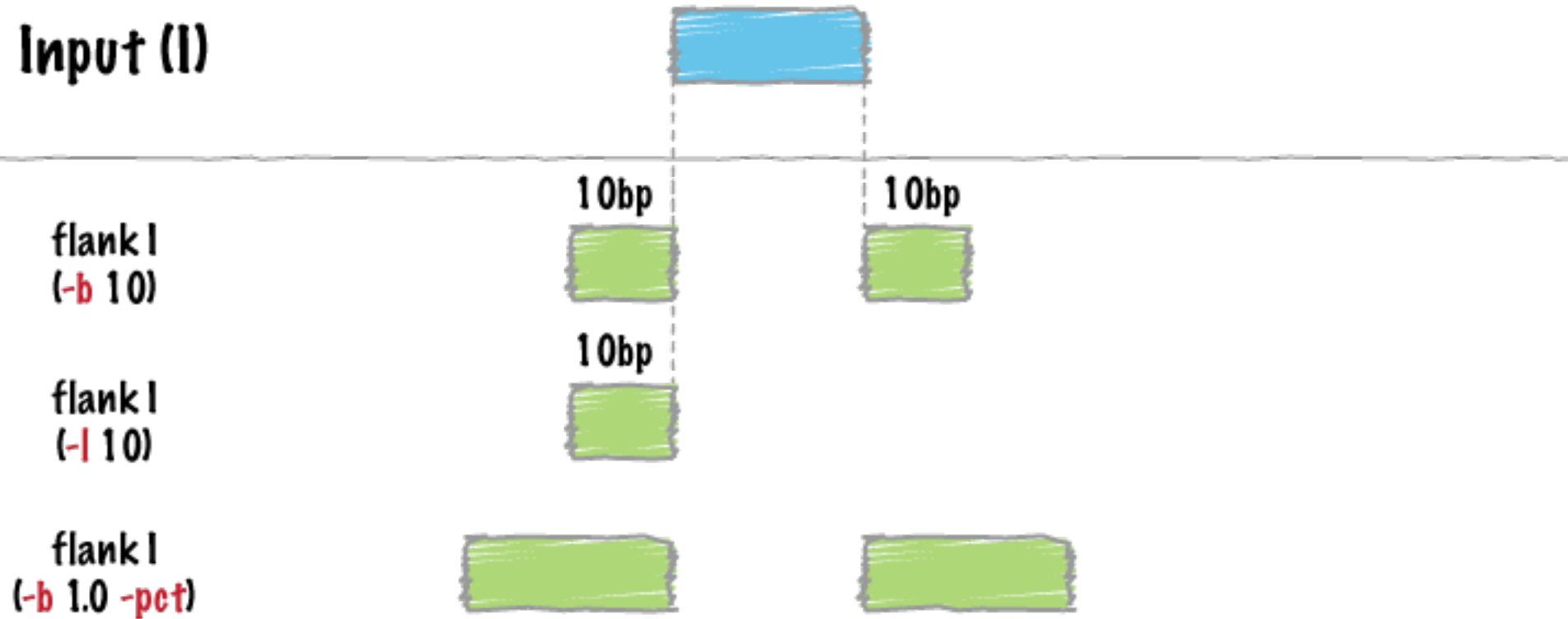




BEDTools: flank

```
bedtools flank [OPTIONS] -i <BED/GFF/VCF> -g <GENOME> [-b or (-l and -r)]
```

Input (I)





BEDTools: complement

```
bedtools complement -i <BED/GFF/VCF> -g <GENOME>
```

Input (I)



complement
(I)





BEDTools: 2 or more input interval files

Compare 2 or more interval files

[intersect](#)

Find overlapping intervals in various ways.

[closest](#)

Find the closest, potentially non-overlapping interval.

[subtract](#)

Remove intervals based on overlaps b/w two files.

[window](#)

Find overlapping intervals within a window around an interval.

[overlap](#)

Computes the amount of overlap from two intervals.



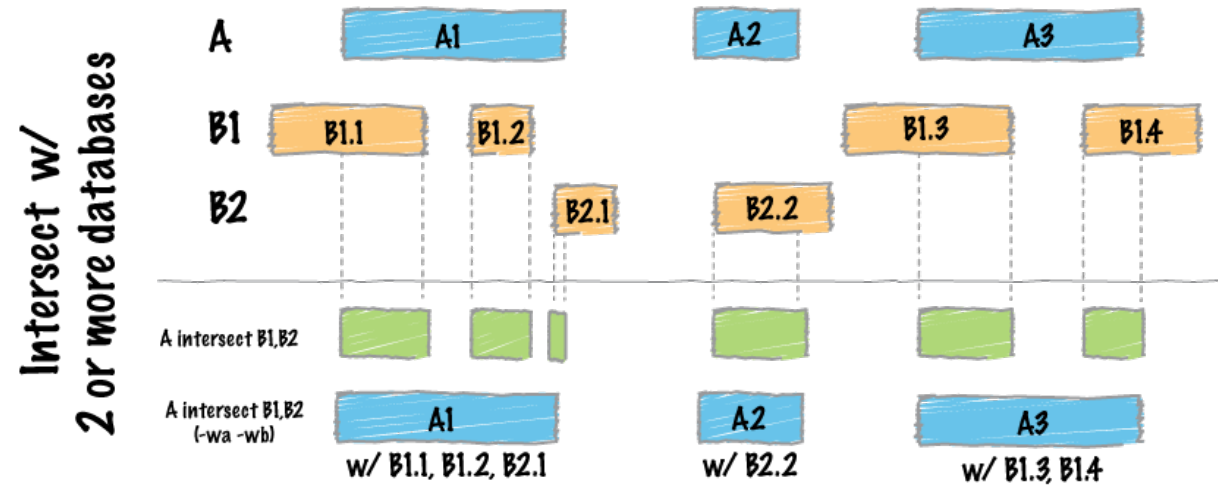
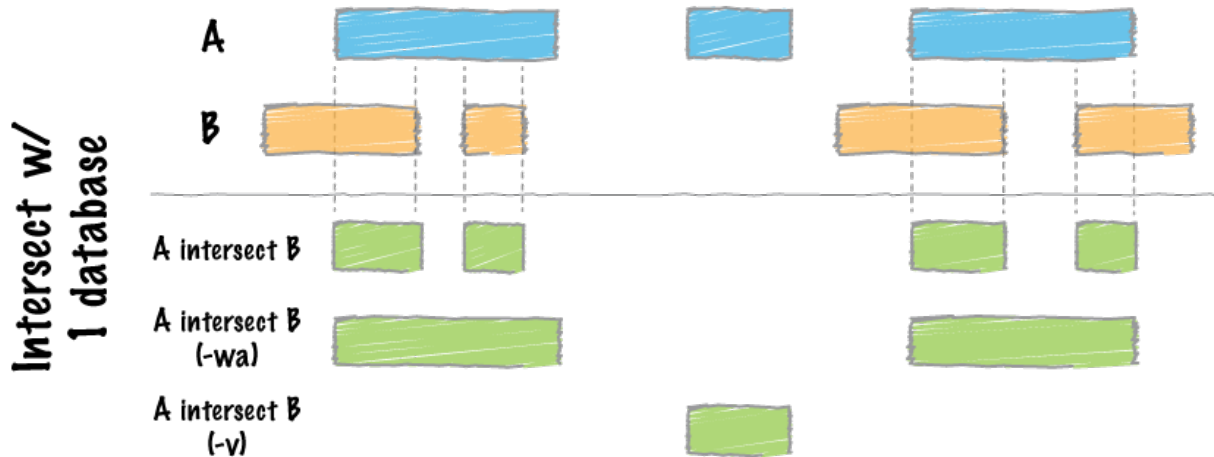
BEDTools: intersect

```
bedtools intersect [OPTIONS] -a <FILE> -b <FILE1, FILE2, ..., FILEN>
```

-a is the "query" file

-b is the "database" file(s)

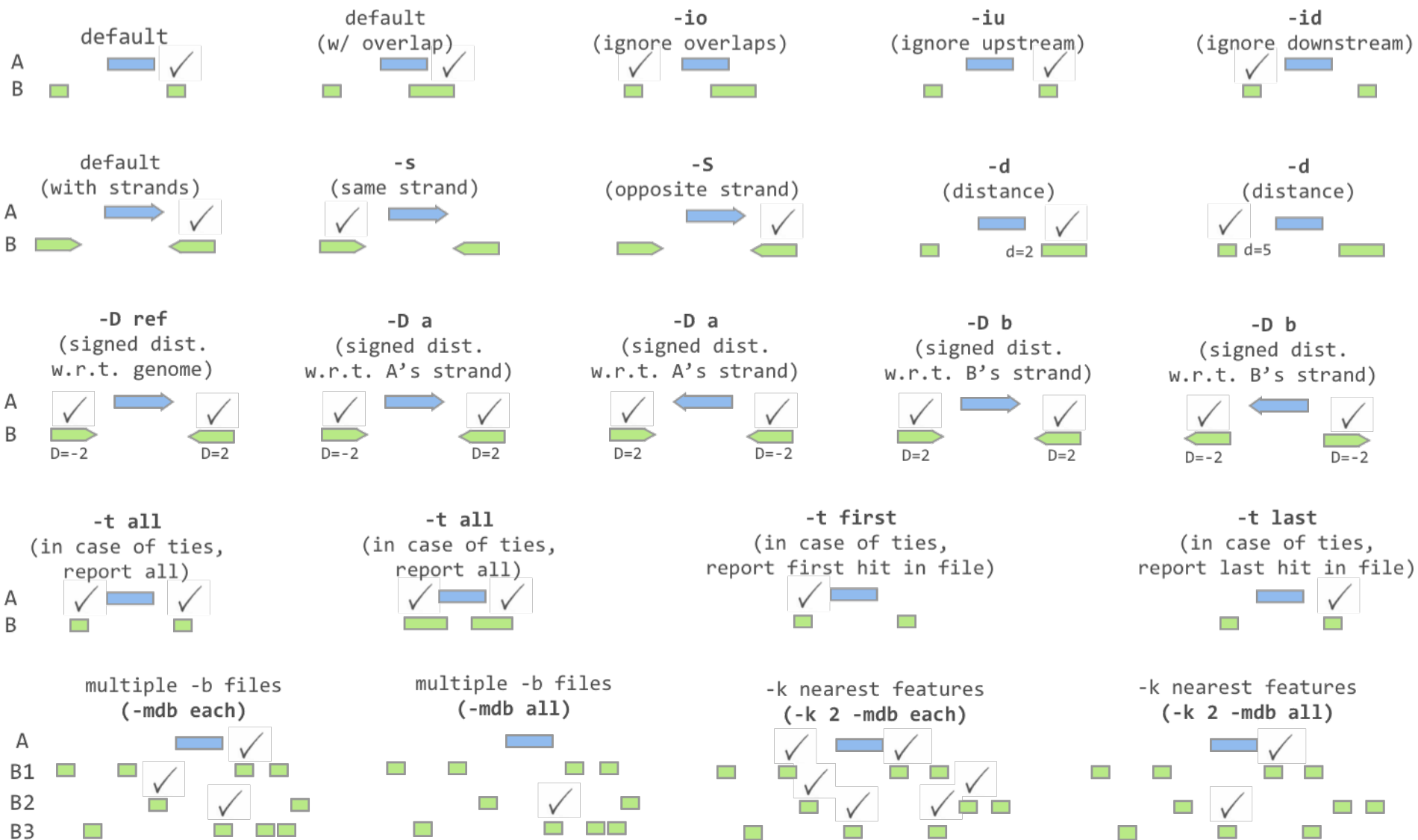
multiple -b files can usually be specified: -b <FILE1, FILE2, ..., FILEN>





BEDTools: closest

```
bedtools closest [OPTIONS] -a <FILE> -b <FILE1, FILE2, ..., FILEN>
```





BEDTools: subtract

```
bedtools subtract [OPTIONS] -a <BED/GFF/VCF> -b <BED/GFF/VCF>
```

A



B



A subtract B



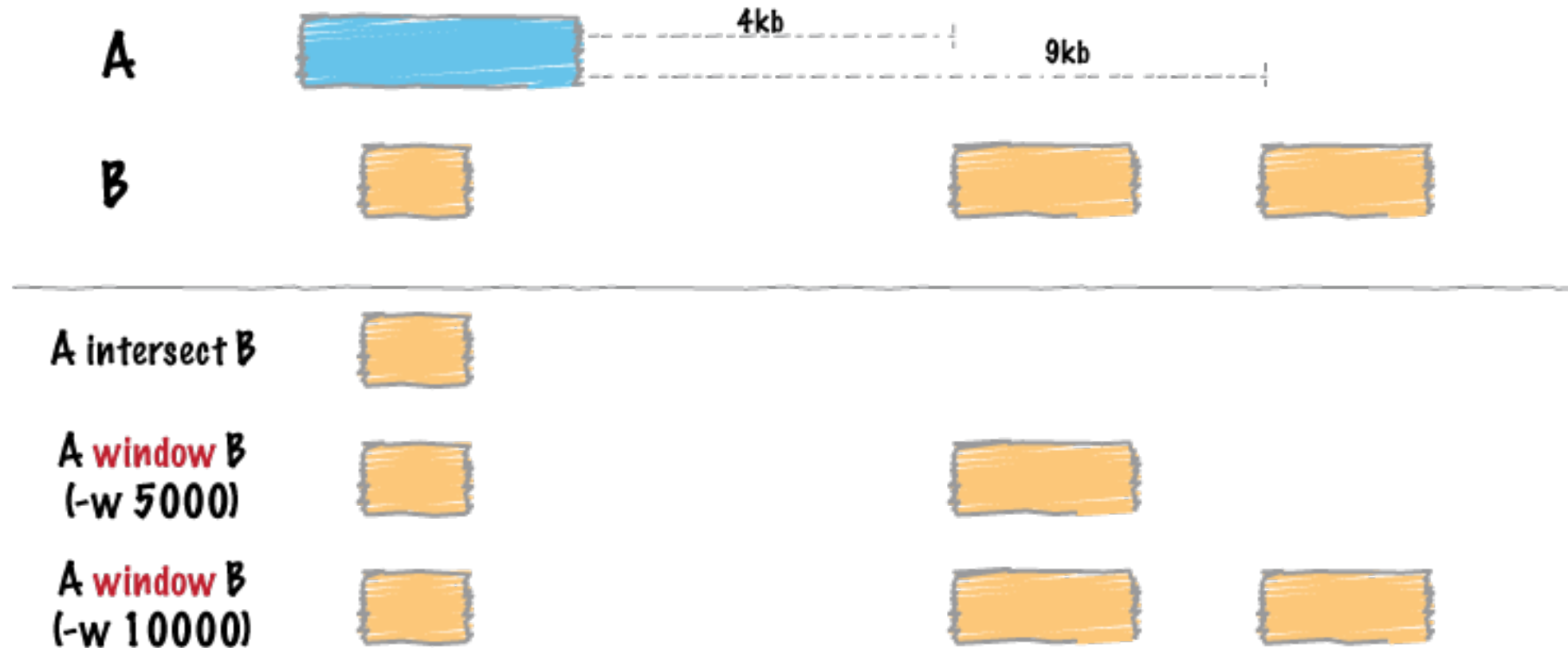
A subtract B
(-A)





BEDTools: window

```
bedtools window [OPTIONS] [-a|-abam] -b <BED/GFF/VCF>
```





BEDTools: overlap

```
bedtools overlap [OPTIONS] -i <input> -cols s1,e1,s2,e2
```

- How much overlap or how far away?
- BEDTools `overlap` works with the output of BEDTools `window`:

```
bedtools window -a A.bed -b B.bed -w 10  
chr1 10 20 A chr1 15 25 B  
chr1 10 20 C chr1 25 35 D
```

```
bedtools window -a A.bed -b B.bed -w 10 | bedtools overlap -i stdin -cols 2,3,6,7  
chr1 10 20 A chr1 15 25 B 5  
chr1 10 20 C chr1 25 35 D -5
```






BEDTools: Genome & FASTA utilities





Genome & FASTA utilities

makewindows	Make interval “windows” across a genome.
nuc	Profile the nucleotide content of intervals in a FASTA file.
getfasta	Use intervals to extract sequences from a FASTA file.
maskfasta	Use intervals to mask sequences from a FASTA file.

getfasta

FASTA	ACAGACTGGTATGAAGGTGGCCACAATTCAGAAAGAAAAAGAAGAGC			
BED				
getfasta	GACT	TGAAGGT	AAAAAAG	

maskfasta

FASTA	ACAGACTGGTATGAAGGTGGCCACAATTCAGAAAGAAAAAGAAGAGC			
BED				
FASTA'	ACANNNNGGTANNNNNN	GGCCACA	NNNNNNNAAGA	NNNNNNAGAGC



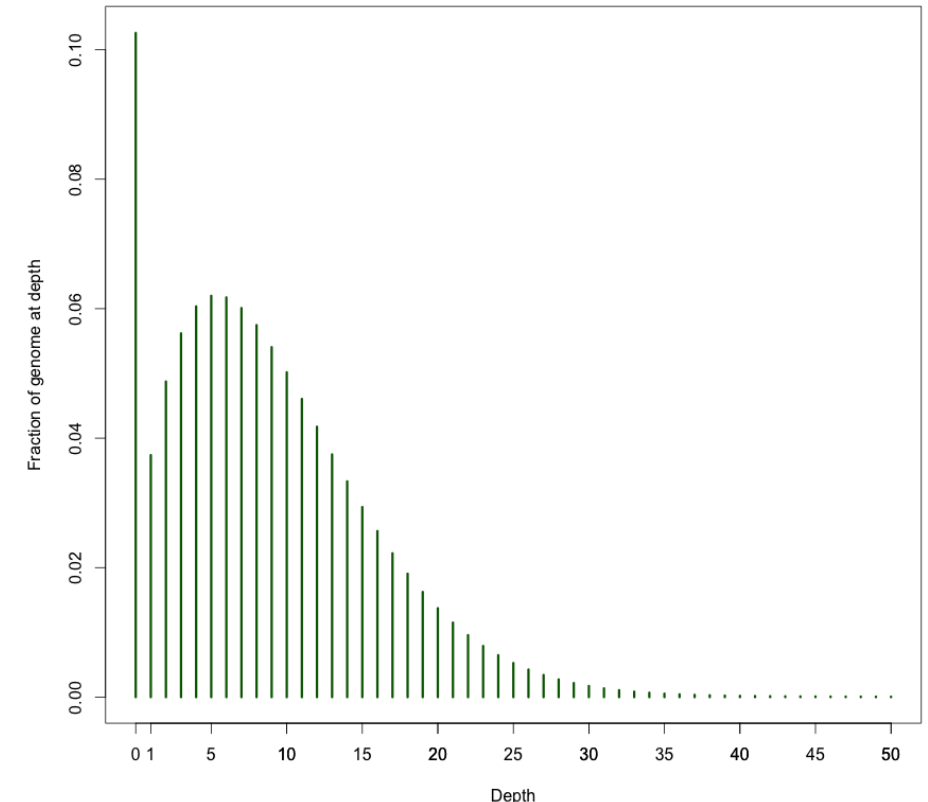
Compute the coverage over defined intervals.

Compute the coverage over an entire genome.

Counts coverage from multiple BAMs at specific intervals.

Convert BAM alignments to BED (& other) formats.

Convert BAM records to FASTQ records.





BEDTools: Database-style summaries

[map](#)

Apply a function to a column for each overlapping interval.

[groupby](#)

Group by common cols. & summarize oth. cols. (~ SQL “groupBy”)

[expand](#)

Replicate lines based on lists of values in columns.

[map](#)

B score = 3 1 5 4 6

A [blue bar] [blue bar]

B_{score} map A
(mean) mean(3, 1, 5) = 3 mean(4, 6) = 5

B_{score} map A
(max) max(3, 1, 5) = 5 max(4, 6) = 6

Valid operations:

sum, min, max, absmin, absmax

mean, median, mode, antimode

stdev, sstdev

collapse

distinct

distinct_sort_num

distinct_sort_num_desc

count

count_distinct

first (i.e., just the first value in the column)

last (i.e., just the last value in the column)

Default: sum

Multiple operations can be specified in a comma-delimited list.



BEDTools: Database-style summaries

[map](#)

Apply a function to a column for each overlapping interval.

[groupby](#)

Group by common cols. & summarize oth. cols. (~ SQL “groupBy”)

[expand](#)

Replicate lines based on lists of values in columns.

bedtools expand

```
$ cat test.txt
```

chr1	10	20	1,2,3	10,20,30
chr1	40	50	4,5,6	40,50,60

```
$ bedtools expand test.txt -c 5
```

chr1	10	20	1,2,3	10
chr1	10	20	1,2,3	20
chr1	10	20	1,2,3	30
chr1	40	50	4,5,6	40
chr1	40	50	4,5,6	50
chr1	40	50	4,5,6	60

```
$ bedtools expand test.txt -c 4,5
```

chr1	10	20	1	10
chr1	10	20	2	20
chr1	10	20	3	30
chr1	40	50	4	40
chr1	40	50	5	50
chr1	40	50	6	60



BEDTools: Statistics & Hypothesis Testing

[jaccard](#)

Calculate the Jaccard statistic b/w two sets of intervals.

[fisher](#)

Calculate Fisher statistic b/w two feature files.

[reldist](#)

Calculate the distribution of relative distances b/w two files.

[shuffle](#)

Randomly redistribute intervals in a genome.

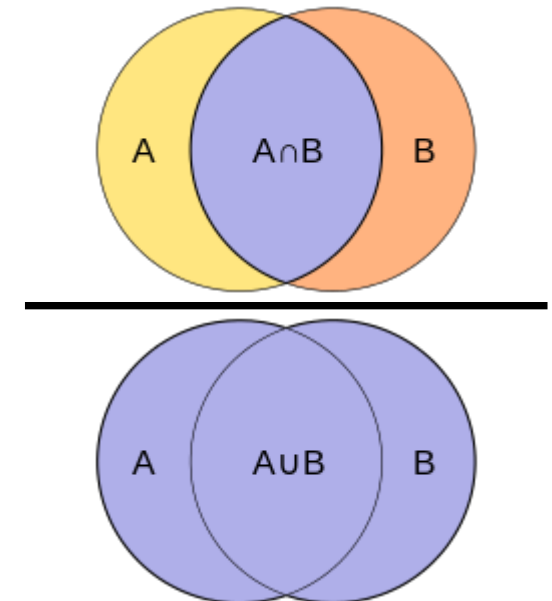
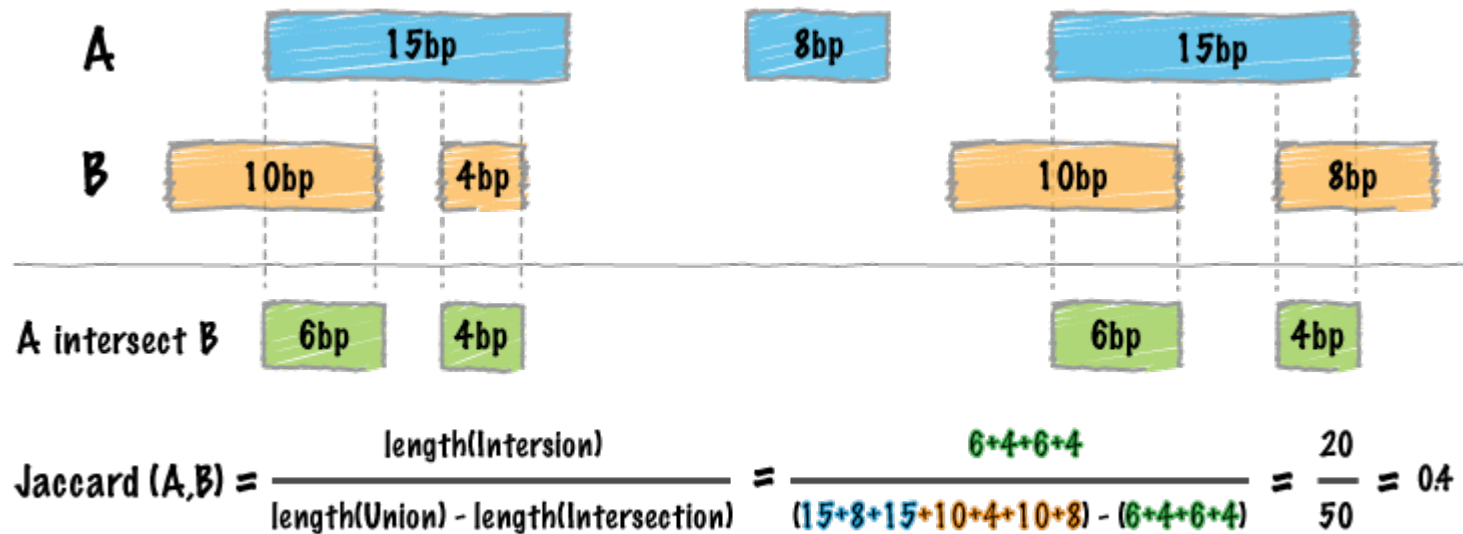
[random](#)

Generate random intervals in a genome.

[sample](#)

Sample random records from file using reservoir sampling.

[jaccard](#)





BEDTools: Statistics & Hypothesis Testing

[jaccard](#)

Calculate the Jaccard statistic b/w two sets of intervals.

[fisher](#)

Calculate Fisher statistic b/w two feature files.

[reldist](#)

Calculate the distribution of relative distances b/w two files.

[shuffle](#)

Randomly redistribute intervals in a genome.

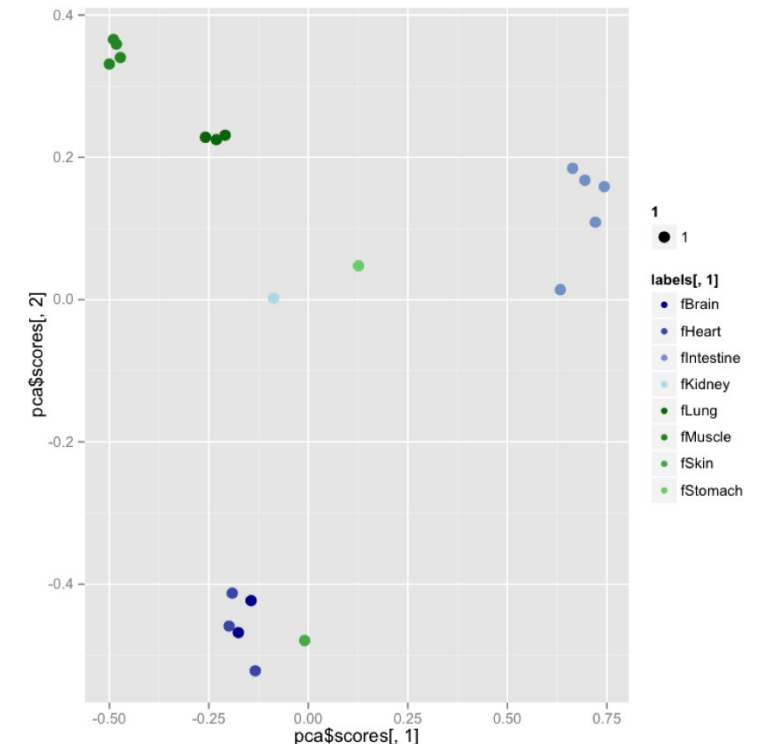
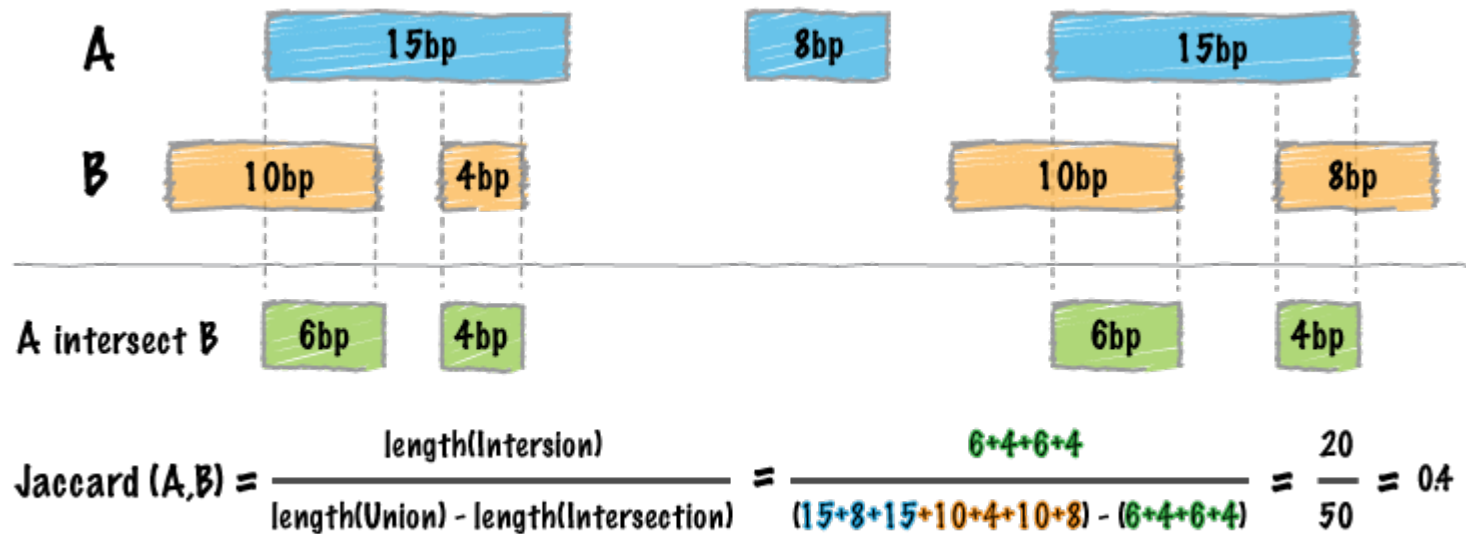
[random](#)

Generate random intervals in a genome.

[sample](#)

Sample random records from file using reservoir sampling.

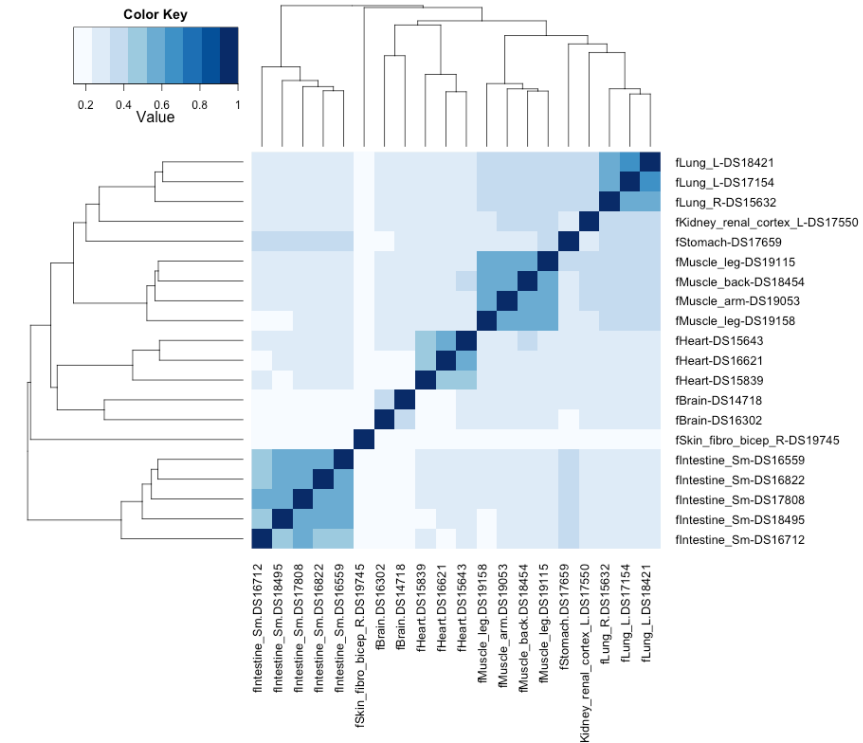
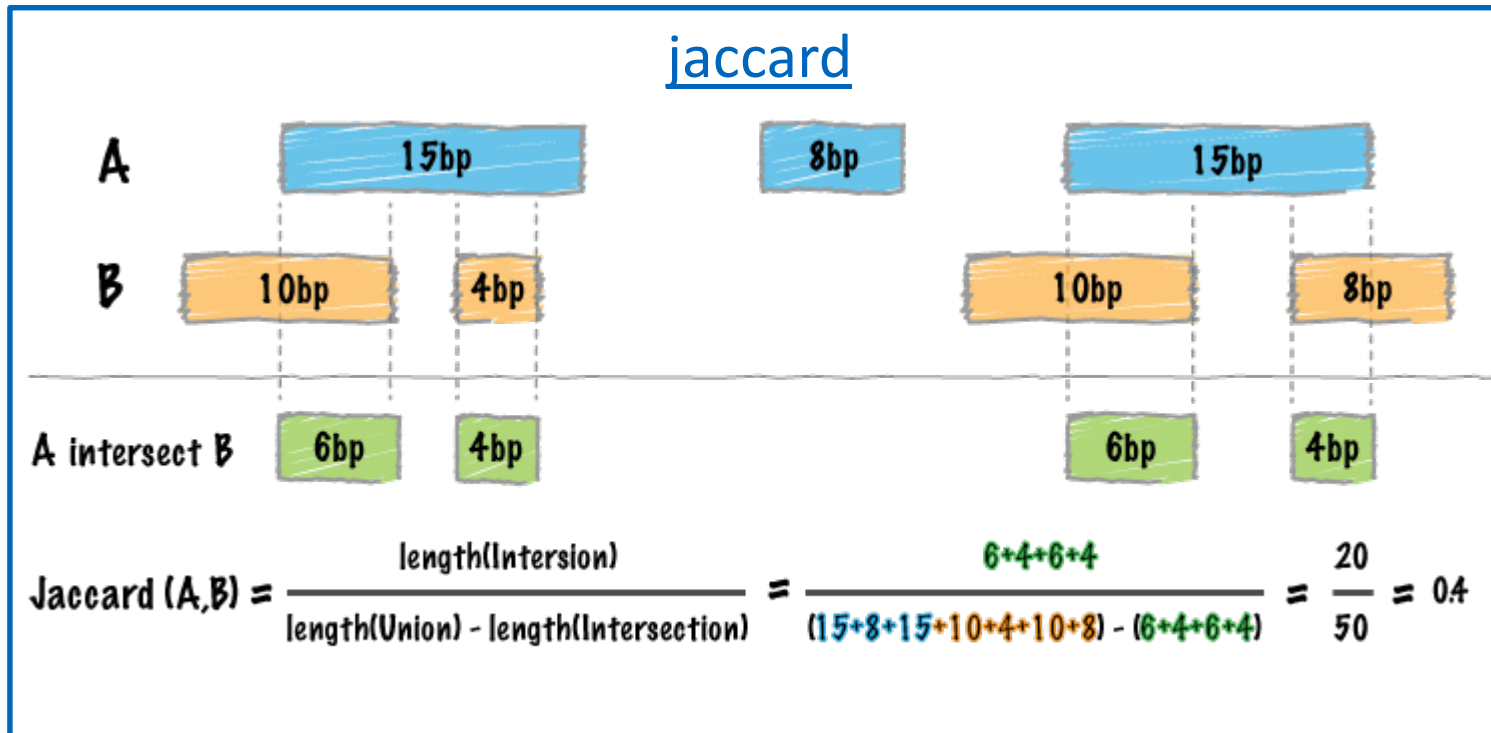
[jaccard](#)





BEDTools: Statistics & Hypothesis Testing

jaccard	Calculate the Jaccard statistic b/w two sets of intervals.
fisher	Calculate Fisher statistic b/w two feature files.
reldist	Calculate the distribution of relative distances b/w two files.
shuffle	Randomly redistribute intervals in a genome.
random	Generate random intervals in a genome.
sample	Sample random records from file using reservoir sampling.

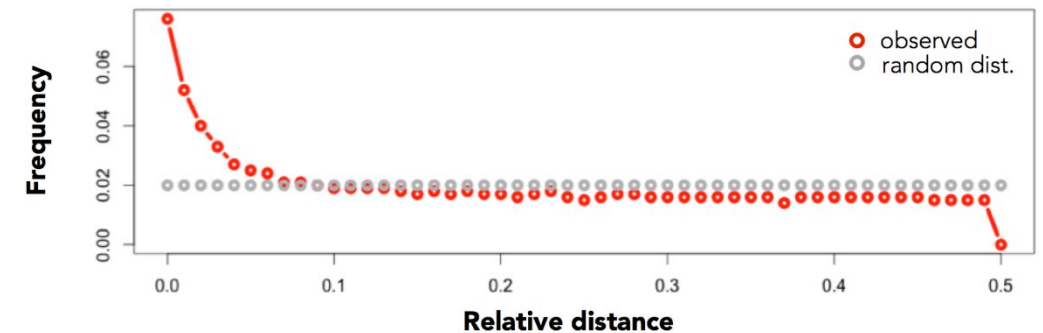
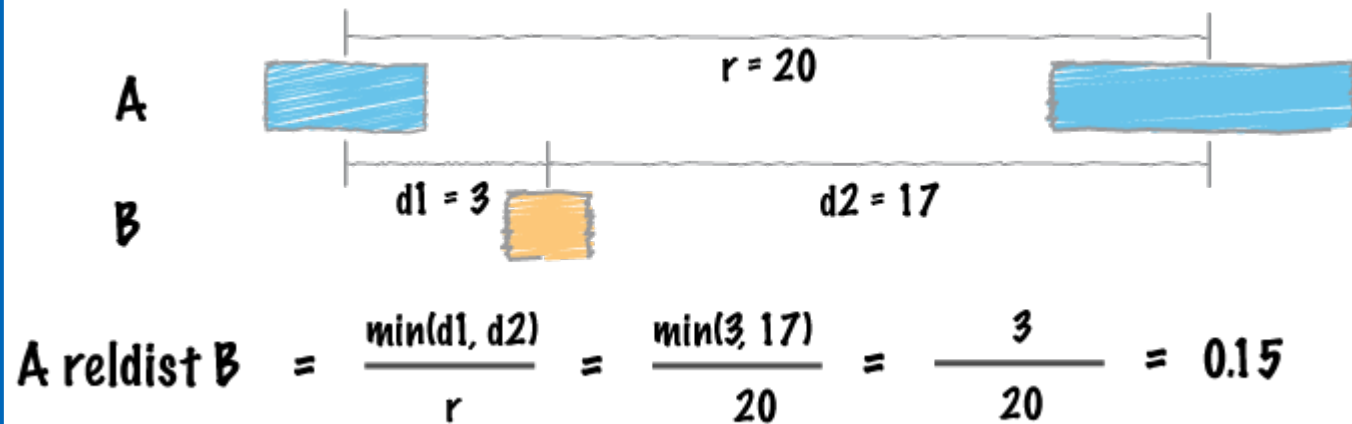




BEDTools: Statistics & Hypothesis Testing

<u>jaccard</u>	Calculate the Jaccard statistic b/w two sets of intervals.
<u>fisher</u>	Calculate Fisher statistic b/w two feature files.
<u>reldist</u>	Calculate the distribution of relative distances b/w two files.
<u>shuffle</u>	Randomly redistribute intervals in a genome.
<u>random</u>	Generate random intervals in a genome.
<u>sample</u>	Sample random records from file using reservoir sampling.

reldist





BEDTools: Combining multiple tools

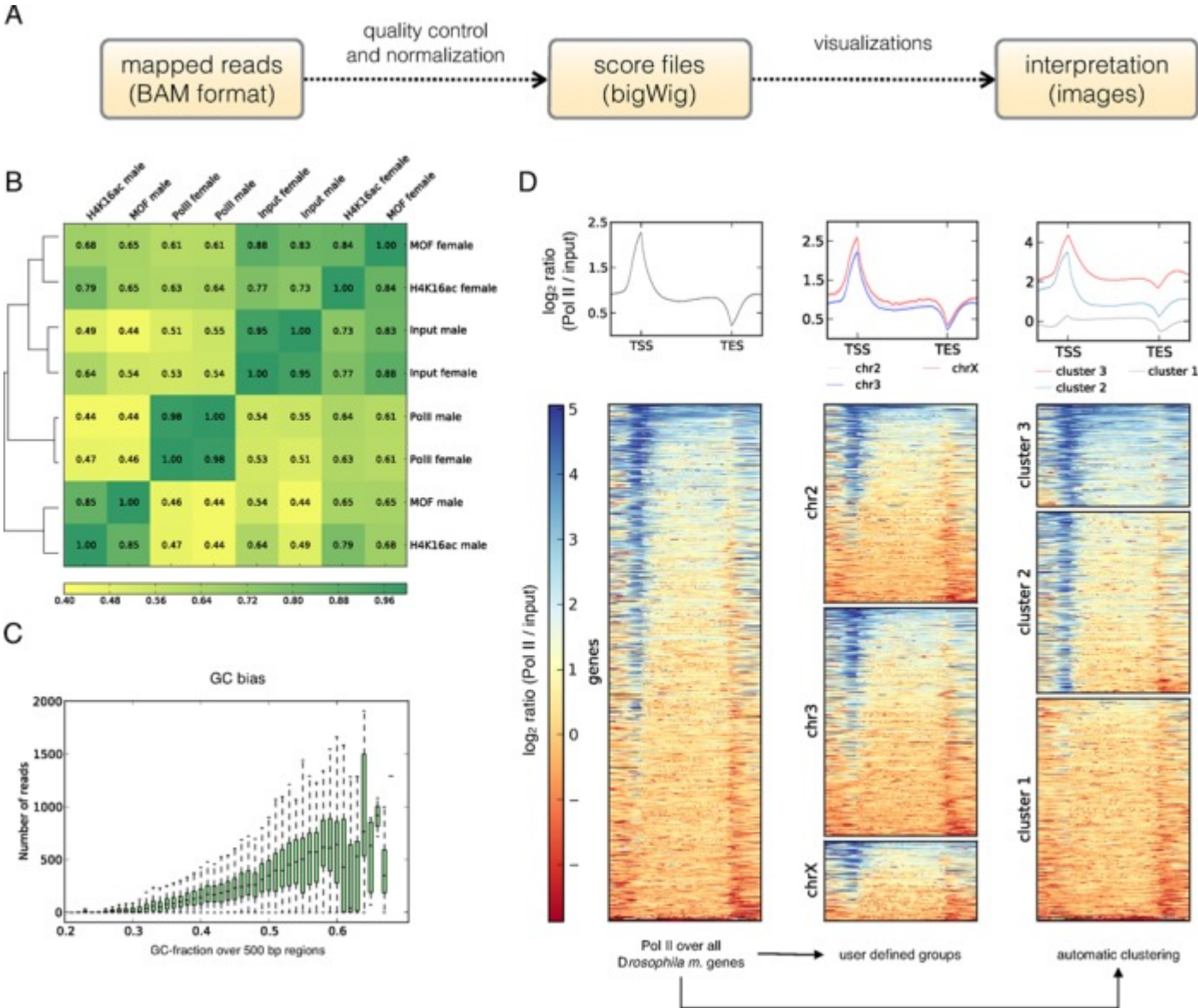
- Multiple BEDTools commands can be piped together
- Most powerful combined with linux, bash/awk scripts, R, &/or python

```
bedtools window -a A.bed -b B.bed -w 10 | bedtools overlap -i stdin -cols 2,3,6,7
chr1 10 20 A chr1 15 25 B 5
chr1 10 20 C chr1 25 35 D -5
```

```
bedtools makewindows -g sacCer3.chrom.sizes -w 50000 \
| sed 's/^chr//' \
| bedtools coverage -a - -b sacCer3_autosomal_genes.bed \
> sacCer3.numGenes.50KwinV2.bed
```

Deeptools

Sample correlation



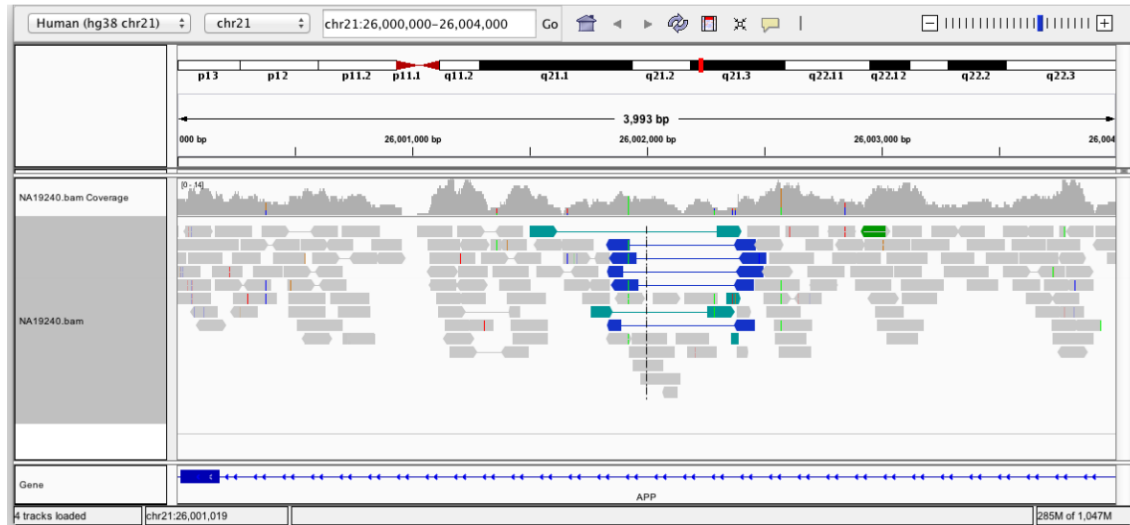
Scale-regions

DeepTools work with both bam file and bigWig file

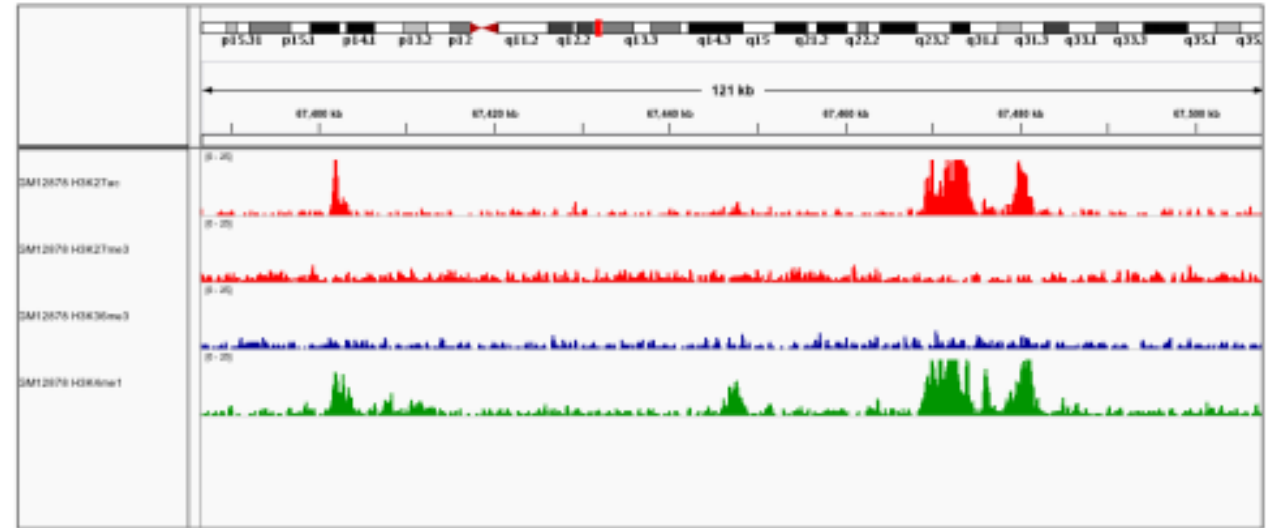
multiBamSummary
multiBigwigSummary

bamCompare
bigwigCompare

bam



bigwig



Convert bam to bigWig

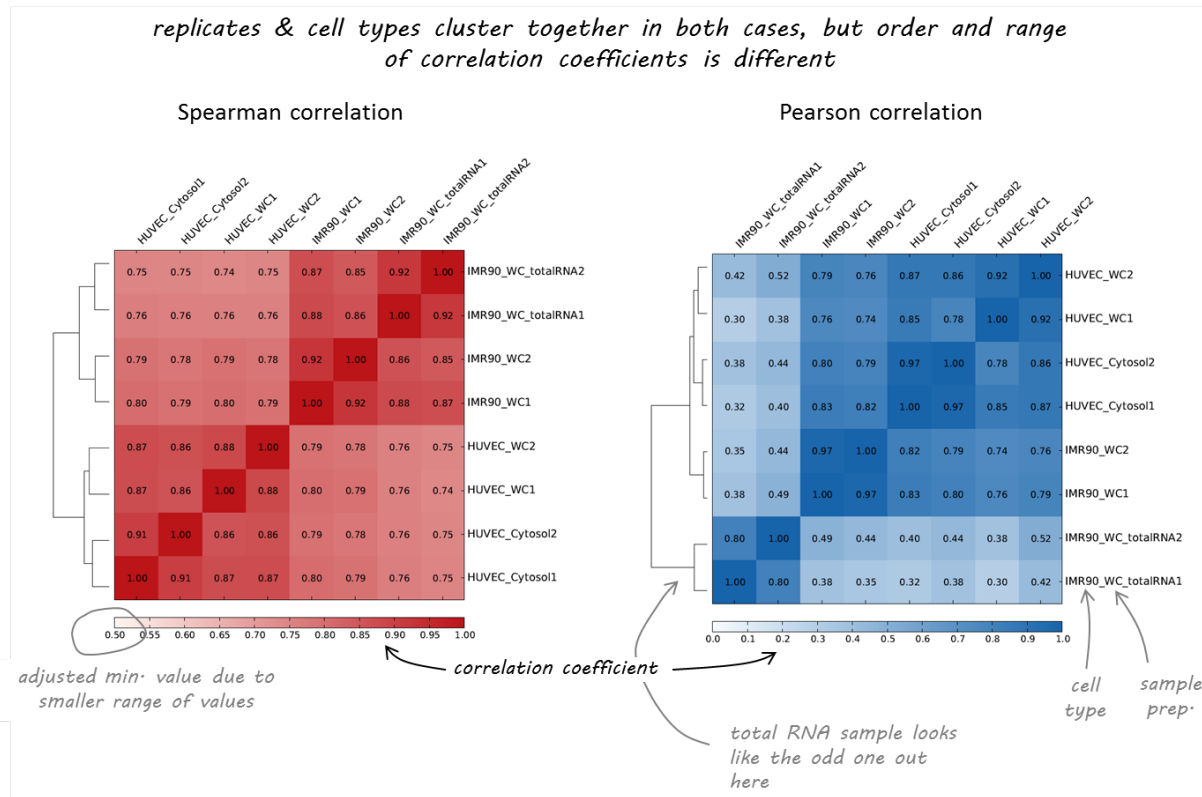
Advanced options:

- Normalized;
- Reads extension to fragment;
- GC-bias correction;

```
bamCoverage --bam a.bam -o a.SeqDepthNorm.bw \
--binSize 10 \
--normalizeUsing RPGC \
--effectiveGenomeSize 2150570000 \
--ignoreForNormalization chrX \
--extendReads \
```

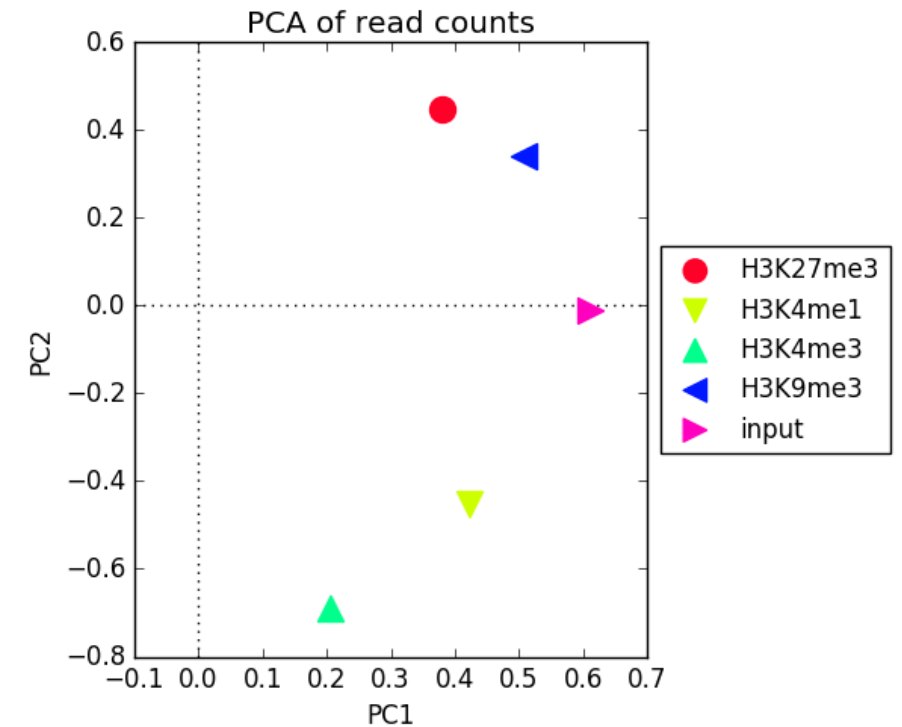
Smaller bin size =
higher resolution & larger files

Correlations between samples



Heatmap (Pearson or Spearman)

plotCorrelation



PCA

plotPCA

QC Plots

multiBamSummary
multiBigwigSummary

Read coverage matrix for correlation plots

Sliding windows

```
multiBigwigSummary bins -b file1.bw file2.bw -o results.npz
```

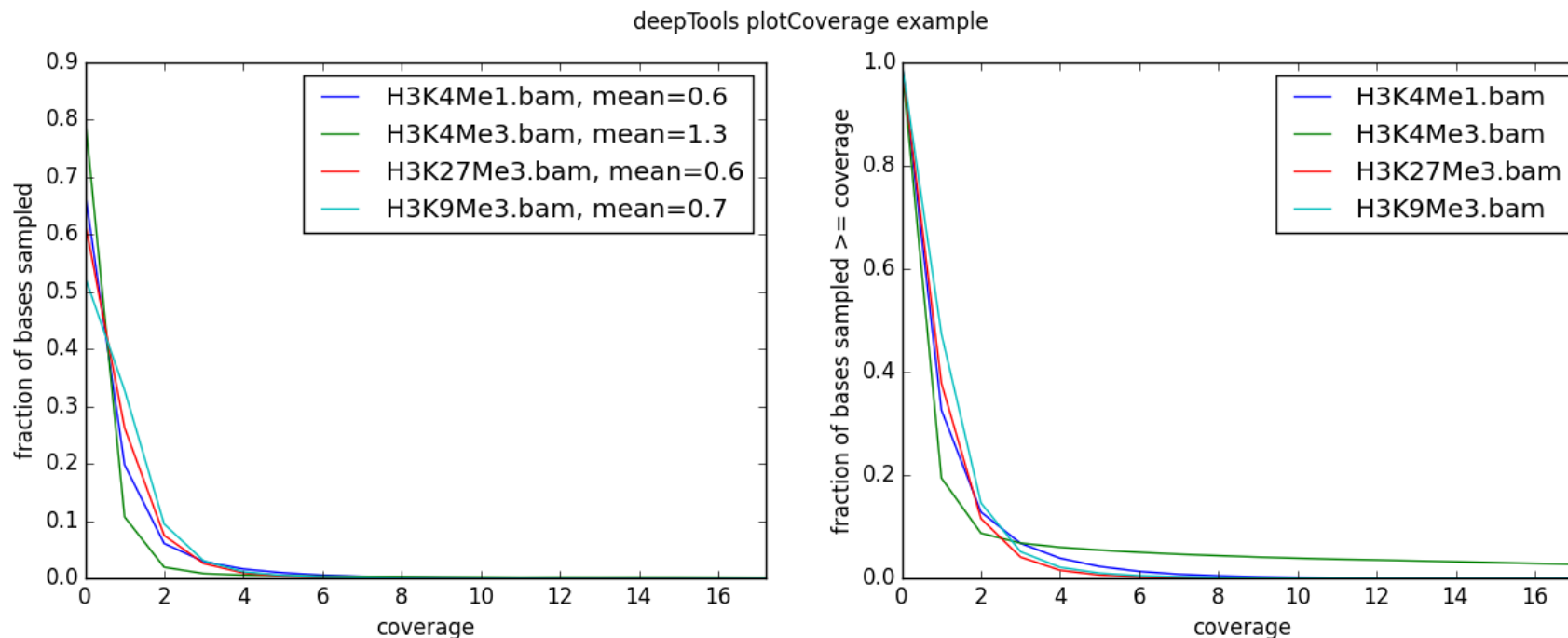
Custom genomic regions (e.g. gene bodies)

```
multiBigwigSummary BED-file -b file1.bw file2.bw -o results.npz --BED selection.bed
```

Some options:

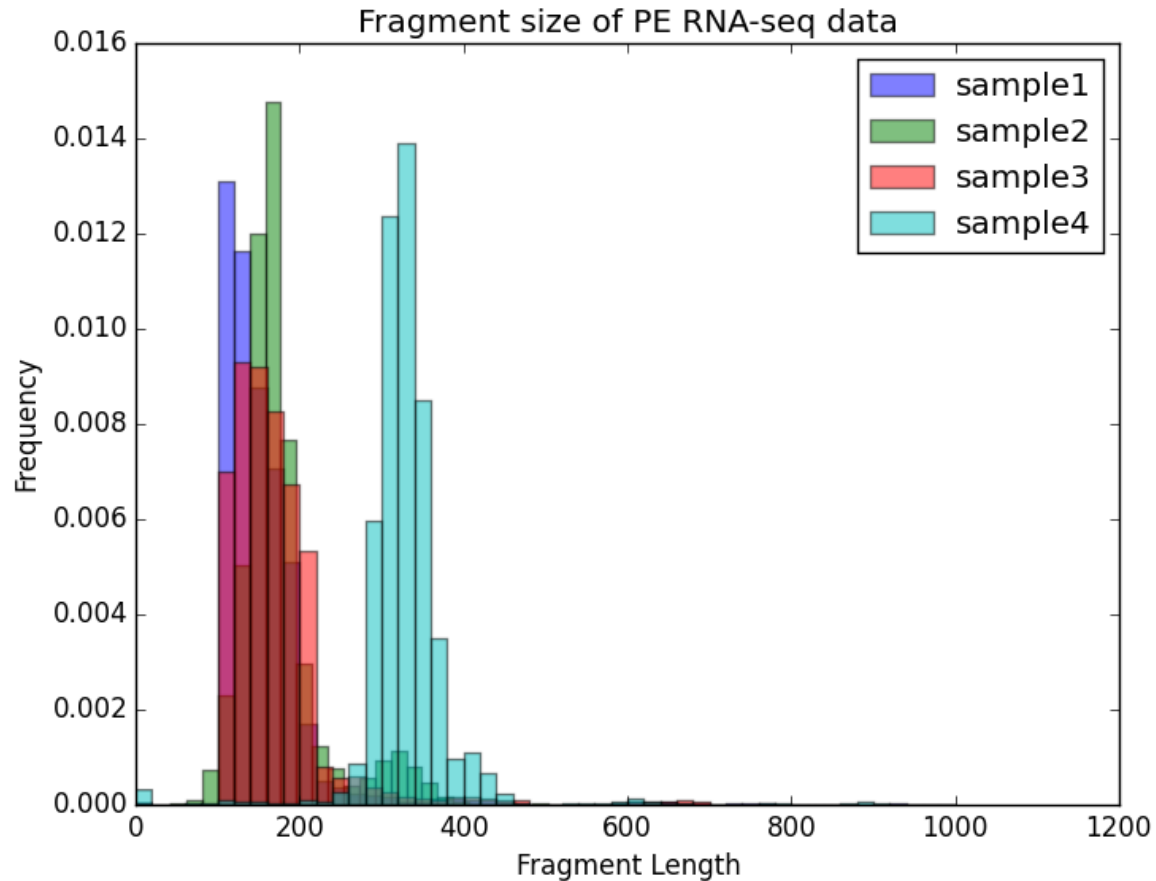
--region chr10:1:891000 :	restrict to selected regions to get quick estimate;
--numberOfProcessors 24:	number of processors;
--outRawCounts:	output raw count in a tab-delimited text file;

Plot histogram of read depth



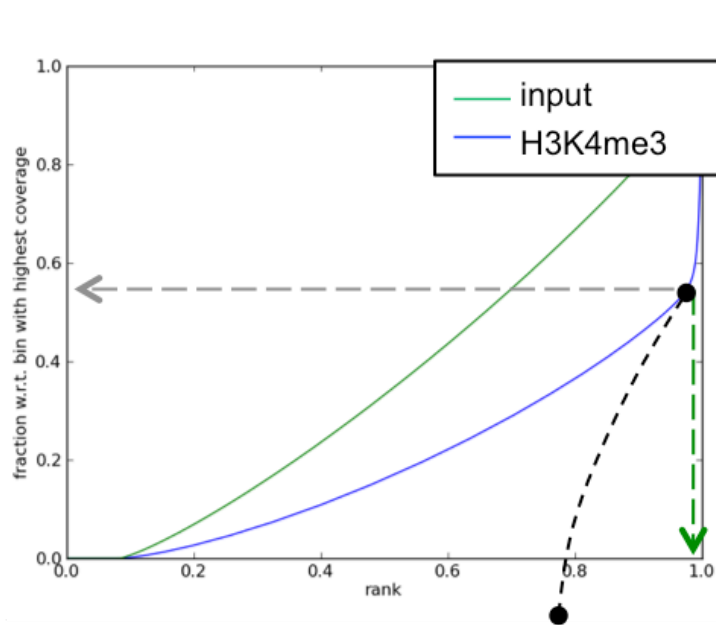
```
plotCoverage -b H3K4Me1.bam H3K4Me3.bam H3K27Me3.bam H3K9Me3.bam
--plotFile example_coverage
-n 1000000
--plotTitle "example_coverage" \
--outRawCounts coverage.tab \
--ignoreDuplicates \
--minMappingQuality 10 \
--region 19
```

QC Plots



```
bamPEFragmentSize \  
-hist fragmentSize.png \  
-T "Fragment size of PE RNA-seq data" \  
--maxFragmentLength 1000 \  
-b testFiles/RNAseq_sample1.bam  
testFiles/RNAseq_sample2.bam \  
testFiles/RNAseq_sample3.bam  
testFiles/RNAseq_sample4.bam \  
-samplesLabel sample1 sample2 sample3 sample4
```

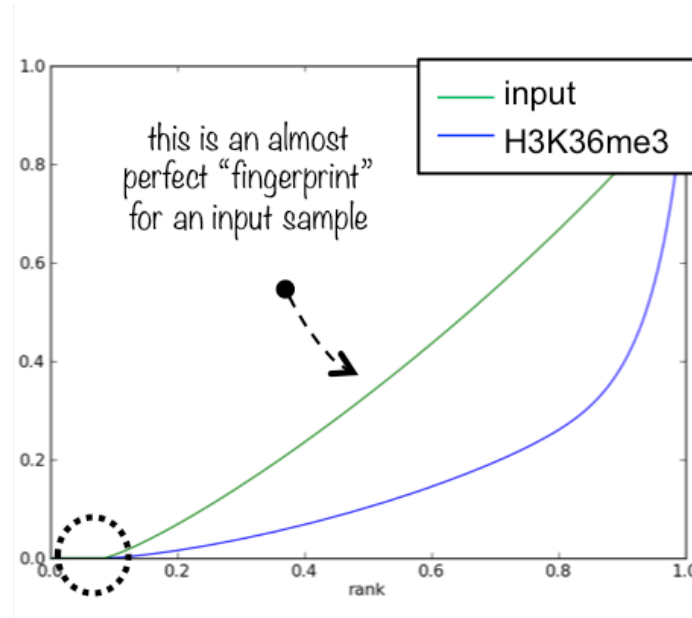
QC Plots



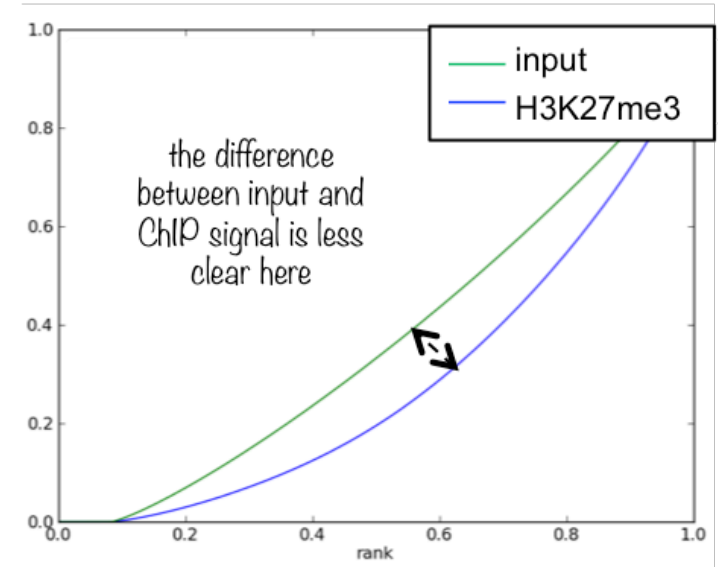
when counting the reads contained in **97%** of all genomic bins, only ca. **55%** of the maximum number of reads are reached, i.e. 3% of the genome contain a very large fraction of reads!

→ this indicates very localized, very strong enrichments!
(as every biologist hopes for in a ChIP for H3K4me3)

plotFingerprint



pay attention to where the curves start to rise – this already gives you an assessment of how much of the genome you have not sequenced at all (i.e. bins containing zero reads – for this example, ca. 10% of the entire genome do not have any read)

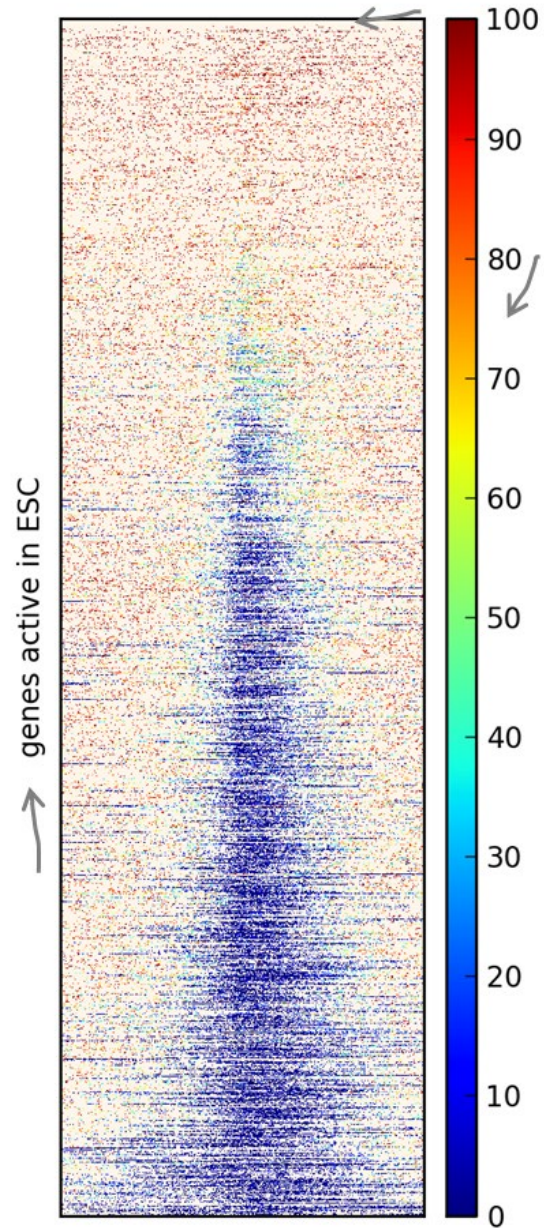


H3K27me3 is a mark that yields broad domains instead of narrow peaks

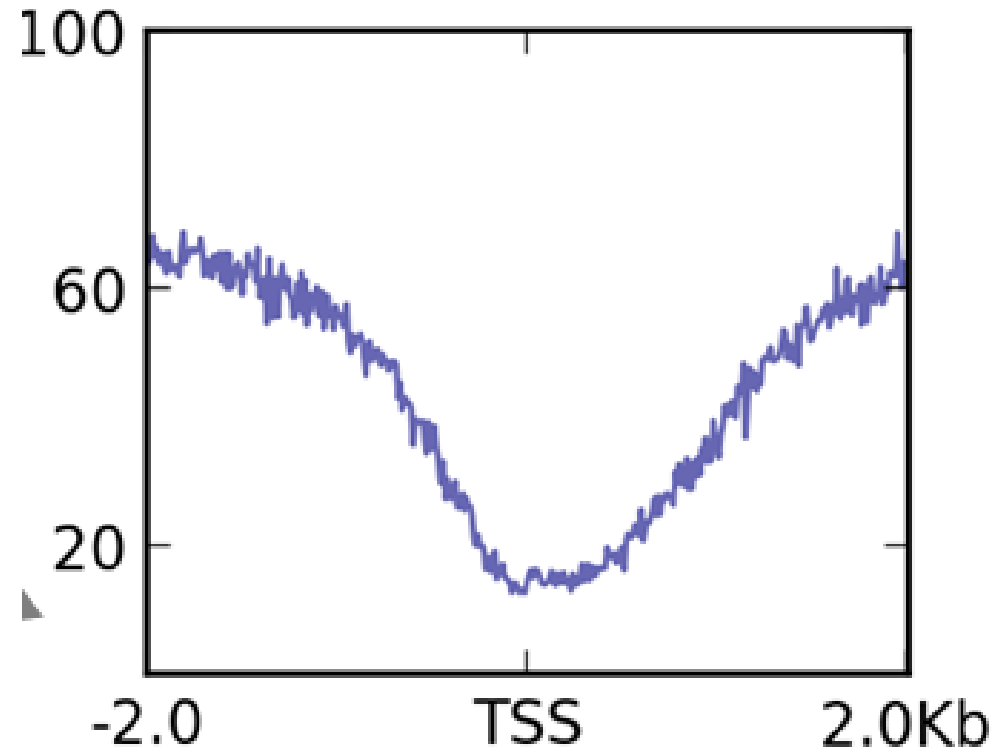
→ it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed

computeMatrix

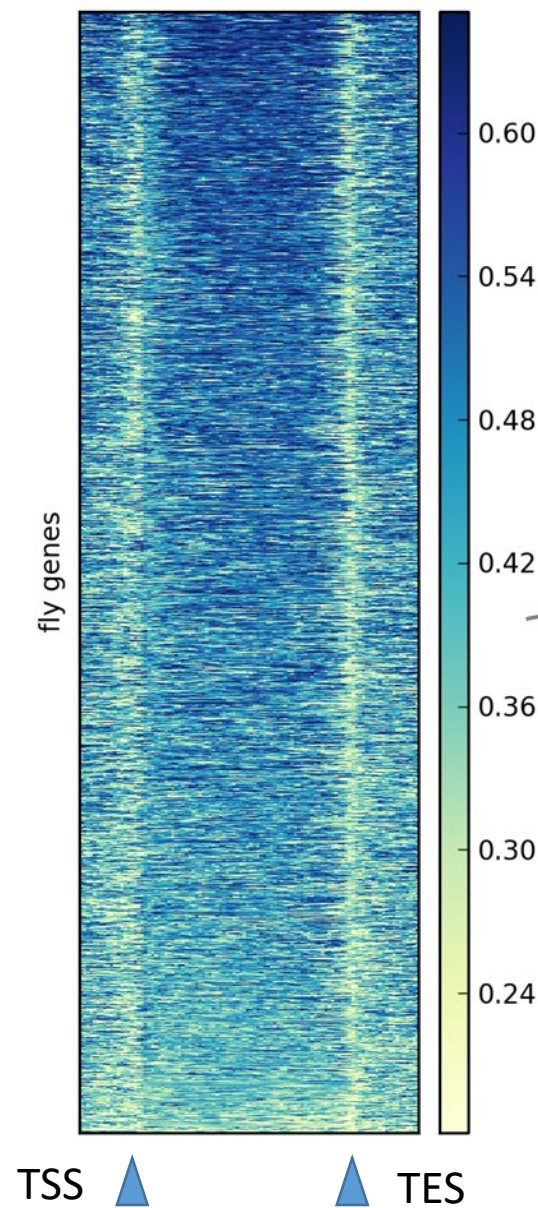
Mode 1: Reference point



▲ Transcription Start Site (TSS)

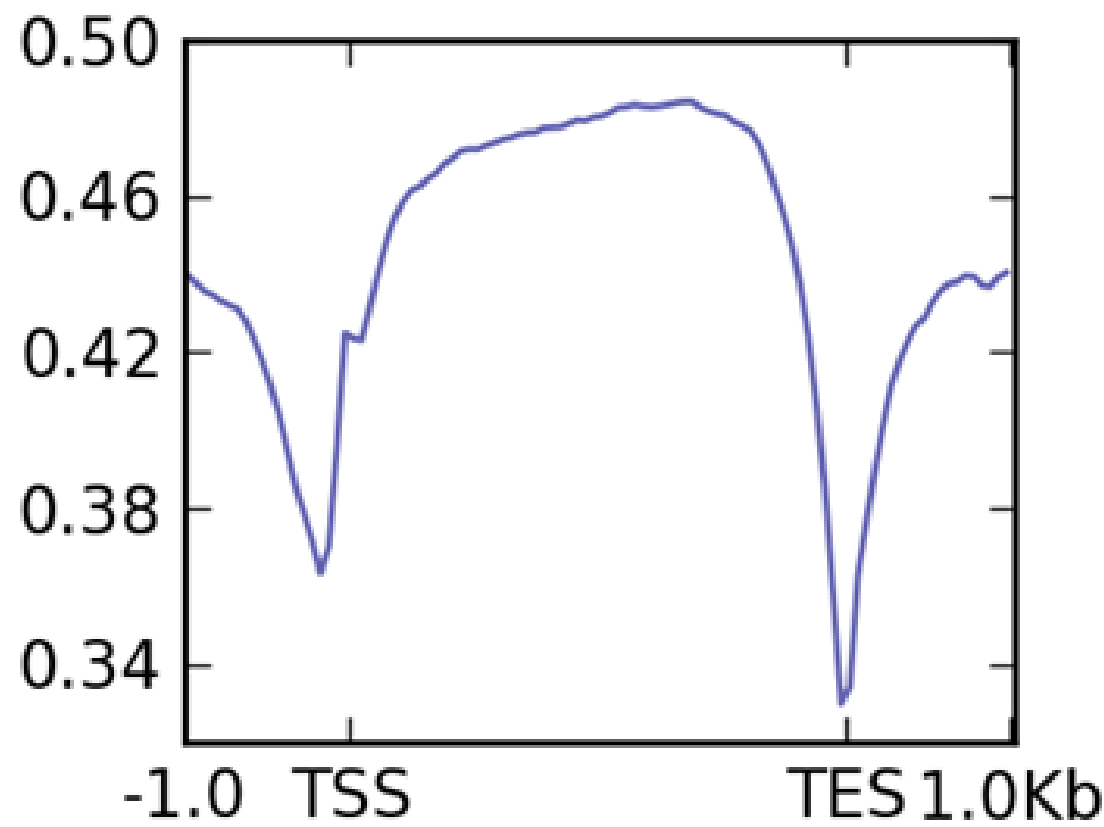


All genes are scale to
the same size



computeMatrix

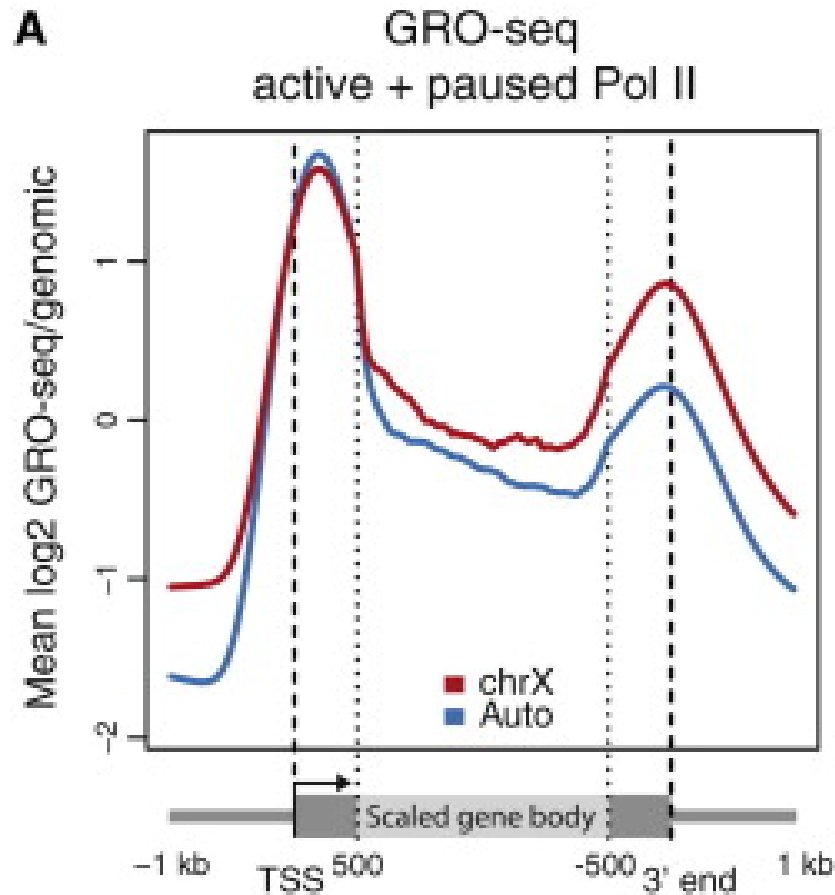
Mode 1: Scaled-region



Advanced features in scaled-region

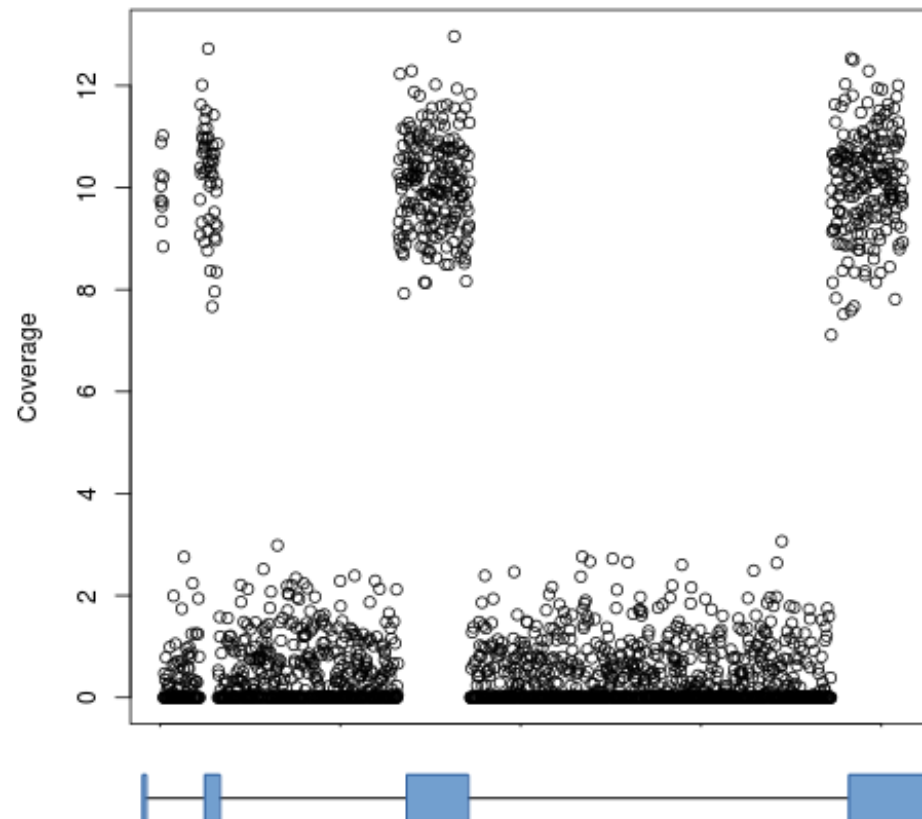
Unscaled 5' and/or 3' regions

-unscaled5prime and *-unscaled3prime*

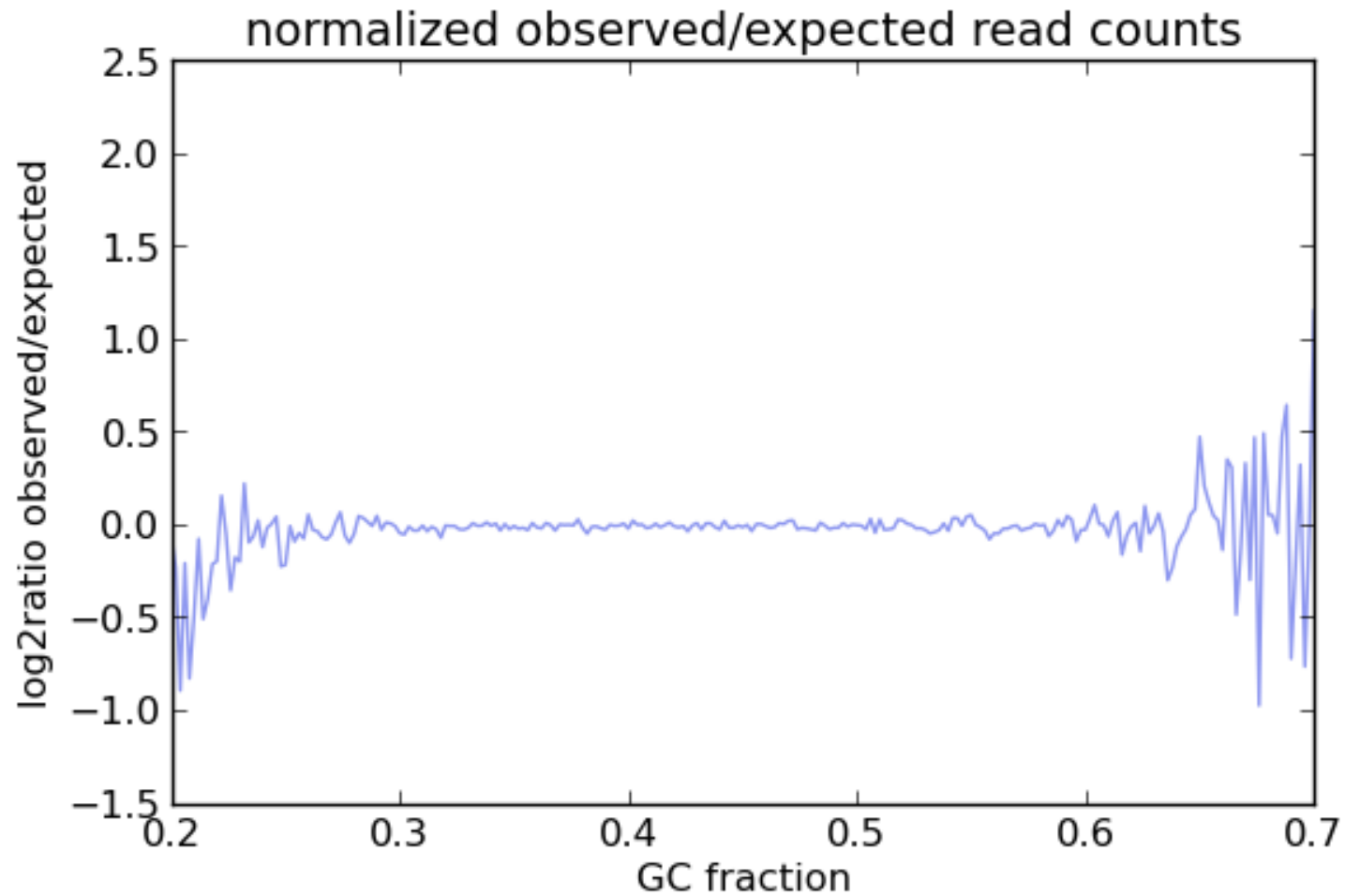


Ignore intron

-metagene



computeGCBias & correctGCBias



HOMER Motif Analysis

<http://homer.ucsd.edu/homer/motif/index.html>



Homer *de novo* Motif Results

[Known Motif Enrichment Results](#)

[Gene Ontology Enrichment Results](#)

If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into [STAMP](#)

More information on motif finding results: [HOMER](#) | [Description of Results](#) | [Tips](#)

Total target sequences = 15212

Total background sequences = 34047

* - possible false positive

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-1835	-4.228e+03	28.11%	5.16%	37.7bp (63.1bp)	NFkB-p65(RHD)/GM12787-p65-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)
2		1e-1716	-3.953e+03	34.50%	8.65%	47.8bp (62.6bp)	PB0058.1_Sfpi1_1 More Information Similar Motifs Found	motif file (matrix)
3		1e-1585	-3.651e+03	28.85%	6.39%	41.8bp (62.8bp)	MA0102.1_Cebpa More Information Similar Motifs Found	motif file (matrix)
4		1e-1004	-2.314e+03	25.07%	7.22%	49.2bp (61.0bp)	NF-E2(bZIP)/K562-NFE2-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)
5		1e-262	-6.039e+02	5.52%	1.27%	47.8bp (59.7bp)	Oct2(POU/Homeobox)/Bcell-Oct2-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)
6		1e-198	-4.565e+02	6.42%	2.09%	49.9bp (53.2bp)	c-Jun-CRE(bZIP)/K562-cJun-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)
7		1e-146	-3.370e+02	7.66%	3.30%	51.7bp (58.9bp)	RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq/Homer More Information Similar Motifs Found	motif file (matrix)

1. Extraction of Sequences;

2. Background Selection;

3. GC Normalization;

4. Enriched motifs

Transcription factor binding site motif databases

Public

JASPAR

<http://jaspar.genereg.net/>

CIS-BP

<http://cisbp.cabr.utoronto.ca>

Commercial

TRANSFAC

<https://genexplain.com/transfac/>

HOMER

Match to known motif

Matches to Known Motifs

NFkB-p65(RHD)/GM12787-p65-ChIP-Seq/Homer

Match Rank: 1
Score: 0.93
Offset: -2
Orientation: forward strand
Alignment: --GGAATTYCCC
 NGGGGATTTC



MA0061.1_NF-kappaB

Match Rank: 2
Score: 0.87
Offset: -1
Orientation: forward strand
Alignment: -GGAATTYCCC
 GGGAATTTC-



MA0105.1_NFKB1

Match Rank: 3
Score: 0.84
Offset: -1
Orientation: forward strand
Alignment: -GGAATTYCCC
 GGGGATTCCCC

