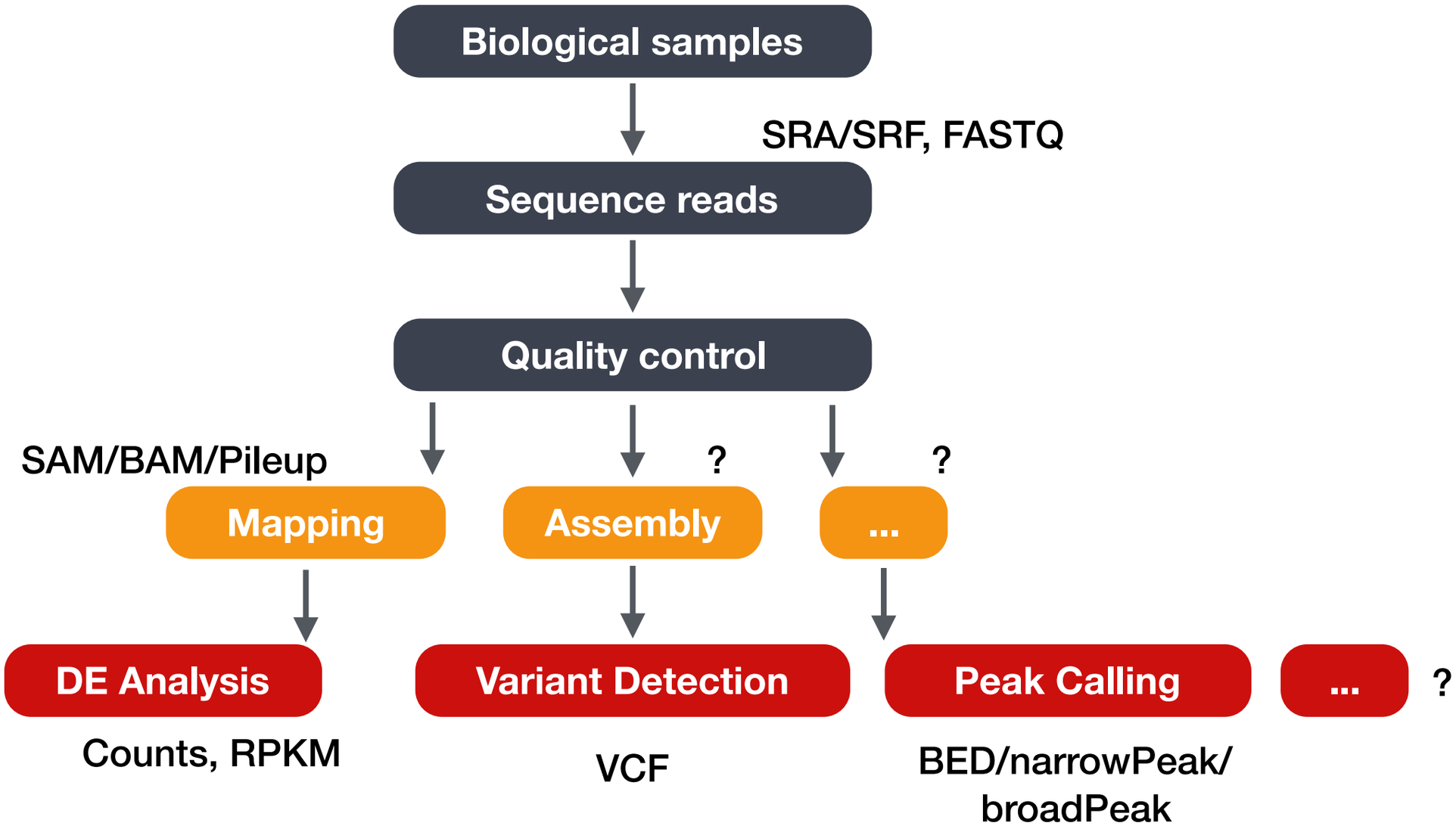
ChIP-seq (NGS) Data Formats



NGS analysis workflows

Common data types and file formats

- You will encounter 3 major types of data, with several associated file formats:
  - Sequence data - FASTA, FASTQ
  - Alignment data - SAM, BAM
  - Genome feature data - BED, Wiggle, GTF, GFF
- Some file formats are not human-readable (***binary***).
- Many are human readable, but extremely large! *(Never use Word or Excel to open these!)*

Common data types and file formats

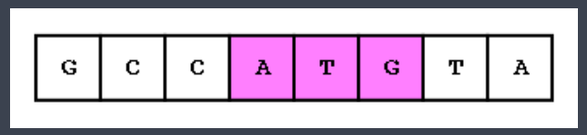
- You will encounter 3 major types of data, with several associated file formats:
  - Sequence data - FASTA, FASTQ
  - Alignment data - SAM, BAM
  - **Genome feature data - BED, Wiggle, GTF, GFF**
- Some file formats are not human-readable (***binary***).
- Many are human readable, but extremely large! *(Never use Word or Excel to open these!)*

Formats for “Genome Features/Coordinates” data

- 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)

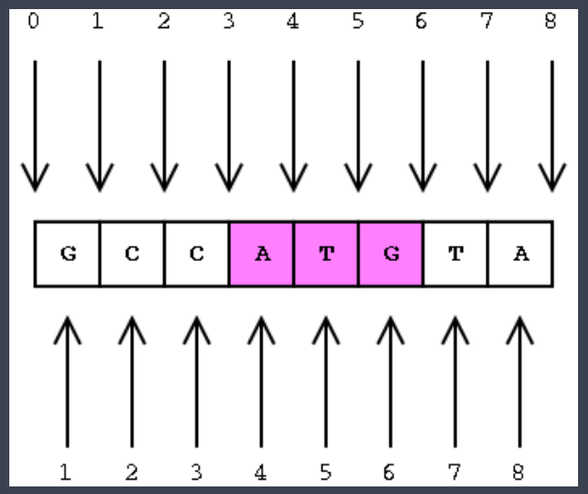
Genomic coordinates can be represented in 2 ways

Where is base 1 and where is base 8?



Genomic coordinates can be represented in 2 ways

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



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. BED 6+4 format = the first 6 columns of a BED file + 4 other columns)*

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   +
```

Chromosome ID →

Start location

End location

Name

Phase (reading frame)

Strand

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

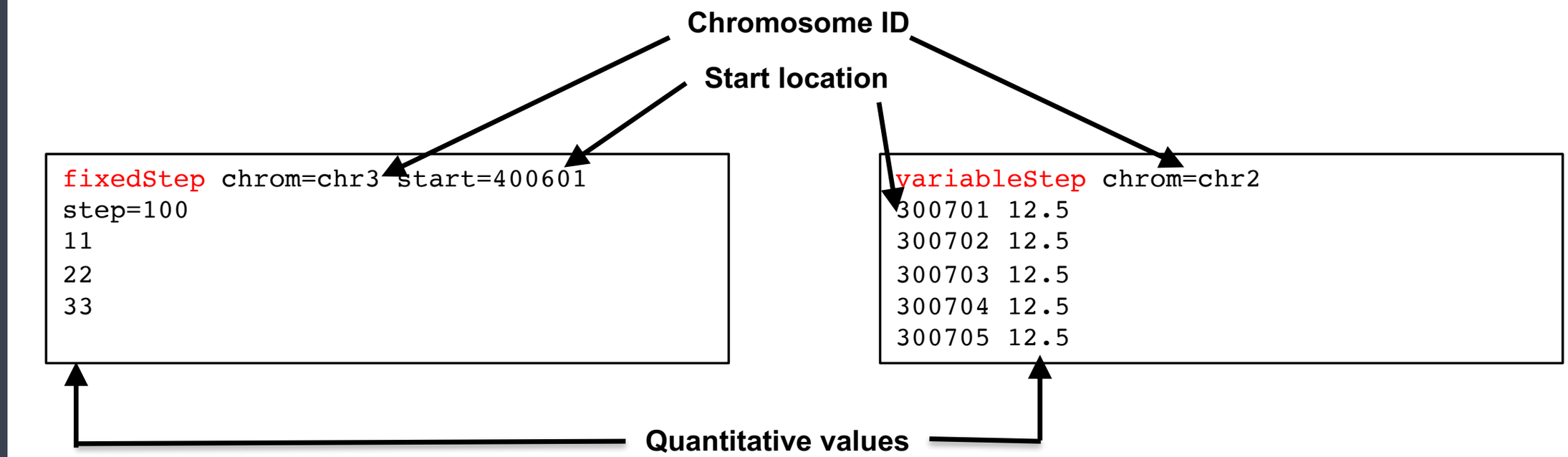
BedGraph format

```
track type=bedGraph name="BedGraph Format" description="BedGraph format" visibility=full
chr19    49302000    49302300    -1.0
chr19    49302300    49302600    -0.75
chr19    49302600    49302900    -0.50
chr19    49302900    49303200    -0.25
```

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



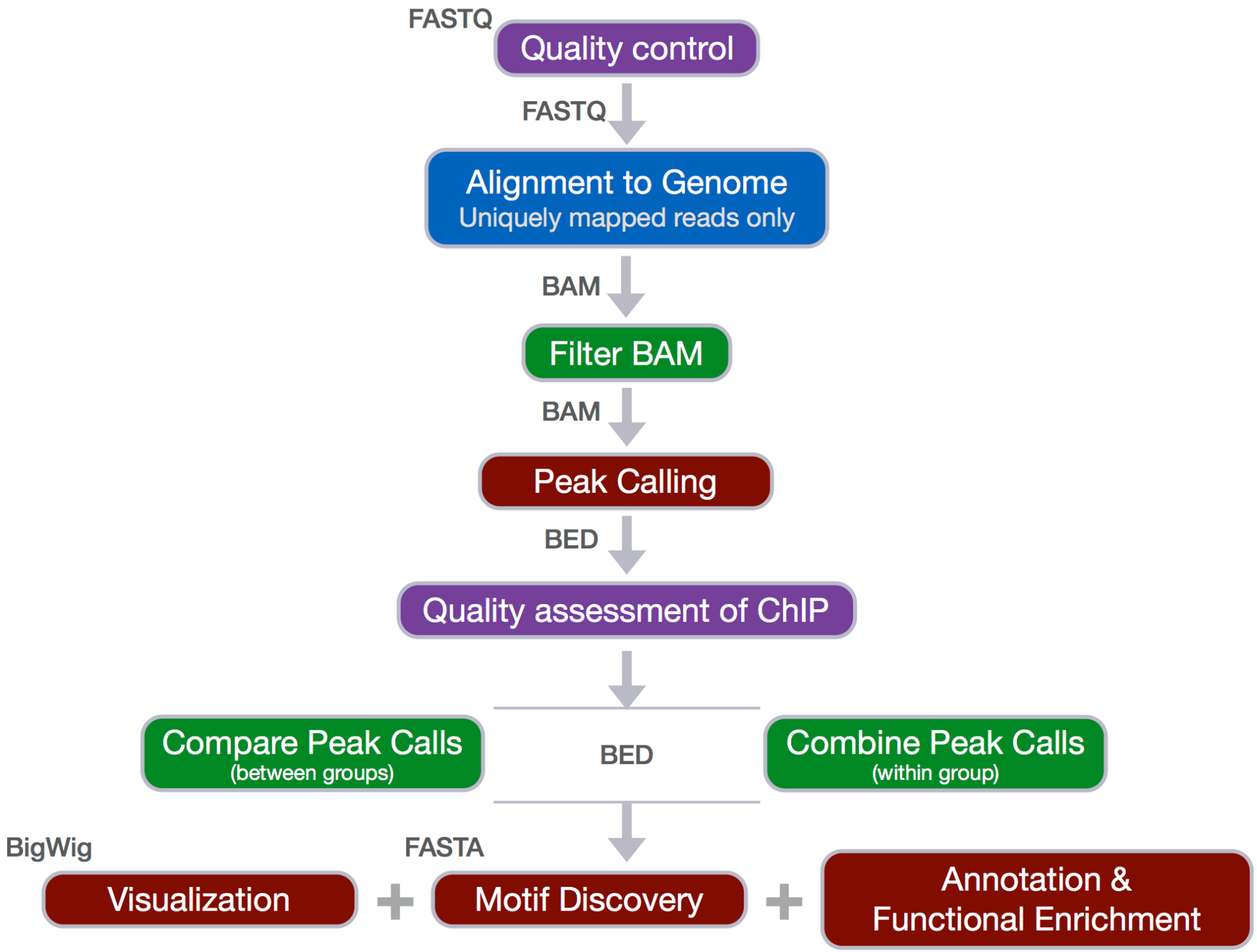
bigWig format

- 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*
- Faster than the wiggle or bedGraph formats; good for large datasets
- **1-based coordinates**

Commonly used file formats for ChIP-seq

- *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



ChIP-seq workflow