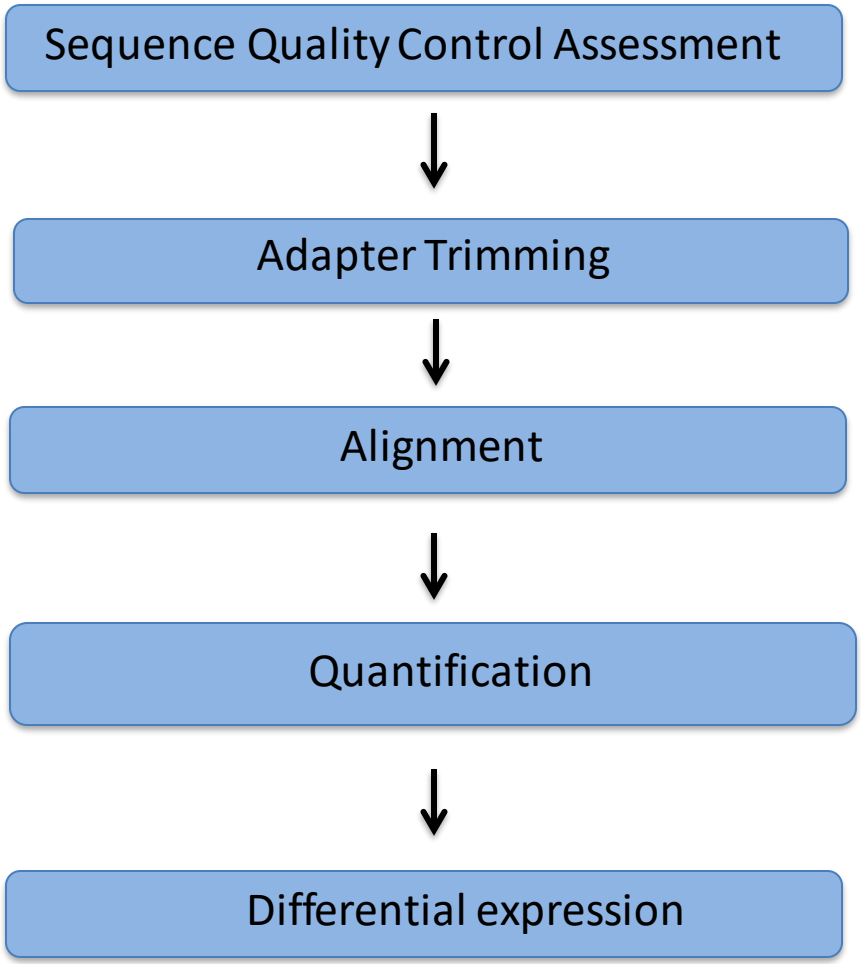
RNA-Seq Data Analysis

UND Genomics Core

RNA-Seq Analysis Pipeline



Software For RNA-Seq Analysis

Step	Software Option	
Sequence Quality Asesement	FastQC	
Adapter Trimming	Trim_galore	FastX
	Cutadapt	Trimmomatic
	Scythe	
Alignment	Hisat2	
	TopHat	
	STAR	
Quantification	FeatureCounts	Stringtie
	HTSeq-Count	Cufflinks
Differential Expression	DESeq2	Ballgown
	edgeR	CuffDiff
	DEXSeq	NOISeq

Fastq files

- Contains sequence and quality information

```
@HWI-D00635:65:C7U1DANXX:7:1101:1448:1950 1:N:0:GCCAAT
NCCATTTGTTTGATATTTTTCTAGAGCAGTAATGTTAAGAAAAAGGTATCT
+
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

HWI-D00635	Machine id
65	Run id
C7U1DANXX	Flow cell id
7	Lane number
1101	Tile number
1448	X coord
1950	Y coord
1	1st in pair
N	Not filtered
0	Control bit
GCAAT	index

Q-score

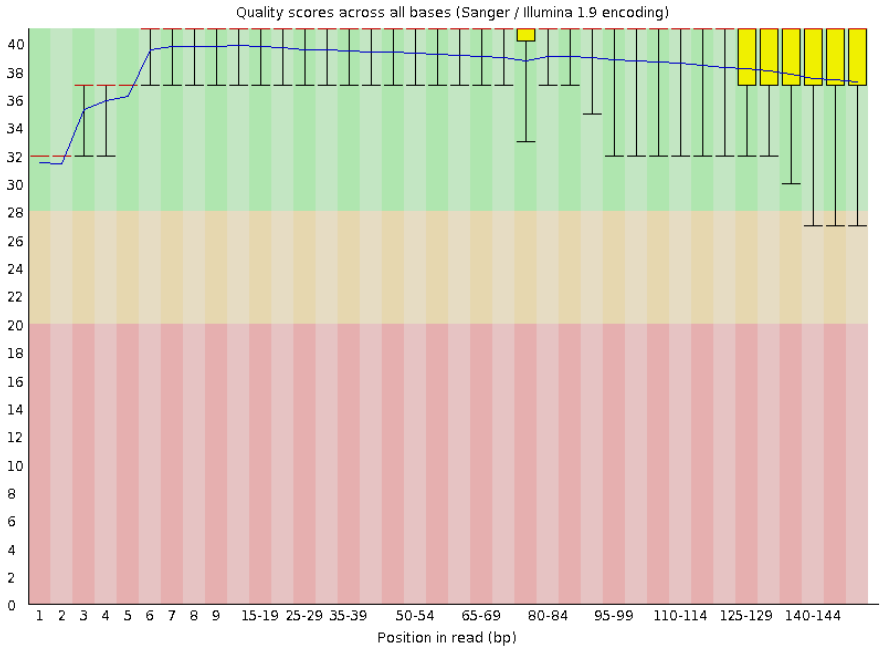
- Q-score is a metric to assess the accuracy of sequencing
- Relates to the probability of a wrong base call via logarithmic function.
- $-10\log_{10}(P)$

Q-score	Error rate	Accuracy
40	1/10,000	99.99%
30	1/1000	99.9%
20	1/100	99%
10	1/10	90%

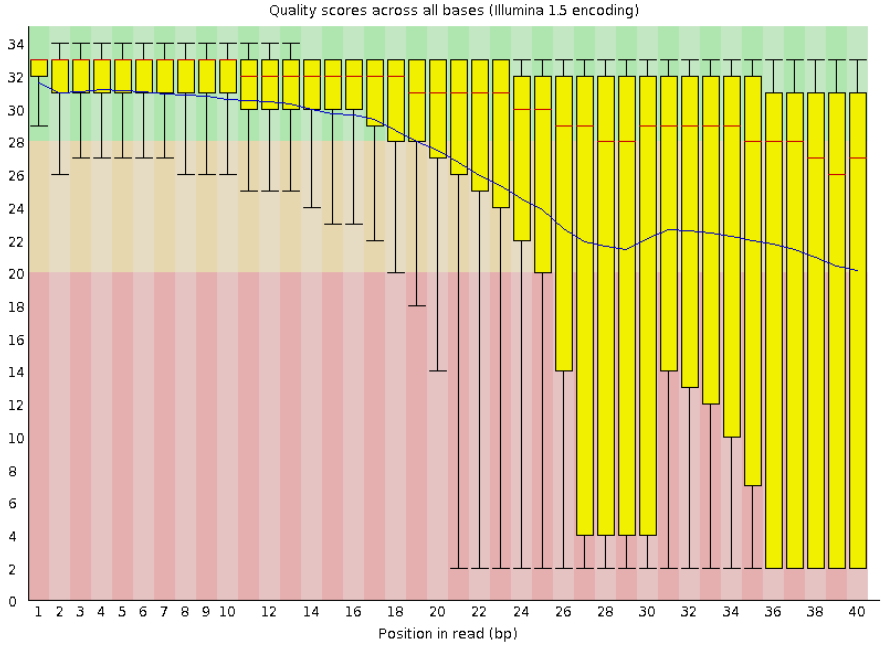
Sequence Quality Assessment

```
Fastqc -o FastQC *fastq.gz
```

Good data

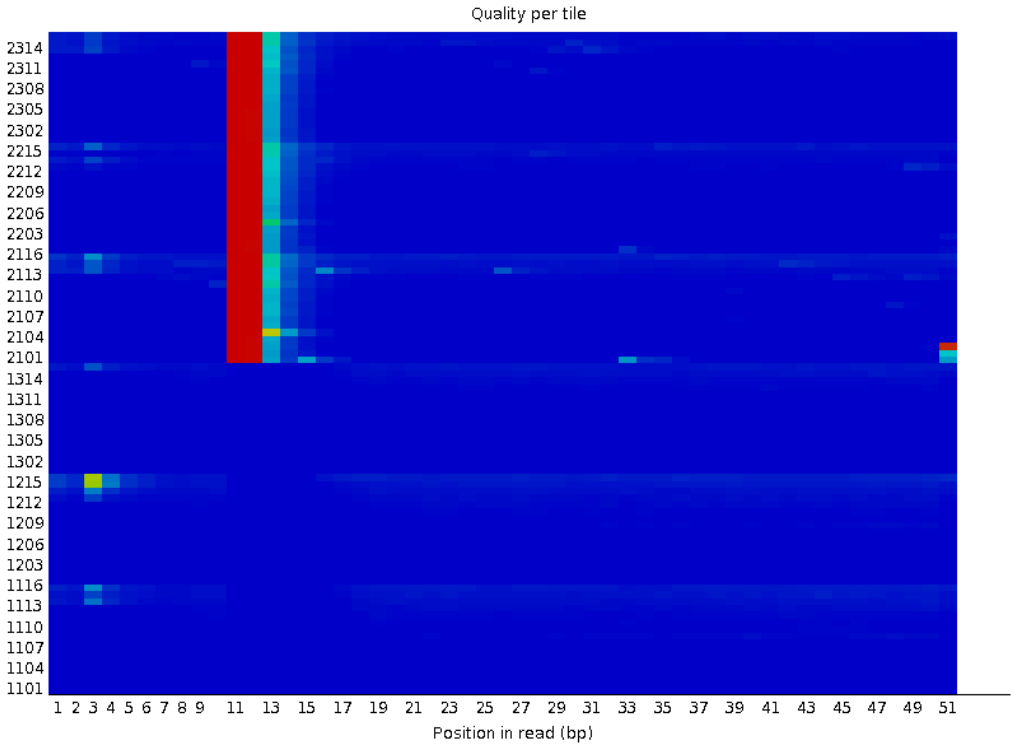


Bad data



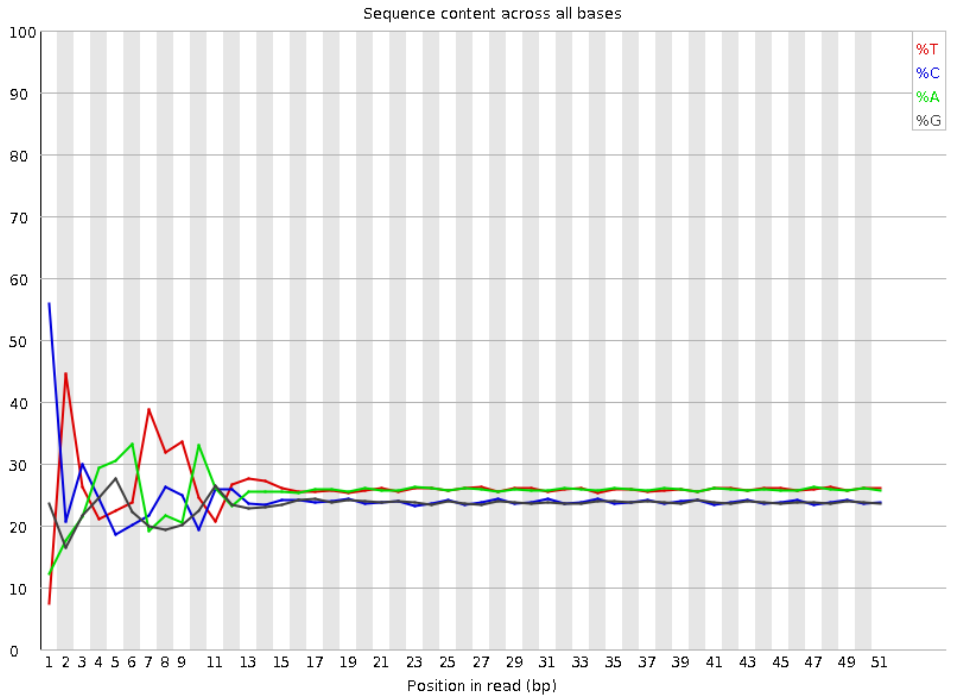
Sequence Quality Assessement

- Q-scores based with respect to the location on flow cell
- Ideally should be all blue

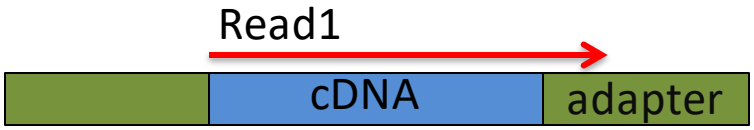


Sequence Quality Assessement

- Common for RNA-Seq
- Not all FastQC warnings apply



Adapter contamination

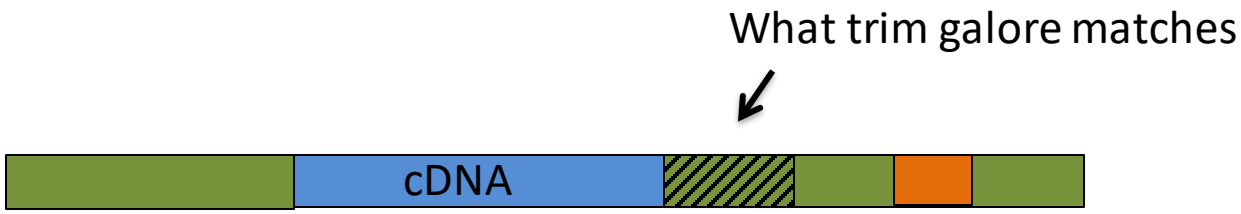


Overrepresented sequences

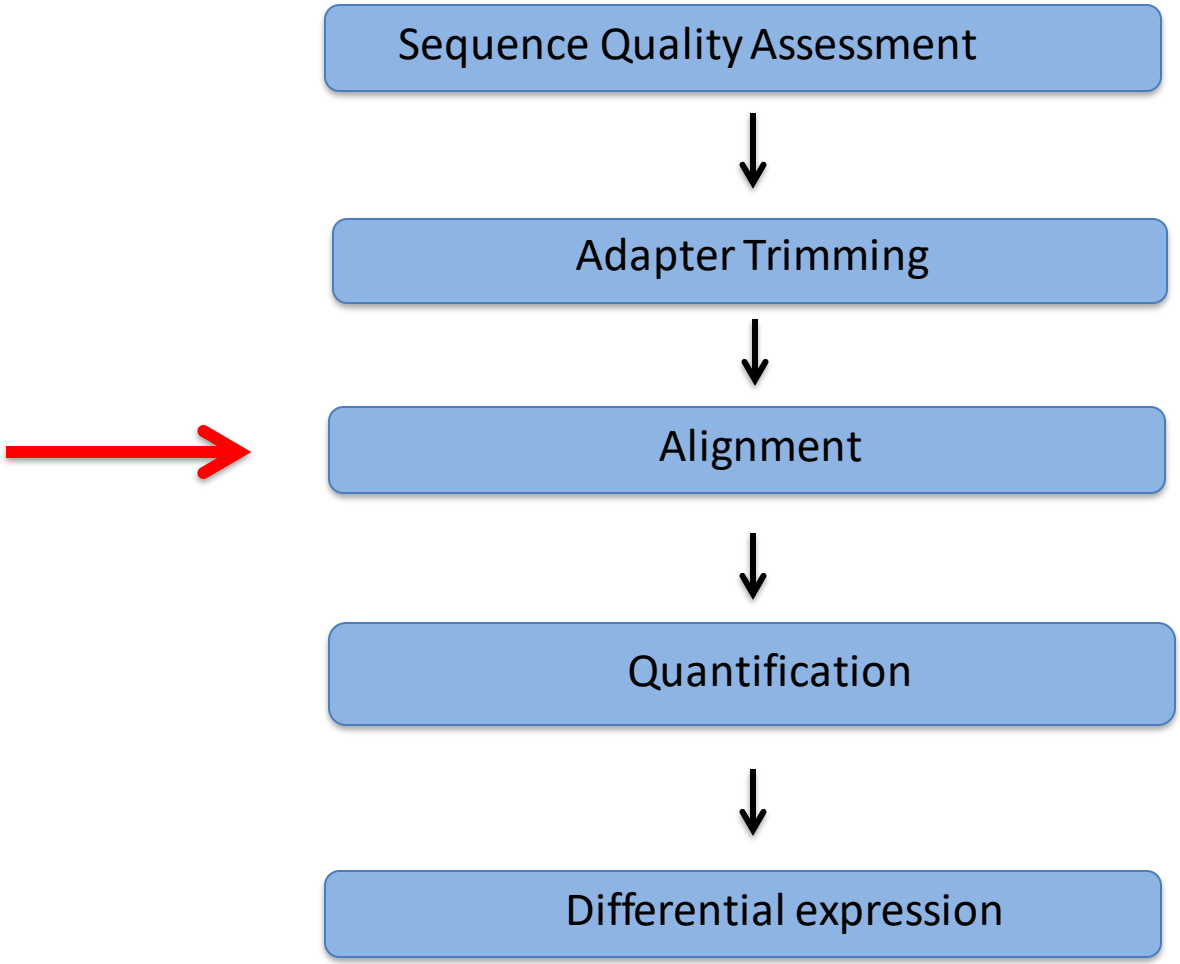
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGC	202000	4.795255640538142	TruSeq Adapter, Index 7 (100% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATATCGTATGC	5795	0.13756686354910164	TruSeq Adapter, Index 7 (98% over 50bp)

Adapter Trimming

- Trim galore
- Tries to automatically detect adapter
- For Illumina adapters, the first 13 bases of the Illumina indexed adapters

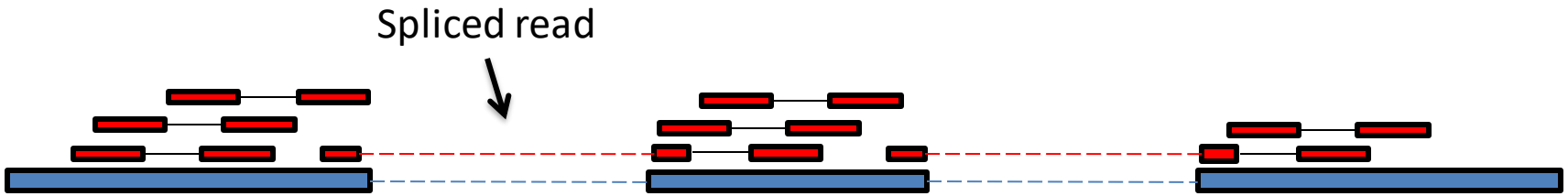


RNA-Seq Analysis Pipeline



Alignment for RNA-Seq

- For eukaryotic genomes, splice- aware aligners can align reads across exon – intron boundaries.



Splice – Aware Alignment Programs

- Tophat
  - https://ccb.jhu.edu/software/tophat/index.shtml
- Hisat2
  - https://ccb.jhu.edu/software/hisat2/manual.shtml
- STAR
  - https://github.com/alexdobin/STAR

Non Splice-aware Aligners

- Bowtie
  - http://bowtie-bio.sourceforge.net/index.shtml
- Bowtie2
  - http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- BWA
  - http://bio-bwa.sourceforge.net/

Reference Files for Alignment

- fasta
- gtf (gene transfer format)

Where to find reference files?

- UCSC
  - http://hgdownload.cse.ucsc.edu/downloads.html
- Ensembl
  - https://uswest.ensembl.org/downloads.html)
- iGenomes
  - https://support.illumina.com/sequencing/sequencing_software/igenome.html

More File Formats

- Fasta

```
>chrM
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTG
GAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATT
CTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACCTACTA
AAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAAT
GTCTGCACAGCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCCTCCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGC
```

- Indexes (*ht.1)
  - Binary version of the genome which allows for quick reading of the genome

GTF file

```
chr1    unknown exon    11874   12227   .       +       .       gene_id "DDX11L1"; gene_name "DDX11L1"; transcript_ic
chr1    unknown exon    12613   12721   .       +       .       gene_id "DDX11L1"; gene_name "DDX11L1"; transcript_ic
chr1    unknown exon    13221   14409   .       +       .       gene_id "DDX11L1"; gene_name "DDX11L1"; transcript_ic
chr1    unknown exon    14362   14829   .       -       .       gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "
chr1    unknown exon    14970   15038   .       -       .       gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "
chr1    unknown exon    15796   15947   .       -       .       gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "
chr1    unknown exon    16607   16765   .       -       .       gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "
chr1    unknown exon    16858   17055   .       -       .       gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "
chr1    unknown exon    17233   17368   .       -       .       gene_id "WASH7P"; gene_name "WASH7P"; transcript_id "
chr1    unknown exon    17369   17436   .       -       .       gene_id "MIR6859-2"; gene_name "MIR6859-2"; transcrip
```

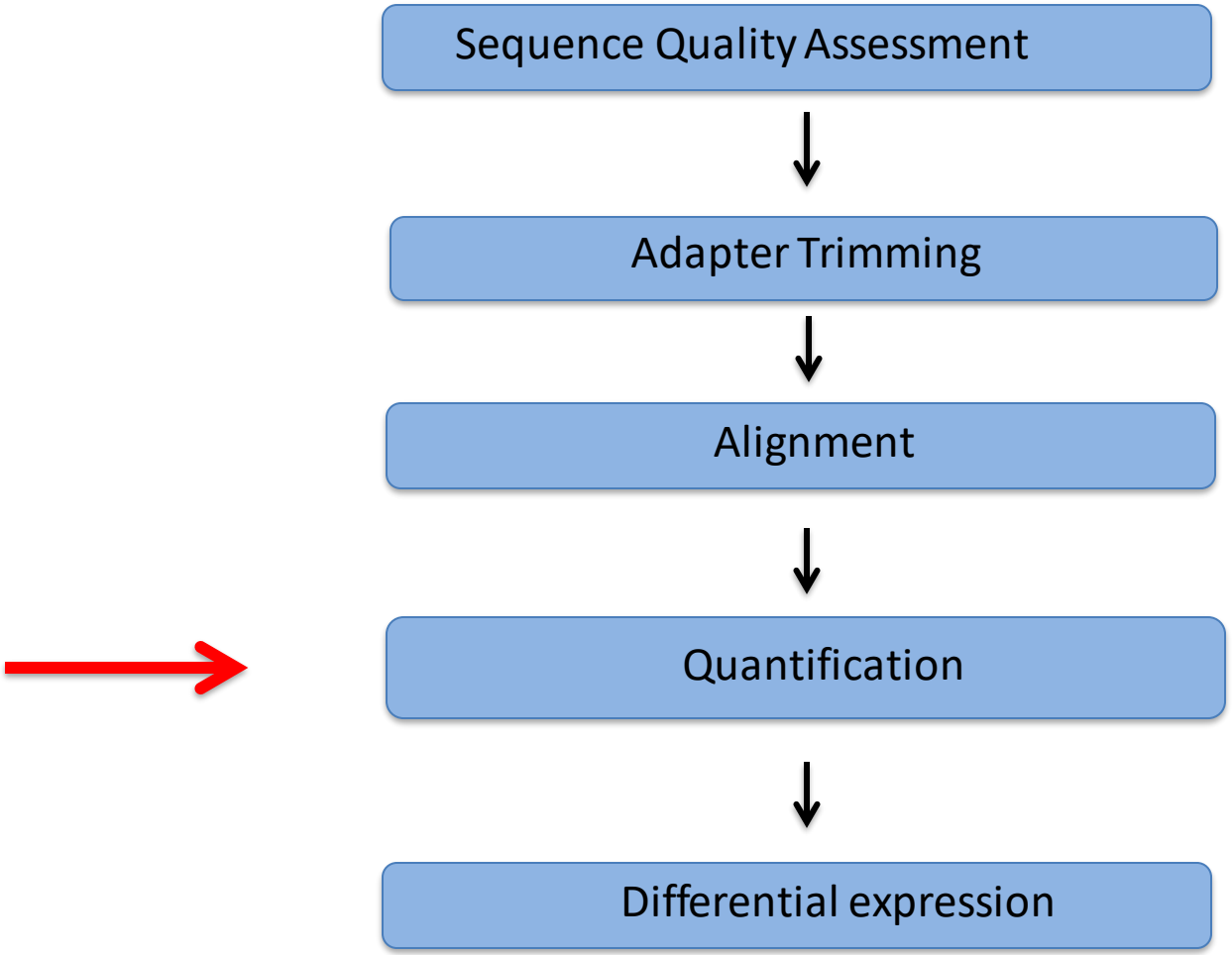
Column	Description
seqname	chromosome
source	source of annotation
feature	type of feature, ex exon
start	start of feature
end	end of feature
score	Confidence of assembled transcript
strand	strand of feature
frame	frame of feature relative to start of coding sequence
attributes	names of feature

Alignment Output

- Sam (Sequence alignment format)
- Bam (binary sam file)

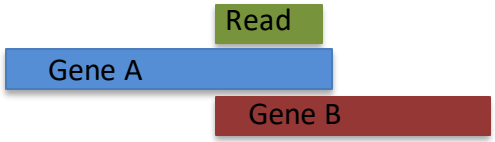
```
[danielle.perley@buddy Alignments]$ samtools view -H 17-110-002.bam
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr10        LN:130694993
@SQ     SN:chr11        LN:122082543
@SQ     SN:chr12        LN:120129022
@SQ     SN:chr13        LN:120421639
@SQ     SN:chr14        LN:124902244
@SQ     SN:chr15        LN:104043685
@SQ     SN:chr16        LN:98207768
@SQ     SN:chr17        LN:94987271
@SQ     SN:chr18        LN:90702639
@SQ     SN:chr19        LN:61431566
@SQ     SN:chr1 LN:195471971
@SQ     SN:chr2 LN:182113224
@SQ     SN:chr3 LN:160039680
@SQ     SN:chr4 LN:156508116
@SQ     SN:chr5 LN:151834684
@SQ     SN:chr6 LN:149736546
@SQ     SN:chr7 LN:145441459
@SQ     SN:chr8 LN:129401213
@SQ     SN:chr9 LN:124595110
@SQ     SN:chrM LN:16299
@SQ     SN:chrX LN:171031299
@SQ     SN:chrY LN:91744698
@PG     ID:hisat2       PN:hisat2       VN:2.0.5        CL:"/usr/share/pathing/../hisat2-2.0.5/hisat2-align-s --wrapper basic-0 -x genome -p 14 --no-spliced-alignment -I 10 -X 2000 -1 /tmp/22750.inpipe1 -
2 /tmp/22750.inpipe2"
```

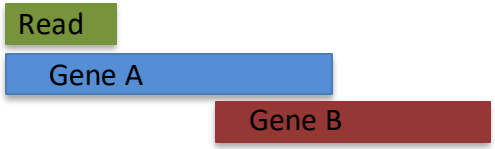
RNA-Seq Analysis Pipeline



Counting Reads

- Count reads that unambiguously align to a gene
- Programs that will count reads:
  - FeatureCounts
  - HTSeq





Transcriptome assembly

- Assemble a map of the genes in the transcriptome using the reads in present sample
- Can be guided with a gtf, if using a well –annotated genome
- Assembles transcriptome and estimate the abundance of transcript at the same time
- Programs that will assemble transcriptomes:
  - Cufflinks
  - Stringtie

- Your Turn!
- Work through a RNA-Seq analysis described in the file Day2_HandsOn.docx in the Day2 workshop material folder