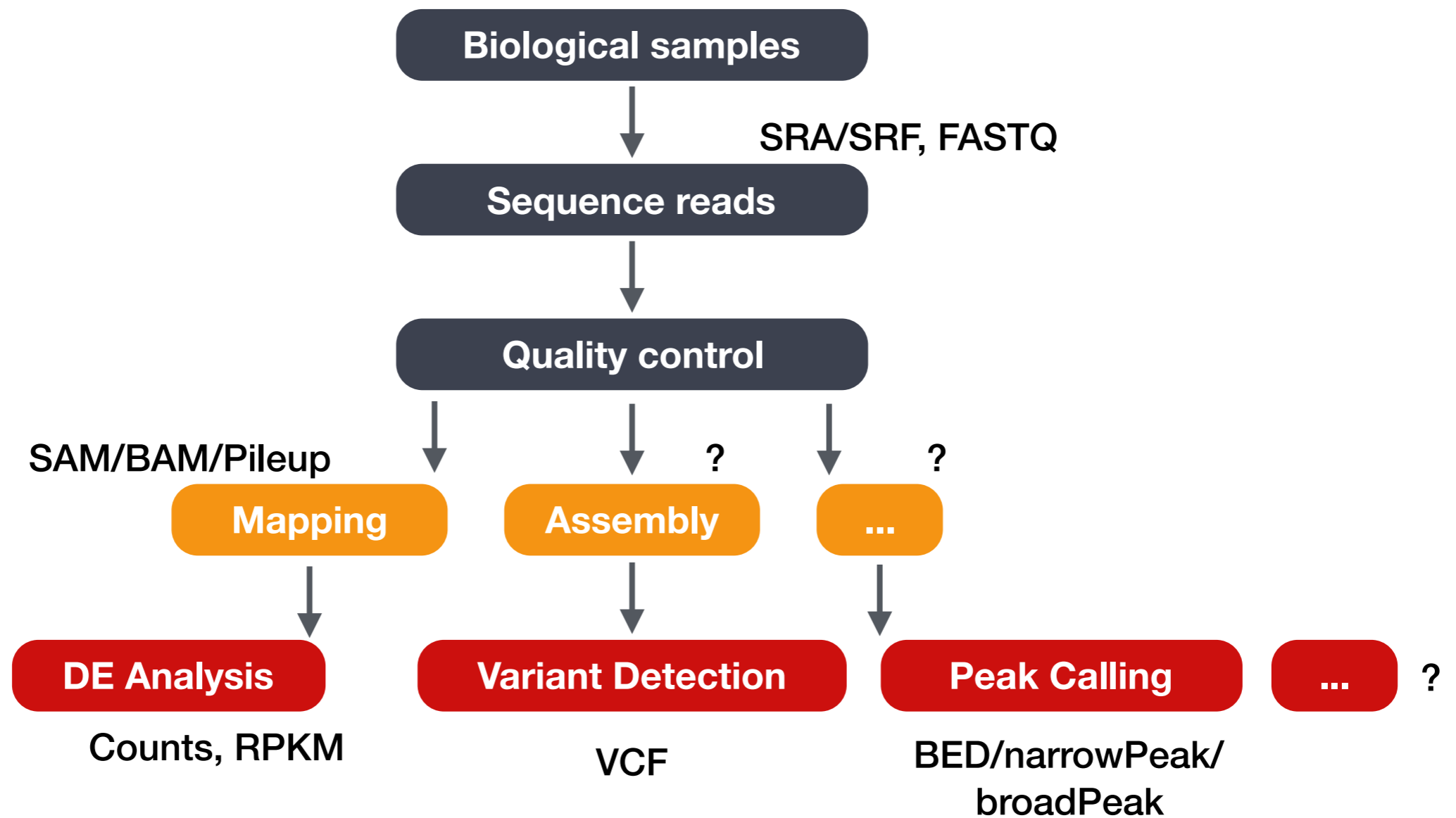


ChIP-seq (NGS) Data Formats



NGS analysis workflows

Common data types and file formats

- You will encounter 3 major types of data, with several associated file formats:
 - ◇ Sequence data
 - ◇ Alignment data
 - ◇ Genome feature data

Common data types and file formats

- You will encounter 3 major types of data, with several associated file formats:
 - ◇ Sequence data
 - ◇ Alignment data
 - ◇ Genome feature data
- Specialized file formats represent these data types in a structured manner, and can combine multiple data types in one file.

Common data types and file formats

- You will encounter 3 major types of data, with several associated file formats:
 - ◇ Sequence data
 - ◇ Alignment data
 - ◇ Genome feature data
- 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**).

Common data types and file formats

- You will encounter 3 major types of data, with several associated file formats:
 - ◇ Sequence data
 - ◇ Alignment data
 - ◇ Genome feature data
- 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!

Simple sequence formats

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

FASTA

```
>SRR014849.1 EIXKN4201CFU84 length=93  
GGGGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAAACCGAAAGGGTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAAAGCAATGCC  
AATA
```

```
>gi|340780744|ref|NC_015850.1| Acidithiobacillus caldus SM-1 chromosome, complete genome  
ATGAGTAGTCATTCAGCGCCGACAGCGTTGCAAGATGGAGCCGCGCTGTGGTCCGCCCTATGCGTCCAACCTGGAGCTCGTCACGAG  
TCCGCAGCAGTTCAATACCTGGCTGCGGCCCTGCGTGGCGAATTGCAGGGTCATGAGCTGCGCCTGCTCGCCCCAATCCCTTCG  
TCCGCGACTGGGTGCGTGAACGCATGGCCGAACTCGTCAAGGAACAGCTGCAGCGGATCGCTCCGGGTTTTGAGCTGGTCTTCGCT  
CTGGACGAAGAGGCAGCAGCGGCGACATCGGCACCGACCGCGAGCATTGCGCCCAGCGCAGCAGCGCACCCGGTGGTCACCGCCT  
CAACCCAGCCTTCAACTTCCAGTCCTACGTCGAAGGGAAGTCCAATCAGCTCGCCCTGGCGGCAGCCCGCCAGGTTGCCCAGCATC  
CAGGCAAATCCTACAACCCACTGTACATTTATGGTGGTGTGGGCCTCGGCAAGACGCACCTCATGCAGGCCGTGGGCAACGATATC  
CTGCAGCGGCAACCCGAGGCCAAGGTGCTCTATATCAGCTCCGAAGGCTTCATCATGGATATGGTGGCCTCGCTGCAACACAATAC  
CATCAACGACTTCAAACAGCGTTATCGCAAGCTGGACGCCCTGCTCATCGACGACATCCAGTTCTTTGCGGGCAAGGACCGCACCC
```

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)  
QIKDLLVSSSTDLDTTLLVLVNAIYFKGMWKTAFAEDTREMPPFHVTKQESKPVQMMCMNNSFNVATLPAE
```


Feature formats

- Tab-delimited (text file separated by tabs)

Feature formats

- Tab-delimited (text file separated by tabs)
- Contain specific information about **genomic coordinates** of various genomic “features” (e.g. exon, UTRs, etc.)

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

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:

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

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.)

Feature formats

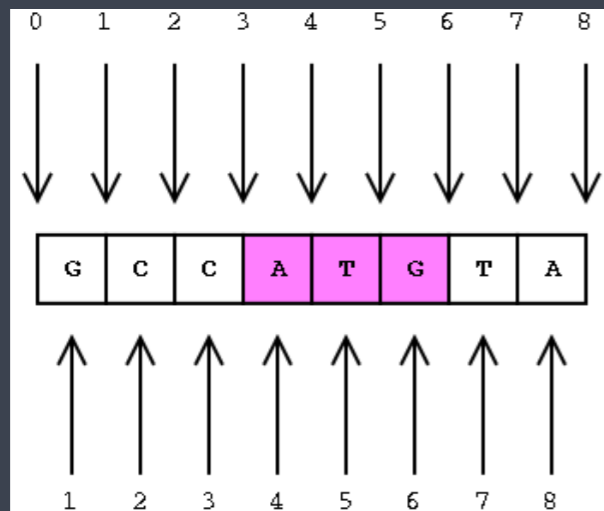
- Tab-delimited (text file separated by tabs)
- Contain specific information about **genomic coordinates** of various genomic “features” (e.g. exon, UTRs, etc.)
- May or may not include sequence data
- Some examples include:
 - ◇ SAM/BAM
 - ◇ UCSC formats (BED, WIG, etc.)
 - ◇ GTF/GFF (GTF v2, and GFF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

Genomic coordinates can be represented in 2 ways

Where is base 1 and where is base 8?

G	C	C	A	T	G	T	A
---	---	---	---	---	---	---	---

Genomic coordinates can be represented in 2 ways

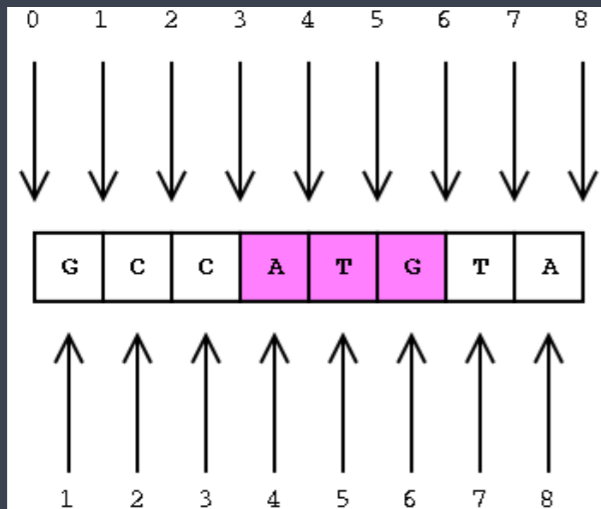


Genomic coordinates can be represented in 2 ways

Coords

Where is ATG?

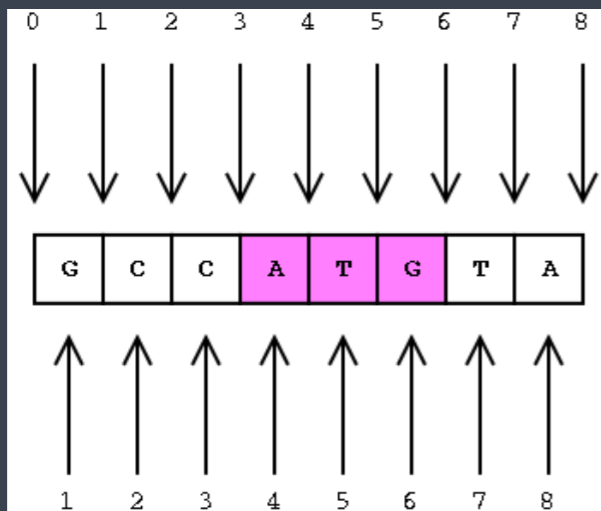
Length



Genomic coordinates can be represented in 2 ways

Coords

0-based (half-open)
preferred by programmers



Where is ATG?

(3, 6]

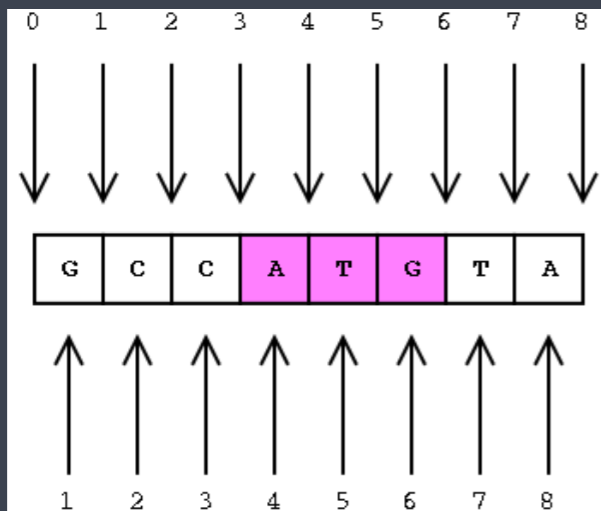
Length

Len = end - start

Genomic coordinates can be represented in 2 ways

Coords

0-based (half-open)
preferred by programmers



1-based (closed)
preferred by biologists

Where is ATG?

(3, 6]

Length

Len = end - start

[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

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

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!

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’

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

Alignment file: SAM

- SAM – Sequence Alignment/Map format

Alignment file: SAM

- SAM – Sequence Alignment/Map format
- SAM file format stores alignment information, including read name, alignment coordinates, mismatches, etc.

Alignment file: SAM

- SAM – Sequence Alignment/Map format
- SAM file format stores alignment information, including read name, alignment coordinates, mismatches, etc.
- Plain text

Alignment file: SAM

- SAM – Sequence Alignment/Map format
- SAM file format stores alignment information, including read name, alignment coordinates, mismatches, etc.
- Plain text
- **1-based coordinates**

Alignment file: SAM

- SAM – Sequence Alignment/Map format
- SAM file format stores alignment information, including read name, alignment coordinates, mismatches, etc.
- Plain text
- **1-based coordinates**
- Files can be very large: Many 100's of GB or more

Alignment file: BAM

- BAM – BGZF compressed SAM format

Alignment file: BAM

- BAM – BGZF compressed SAM format
- Binary (compressed) version of SAM and is therefore not human readable

Alignment file: BAM

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

Alignment file: BAM

- BAM – BGZF compressed SAM format
- Binary (compressed) version of SAM and is therefore not human readable
- **0-based coordinates**
- Makes the alignment information easily accessible to downstream applications (SAM format is mostly useless for downstream analyses)

Alignment file: BAM

- BAM – BGZF compressed SAM format
- Binary (compressed) version of SAM and is therefore not human readable
- **0-based coordinates**
- Makes the alignment information easily accessible to downstream applications (SAM format is mostly useless for downstream analyses)
- Files are typically very large: ~ 1/5 of SAM, but still very large

Genome interval file: BED

- Tab- or whitespace-delimited text file; consists of one line per feature

Genome interval file: BED

- Tab- or whitespace-delimited text file; consists of one line per feature
- **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

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.

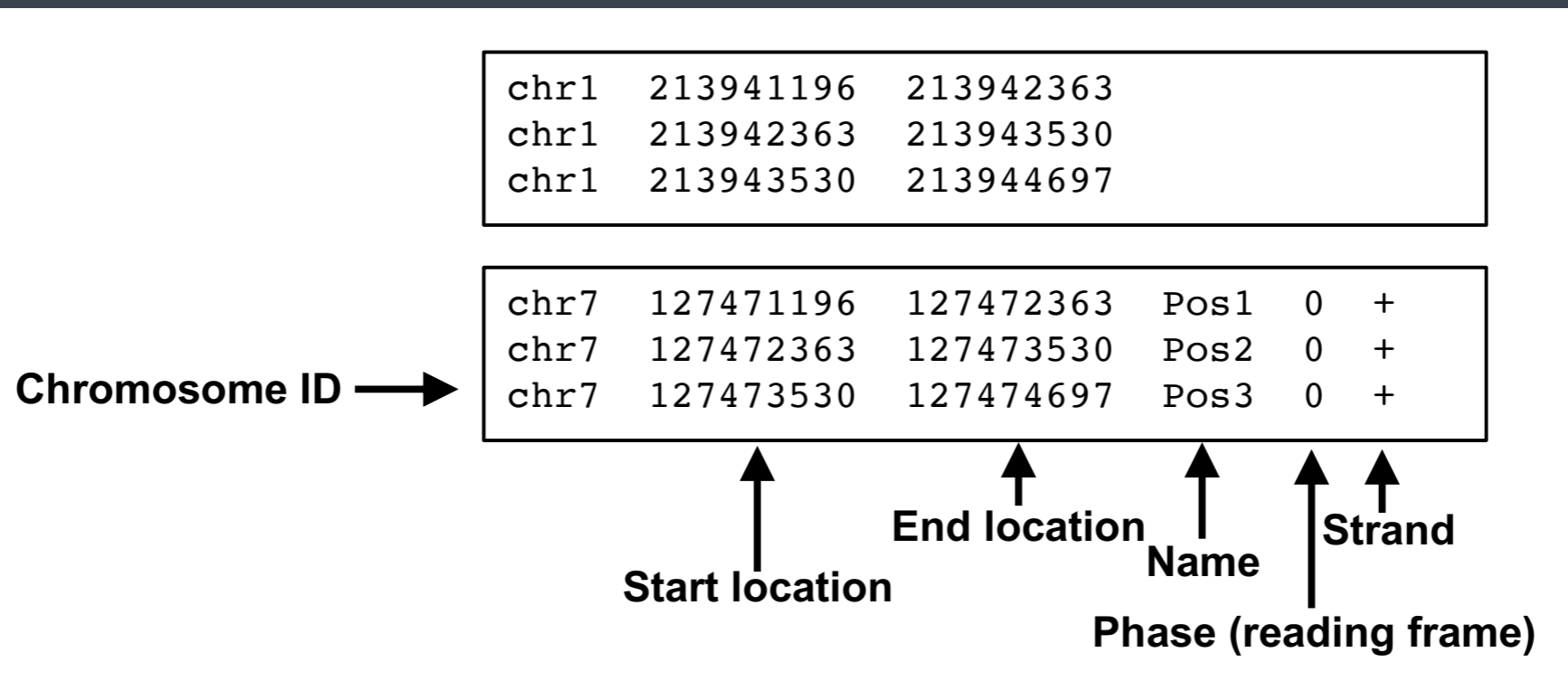
Genome interval file: BED

- 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

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



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

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)

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**

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)

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

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)

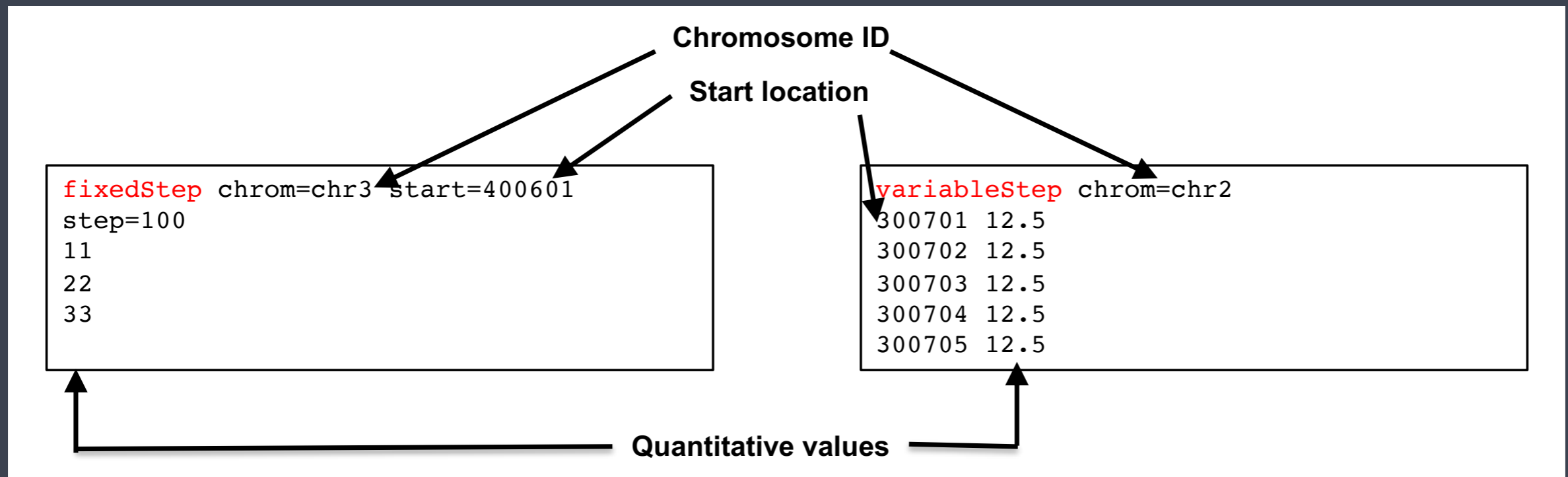
Wiggle format

- 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

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*

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

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

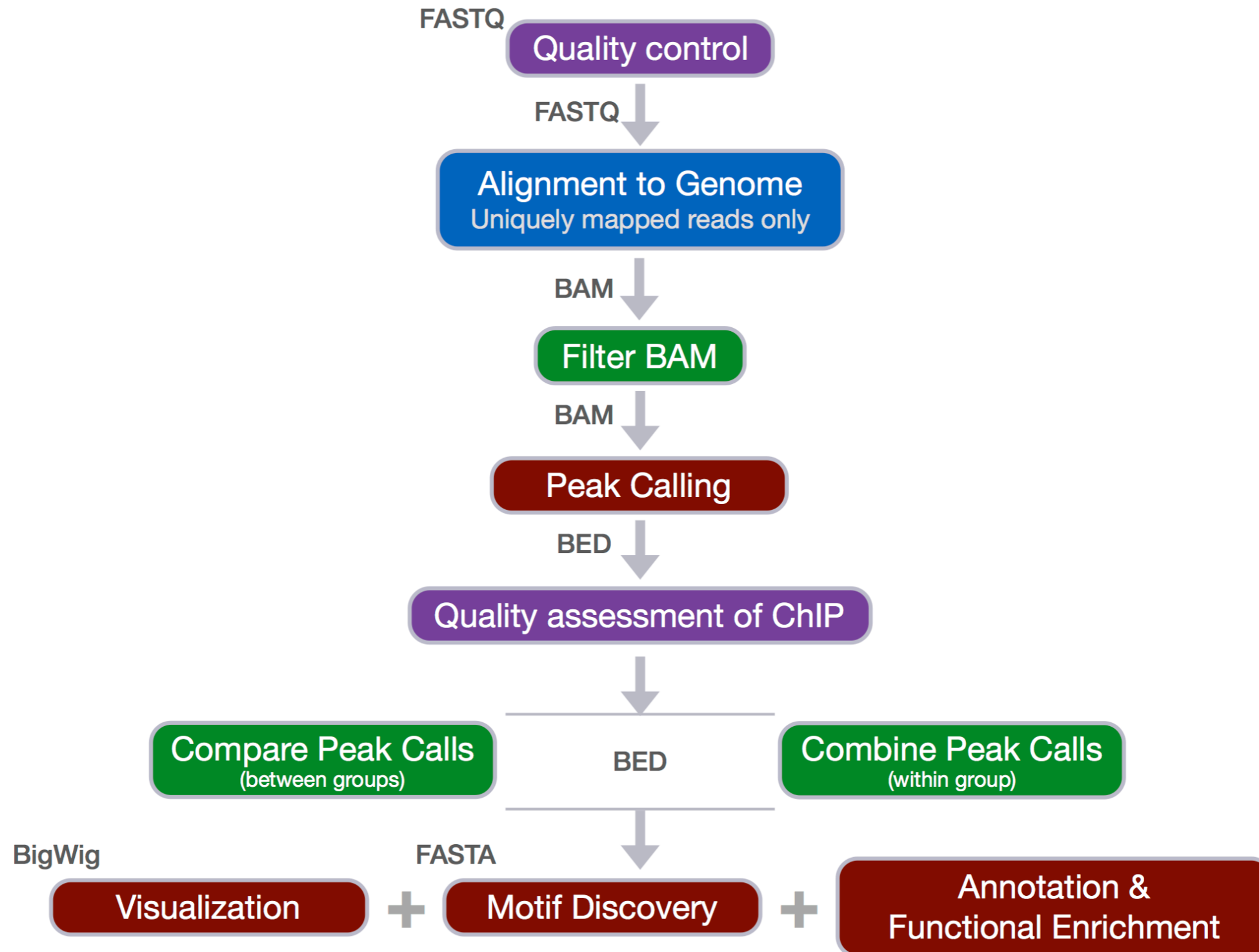
bigWig format

- 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
- *GFF3 – Gene feature format (genome interval ++)*
- *GTF – Gene transfer format (genome interval ++)*

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



ChIP-seq workflow

These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

