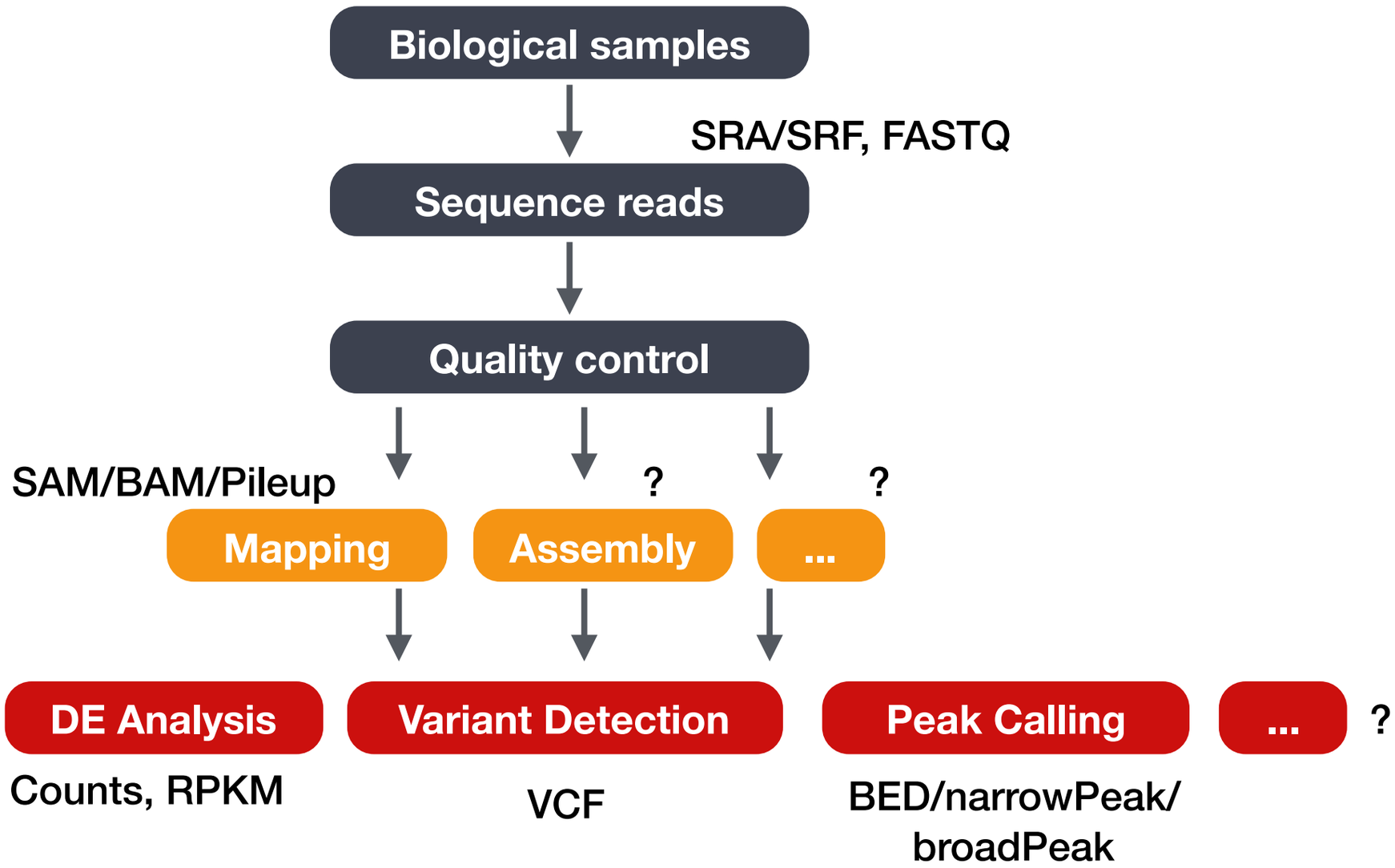
RNA-seq workflow and data standards



NGS analysis workflows

Common data types and file formats

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

```
>gi|340780744|ref|NC_015850.1| Acidithiobacillus caldus SM-1 chromosome, complete genome
ATGAGTAGTCATTCAGCGCCGACAGCGTTGCAAGATGGAGCCGCGCTGTGGTCCGCCCTATGCGTCCAACTGGAGCTCGTCACGAG
TCCGCAGCAGTTCAATACCTGGCTGCGGCCCCTGCGTGGCGAATTGCAGGGTCATGAGCTGCGCCTGCTCGCCCCCAATCCCTTCG
TCCGCGACTGGGTGCGTGAACGCATGGCCGAACTCGTCAAGGAACAGCTGCAGCGGATCGCTCCGGGTTTTGAGCTGGTCTTCGCT
CTGGACGAAGAGGCAGCAGCGGCGACATCGGCACCGACCGCGAGCATTGCGCCCGAGCGCAGCAGCGCACCCGGTGGTCACCGCCT
CAACCCAGCCTTCAACTTCCAGTCCTACGTCGAAGGGAAGTCCAATCAGCTCGCCCTGGCGGCAGCCCGCCAGGTTGCCCAGCATC
CAGGCAAATCCTACAACCCACTGTACATTTATGGTGGTGTGGGCCTCGGCAAGACGCACCTCATGCAGGCCGTGGGCAACGATATC
CTGCAGCGGCAACCCGAGGCCAAGGTGCTCTATATCAGCTCCGAAGGCTTCATCATGGATATGGTGCGCTCGCTGCAACACAATAC
CATCAACGACTTCAAACAGCGTTATCGCAAGCTGGACGCCCTGCTCATCGACGACATCCAGTTCTTTGCGGGCAAGGACCGCACCC
```

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

FASTQ: FASTA with Quality scores

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAAAGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""7F@71,'';C?,B;?6B;:EA1EA1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@/=<?7=9<2A8==
```

Line	Description
1	Always begins with '@' and then information about the read
2	The actual DNA sequence
3	Always begins with a '+' and sometimes the same info in line 1
4	Has a string of characters which represent the quality score

Feature formats

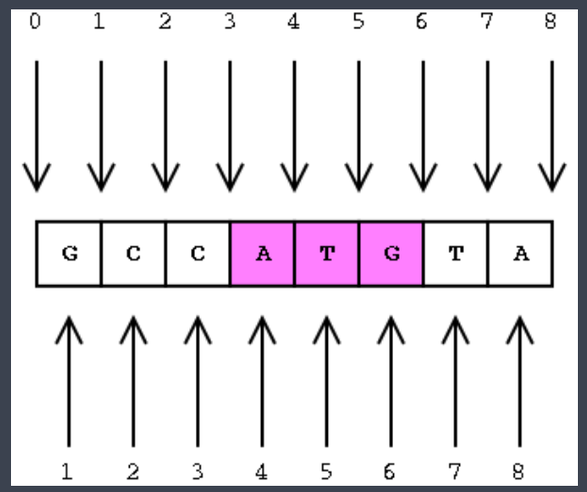
- Tab-delimited (Text file separated by tabs)
- Contain specific information about genome (or assembly) coordinates
- May or may not include sequence data
- Some examples include:
    - GTF/GFF (GTF v2, and GFF v3)
    - SAM/BAM
    - UCSC formats (BED, WIG, etc.)

Genomic coordinates can be represented in 2 ways

Where is 1 and where is 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



Feature formats: GTF (Gene Transfer Format)

- Evolved from Sanger Centre GFF (gene feature format) originally, but repeatedly modified
- Differences in representation of information make it distinct from GFF
- **1-based coordinates**
- Source of the GTF is important, subtle differences between an Ensembl version and a UCSC version can cause issues.

```
chr1 unknown   exon          113217048  113217252  .   +    .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown   exon          113217048  113217351  .   +    .   gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown   exon          113217470  113217671  .   +    .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown   CDS           113217535  113217671  .   +    0   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown   start_codon   113217535  113217537  .   +    .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

Chromosome ID
Source
Gene feature
Start location
End location
Score (user defined)
Strand
Reading frame
Attributes

Feature formats: GFF3 (Gene Feature Format)

- Tab-delimited file to store genomic features, e.g. genomic intervals of genes and gene structure
- Attributes are hierarchical
- Meant to be unified replacement for GFF/GTF (includes specification)
- **1-based coordinates**

Feature formats: GFF3 versus GTF

GFF3 – Gene feature format

chr1	ensembl_havana	transcript	112674487	112700739	.	+	.	ID=transcript:ENST00000369645;Parent=gene:ENSG00000155363;Name=MOV10-006;biotype=protein_coding;ccdsid=CCDS853.1;havana_transcript=OTTHUMT00000032911;havana_version=1;tag=basic;transcript_id=ENST00000369645;transcript_support_level=5 (assigned to previous version 4);version=5
chr1	havana	exon	112674487	112674729	.	+	.	Parent=transcript:ENST00000369645;Name=ENSE00001450533;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_id=ENSE00001450533;rank=1;version=1
chr1	havana	five_prime_UTR	112674487	112674729	.	+	.	Parent=transcript:ENST00000369645
chr1	havana	five_prime_UTR	112674848	112674912	.	+	.	Parent=transcript:ENST00000369645
chr1	havana	exon	112674848	112675049	.	+	.	Parent=transcript:ENST00000369645;Name=ENSE00003676444;constitutive=0;ensembl_end_phase=2;ensembl_phase=-1;exon_id=ENSE00003676444;rank=2;version=1

GTF – Gene transfer format

chr1	havana	transcript	112674487	112700739	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	exon	112674487	112674729	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; exon_number "1"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; exon_id "ENSE00001450533"; exon_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	five_prime_utr	112674487	112674729	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	five_prime_utr	112674848	112674912	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	exon	112674848	112675049	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; exon_number "2"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; exon_id "ENSE00003676444"; exon_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";

Always check which of the two formats is accepted by the application you're using

Alignment file: SAM

- SAM – Sequence Alignment/Map format
- SAM file format stores alignment information
- Plain text
- **1-based coordinates**
- Files can be very large: Many 100's of GB or more
- Normally converted into BAM to save space (and text format is mostly useless for downstream analyses)

Alignment file: BAM

- BAM – BGZF compressed SAM format
- Compressed/binary version of SAM and is not human readable. Uses a specialize compression algorithm optimized for indexing and record retrieval (bgzip)
- **0-based coordinates**
- Makes the alignment information easily accessible to downstream applications
- Files are typically very large: ~ 1/5 of SAM, but still very large

Feature format

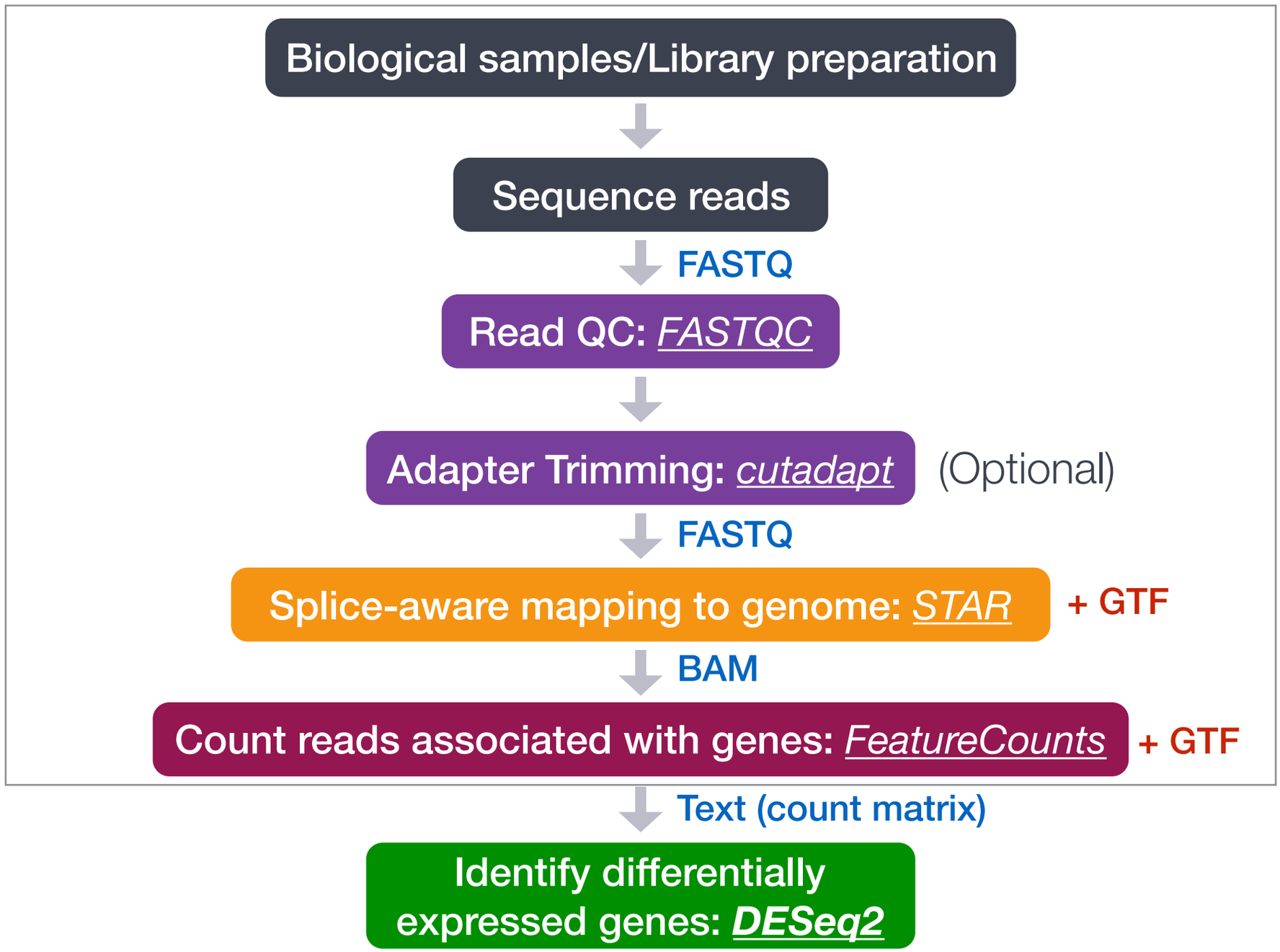
- The chromosome (or contig) names in feature format files MUST match the reference sequence name
  - Tied to a specific version (assembly/release) of a reference genome
  - Not all reference genomes are the represented the same!
  - E.g. human chromosome 1
    - **UCSC – ‘chr1’** versus **Ensembl/NCBI – ‘1’**

- Best practice: get feature format files from the same source (i.e UCSC, Ensembl, NCBI) as the reference genome

Commonly used file formats

- FASTA
- FASTQ – Fasta with quality
- GFF3 – Gene feature format (genome interval ++)
- GTF – Gene transfer format (genome interval ++)
- SAM – Sequence Alignment/Map format
- BAM – Binary Sequence Alignment/Map format
- *Bed – Basic genome interval (0-based coordinates)*
- *Wiggle (wig, bigwig) – tab-limited format to represent values, usually associated with a set of genomic coordinates (0-based coordinates)*

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



RNA-seq workflow