Useful software utilities for computational genomics

Shamith Samarajiwa

CRUK Summer School in Bioinformatics
July 2019

UNIVERSITY OF CAMBRIDGE

MRC Cancer Unit

Overview

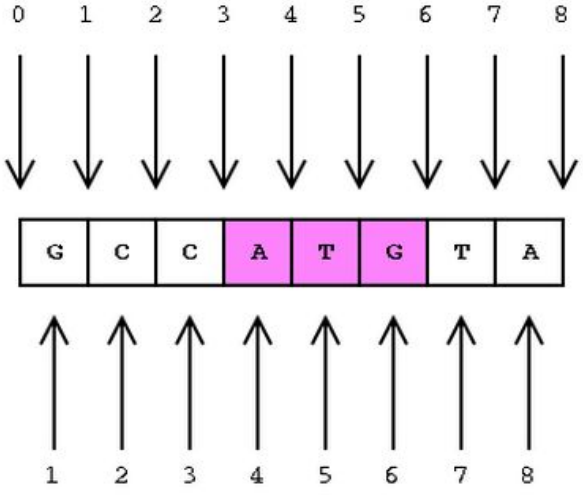
- Genomic coordinate systems
- Search and download genomic datasets: GEOquery, GEOmetadb, and SRAdb
- Getting annotation with UCSC Table browser, Utilities (liftover)
- Manipulate genomic range data with
  - bedtools or bedops
  - IRanges, GenomicRanges and GRange container objects
  - Plyranges - tidy operations on genomics data
- Annotation and visualization with ChIPpeakAnno, ChIPseeker, Rtracklayer and Deeptools2
- Gene set enrichment with chipenrich

Genomic Coordinate Systems

- There are two major coordinate systems in genomics.
- **Base coordinate system** anchors genomic feature to nucleotide positions while the **Interbase coordinate system** anchor genomic feature between nucleotide positions.
- Most genome annotation portals (e.g. **NCBI or Ensembl**), bioinformatics software (e.g. BLAST) and annotation file formats (e.g. **SAM, VCF, GFF and Wiggle**) use the base coordinate system, which represents a feature starting at the first nucleotide as **position 1**
- Other systems (e.g. **UCSC, Chado, DAS2**) and formats **(BAM, BCFv2, BED, and PSL)** use the interbase coordinate system, whereby a feature starting at the first nucleotide is represented as **position 0**

Genomic Coordinate Systems

- The UCSC genome browser uses both systems and refer to the base coordinate system as "**one-based, fully-closed**" (used in the UCSC genome browser display) and interbase coordinate system as "**zero-based, half-open**" (used in their tools and file formats).
- The interbase coordinate system is also referred to as "space-based".

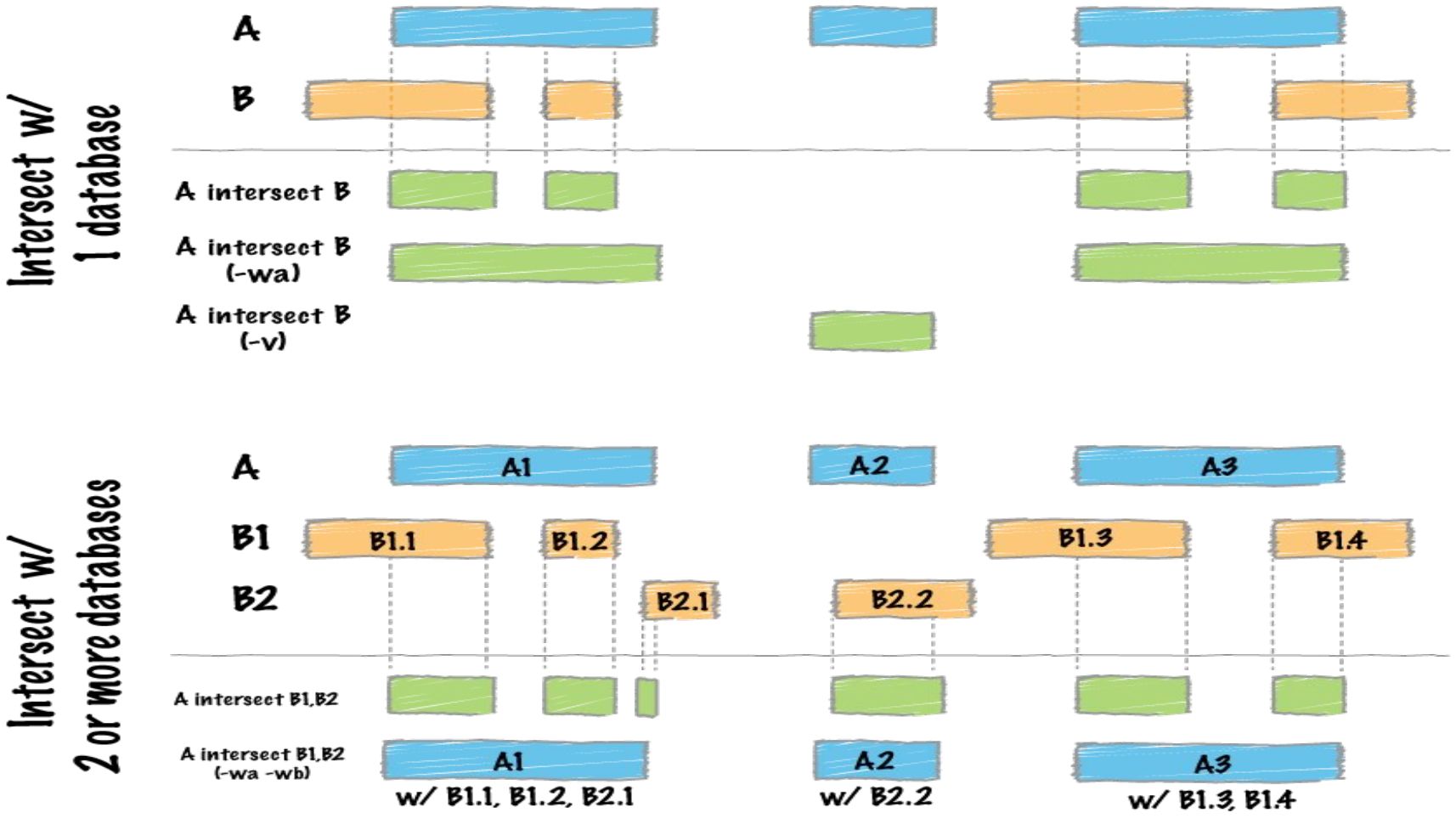


There are several advantage for using the interbase coordinate system including:

1. the ability to represent features that occur between nucleotides (like a splice site),
2. simpler arithmetic for computing the length of features (length=end-start) and overlaps (max(start1,start2), min(end1,end2))
3. more rational conversion of coordinates from the positive to the negative strand

http://bergmanlab.genetics.uga.edu/?s=coordinate

Genomic interval data with bedtools



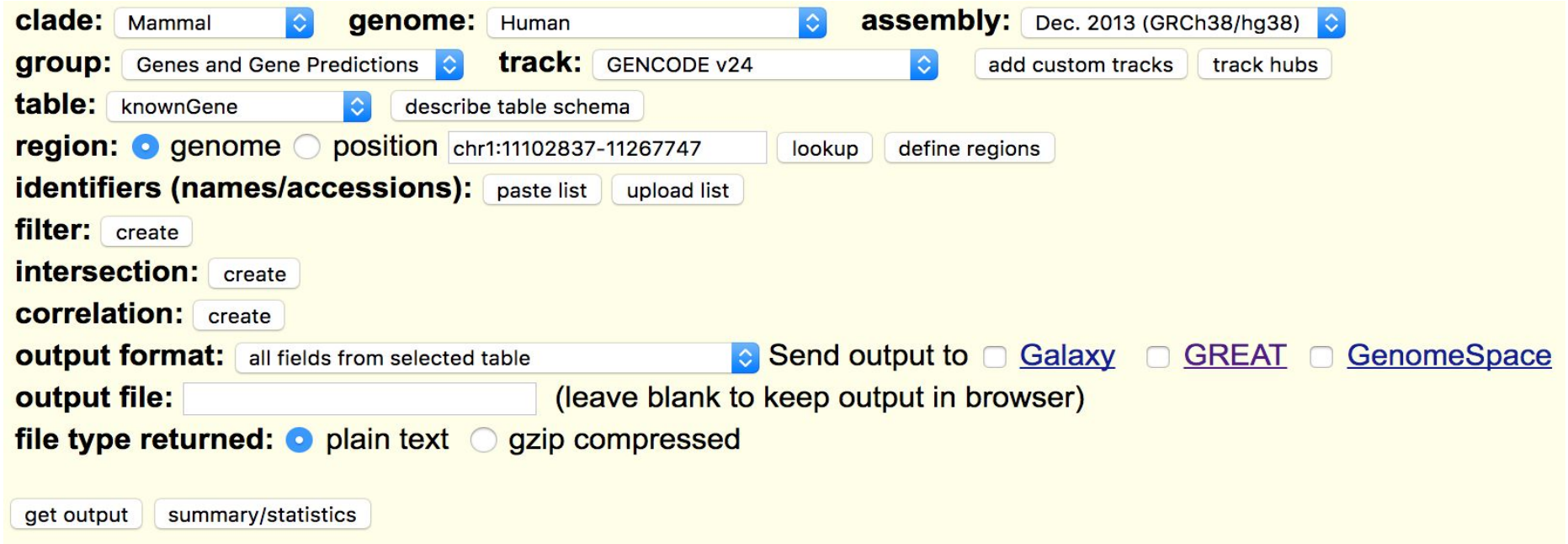
Search and download genomic data

- **GEOmetadb** is an SQLite database which stores GEO metadata and annotations. Provides fast search access to GEO annotation and dataset information.
- **GEOquery** is an interface to GEO platform, sample, series and dataset information
- **SRAdb** enables searching a local sqlite database for SRA (NCBI Sequence Read Archive) annotation
- **NCBI SRA Toolkit**

UCSC Table Browser

- Search for genes and annotation
- Setup and filters
- Join tables
- Retrieve sequences
- Intersecting tracks
- Export to external resources

UCSC Table browser interface



UCSC Table browser usage

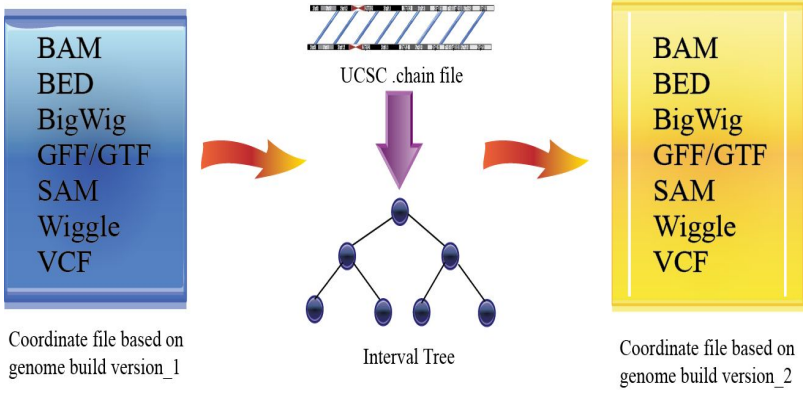
- Retrieve the DNA sequence data or annotation data underlying Genome Browser tracks for the entire genome, a specified coordinate range, or a set of accessions
- Apply a filter to set constraints on field values included in the output
- Generate a custom track and automatically add it to your session so that it can be graphically displayed in the Genome Browser
- Conduct both structured and free-from SQL queries on the data
- Combine queries on multiple tables or custom tracks through an intersection or union and generate a single set of output data
- Display basic statistics calculated over a selected data set
- Display the schema for table and list all other tables in the database connected to the table
- Organize the output data into several different formats for use in other applications, spreadsheets, or databases

UCSC utilities

- Useful UCSC utilities
  - Dusters -DNA and Protein (removes non sequence and format characters from fasta files)
  - LiftOver tool convert between genome build coordinates.
  - bedCoverage analyses coverage by bed files - chromosome by chromosome and genome-wide usage: bedCoverage *database bedFile*
  - bedGraphToBigWig converts a bedGraph file to bigWig format.
    usage:
    bedGraphToBigWig *in.bedGraph chrom.sizes out.bw*
  - bigWigToWig Convert bigWig to wig. This will keep more of the same structure of the original wig than bigWigToBedGraph does, but still will break up large stepped sections into smaller ones.
    Usage bigWigToWig *in.bigWig out.wig*

Liftover of genome coordinates

- Reference genome assemblies are subject to change and refinement from time to time. Examples:
  - You want to convert Encode peak files from hg19 to GRCh38
  - Convert your hg38 coordinates to hg19, so rGREAT can be used.
- Generally, researchers need to convert results that have been analyzed using old assemblies to newer versions or *vice versa*, to facilitate meta-analysis, direct comparison as well as data integration and visualization.
- There a number of liftover utilities:
  - UCSC liftOver tool: BED
  - Pyliftover: conversion of point coordinates
  - NCBI remap: BED, GFF, GTF, VCF, etc
  - Galaxy: BED, GFF, GTF
  - CrossMap: SAM/BAM, Wiggle/BigWig, BED, GFF/GTF, VCF



- Usually requires a “chain file” that describes pairwise alignments between two genomes.

bedtools

- Bedtools documentation
- 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 chaining multiple bedtools operations on the UNIX command line.

bedops

SET OPERATIONS

- **bedops** - apply set operations on any number of BED inputs
- **bedextract** - efficiently extract BED features
- **closest-features** - matches nearest features between BED files

STATISTICS

- **bedmap** - map overlapping BED elements onto target regions and optionally compute any number of common statistical operations

FILE MANAGEMENT

- **sort-bed** - apply lexicographical sort to BED data
- **starch** and **unstarch** - compress and extract BED data
- **starchcat** - merge compressed archives
- **starchstrip** - filter archive by chromosomes
- Conversion tools - convert common genomic formats to BED

PERFORMANCE

- Parallel **bam2bed** and **bam2starch** - parallelized conversion and compression of BAM data
- Set operations with **bedops**
- Compression characteristics of **starch**
- Independent testing

OTHER

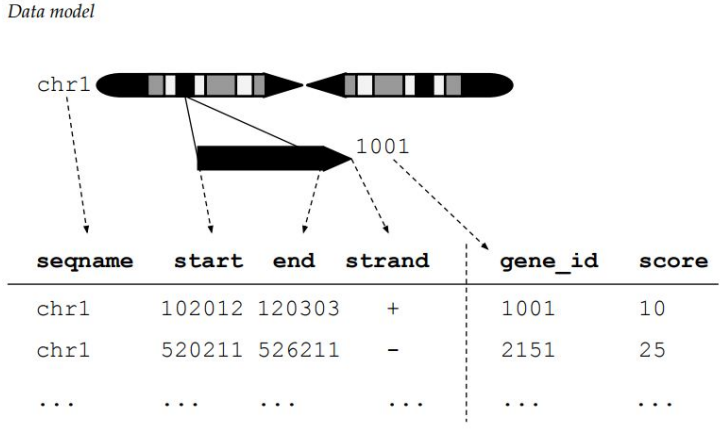
- Table summary of **BEDOPS** toolkit
- Starch v2.2 format specification
- About nested elements
- Revision history
- Github release instructions
- Github repository

SUPPORT RESOURCES

- How to install **BEDOPS**
- Usage examples of **BEDOPS** tools in action
- **BEDOPS** user forum
- **BEDOPS** discusssion mailing list

Bioconductor: IRanges and GRanges

- IRanges (collections of integer intervals) and GRanges (IRanges + associated genomic annotation) are fast and efficient data structures for genomic data
- A large number of objects/tasks in computational genomics can be formulated in terms of;
  - integer intervals
  - manipulation of integer intervals
  - overlap of integer intervals
- Objects:
  - A gene or transcript (a union of int
  - A collection of SNPs (intervals of w
  - TF binding sites (a collection of aligned short reads)
- Tasks:
  - Which TF binds to promoters of genes (overlap between intervals)
  - Which SNPs maps to a collection of exons

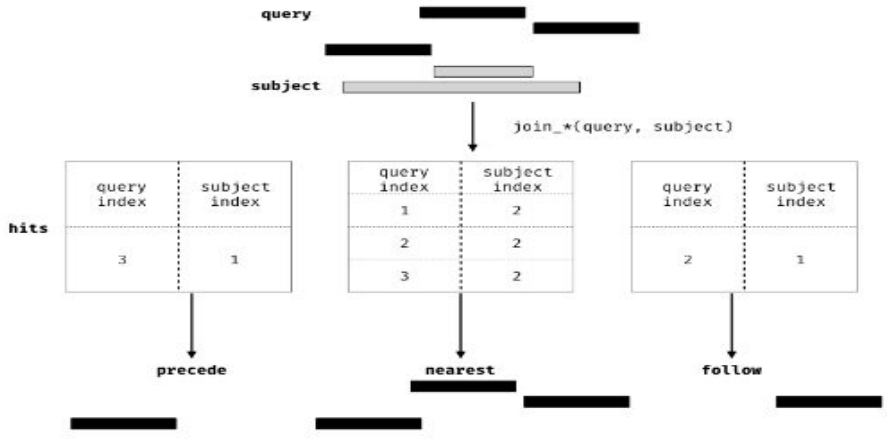


The GRanges class

- **GRanges** container stores a set of genomic ranges (a.k.a. genomic regions or genomic intervals).
- Each genomic range is described by **chromosome name**, **start**, **end**, and **strand**.
- **start** and **end** are both **1-based** positions relative to the 5' end of the plus strand of the chromosome, even when the range is on the minus strand.
- **start** and **end** are both considered to be included in the interval (except when the range is empty).
- The width of the range is the number of genomic positions included in it. So *width = end - start + 1*.
- **end** is always >= **start**, except for empty ranges (a.k.a. zero-width ranges) where end = start - 1. Note that the start is always the leftmost position and the end the rightmost, even when the range is on the minus strand.
- In features (genes, transcripts) located on the minus strand, the TSS is at the end of the range.

plyranges

- A dplyr-based API for computing on genomic ranges
- Similar to dplyr but for genomic range data
  - mutate, stretch, anchor, shift, flank
  - group_by
  - filter
  - reduce, summarize
  - intersect, union,
  - overlap

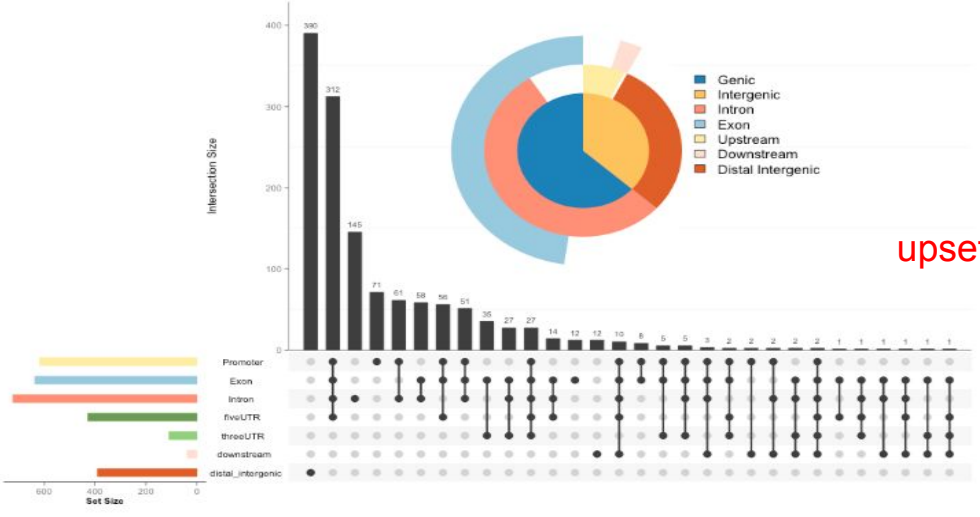


ChIPpeakAnno

- Enables annotation of peaks to genomic features (Genes, TSSs, Promoters, Enhancers, CpG islands, Repeats etc) and custom annotation
- Can perform overlap analysis
- Examples:
  - Combine with chromosomal conformation capture to find peaks associated with enhancer interactions
  - Combine profiles from many TFs
  - Visualize binding patterns
  - Identify bidirectional promoters

ChIPseeker

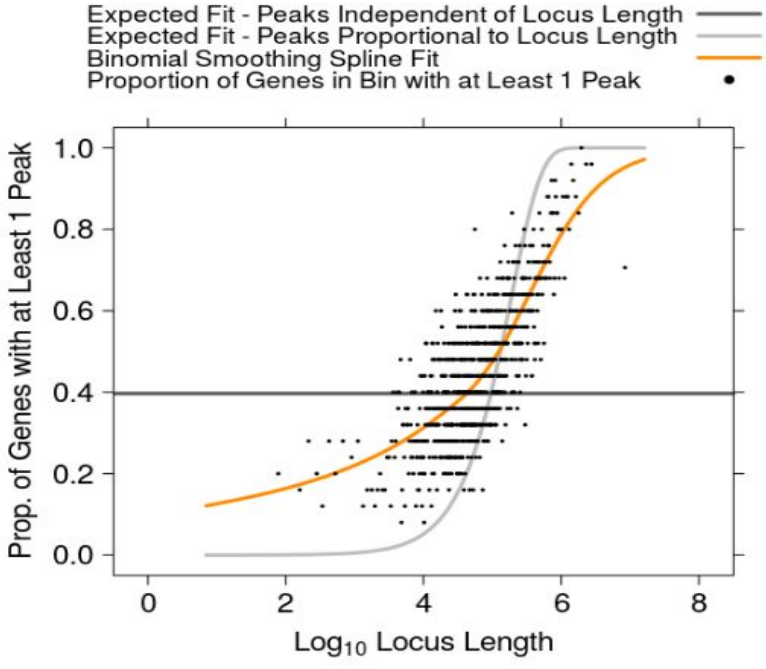
- Provides methods to identify overlap with annotated genomic features, other ChIP-seq datasets and biological replicates.
- Useful for detecting co-operative binding of different factors.
- Contains a variety of visualization options
- Peak Annotation is performed by ***annotatePeak***. User can define TSS (transcription start site) region, by default TSS is defined from -3kb to +3kb.
- Peaks are annotated with the following features:
  - Promoter
  - 5’ UTR
  - 3’ UTR
  - Exon
  - Intron
  - Downstream
  - Intergenic



upset plot

chipenrich

- Gene Set Enrichment Analysis for ChIP-seq.
- Takes into account length or coverage biases (length of a gene's regulatory region affects the probability that a peak will be assigned to it, the number of peaks that will be assigned to it, or the proportion of it covered by peaks.)
  - Broadenrich is designed for use with broad peaks.
  - chipenrich is designed for use with 1,000s or 10,000s of narrow peaks.
  - Polyenrich is also designed for narrow peaks, but where there are 100,000s of narrow peaks.
- Built in locus definitions, gene sets and mappability.
- Enrichment methods with FP minimization

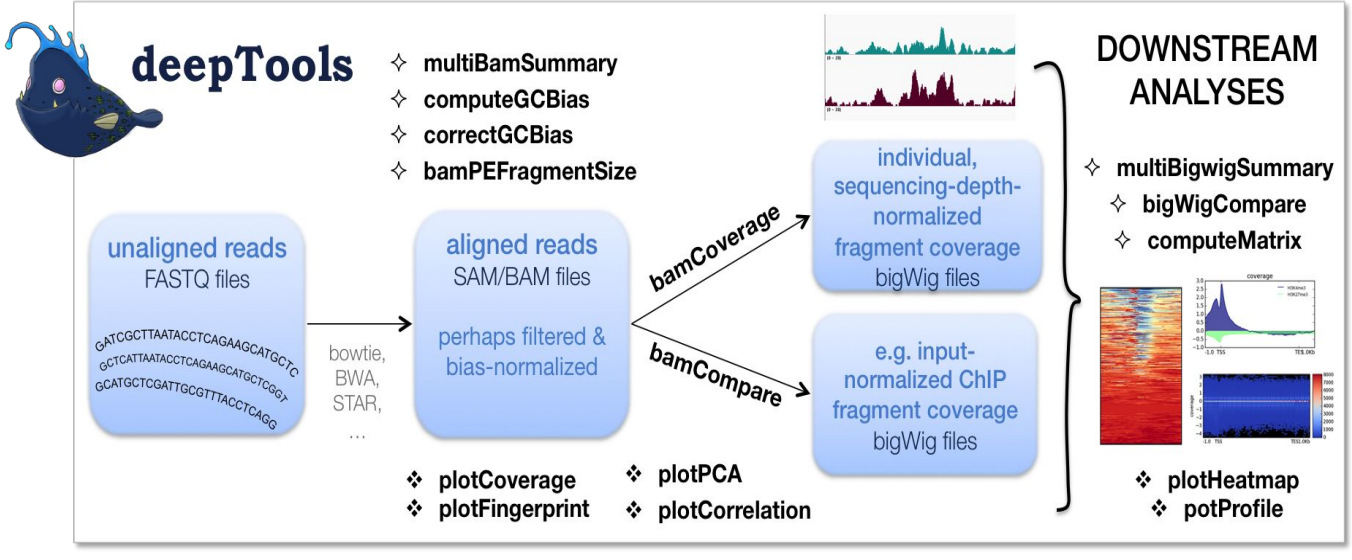


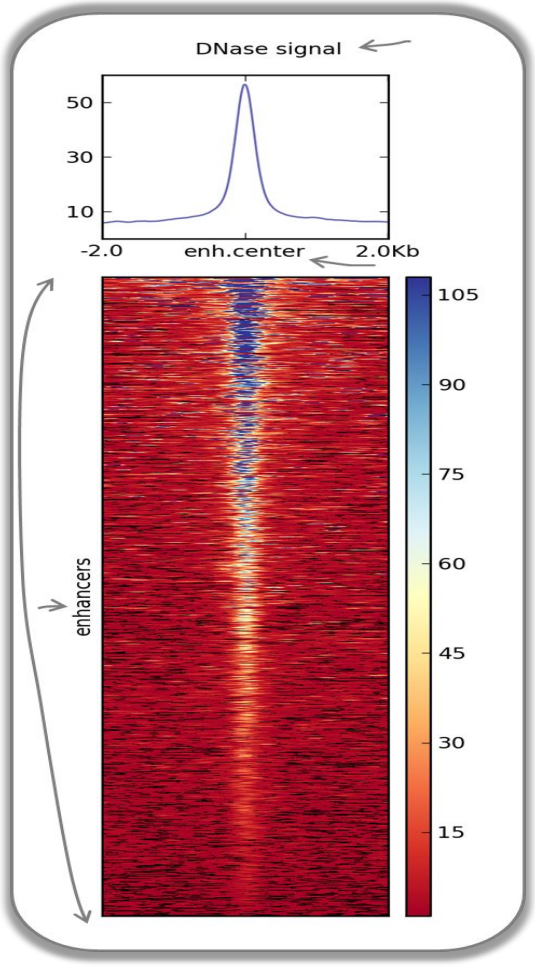
Rtracklayer

- The ***rtracklayer*** package is an interface (or layer) between R and genome browsers
- Its main purpose is the visualization of genomic annotation tracks, whether generated through experimental data analysis performed in R or loaded from an external data source.
- The features of ***rtracklayer*** may be divided into two categories:
  - The import/export of track data
  - The control and querying of external genome browser sessions and views.
- The basic track data structure in Bioconductor is the *GRanges* class, defined in the ***GenomicRanges*** package.
- All positions in a GRanges should be 1-based.
- With ***rtracklayer***, the user may start a genome browser session, create and manipulate genomic views, and import/export tracks and sequences to and from a browser.

deepTools2

Plotting coverage, GC bias detection, Normalization, clustering and visualization of aligned sequencing data





Ramirez et al., Nuc. Acids Res. 2016

References

- G Yu et al., ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015, 31(14):2382-2383
- Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014 Sep 8;47:11.12.1-34
- Neph et al., BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012 Jul 15;28(14):1919-20.
- Lawrence M et al., Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8)
- Lawrence M et al., rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009 Jul 15;25(14):1841-2
- Hung JH et al., Visualizing Genomic Annotations with the UCSC Genome Browser. Cold Spring Harb Protoc. 2016 Nov 1;2016(11):pdb.prot093062.
- Ramírez et al., deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016 Jul 8;44