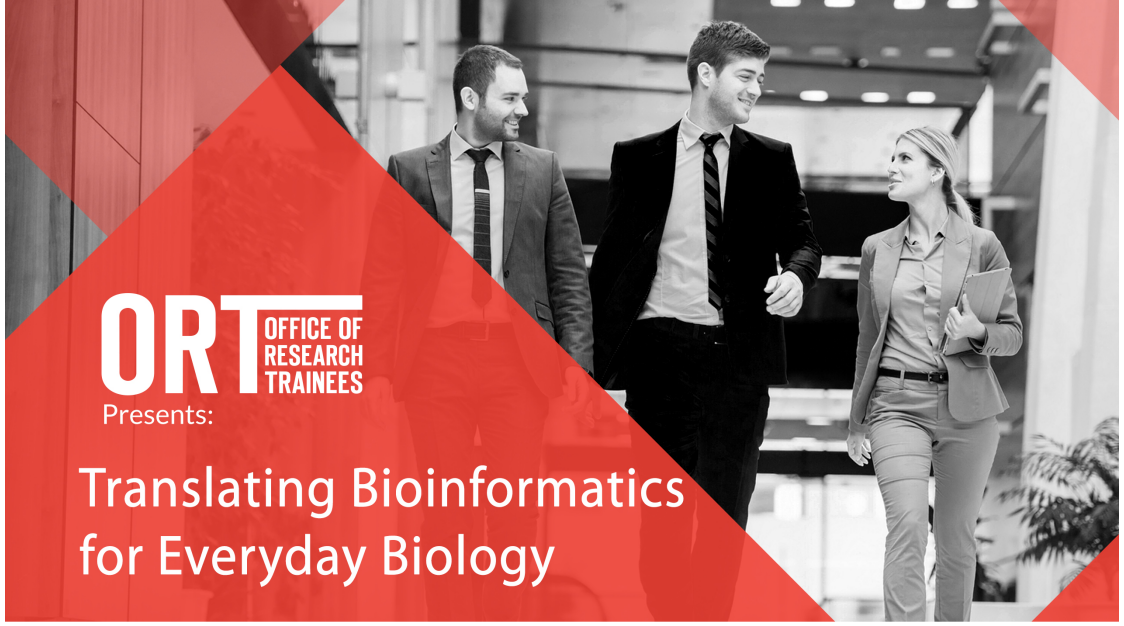


RNA-seq analysis

Musa Ahmed

Nov 15, 2019

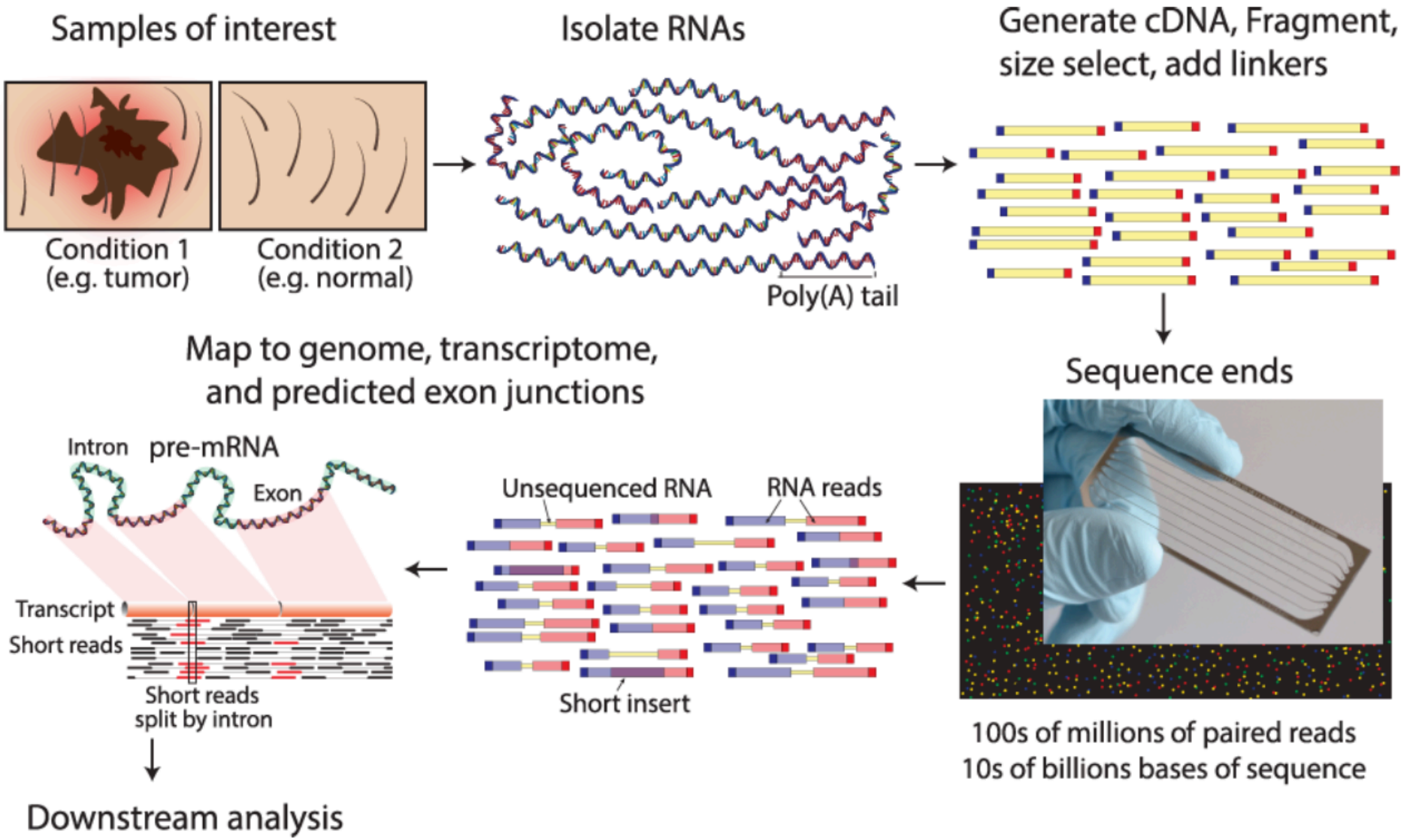
GOAL

- Basics of RNA-seq analysis
- Applications
- Challenges
- Practical
  - Alignment
  - DGE analysis

What is RNA-seq

- RNA-seq works by sequencing every RNA molecule and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced.

RNA sequencing



Design Experiment

1. Carefully design the experiment.

Purify RNA

2. Isolate and purify input RNA.

Prepare Libraries

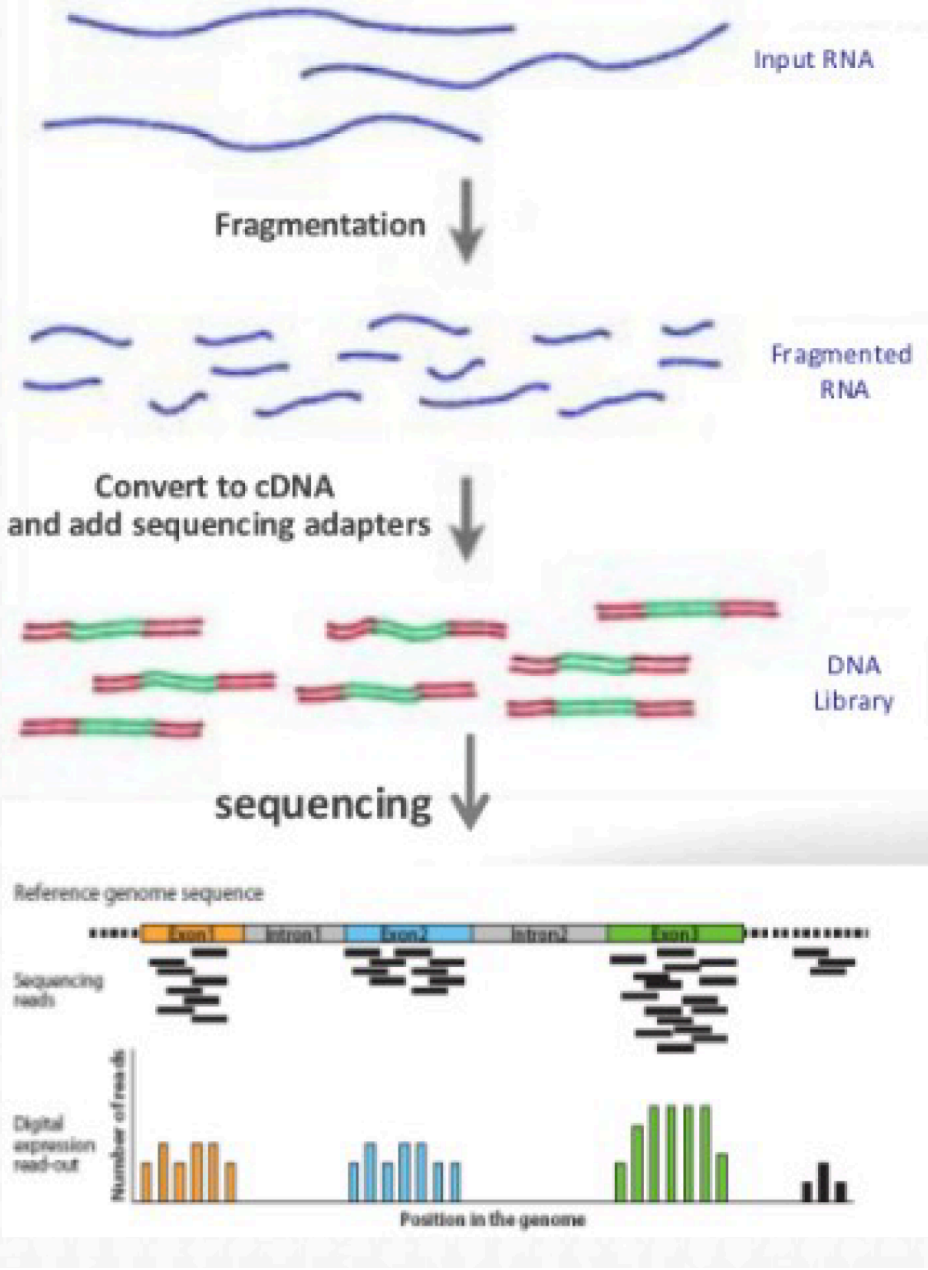
3. Convert the RNA to cDNA and add sequencing adapters.

Sequence

4. Sequence cDNAs using one of the available NGS platforms.

Analysis

5. Analyze the resulting short-read sequences.

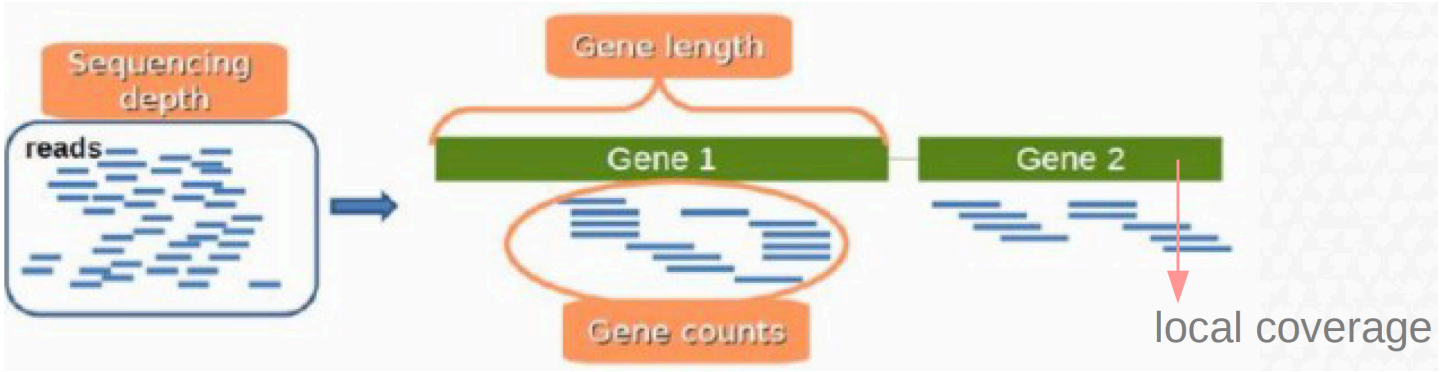


Why RNA-seq?

- Gene expression
- Splicing/isoforms
- de-novo assembly of transcriptome
- Allele-specific expression

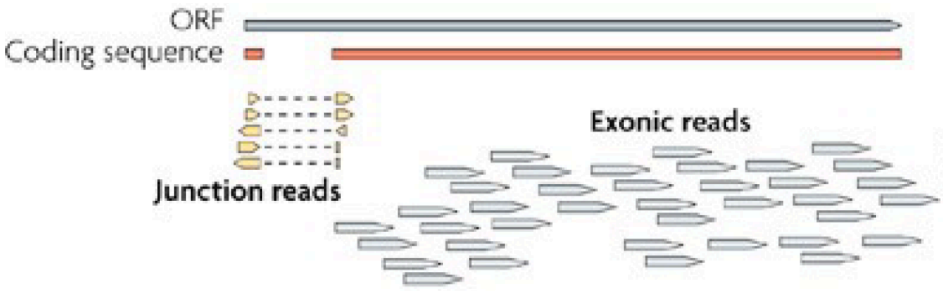
Key concepts

- **Sequencing depth**
  - Total number of reads mapped to the genome. (Library size) Could also be applied to samples.
- **Coverage**
  - Number of reads mapped to a specic region (average of them if we are talking about the whole genome...) . Not typically used for RNA-seq
- **Gene length**
  - Number of bases that a gene has.

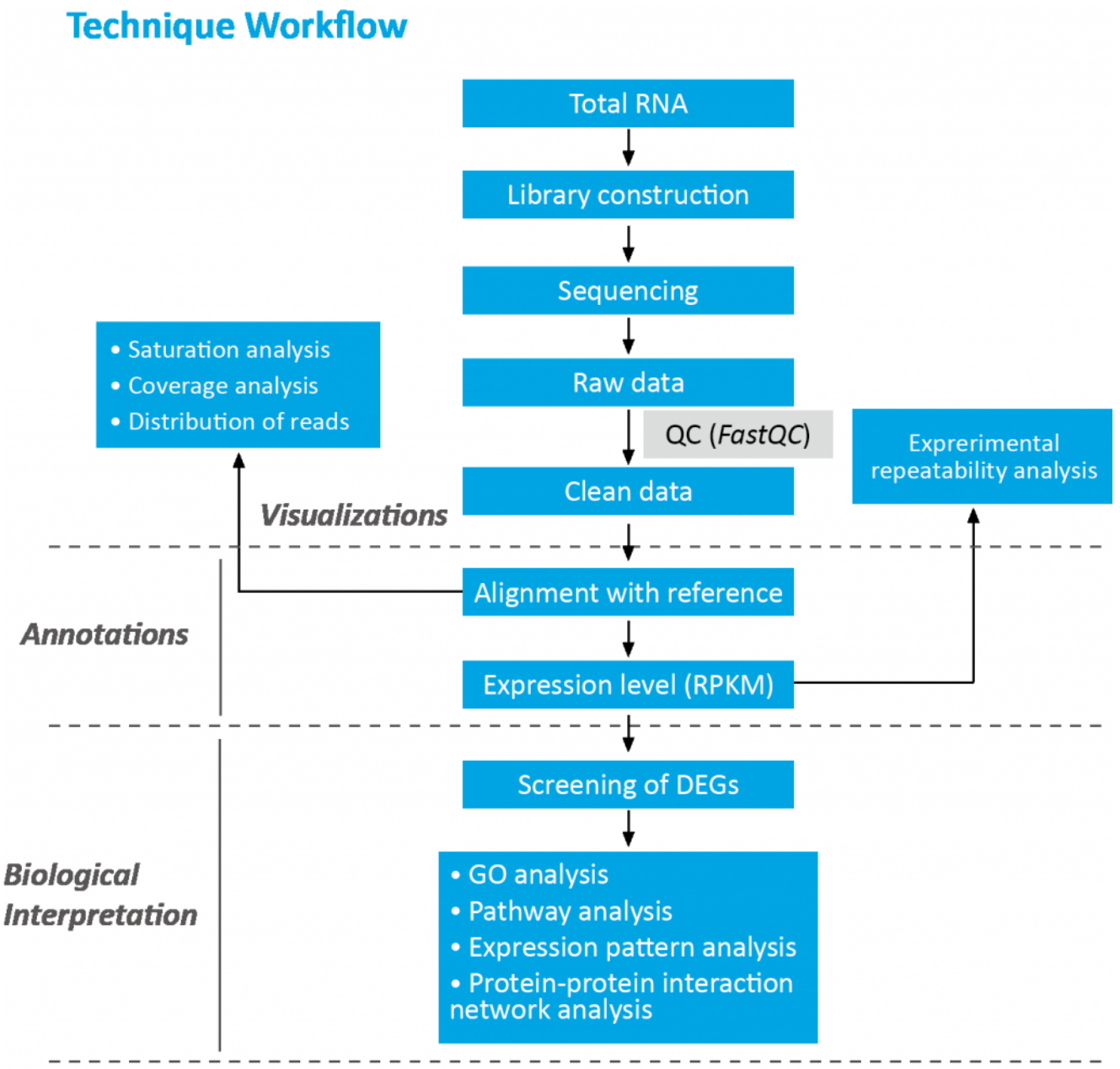


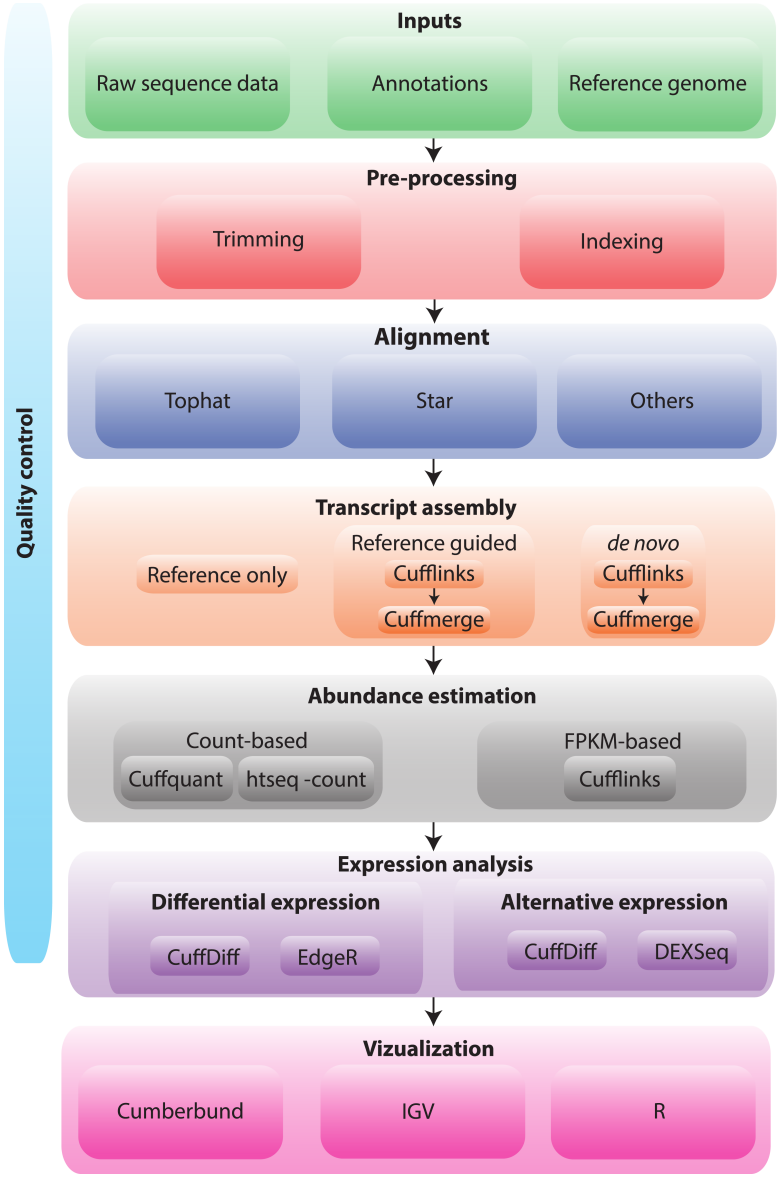
Key concepts

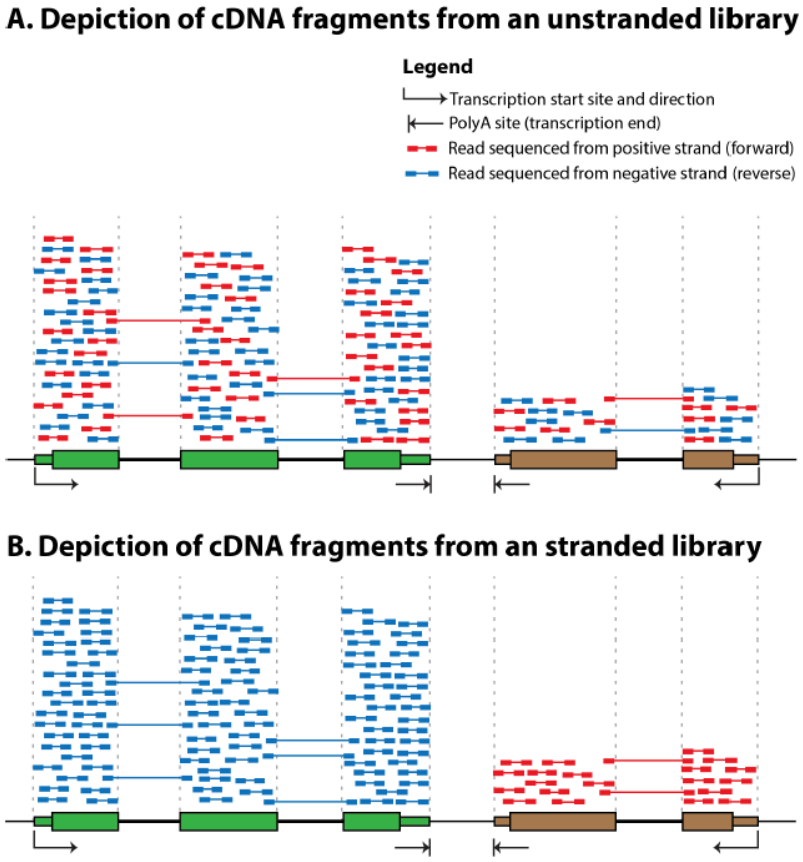
- **Exonic reads:** Reads within exons
- **Junction reads:** Reads spanning exon junctions



BGI workflow

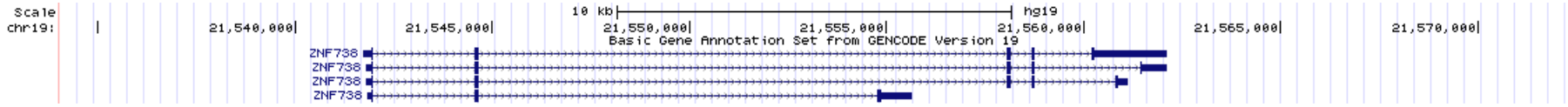



Quality control
Inputs
Raw sequence data
Annotations
Reference genome
Pre-processing
Trimming
Indexing
Alignment
Tophat
Star
Others
Transcript assembly
Reference only
Reference guided
Cufflinks
Cuffmerge
de novo
Cufflinks
Cuffmerge
Abundance estimation
Count-based
Cuffquant
htseq -count
FPKM-based
Cufflinks
Expression analysis
Differential expression
CuffDiff
EdgeR
Alternative expression
CuffDiff
DEXSeq
Vizualization
Cumberbund
IGV
R


A. Depiction of cDNA fragments from an unstranded library
Legend
Transcription start site and direction
PolyA site (transcription end)
Read sequenced from positive strand (forward)
Read sequenced from negative strand (reverse)
B. Depiction of cDNA fragments from an stranded library

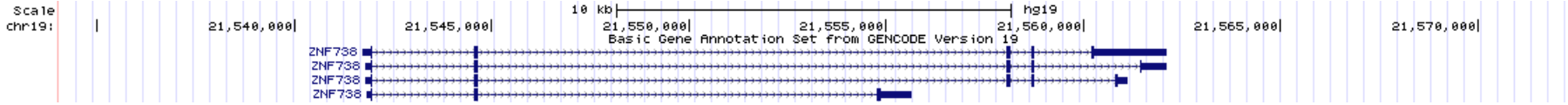
Reference genomes by Genome Reference Consortium

- Assembled whole genome using a bunch of individual genome sequences
  - Human – Hg38 (GRCh38), hg19 (GRCh37), b37 etc
  - Mouse – mm10, mm9 etc

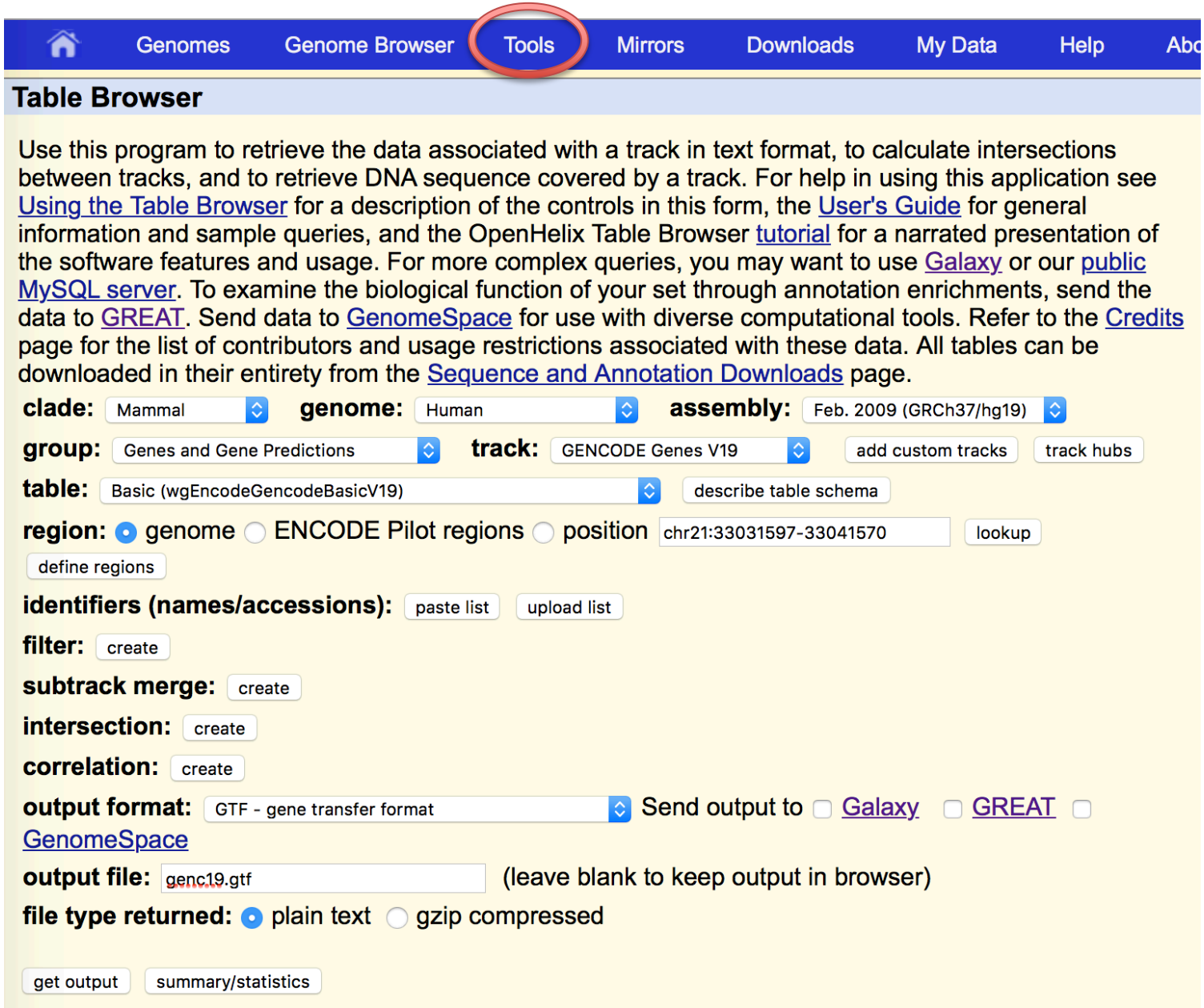


Gene annotations

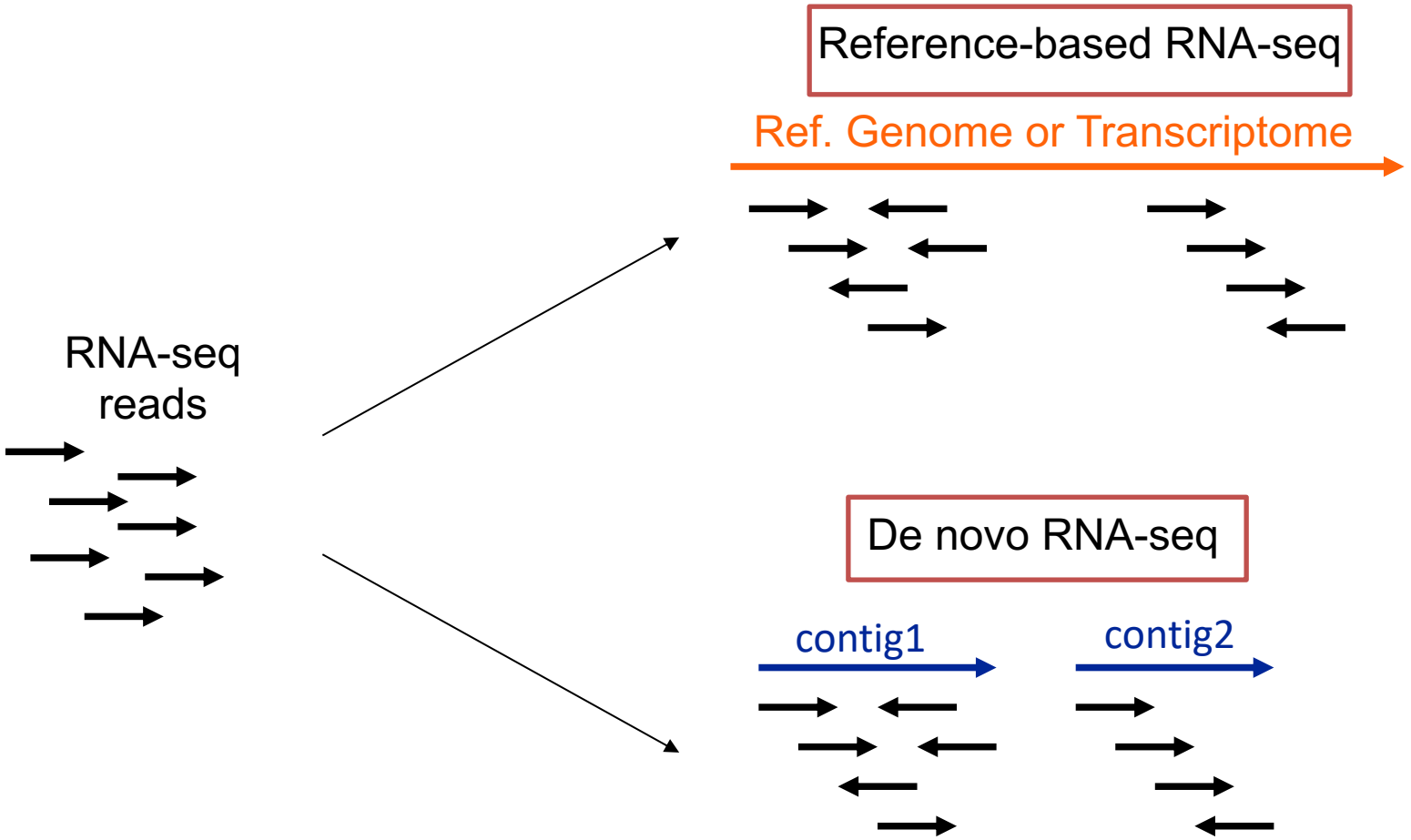
- Annotation of the whole genome (protein coding genes, noncoding RNAs etc)
  - RefSeq (good for general analysis)
  - GENCODE/ENSEMBL (good for noncoding genes)
  - miTranscriptome (good for noncoding genes, suitable for really deep sequencing)



Annotations can be downloaded from UCSC genome browser



RNA-seq alignment can be annotation-dependent or de novo



RNA-seq analysis: overview

Raw fastq

trimming (cutadapt), qc (fastqc, htseq) etc

fastq

Mapping

bigwig

Signal track for visualization

sam/bam

Quantification

read count

Normalization and Comparative analysis

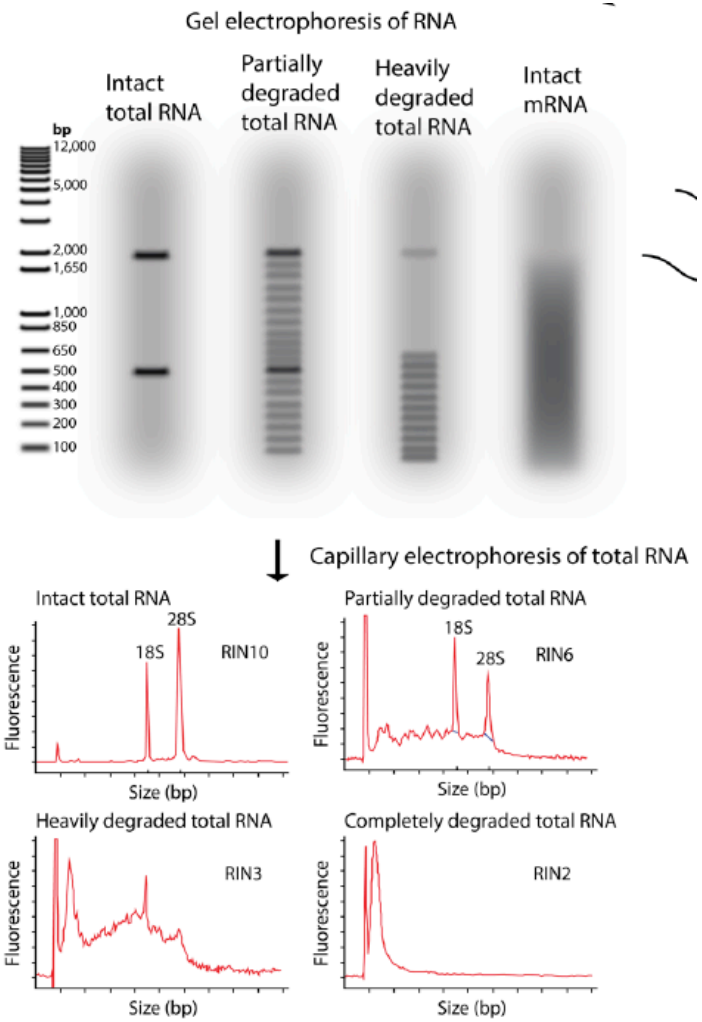
differential expression

RNA-seq questions during library preparation

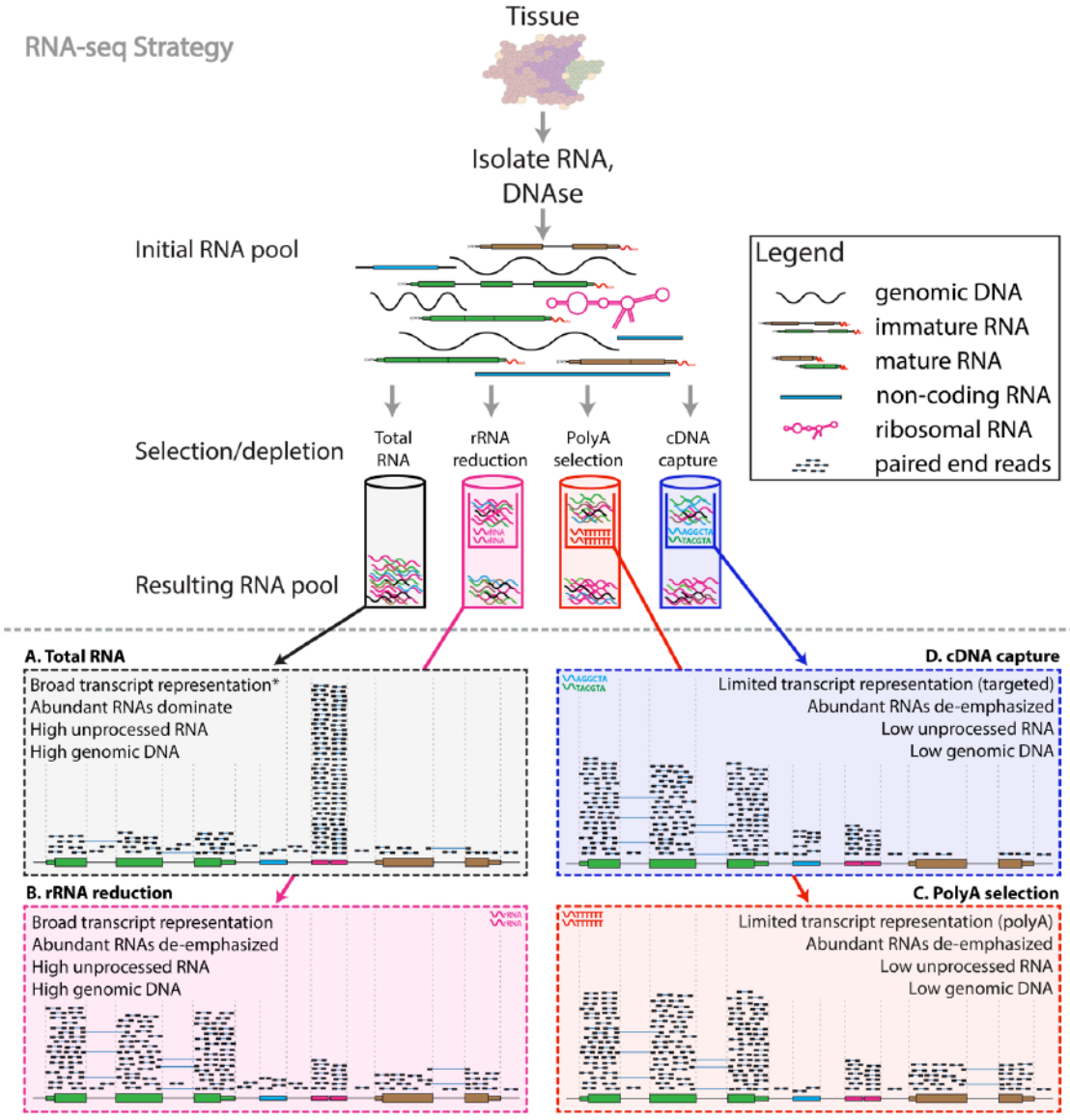
- Construction strategies
  - total RNA or polyA+ RNA?
  - Ribo minus?
  - Stranded or unstranded?
  - Read length?
  - Single end vs paired end?
  - size selection – microRNA?
  - Depth?
- RNA quantity
- RNA quality
  - RNA is fragile and easily degraded
  - Low quality material can bias the data
- Replicates

Agilent

- https://github.com/griffithlab/rnaseq_tutorial/wiki/Resources/Agilent_Trace_Examples.pdf
- ‘RIN’ = RNA integrity number
  - 0 (bad) to 10 (good)

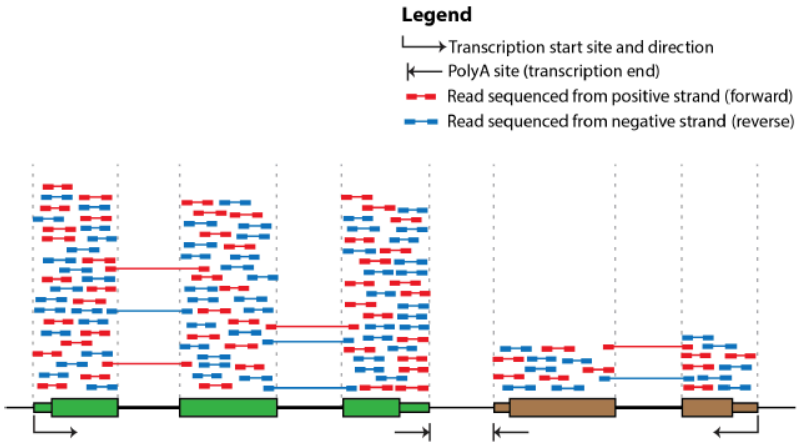


Strategies

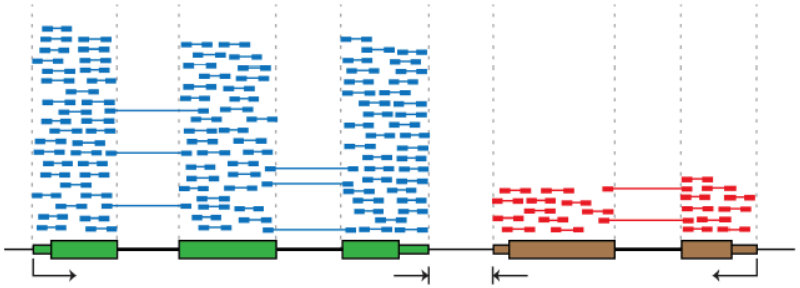

RNA-seq Strategy
Tissue
Isolate RNA, DNAse
Initial RNA pool
Selection/depletion
Total RNA
rRNA reduction
PolyA selection
cDNA capture
Resulting RNA pool
Legend
genomic DNA
immature RNA
mature RNA
non-coding RNA
ribosomal RNA
paired end reads
A. Total RNA
Broad transcript representation*
Abundant RNAs dominate
High unprocessed RNA
High genomic DNA
D. cDNA capture
Limited transcript representation (targeted)
Abundant RNAs de-emphasized
Low unprocessed RNA
Low genomic DNA
B. rRNA reduction
Broad transcript representation
Abundant RNAs de-emphasized
High unprocessed RNA
High genomic DNA
C. PolyA selection
Limited transcript representation (polyA)
Abundant RNAs de-emphasized
Low unprocessed RNA
Low genomic DNA

Stranded vs unstranded

A. Depiction of cDNA fragments from an unstranded library



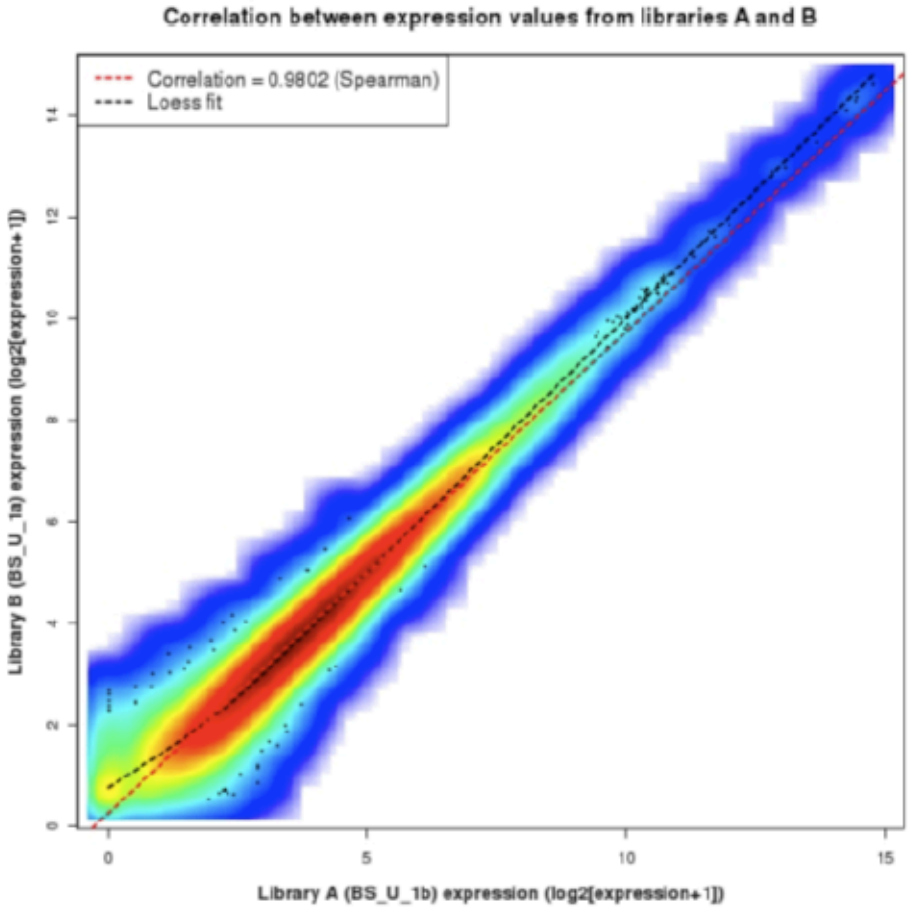
B. Depiction of cDNA fragments from an stranded library



Library Kit	Stranded	5' to 3' IGV	TopHat (--library-type parameter)	HTSeq (--stranded/-s)	Picard (STRAND_SPECIFICITY option of CollectRnaSeqMetrics)
TruSeq Strand Specific Total RNA	Yes	F2R1	fr-firststrand	reverse	SECOND_READ_TRANSCRIPTION_STRAND
NuGEN Encore	Yes	F1R2	fr-secondstrand	yes	FIRST_READ_TRANSCRIPTION_STRAND
NuGEN OvationV2	No	F2R1 or F1R2	fr-unstranded	no	NONE

Replicates

- Technical Replicate
  - Multiple instances of sequence generation
    - Flow Cells, Lanes, Indexes
- Biological Replicate
  - Multiple isolations of cells showing the same phenotype, stage or other experimental condition
  - Some example concerns/ challenges:
    - Environmental Factors, Growth Conditions, Time
  - Correlation Coefficient 0.92-0.98



RNA-seq questions during mapping

- Reference genome version – the latest version may have compatibility issues with other analysis
- Annotation – refseq or gencode or ENSEMBL
- Want junction read or not
- Remove duplicates? (No!)
- How many missmatches to allow?

RNA-seq questions during quantification

- Keep reads mapping to multiple loci?
- Keep reads overlapping multiple genes?

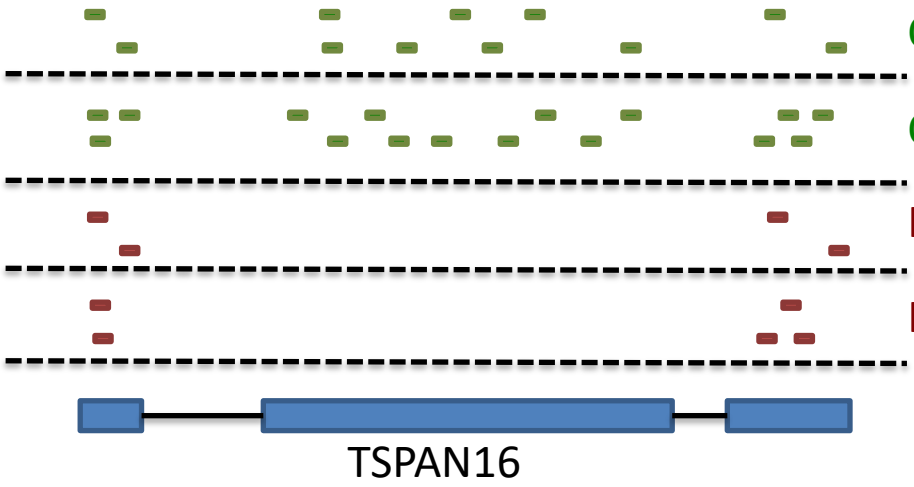
Alignment/Mapping tools

- TopHat2 (widely used, slow)
- STAR (super fast, very popular now, very demanding)
- RSEM (getting popular, used by ENCODE)
- Rsubread (works in R, not very popular)
- Sailfish (good for isoforms, less popular)

Quantification

- TopHat2 → Cufflinks2 (provides FPKM) → Cuffdiff (DGE analysis)
- STAR → HTSeq-count, featureCount (provides raw count) → DESeq2, EdgeR (DGE analysis and normalized count)
- RSEM → RSEM (provides expected count, TPM and FPKM) → EBSeq
- Rsubread → featureCount → DESeq2, EdgeR
- Sailfish →Sailfish (provides raw count, TPM) → DESeq2, EdgeR

Summarized RNA-seq



	Control1	Control2	KnockDown1	KnockDown2
TSPAN6	11	16	4	5
TNMD	1	0	0	0
DPM1	435	743	836	739
SCYL3	203	218	416	352
C1orf112	216	643	714	704
FGR	2365	5011	2828	2294
CFH	6	1	4	0
FUCA2	380	865	431	523
…	…	…	…	…
NFYA	888	827	1674	1580

Key concepts

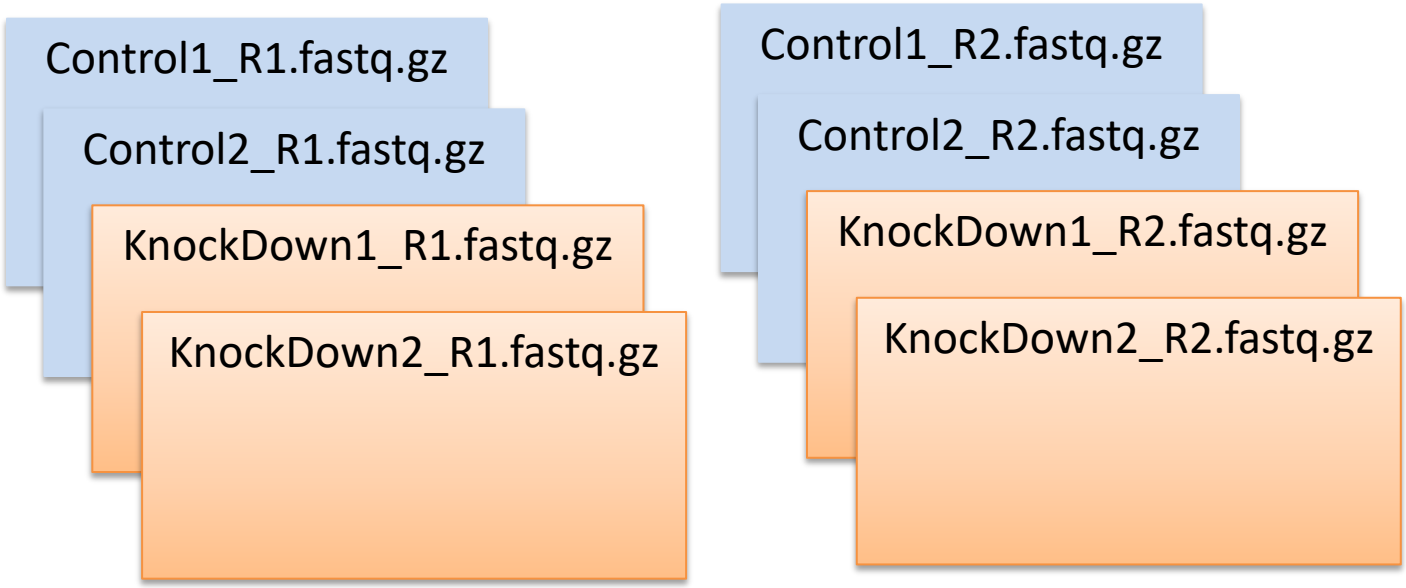
- **Expression units:** There are several expression units available – RPKM/FPKM, CPM, TPM, Normalized expression
- **Fasta file:** Sequence storing file (can be opened in TextWrangler (unix) or Notepad++ (windows))
  - Format:

```
>sequence1
ATCGTGCTGATGCGTGACG
```

- **Fastq file:** Sequence storing file with quality score, what you get from the sequencing centre
- **Bed file:** Standard file format for storing genomic coordinates
  - Format:

```
chr1  1033453 10443542    locus1
chr3  4442235 45235256    locus2
```

First file: fastq



~ 10Gb each sample

```
@ERR127302.1 HWI-EAS350_0441:1:1:1055:4898#0/1
GGCTCATCTTGAACTGGGTGGCGACCGTCCCTGGCCCCTTCTTGACACCCA
+
4=B@D99BDDDDDDD:DD?B<<=?>6B##########################
```

From fastq to sam/bam

Control1.bam

Control2.bam

SRR013667.1 99 19 8882171 60 76M = 8882214 119 NCCAGCAGCCATAACTGGAATGGGAAATAAACACTATGTTCAAAG

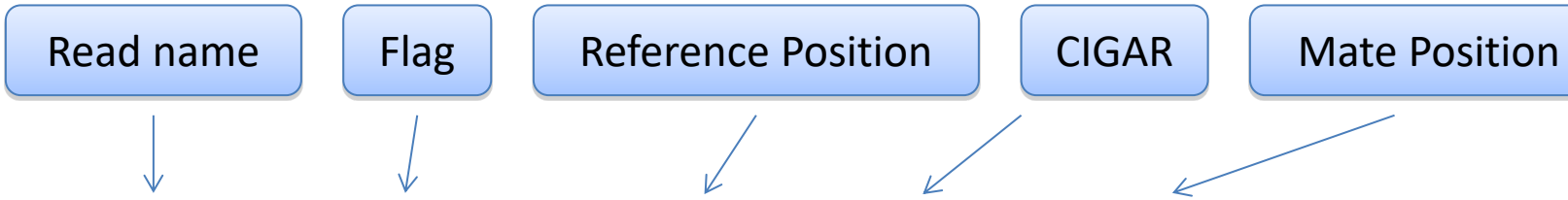
KnockDown1.bam

KnockDown2.bam

SRR013667.1 99 19 8882171 60 76M = 8882214 119 NCCAGCAGCCATAACTGGAATGGGAAATAAACACTATGTTCAAAG

~ 10Gb each bam

- Used to store alignments
- SAM = text, BAM = binary, CRAM=compressed binary




```
SRR013667.1 99 19 8882171 60 76M = 8882214 119
```

SAM: Sequence Alignment/Map format

Trivia time!

- What is the first step after getting the fastq file?
  - a) Contact bioinformatician b) Alignment then QC
    c) QC then alignment d) Quantification and plotting
- Should we always have replicates?
  - a) Yes b) No
- From fastq we make bam files. What do bam files contain?
  - a) Mapped reads b) Treasure map c) Raw reads d) Nothing useful really
- Can I use FPKM in DESeq2?
  - a) Yes b) No
- Which one of below is a major bottleneck in gene expression analysis?
  - a) My confidence b) High performance computers
    c) Repeats in the genome d)ribosomal RNAs
- What data should I use to generate an expression boxplot for *MYC* for 4 samples processed together?
  - a) raw read count b) FPKM c) normalized read count d) TPM
- HTSeq-count or featureCount requires _____
  - a) fastq files b) latest computer c) bam files d) annotation file

Trivia time!

- What does GTF file contain?
  - a) Mapping information b) Genome annotation
    c) Quality information d) Loci information

- What does bed file contain?
  - a) Mapping information b) Genome annotation
    c) Quality information d) Loci information

- You should consider before deciding on sequencing depth -
  - a) single end or paired end b) research purpose
    c) RNA quantity d) Read length

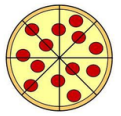
- I want to study alternative splicing. Factors in order of their importance for my study -
  - a) selection, paired-end, depth, read length
    b) Read length, selection, paired-end, depth
    c) Depth, read length, paired-end, selection
    d) Read length, depth, paired-end, selection

Resources

- **https://www.ncbi.nlm.nih.gov/pubmed/26248053**
- http://www.bioconductor.org/help/workflows/rnaseqGene
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/
- https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise

Are there any particular topics you would like to be discussed?

- "I'm sure it will be covered by quality control checks that determine if your RNAseq data is poor or good enough for further use."
  - RNA integrity
  - fastQC
  - Check known genes
  - Compare replicates (correlation, PCA etc)
  - Visualize read distribution in IGV
- "TCGA"
  - cBioPortal
  - Xena browser (https://tcga.xenahubs.net)
- "N/A But I don't need pizza :)"



NEVER TRUST SOMEONE WHO DOESN'T EAT PIZZA