

Short Read Mapping

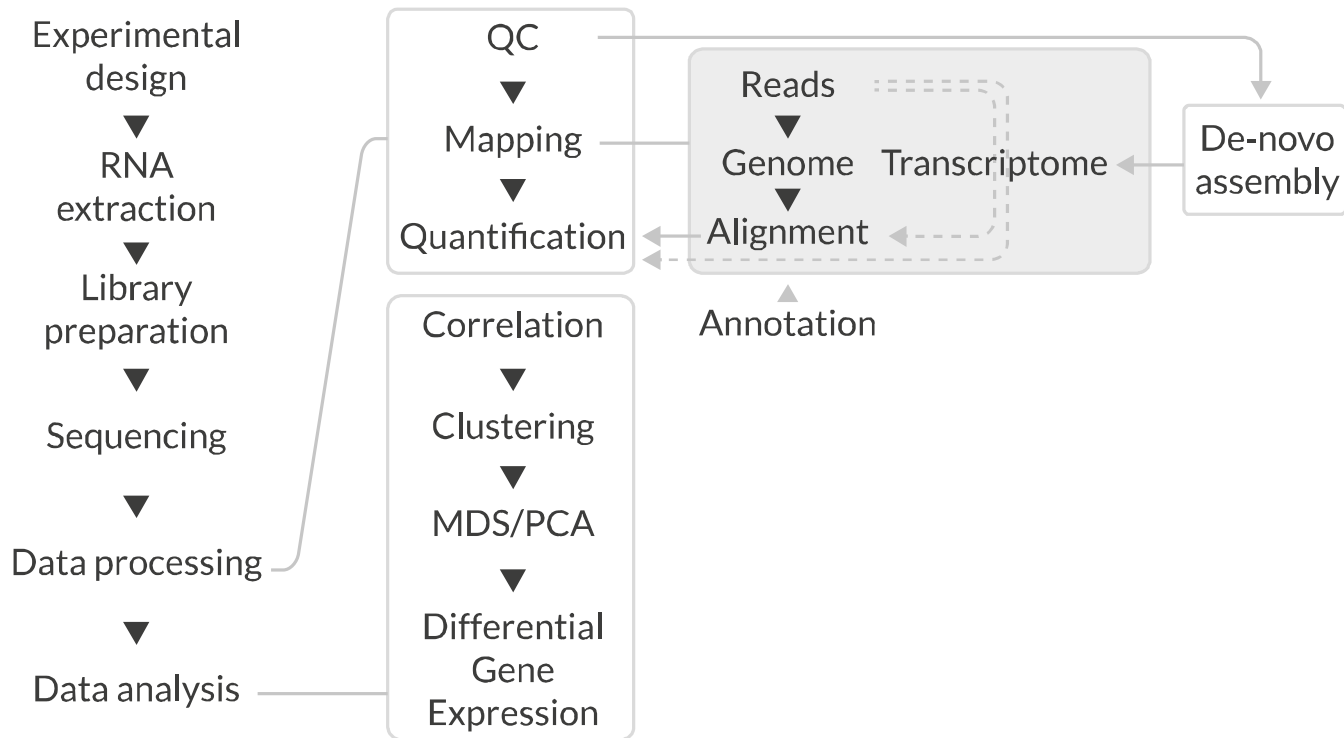
RNA-Seq Analysis Workshop

Roy Francis | 28-Oct-2018

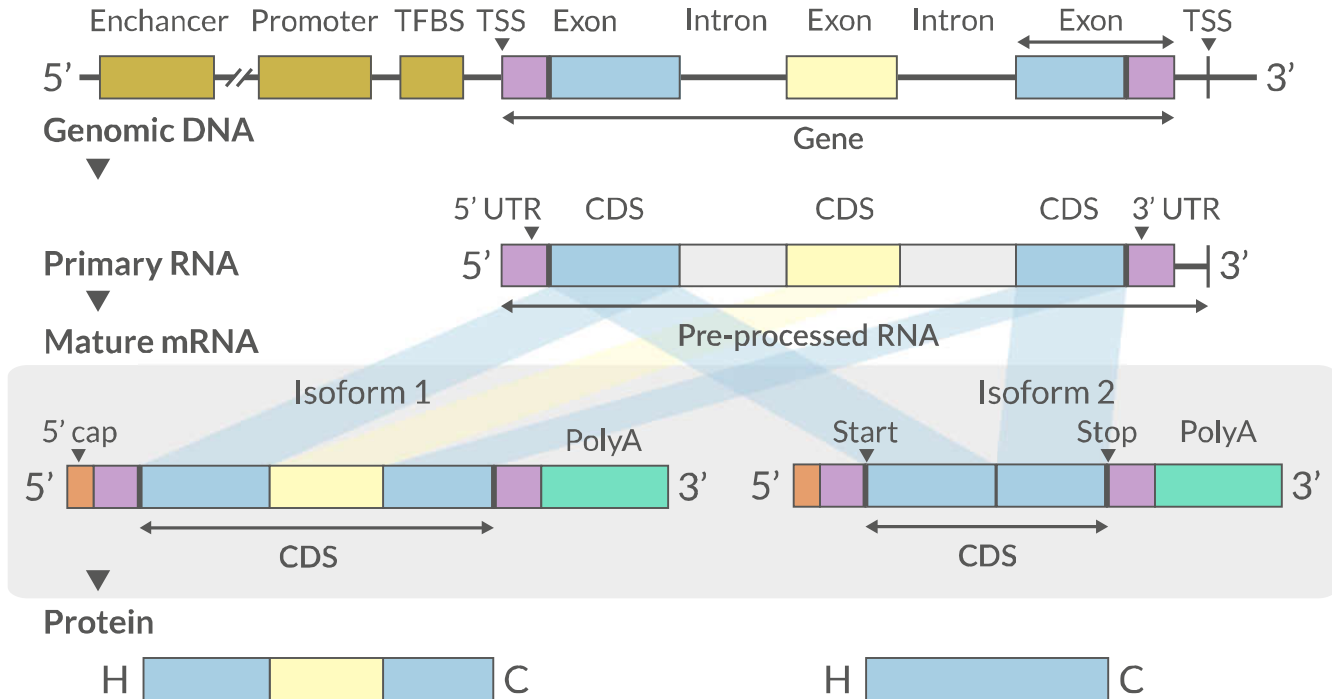
Contents

- Workflow
- Mapping
- Aligners
- Alignment
- Visualisation
- Quantification
- Long reads
- Summary

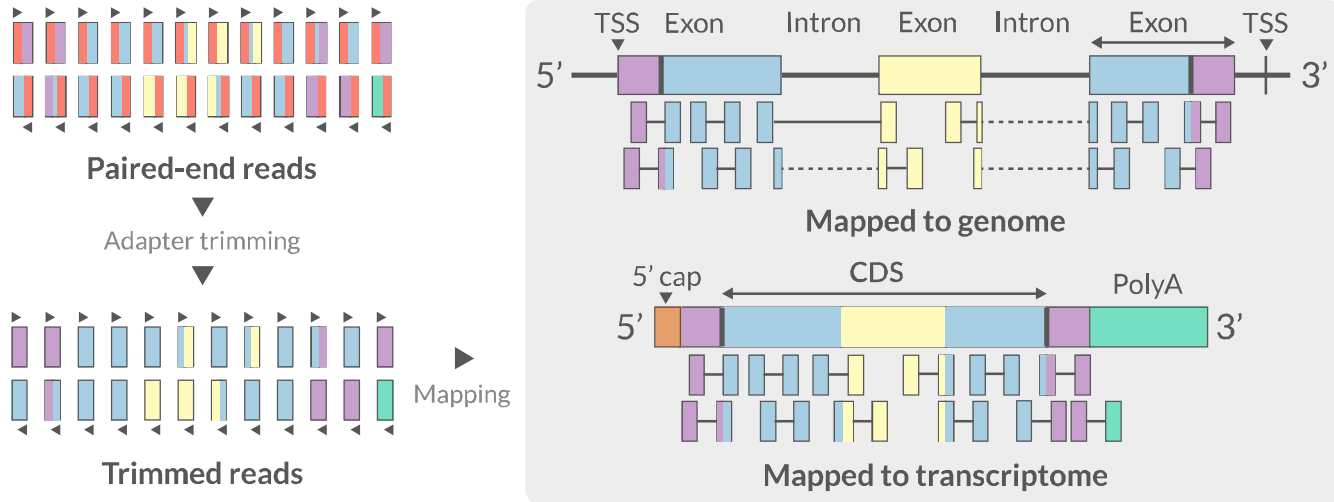
Workflow



Mapping



Mapping



- Aligning reads back to a reference sequence
- Mapping to genome vs transcriptome
- Splice-aware alignment (genome)

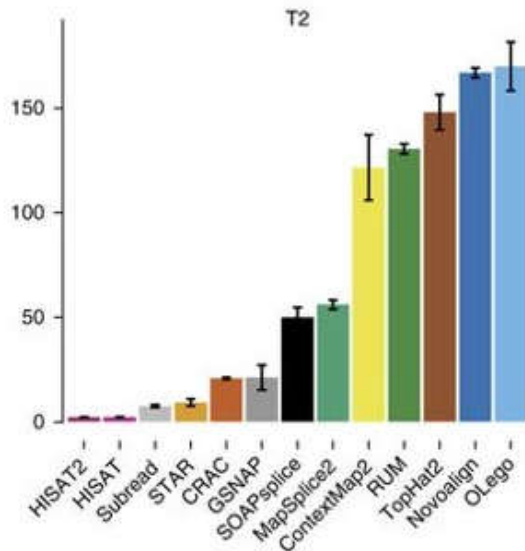
Considerations

- Speed
- Accuracy
- Resources
- Settings
- Purpose (General/Specific)
- Support & Community

Features

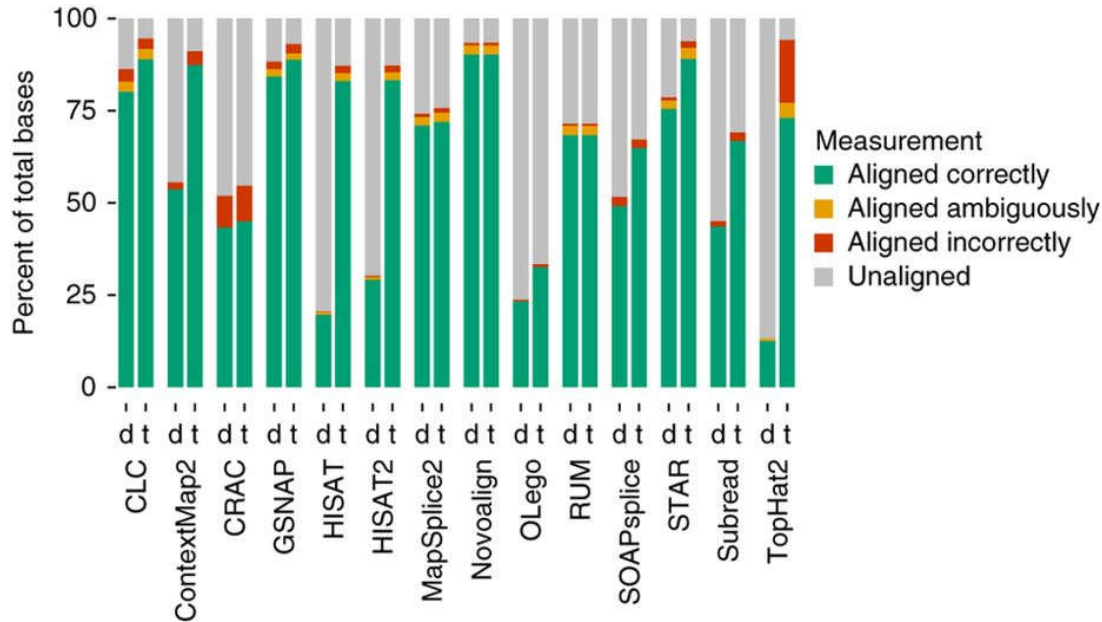
- Reference index
- Read pair alignment
- Consider base quality scores
- Sophisticated indexing to decrease CPU and memory usage
- Resolving multi-mappers
 - Report first X alignments and flag read as multi-mapping
- Use known annotations (junctions)
- 2-pass approach

Aligners | Speed



| Program | Time_Min | Memory_GB |
|---------|----------|-----------|
| HISATx1 | 22.7 | 4.3 |
| HISATx2 | 47.7 | 4.3 |
| HISAT | 26.7 | 4.3 |
| STAR | 25 | 28 |
| STARx2 | 50.5 | 28 |
| GSNAP | 291.9 | 20.2 |
| TopHat2 | 1170 | 4.3 |

Aligners | Accuracy



Increasing Accuracy

- Novel variants / RNA editing
- Allele-specific expression
- Genome annotation
- Gene and transcript discovery
- Differential expression

 **STAR, HiSat2, GSNAP, Novoalign** (Commercial)

- Reads (FASTQ)

```
@ST-E00274:179:HHYMLALXX:8:1101:1641:1309 1:N:0:NGATGT
NCATCGTGGTATTTGCACATCTTTTCTTATCAAATAAAAAGTTTAACTACTCAGTTATGCGCATACGTTTTTTGATGG
+
#AAAFafa<-AFFJJJafa-FFJJJJFFFAJJJJ-<FFJJJ-A-F-7--FA7F7-----FFFJfa<FFFFJ<AJ--FF-
```

```
@instrument:runid:flowcellid:lane:tile:xpos:ypos
read:isfiltered:controlnumber:sampleid
```

- Reference Genome/Transcriptome (FASTA)

```
>1 dna:chromosome chromosome:GRCz10:1:1:58871917:1 REF
GATCTTAAACATTTATTCCCCCTGCAAACATTTTCAATCATTACATTGTCATTTCCCTC
CAAATTAATTTAGCCAGAGGCGCACAAACATACGACCTCTAAAAAAGGTGCTGTAACATG
```

- Annotation (GTF/GFF)

```
#!/genome-build GRCz10
#!/genebuild-last-updated 2016-11
4      ensembl_havana  gene      6732      52059      .      -      .      gene_id "ENS
```

```
chr source feature start end score strand frame attribute
```

Alignment

- SAM/BAM (Sequence Alignment Map format)

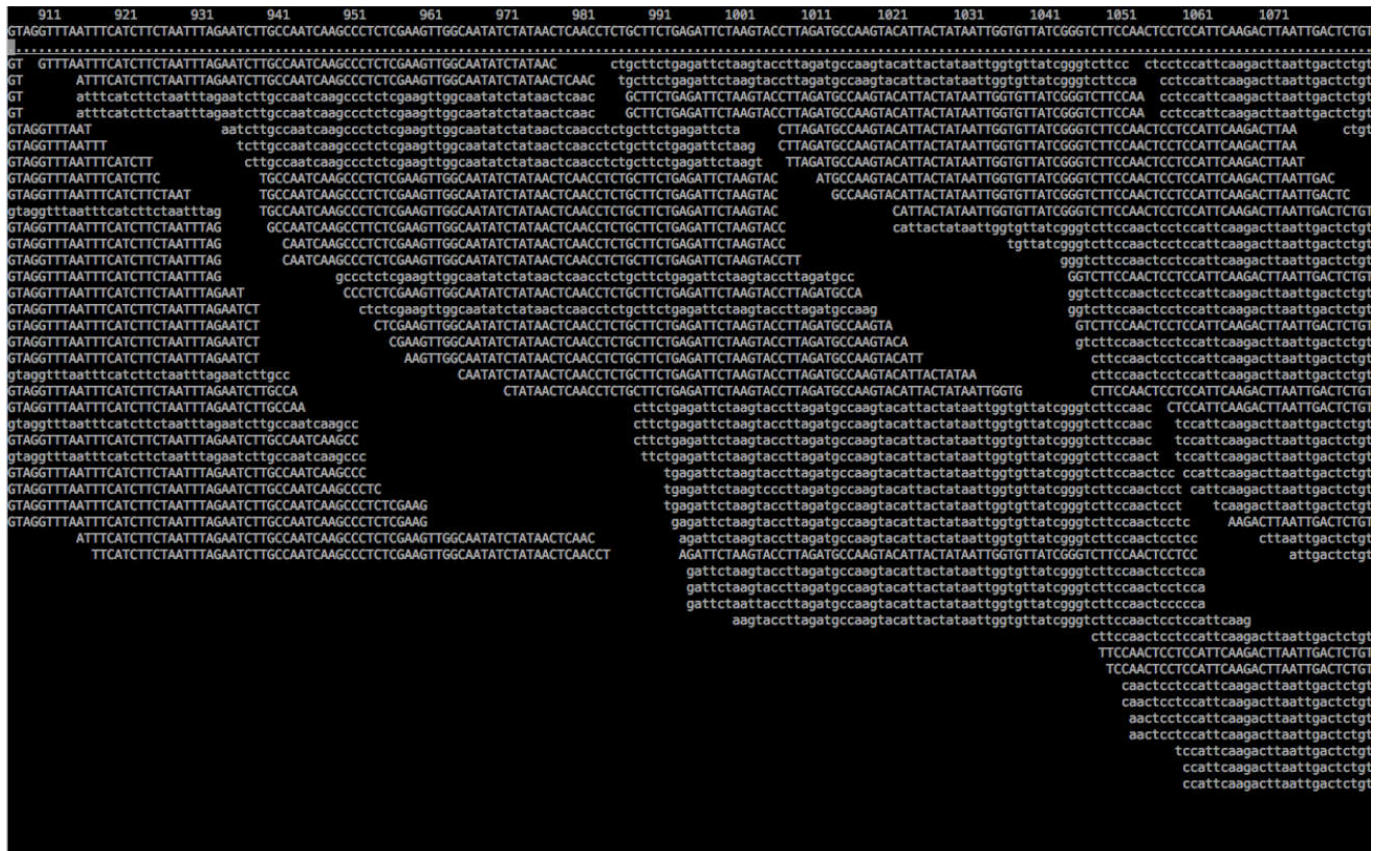
```
ST-E00274:188:H3JWNCCXY:4:1102:32431:49900 163 1 1 60 8S13
```

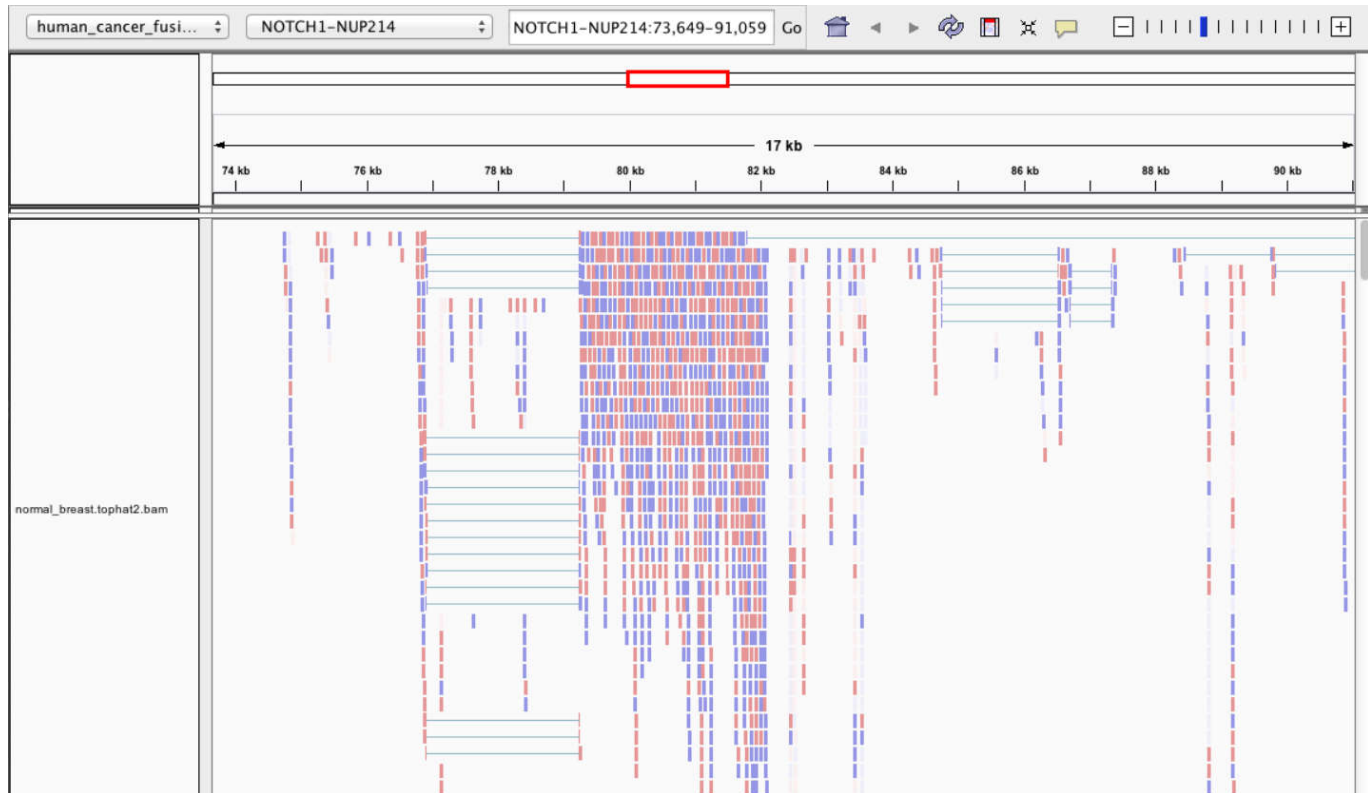
```
query flag ref pos mapq cigar mrnm mpos tlen seq qual opt
```

Alignment formats

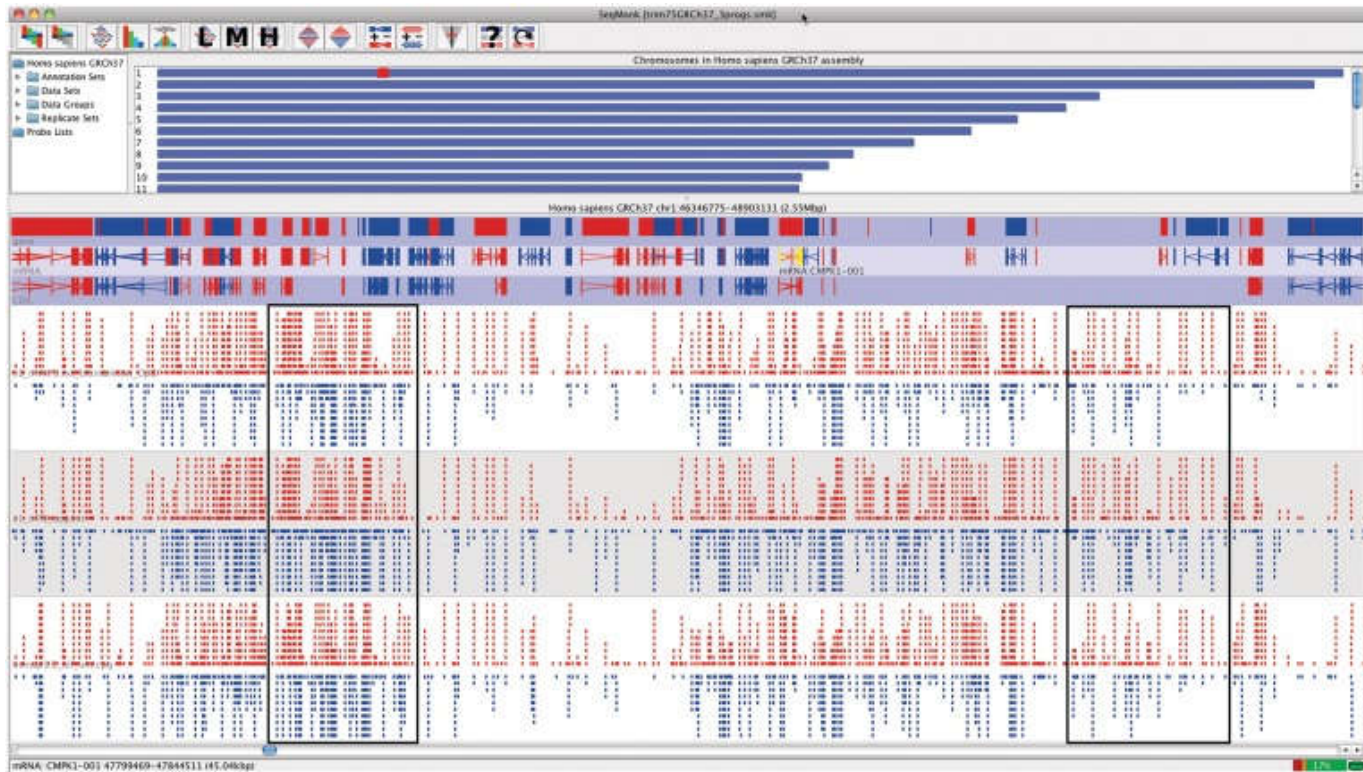
| Format | Size_GB |
|-----------------|---------|
| SAM | 7.4 |
| BAM | 1.9 |
| CRAM lossless Q | 1.4 |
| CRAM 8 bins Q | 0.8 |
| CRAM no Q | 0.26 |

samtools tview alignment.bam genome.fasta



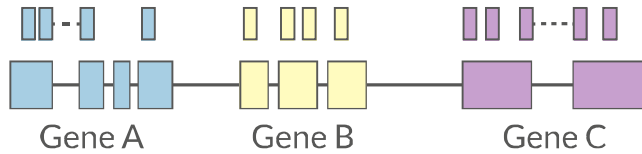


 IGV, UCSC Genome Browser



Quantification | Counts

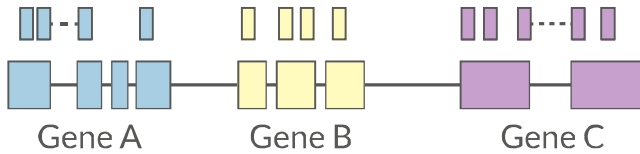
- Read counts = gene expression
- Reads can be quantified on any feature (gene, transcript, exon etc)
- Intersection on gene models
- Gene/Transcript level



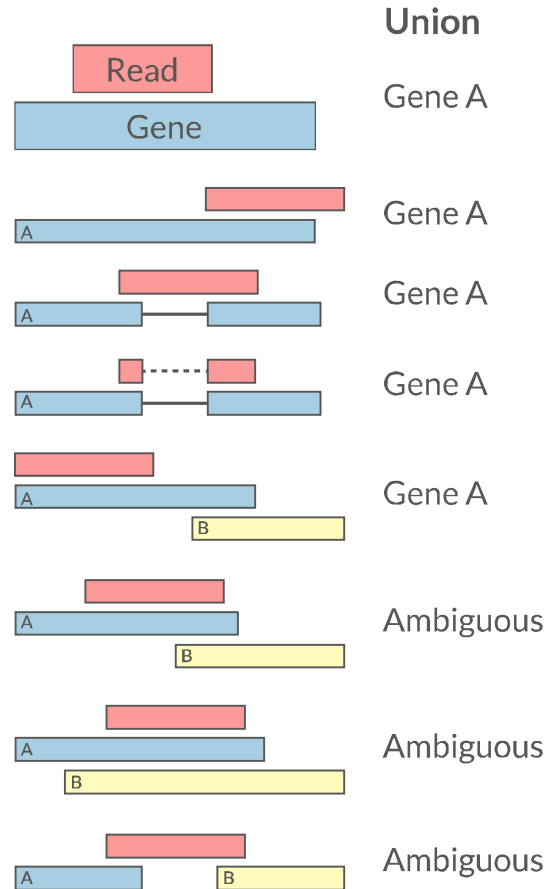
 featureCounts, HTSeq

Quantification | Counts

- Read counts = gene expression
- Reads can be quantified on any feature (gene, transcript, exon etc)
- Intersection on gene models
- Gene/Transcript level



 featureCounts, HTSeq



- Added (BEDTools multicov)
- Discard (featureCounts, HTSeq)
- Distribute counts (Cufflinks)
- Rescue
 - Probabilistic assignment (Rcount, Cufflinks)
 - Prioritise features (Rcount)
 - Probabilistic assignment with EM (RSEM)

🔗 Fu, Yu, *et al.* "Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers." *BMC genomics* 19.1 (2018): 531

🔗 Parekh, Swati, *et al.* "The impact of amplification on differential expression analyses by RNA-seq." *Scientific reports* 6 (2016): 25533

🔗 Klepikova, Anna V., *et al.* "Effect of method of deduplication on estimation of differential gene expression using RNA-seq." *PeerJ* 5 (2017): e3091

- Count methods
 - Provide no inference on isoforms
 - Cannot accurately measure fold change
- Probabilistic assignment
 - Deconvolute ambiguous mappings
 - Transcript-level
 - cDNA reference

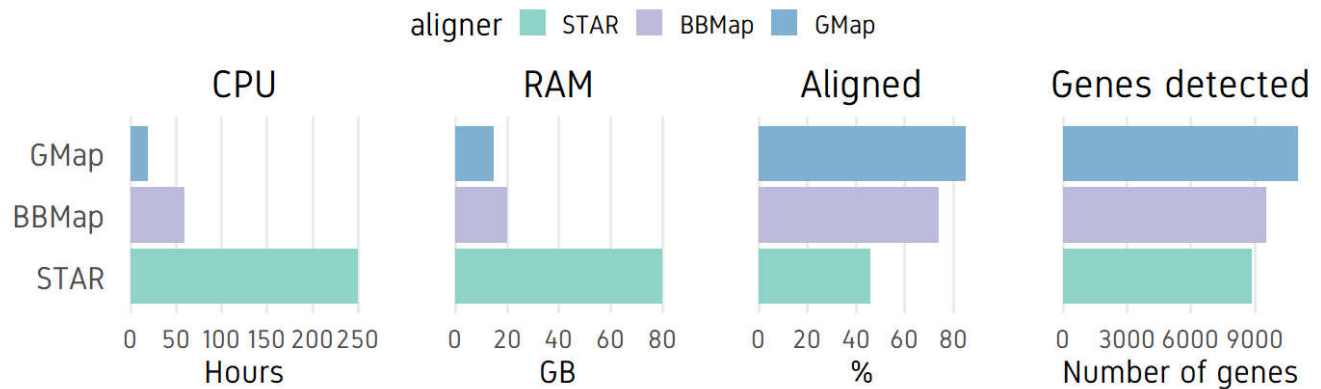
Kallisto, Salmon

- Direct from FastQ to counts
- Ultra-fast & alignment-free
- Uses transcriptome reference
- Subsampling & quantification confidence
- Transcript-level estimates improves gene-level estimates
- Kallisto/Salmon > transcript-counts > `tximport()` > gene-counts

RSEM, Kallisto, Salmon, Cufflinks2

Long-Read RNA-Seq

- PacBio, Nanopore etc
- Long reads, full transcripts
- High error rate
- Expensive



- Results are comparable with MinION data.


 [GMAP](#), [BMAP](#), [STAR](#)

Summary

- STAR, HISAT2 and GSNAP are good general purpose aligners
- Use HISAT2 if RAM is limited
- Consider using 2-pass mapping
- Be stringent with junction discovery criteria
- Map to genome for annotation/discovery
- For well known transcriptomes, Kallisto/Salmon offers ultra-fast quantification
- For long reads, GMAP and BMAP are good choice of aligners



Thank you! Questions?

Built on:  28-Oct-2018 at 15:46:27

2018 Roy Francis | [SciLifeLab](#) | [NBIS](#)