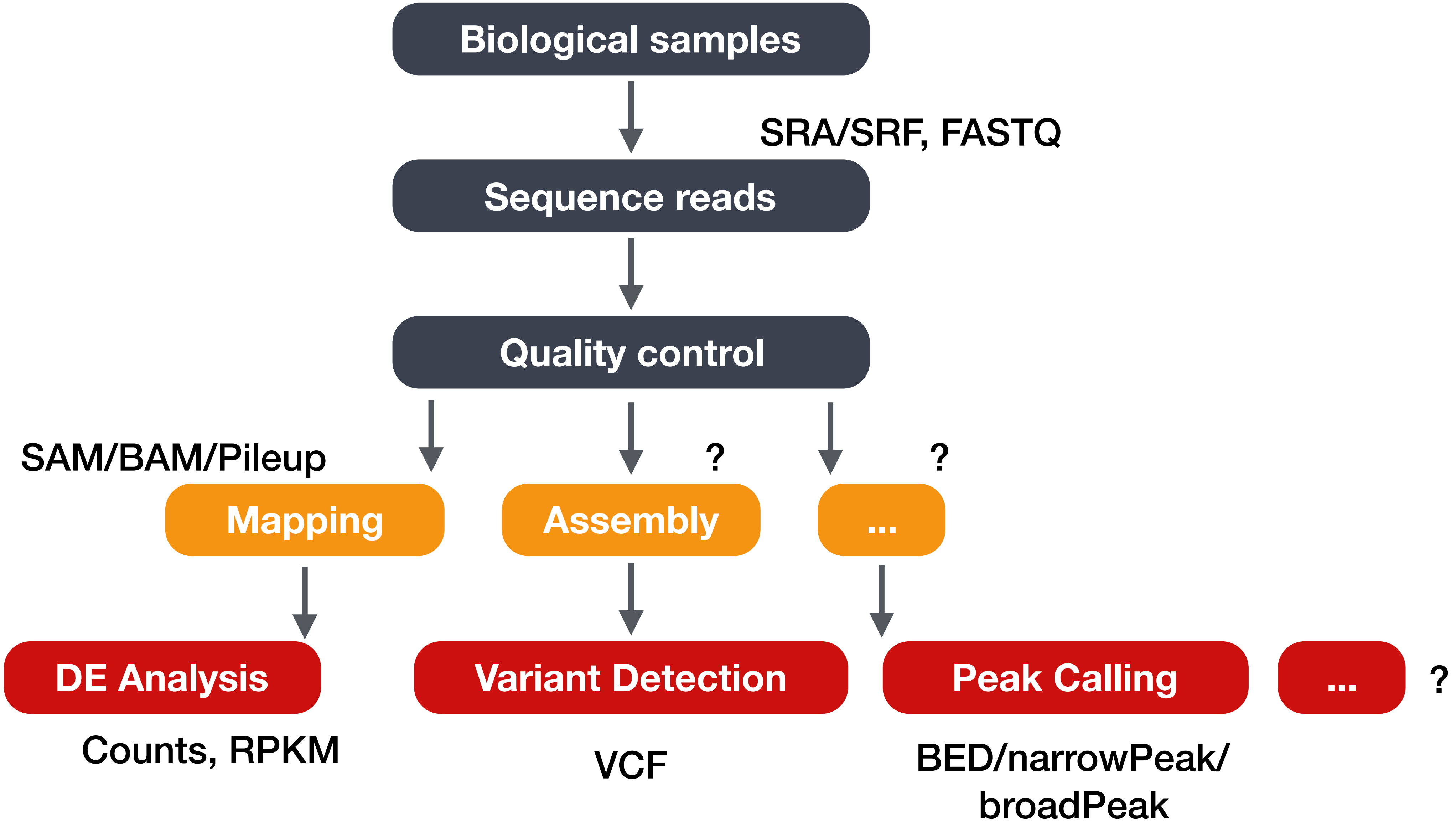
ChIP-seq (NGS) Data Formats



Biological samples

SRA/SRF, FASTQ

Sequence reads

Quality control

SAM/BAM/Pileup

Mapping

?

Assembly

?

...

DE Analysis

Counts, RPKM

Variant Detection

VCF

Peak Calling

BED/narrowPeak/
broadPeak

...

?

NGS analysis workflows

Common data types and file formats

- You will encounter 2 major types of data formats:
    - ◇ Sequence formats
    - ◇ Genome feature formats (information taking genome coordinates into account)
- Specialized file formats represent these data types in a structured manner, and can combine multiple data types in one file.
- Some file formats are not human-readable (***binary***).
- Many are human readable, but extremely large; ***never use Word or Excel to open these!***

Sequence formats

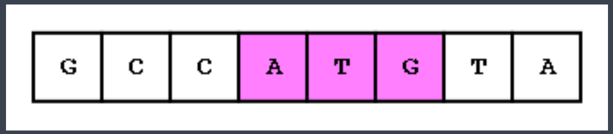
- FASTA (simple representation of sequence data: protein & nucleotide)
- FASTQ (complex, includes data quality information: raw sequencing)

Genome feature formats

- Tab-delimited (text file separated by tabs)
- Contain specific information about **genomic coordinates** of various genomic “features” (e.g. exon, UTRs, etc.)
- May or may not include sequence data
- Some examples include:
    - ◇ SAM/BAM
    - ◇ UCSC formats (BED, WIG, etc.)
    - ◇ GTF/GFF (GTF v2, and GFF v3)

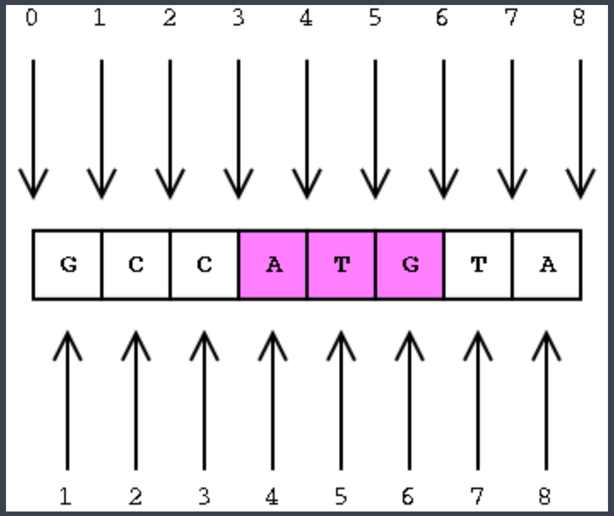
Types of genomic coordinates

Where is base 1 and where is base 8?



Types of genomic coordinates

Coords	***Where is ATG?***	***Length***
0-based (half-open) *preferred by programmers*	`( 3, 6 ]`	Len = end - start
1-based (closed) *preferred by biologists*	`[ 4, 6 ]`	Len = end – start + 1



Feature format

- The chromosome names in a feature format file MUST match the names in the associated reference genome file
  - ◇ Tied to a specific version of a reference genome
  - ◇ Not all reference genomes are the represented the same!
  - ◇ E.g. human chromosome 1
    - ◇ **UCSC – ‘chr1’** versus **Ensembl/NCBI – ‘1’**
- Best practice: get feature format files from the same source (i.e UCSC, Ensembl, NCBI) as the reference genome

Feature formats for alignment

SAM – Sequence Alignment/Map format

- SAM file format stores alignment information, including read name, alignment coordinates, mismatches, etc. in plain text
- **1-based coordinates**

BAM – BGZF compressed SAM format

- Binary (compressed) version of SAM and is therefore not human readable
- **0-based coordinates**

Genome interval file: BED

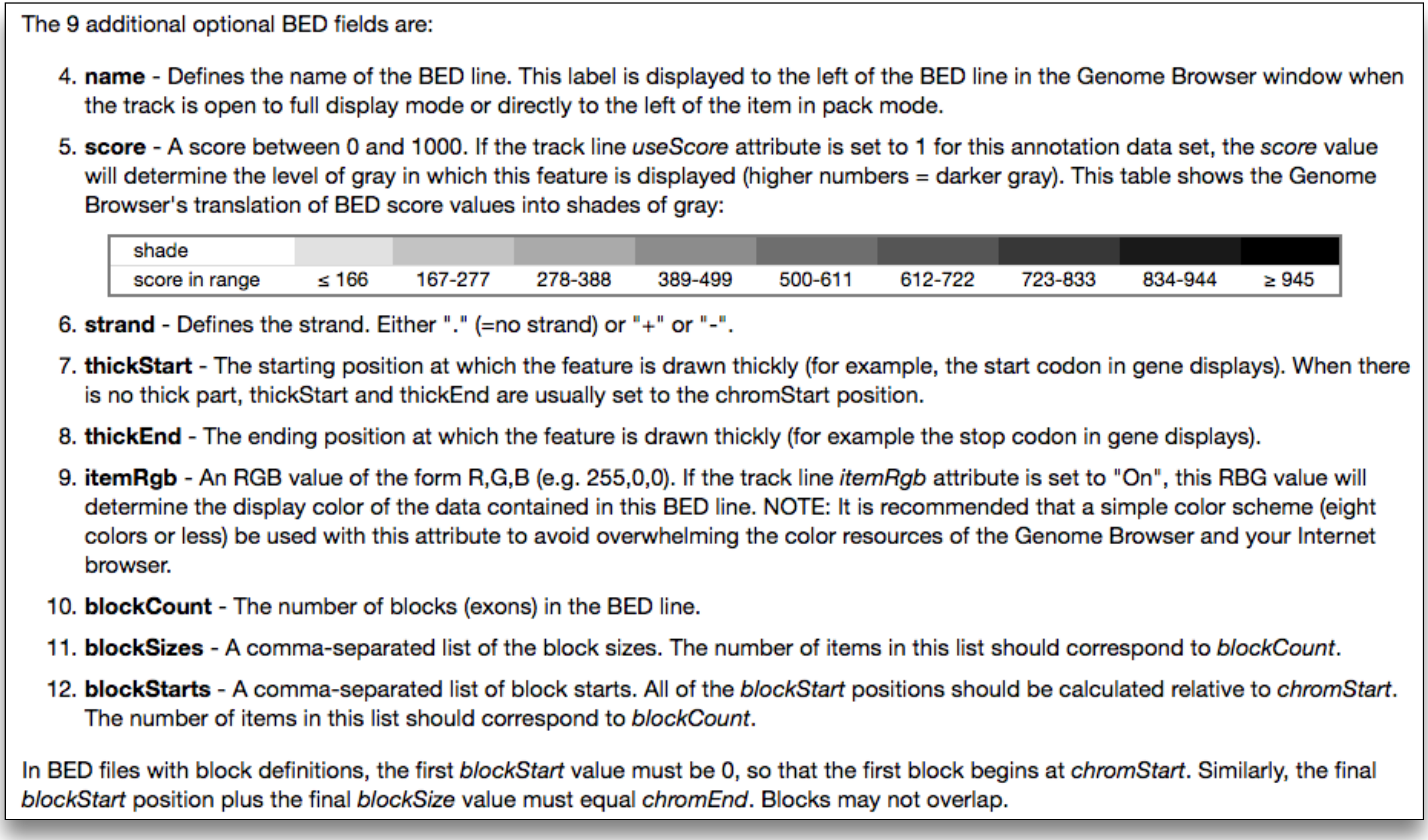
- Tab- or whitespace-delimited text file; consists of one line per feature
- **0-based coordinates**
- The first three fields/columns in each feature line are required:
    - ***chr***: chromosome name/ID
    - ***start***: start position of the feature
    - ***end***: end position of the feature

```
chr1  213941196  213942363
chr1  213942363  213943530
chr1  213943530  213944697
```

Genome interval file: BED

- Tab- or whitespace-delimited text file; consists of one line per feature
- **0-based coordinates**
- The first three fields/columns in each feature line are required:
    - ***chr***: chromosome name/ID
    - ***start***: start position of the feature
    - ***end***: end position of the feature
- There are nine additional fields that are optional.
- Sometimes the BED format is referenced based on the number of additional fields, e.g. BED12 format = BED file with all 12 columns

Genome interval file: BED



The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

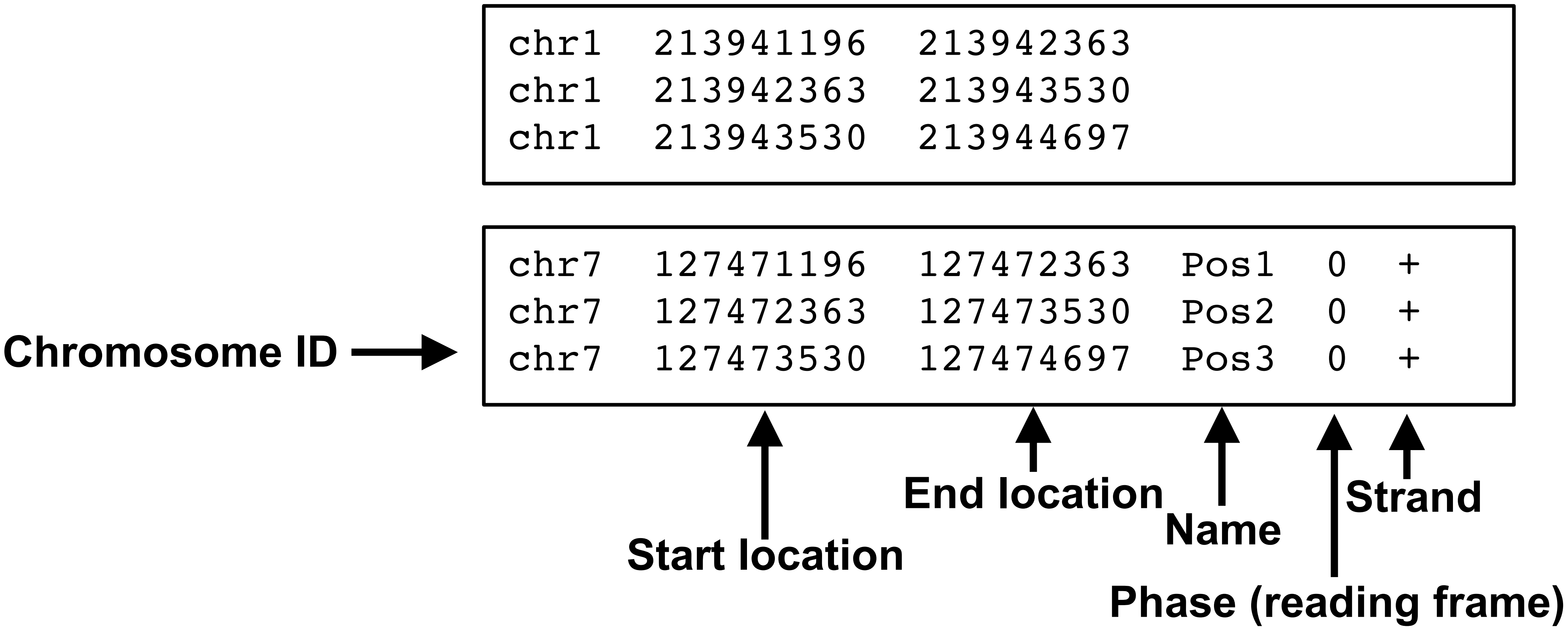
shade	░	░	▒	▒	▓	▓	█	█	█
score in range	≤ 166	167-277	278-388	389-499	500-611	612-722	723-833	834-944	≥ 945

6. **strand** - Defines the strand. Either "." (=no strand) or "+" or "-".
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, thickStart and thickEnd are usually set to the chromStart position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

In BED files with block definitions, the first *blockStart* value must be 0, so that the first block begins at *chromStart*. Similarly, the final *blockStart* position plus the final *blockSize* value must equal *chromEnd*. Blocks may not overlap.

https://genome.ucsc.edu/FAQ/FAQformat.html#format1

Genome interval file: BED



```
chr1   213941196   213942363
chr1   213942363   213943530
chr1   213943530   213944697
```

```
chr7   127471196   127472363   Pos1   0   +
chr7   127472363   127473530   Pos2   0   +
chr7   127473530   127474697   Pos3   0   +
```

BedGraph format

- Allows the display of continuous-valued data in a track format, especially for data that is sparse or contains elements of varying size
- Based on the BED format, but with a few differences:
  - The score is placed in column 4 not 5
  - Track lines must also be included (these are optional in BED files)
- **0-based coordinates**
- Preserve data in original format (no compression)
- **Often used for displaying density or coverage information**

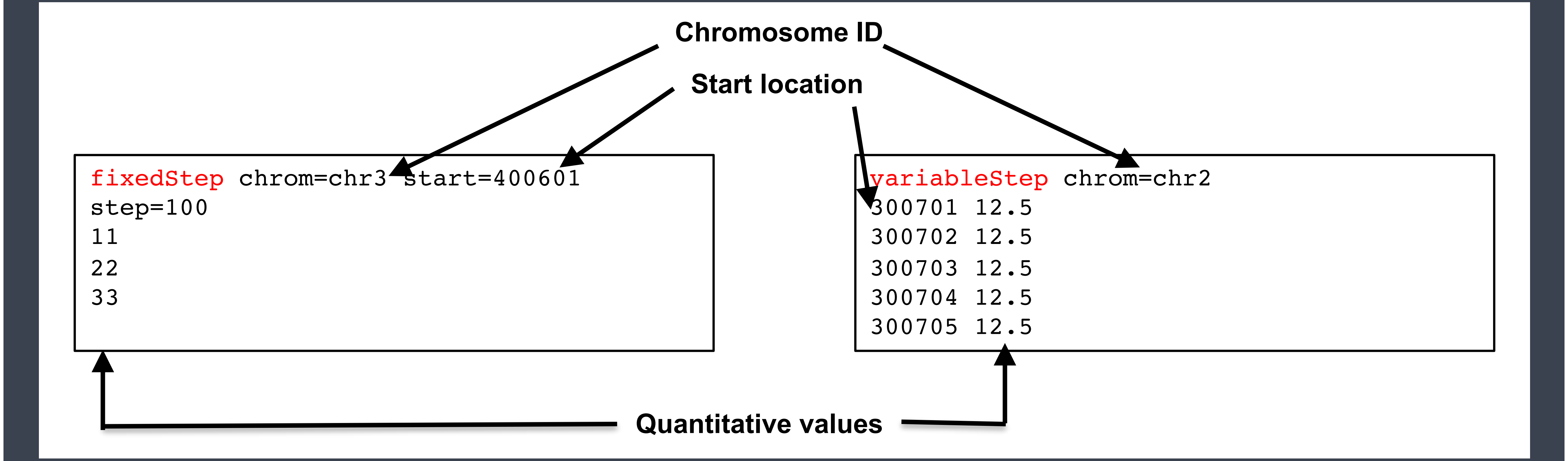
BedGraph format

```
track type=bedGraph name="BedGraph Format" description="BedGraph format”
chr19   49302000   49302300   -1.00
chr19   49302300   49302600   -0.75
chr19   49302600   49302900   -0.50
chr19   49302900   49303200   -0.25
chr19   49303200   49303500    0.00
chr19   49303500   49303800    0.25
chr19   49303800   49304100    0.50
chr19   49304100   49304400    0.75
chr19   49304400   49304700    1.00
```

Wiggle format

- Similar to the bedGraph format but:
    - it’s compressed, and exact data values cannot be recovered from the compression
    - data elements need to be equally sized (i.e bins of specified size)
- Associates a floating point number with positions in the genome, which is plotted on the track’s vertical axis to create a wiggly line
- **1-based coordinates**

Wiggle format



Chromosome ID

Start location

```
fixedStep chrom=chr3 start=400601
step=100
11
22
33
```

```
variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
```

Quantitative values

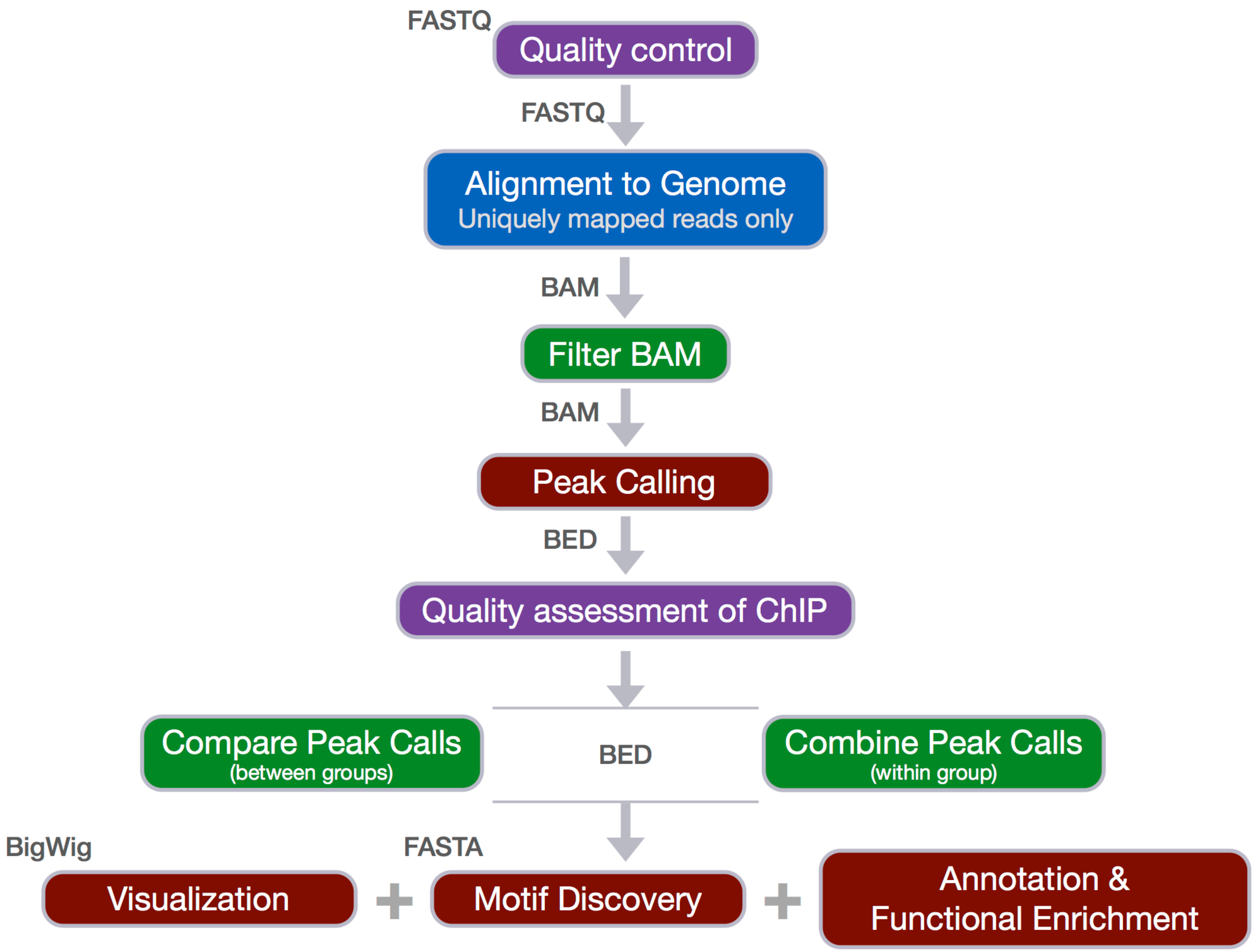
bigWig format

- An indexed binary format derived from the wiggle file
    - *Initially created for the wiggle file, but now bigWig can also be created from bedGraph files*
- Only portions of the file is needed to display are transferred
- Faster than the wiggle or bedGraph formats; good for large datasets
- **1-based coordinates**

Commonly used file formats

- FASTA
- FASTQ – Fasta with quality
- SAM – Sequence Alignment/Map format
- BAM – Binary Sequence Alignment/Map format
- Bed – Basic genome interval
- BedGraph
- Wiggle (wig, bigwig) – tab-limited format to represent continuous values

http://genome.ucsc.edu/FAQ/FAQformat.html



FASTQ

Quality control

FASTQ

Alignment to Genome
Uniquely mapped reads only

BAM

Filter BAM

BAM

Peak Calling

BED

Quality assessment of ChIP

Compare Peak Calls
(between groups)

BED

Combine Peak Calls
(within group)

BigWig

Visualization

+

FASTA

Motif Discovery

+

Annotation &
Functional Enrichment

ChIP-seq workflow