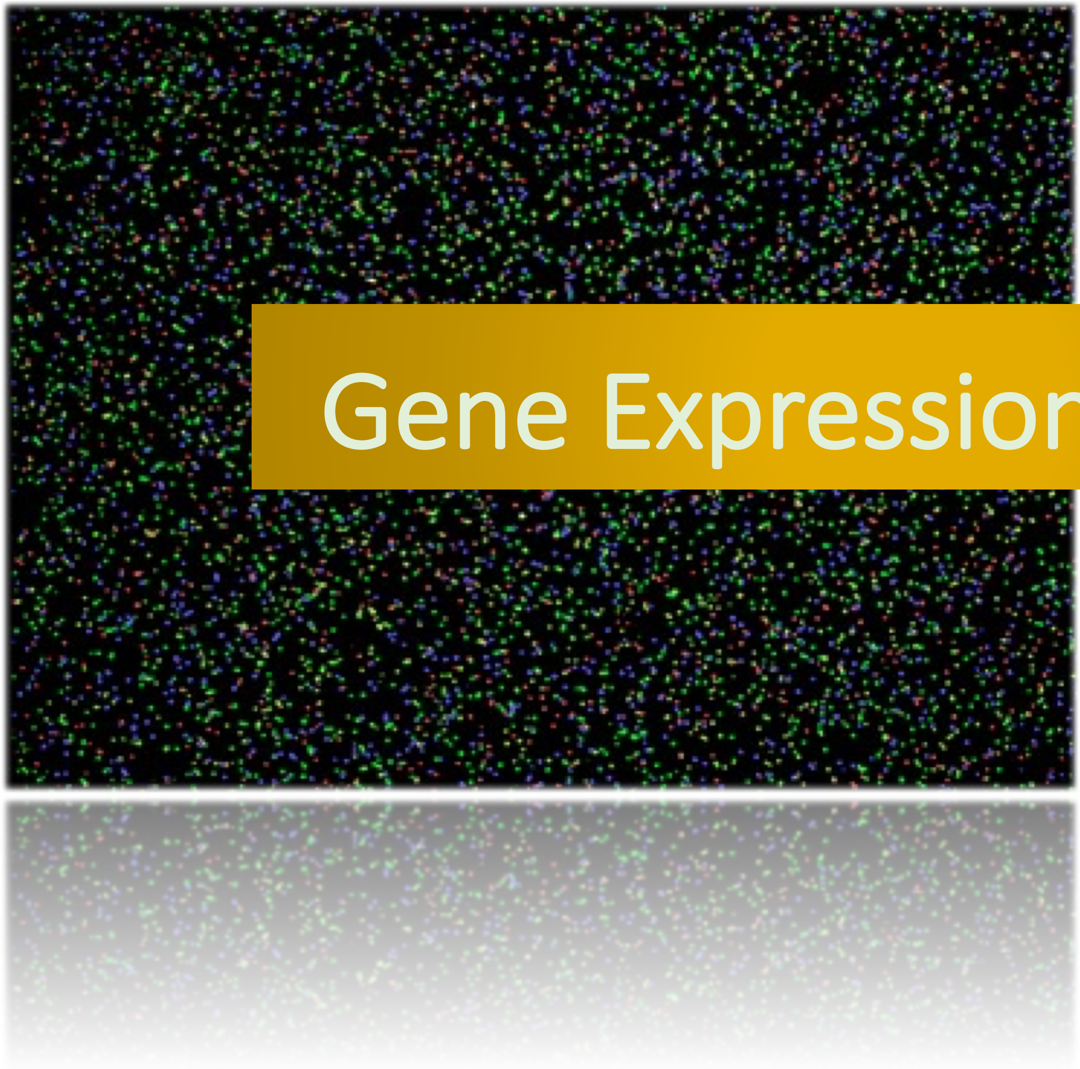


Gene Expression Analysis

P. Tang (鄧致剛); YM Yeh (葉元鳴)

Bioinformatics Center, Chang Gung University.

Central dogma

Gene expression

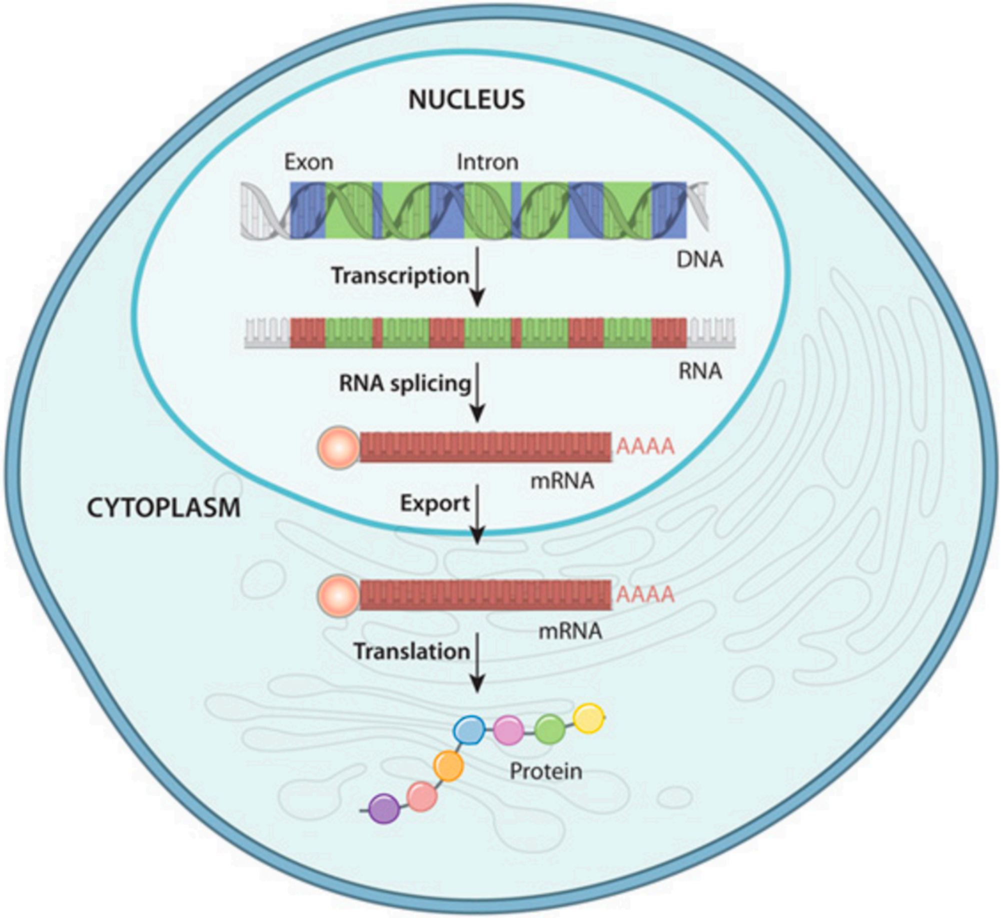
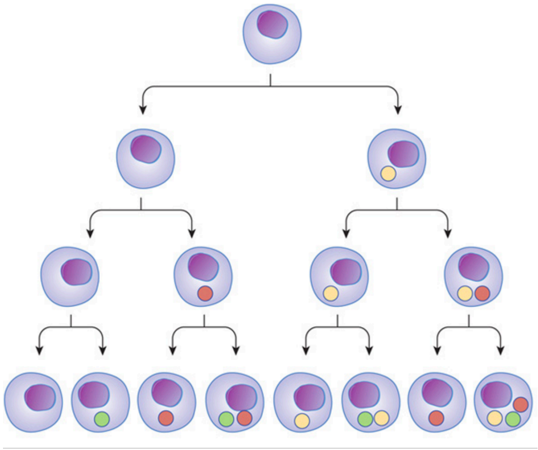


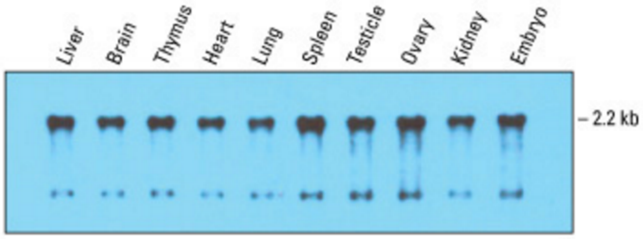
Figure 1: An overview of the flow of information from DNA to protein in a eukaryote
First, both coding and noncoding regions of DNA are transcribed into mRNA. Some regions are removed (introns) during initial mRNA processing. The remaining exons are then spliced together, and the spliced mRNA molecule (red) is prepared for export out of the nucleus through addition of an endcap (sphere) and a polyA tail. Once in the cytoplasm, the mRNA can be used to construct a protein.

(vs. DNA) Why RNA ?

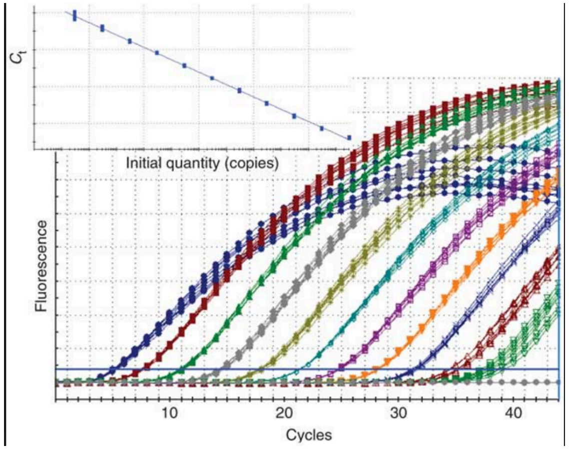
- Differentially expressed – Functional studies
    - Different cell types (muscle cells, fibroblasts)
    - Environmental conditions (heat shock, nutrient deprivation)
    - Developmental phases (embryonic day 12)
    - Cell-cycle stages (S phase)
    - Disease states (tumor cells, virus-infected cells)
- Transcription level – Molecular features
    - Alternative isoforms
    - Fusion transcripts
    - RNA editing
- Prioritizing protein coding somatic mutations (often heterozygous)



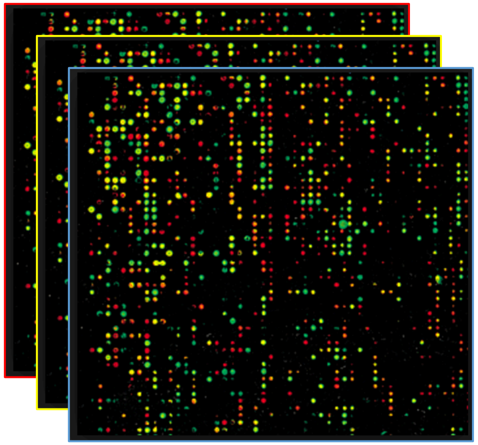
Evolution of transcriptomics technologies



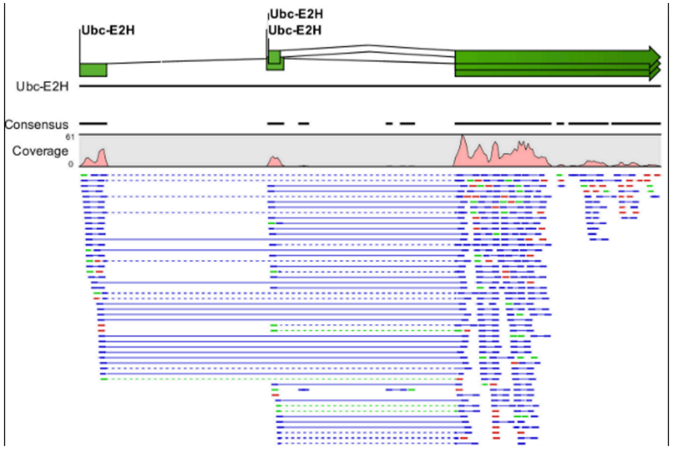
Northern Blot



Real time RT-PCR

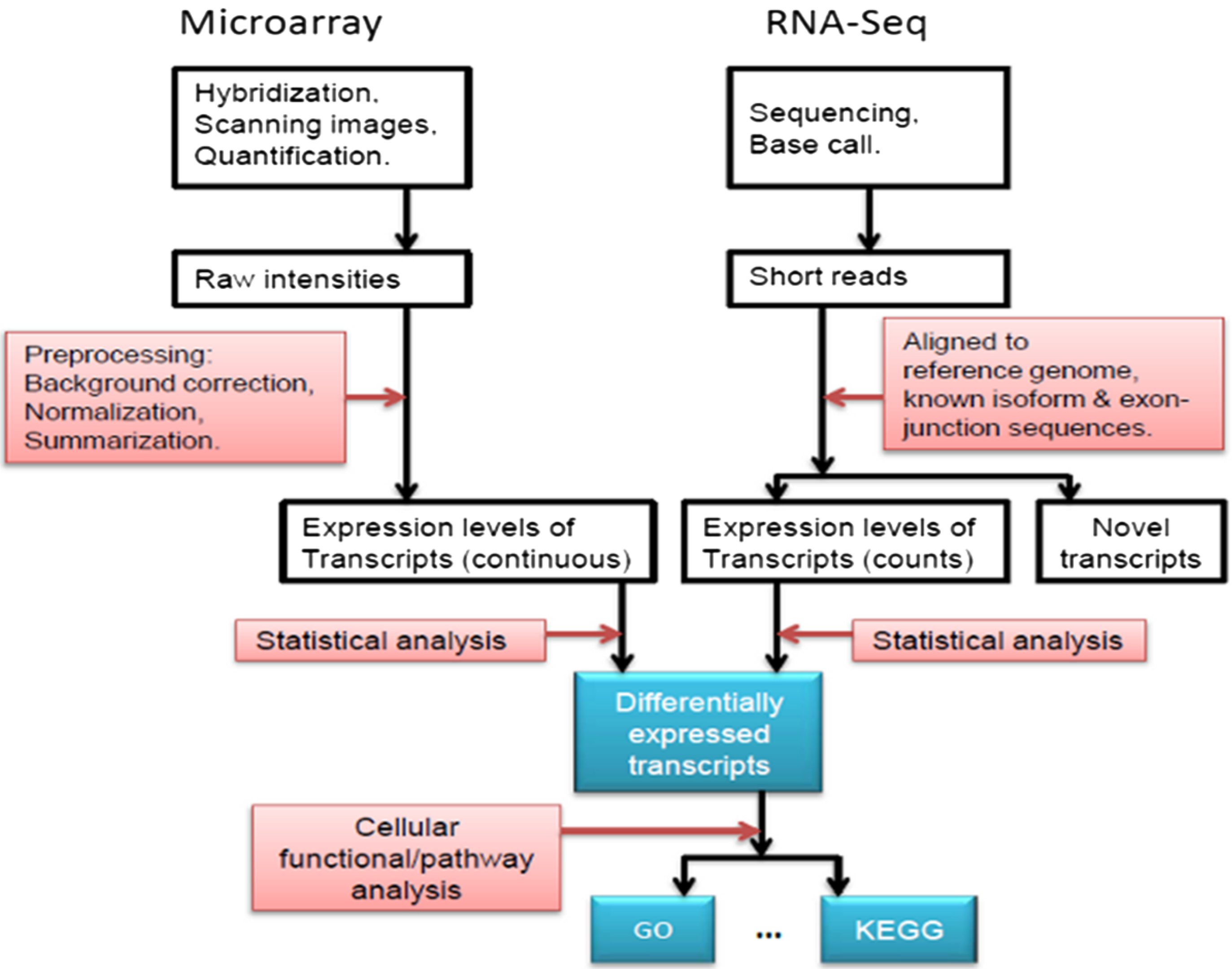


Microarray



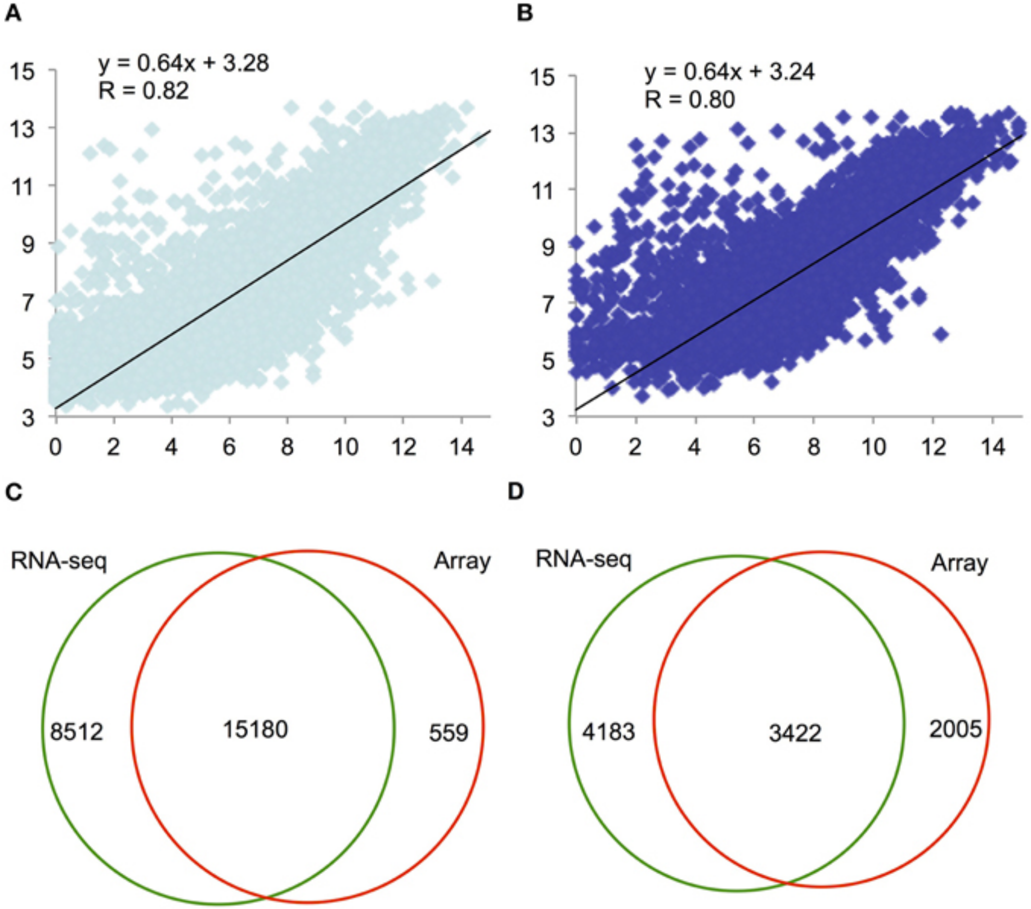
RNA seq

Overview of analysis workflow for microarray and RNA-seq transcriptional profiling


Microarray
RNA-Seq
Hybridization, Scanning images, Quantification.
Sequencing, Base call.
Raw intensities
Short reads
Preprocessing: Background correction, Normalization, Summarization.
Aligned to reference genome, known isoform & exon-junction sequences.
Expression levels of Transcripts (continuous)
Expression levels of Transcripts (counts)
Novel transcripts
Statistical analysis
Statistical analysis
Differentially expressed transcripts
Cellular functional/pathway analysis
GO
...
KEGG

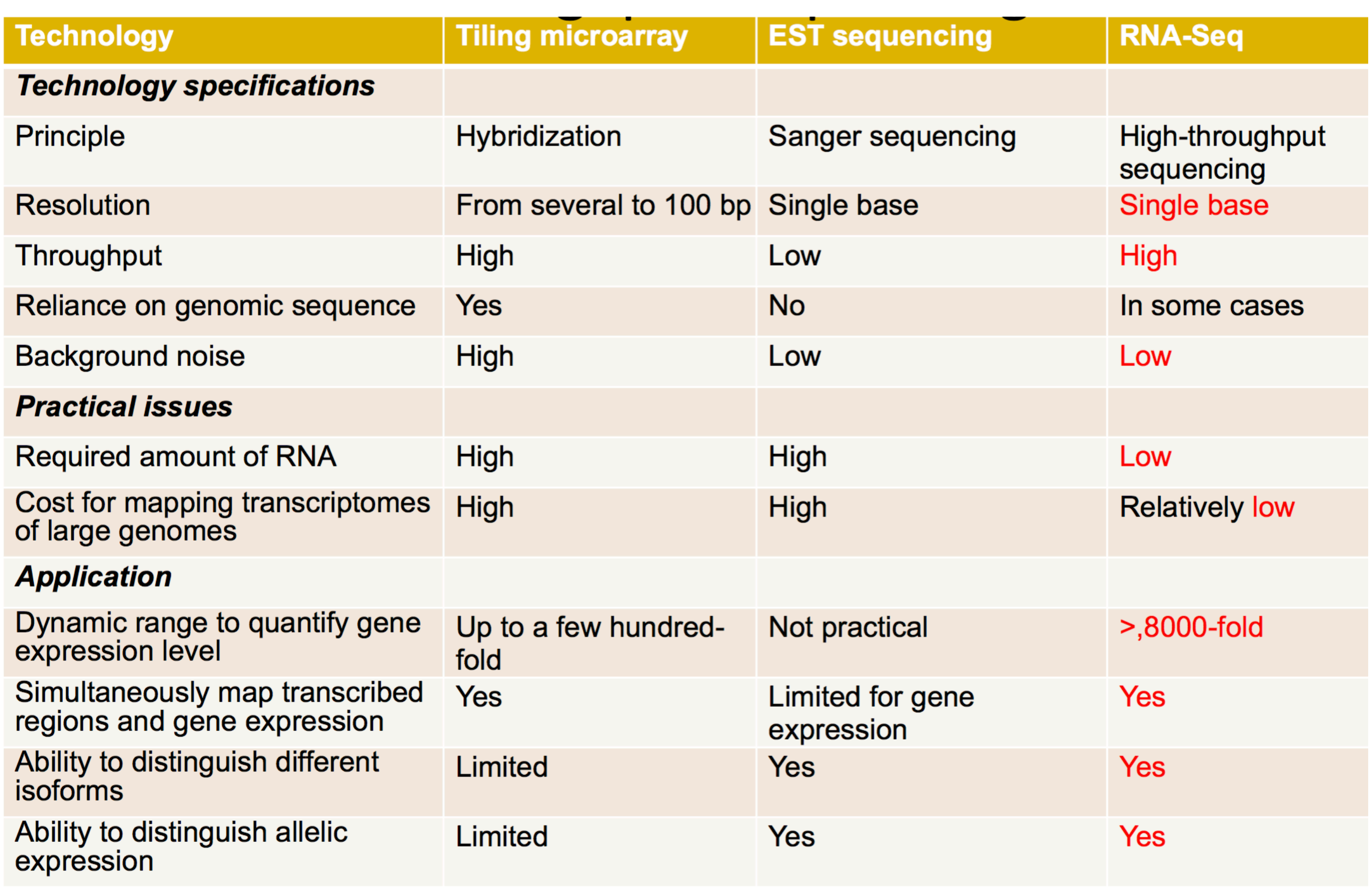
Fang Z et.al. Cell Biosci. 2012;2:26

RNA-seq vs. Microarray yield correlated results



Zhing LS et. Al. *Front. Plant Sci*. **5**:802.

Microarray -> High Throughput Sequencing (HTS)

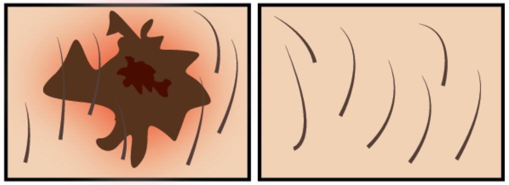


Technology	Tiling microarray	EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low
Application			
Dynamic range to quantify gene expression level	Up to a few hundred-fold	Not practical	>,8000-fold
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes

Wang et al. Nature Rev Genet, 10:57, 2009

RNA sequencing

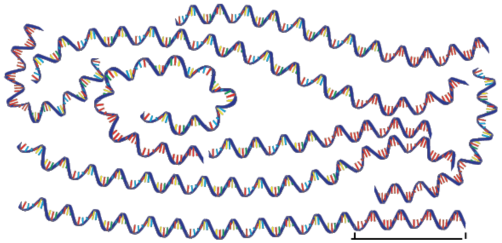
Samples of interest



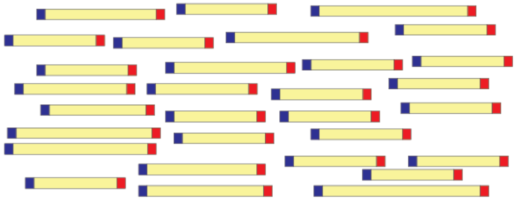
Condition 1
(e.g. tumor)

Condition 2
(e.g. normal)

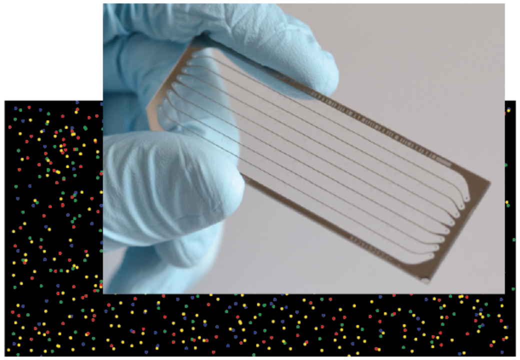
Isolate RNAs



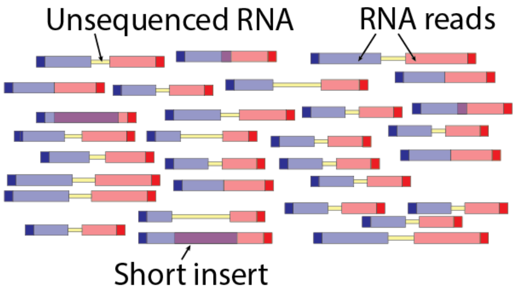
Generate cDNA, Fragment,
size select, add linkers



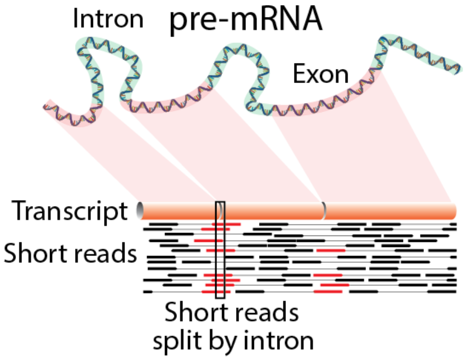
Sequence ends



100s of millions fo paired reads
10s of billions bases of sequence



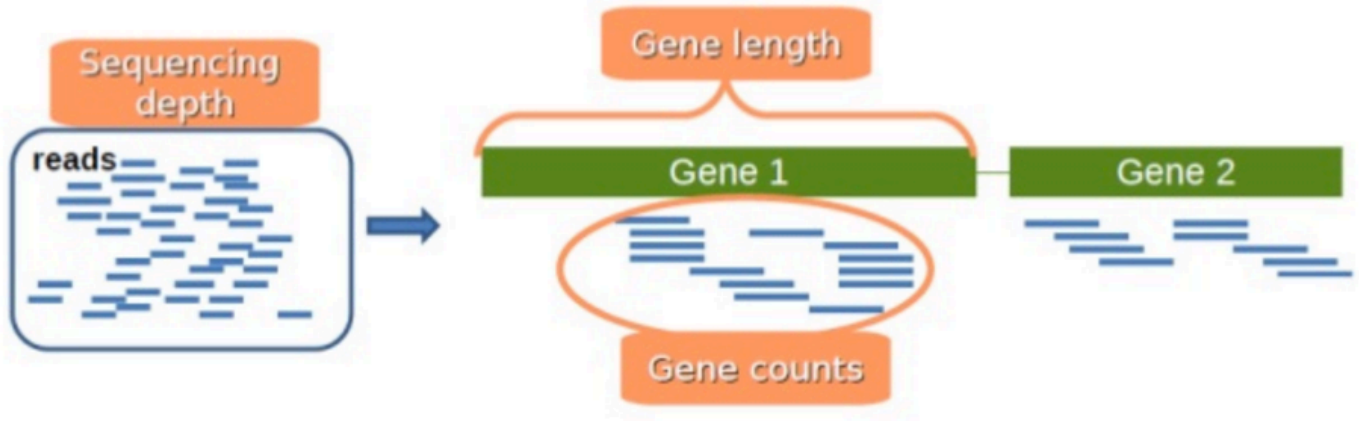
Map to genome, transcriptome,
and predicted exon junctions



Downstream analysis

Important concepts

- Sequencing depth (X)
- Gene length
- Gene counts

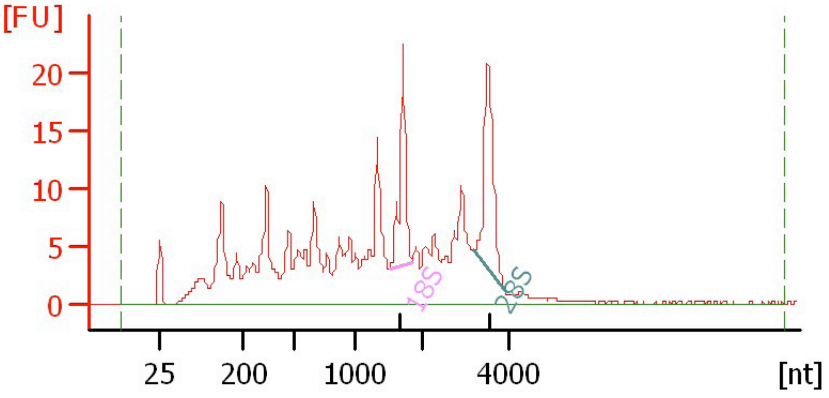


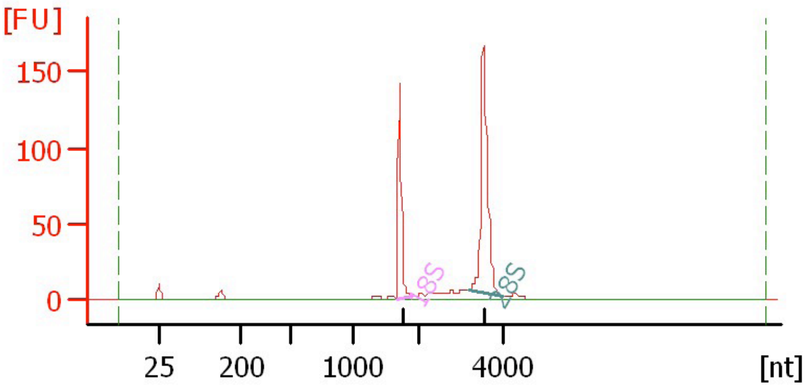
Challenges

- Sample
  - Purity?, quantity?, quality?
- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
  - 10^5 – 10^7 orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction of highly expressed genes may consume the majority of reads
  - Ribosomal and mitochondrial genes
- RNAs come in a wide range of sizes
  - Small RNAs must be captured separately
  - PolyA selection of large RNAs may result in 3' end bias
- RNA is fragile compared to DNA (easily degraded)

Agilent example / interpretation

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



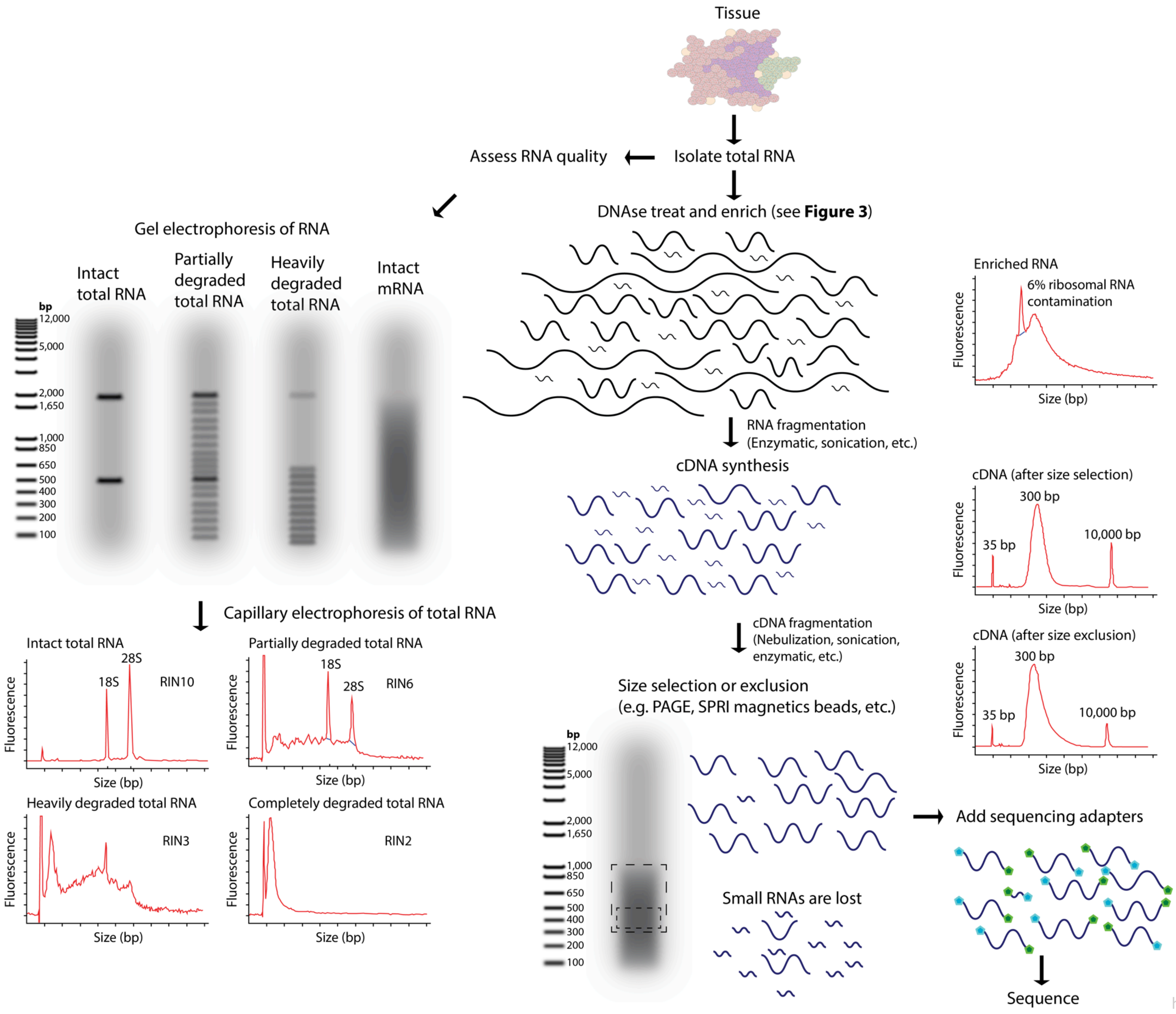


RNA-seq library construction strategies

- Total RNA versus polyA+ RNA?
- Ribo-reduction?
- Size selection (before and/or after cDNA synthesis)
    - Small RNAs (microRNAs) vs. large RNAs?
    - A narrow fragment size distribution vs. a broad one?
- Linear amplification?
- Stranded vs. un-stranded libraries
- Exome captured vs. un-captured
- Library normalization?

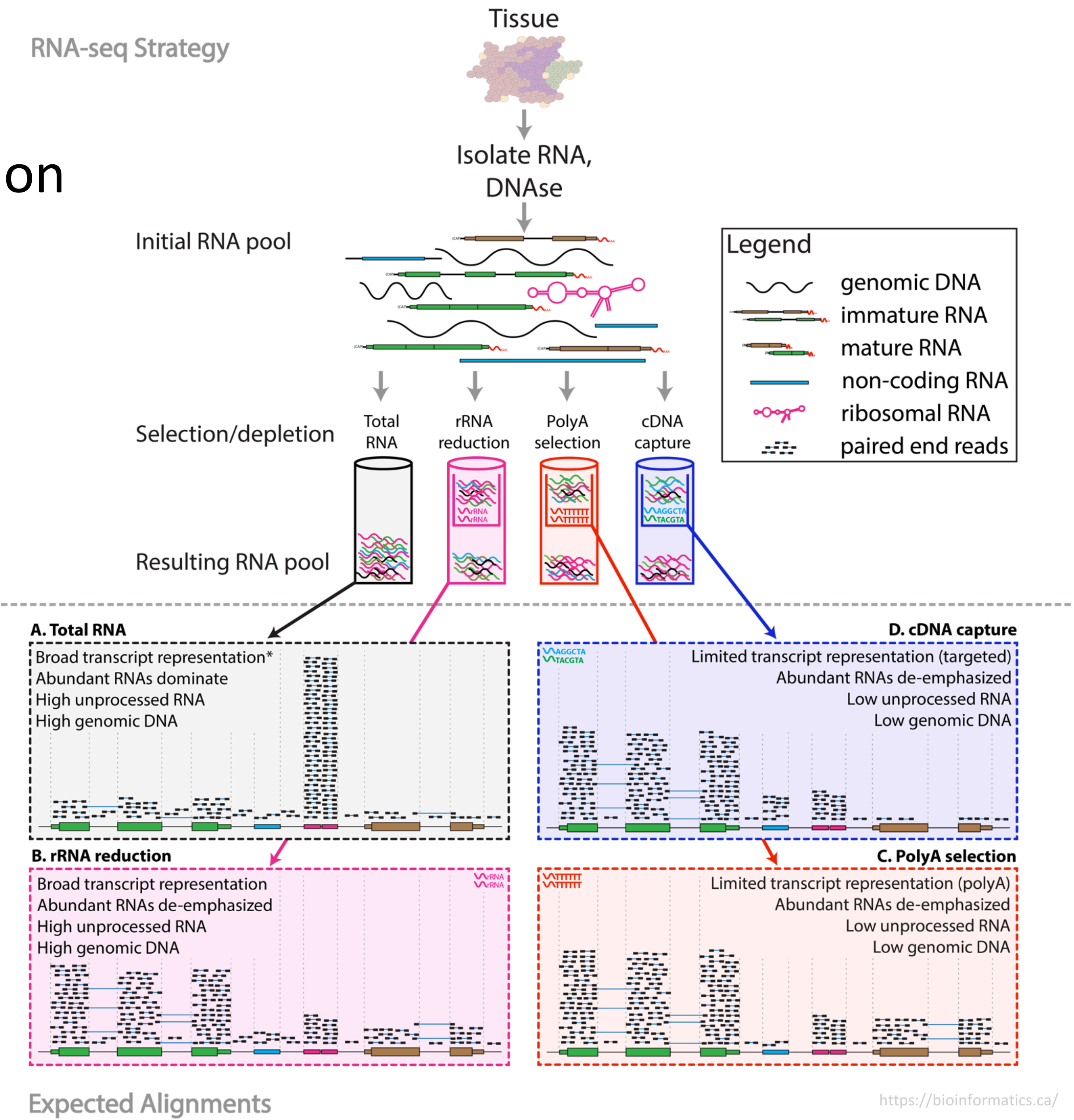
- These details can affect analysis strategy
    - Especially comparisons between libraries

Fragmentation and size selection


Tissue
Isolate total RNA
Assess RNA quality
DNAse treat and enrich (see **Figure 3**)
Gel electrophoresis of RNA
Intact total RNA
Partially degraded total RNA
Heavily degraded total RNA
Intact mRNA
bp
12,000
5,000
2,000
1,650
1,000
850
650
500
400
300
200
100
Capillary electrophoresis of total RNA
Intact total RNA
28S
18S
RIN10
Partially degraded total RNA
18S
28S
RIN6
Heavily degraded total RNA
RIN3
Completely degraded total RNA
RIN2
Fluorescence
Size (bp)
Enriched RNA
6% ribosomal RNA contamination
Fluorescence
Size (bp)
RNA fragmentation (Enzymatic, sonication, etc.)
cDNA synthesis
cDNA (after size selection)
300 bp
35 bp
10,000 bp
Fluorescence
Size (bp)
cDNA fragmentation (Nebulization, sonication, enzymatic, etc.)
cDNA (after size exclusion)
300 bp
35 bp
10,000 bp
Fluorescence
Size (bp)
Size selection or exclusion (e.g. PAGE, SPRI magnetics beads, etc.)
bp
12,000
5,000
2,000
1,650
1,000
850
650
500
400
300
200
100
Small RNAs are lost
Add sequencing adapters
Sequence

https://bioinformatics.ca/

RNA sequence selection/depletion



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

Expected Alignments

https://bioinformatics.ca/

Stranded vs. un-Stranded

A. Depiction of cDNA fragments from an unstranded library

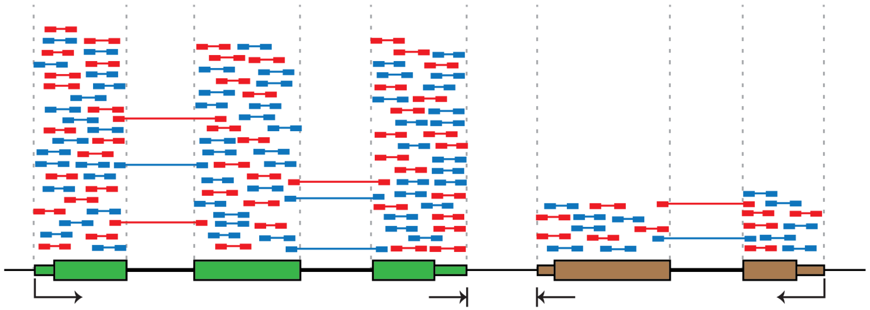
Legend

↳ Transcription start site and direction

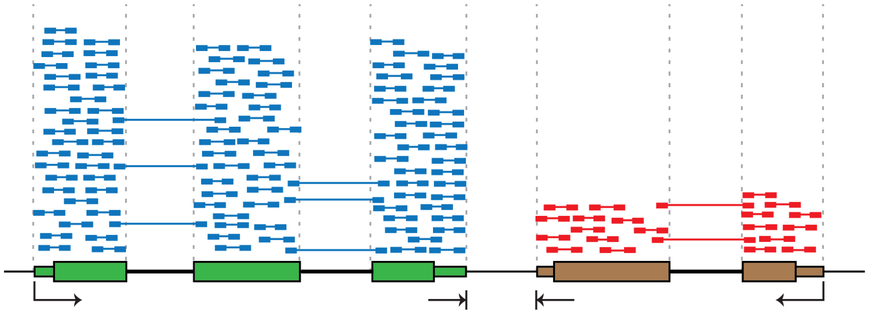
⇤ PolyA site (transcription end)

🟥 Read sequenced from positive strand (forward)

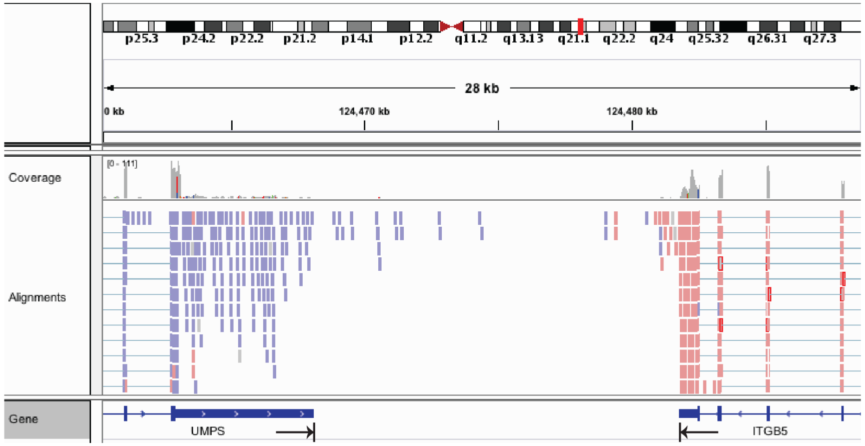
🟦 Read sequenced from negative strand (reverse)



B. Depiction of cDNA fragments from an stranded library

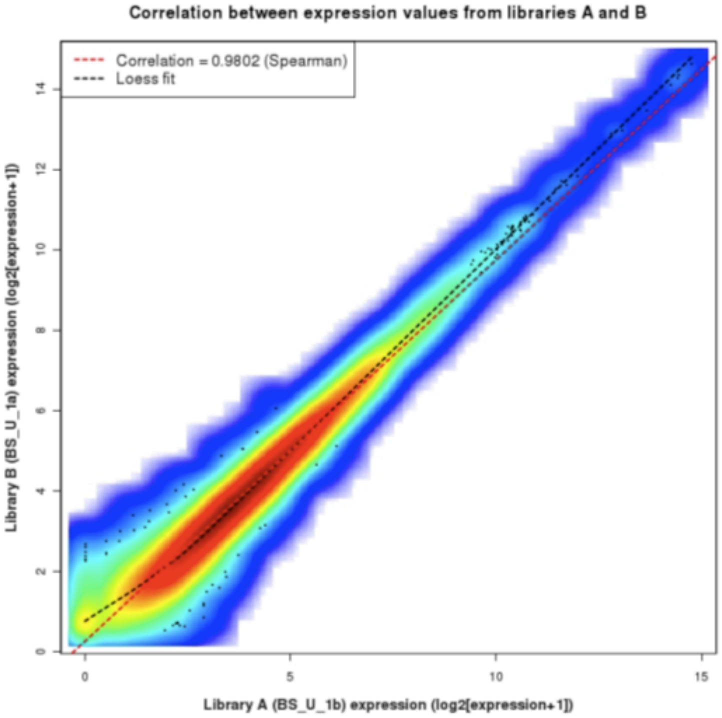


C. Viewing strand of aligned reads in IGV



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



Common analysis goals of RNA-Seq

- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
  - Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

General themes of RNA-seq workflows

- Each type of RNA-seq analysis has distinct requirements and challenges but also a common theme:

1. **Obtain raw data** (convert format)
2. **Align/assemble reads**
3. **Process alignment** with a tool specific to the goal
    - e.g. ‘cufflinks’ for expression analysis, ‘defuse’ for fusion detection, etc.
4. **Post process**
    - Import into downstream software (R, Matlab, Cytoscape, Ingenuity, etc.)
5. **Summarize and visualize**
    - Create gene lists, prioritize candidates for validation, etc.

How much library depth is needed for RNA-seq?

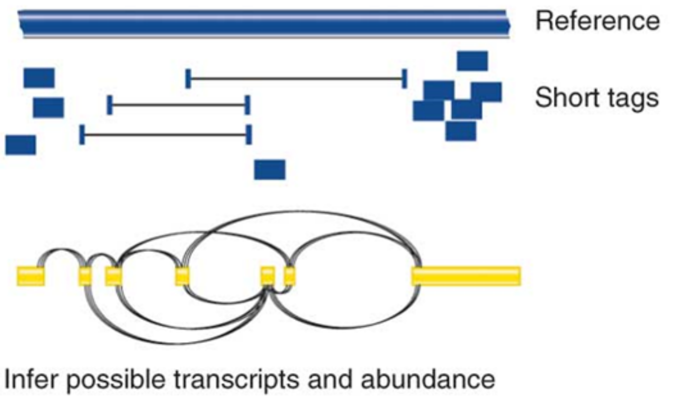
- Depends on a number of factors:
  - Question being asked of the data. **Gene expression**? **Alternative expression**? **Mutation calling**?
  - Tissue type, RNA preparation, quality of input RNA, library construction method, etc.
  - Sequencing type: read length, paired vs. unpaired, etc.
  - Computational approach and resources
- Identify publications with similar goals
- Pilot experiment
- Good news: 1-2 lanes of recent Illumina HiSeq data should be enough for most purposes

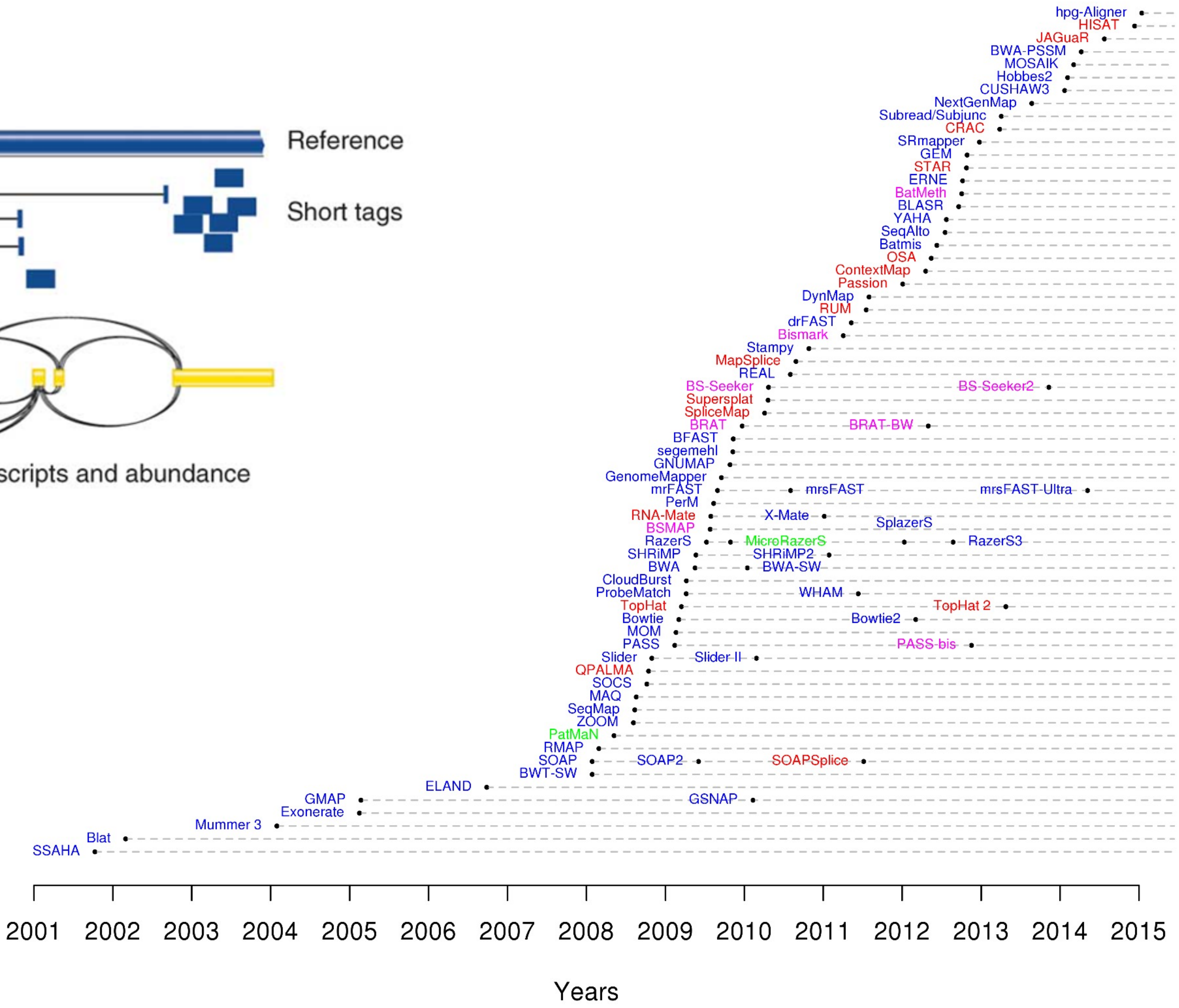
What mapping strategy should I use for RNA-seq?

- Depends on read length
- < 50 bp reads
  - Use aligner like BWA and a genome + junction database
  - Junction database needs to be tailored to read length
    - Or you can use a standard junction database for all read lengths and an aligner that allows substring alignments for the junctions only (e.g. BLAST … slow).
  - Assembly strategy may also work (e.g. Trans-ABySS)
- > 50 bp reads
  - Spliced aligner such as Bowtie/TopHat, STAR, HISAT, etc.

Informatics for RNA-seq: A web resource for analysis on the cloud. 11(8):e1004393. 2015.

Reference Sequence Alignment (Mapping)

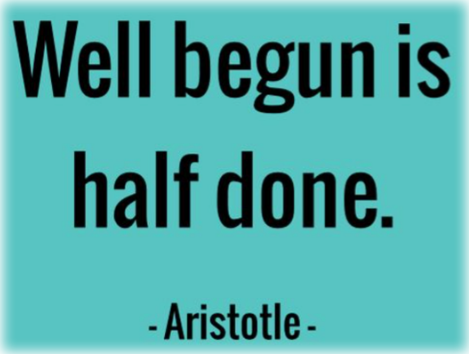




DNA mappers are plotted in blue, RNA mappers in red, miRNA mappers in green and bisulphite mappers in purple.

http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Experiment design



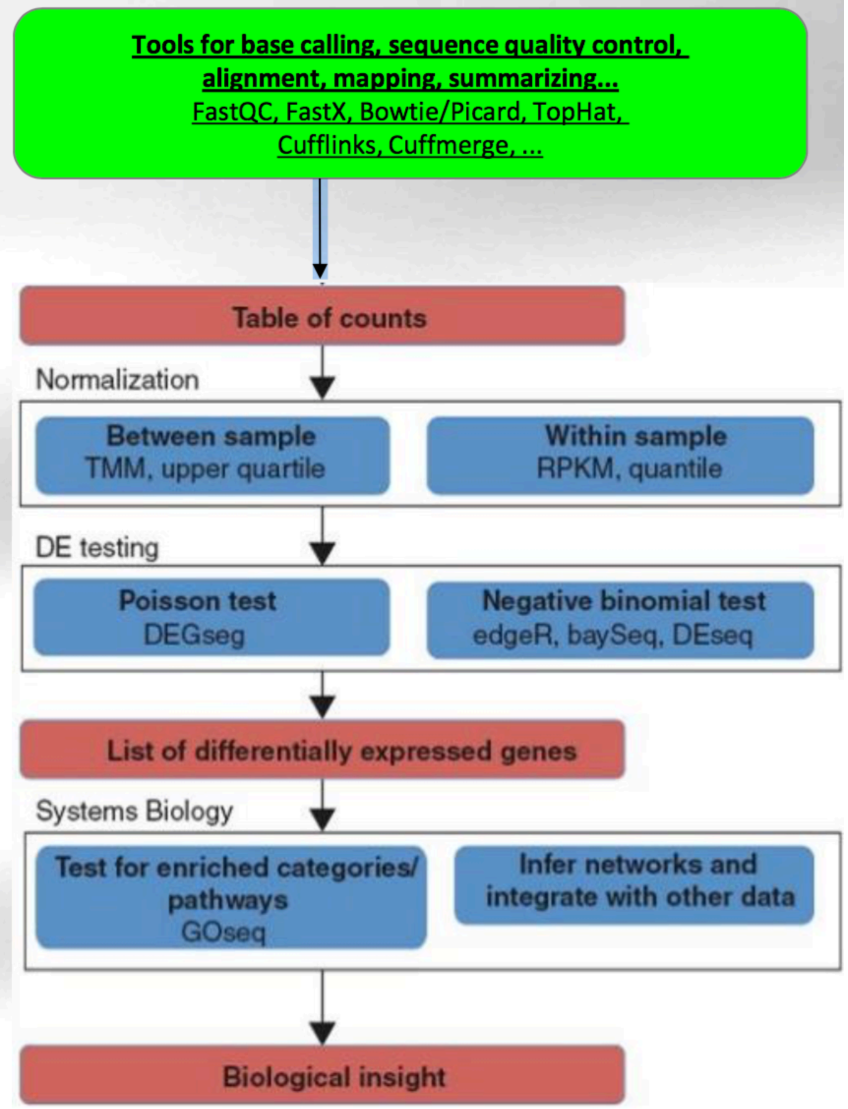
Well begun is half done.

- Aristotle -

- A clearly defined biological question
- Well control of potential sources of variation
- HTS experimental replicates
- Compliance with the standard of ~~microarray~~ (HTS based) information collection (MINSEQE)
    - ➢ http://fged.org/projects/minseqe/

RNA seq analysis workflow

- Reads are mapped to the reference genome or transcriptome
- Mapped reads are assembled into expression summaries (tables of counts, showing how may reads are in coding region, exon, gene or junction)
- Data is normalized
- Statistical testing of differential expression (DE) is performed, producing a list of genes with *p-values* and fold changes.
- Similar downstream analysis than microarray results (Functional Annotations, Gene Enrichment Analysis; Integration with other data...)


Tools for base calling, sequence quality control, alignment, mapping, summarizing...
FastQC, FastX, Bowtie/Picard, TopHat, Cufflinks, Cuffmerge, ...
Table of counts
Normalization
Between sample
TMM, upper quartile
Within sample
RPKM, quantile
DE testing
Poisson test
DEGseg
Negative binomial test
edgeR, baySeq, DEseq
List of differentially expressed genes
Systems Biology
Test for enriched categories/ pathways
GOseq
Infer networks and integrate with other data
Biological insight

Normalization/scaling/transformation: different goals

- **R/FPKM:** (Mortazavi et al. 2008)
  - Correct for: differences in sequencing depth and transcript length
  - Aiming to: compare a gene across samples and diff genes within sample

$$\text{RPKM} = \frac{\dfrac{\text{number of reads in region}}{\text{region length} \times 10^3}}{\text{total reads} \times 10^6}$$

- **TMM**: (Robinson and Oshlack 2010)
  - Correct for: differences in transcript pool composition; extreme outliers
  - Aiming to: provide better across-sample comparability

- **TPM**: (Li et al 2010, Wagner et al 2012)
  - Correct for: transcript length distribution in RNA pool
  - Aiming to: provide better across-sample comparability

- **Limma voom (logCPM)**: (Lawet al 2013)
  - Aiming to: stabilize variance; remove dependence of variance on the mean

Optimal Scaling of Digital Transcriptomes

Gustavo Glusman, Juan Caballero, Max Robinson, Burak Kutlu, Leroy Hood

Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

Differential expression analysis

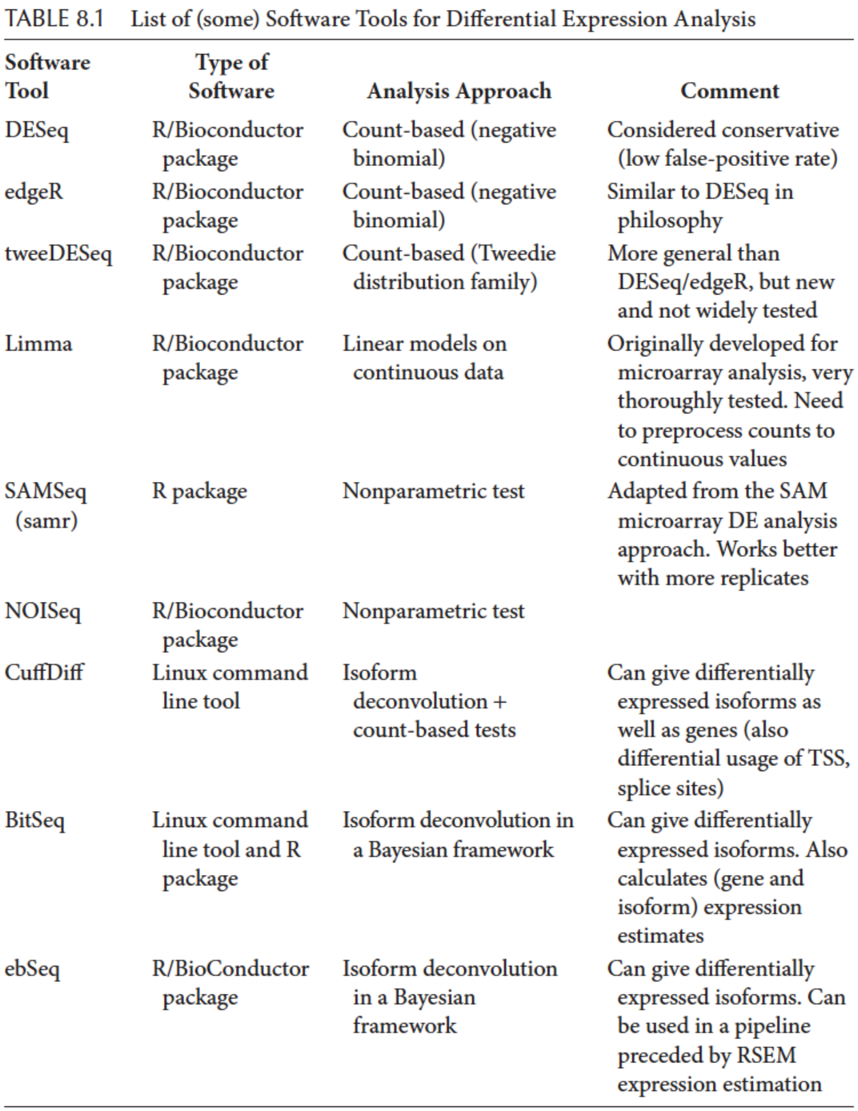
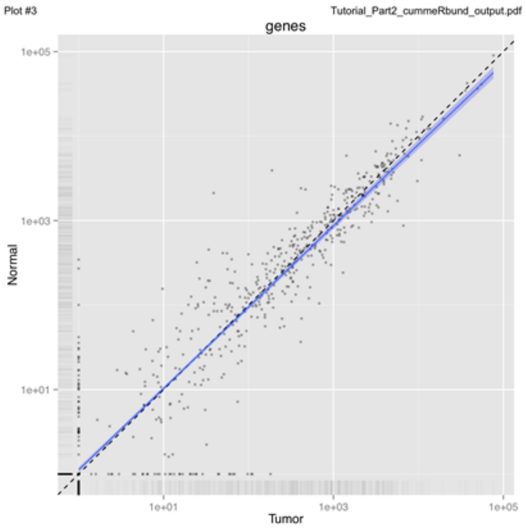
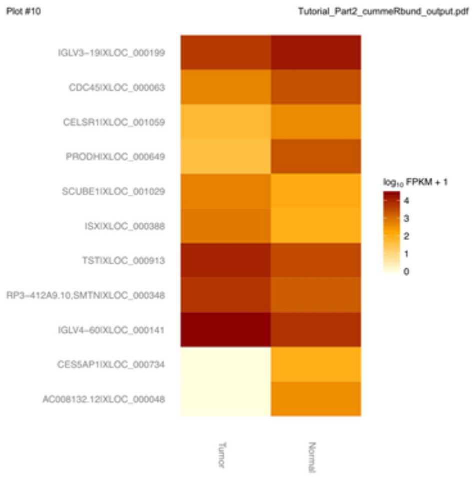


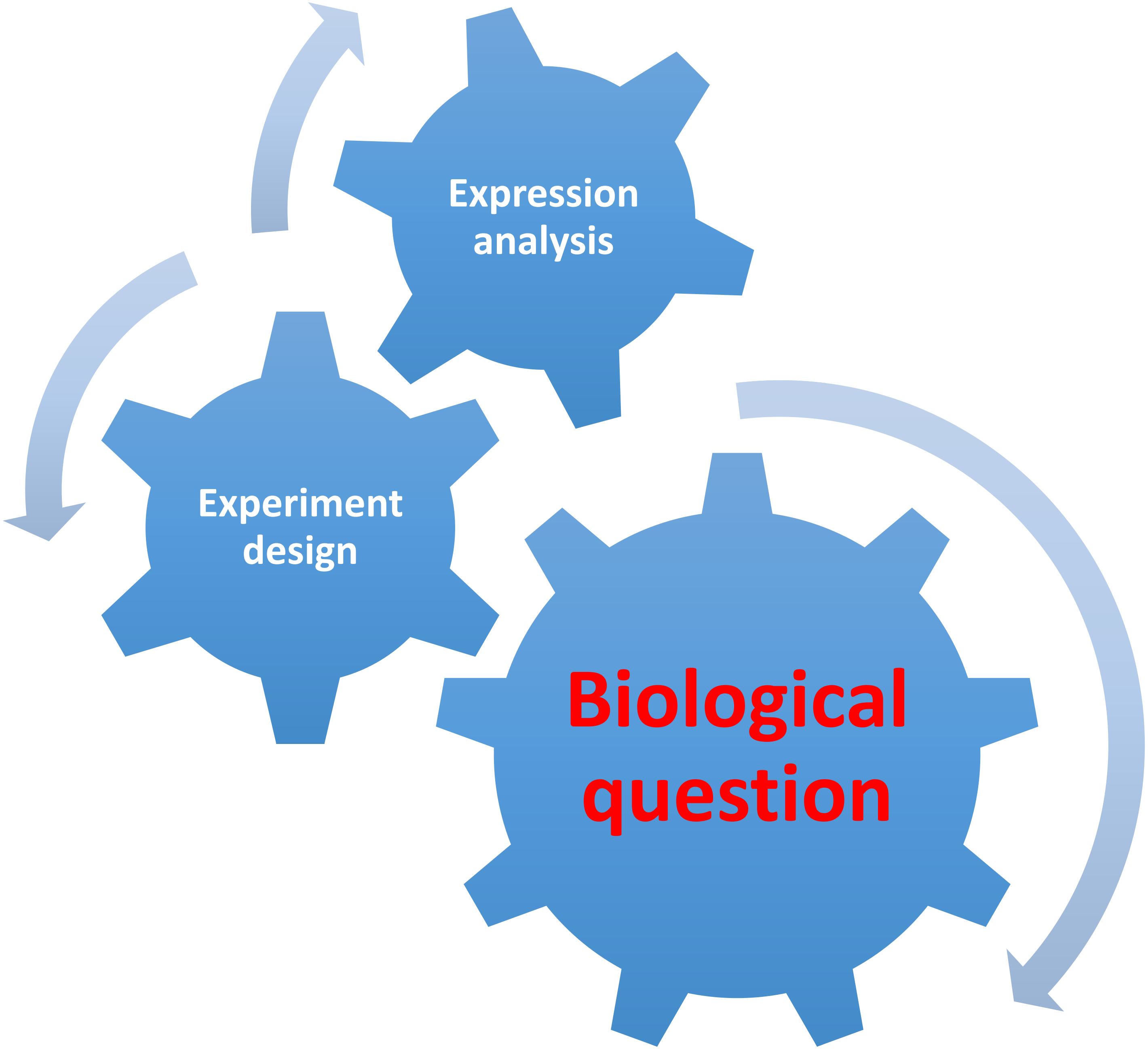
TABLE 8.1 List of (some) Software Tools for Differential Expression Analysis

Software Tool	Type of Software	Analysis Approach	Comment
DESeq	R/Bioconductor package	Count-based (negative binomial)	Considered conservative (low false-positive rate)
edgeR	R/Bioconductor package	Count-based (negative binomial)	Similar to DESeq in philosophy
tweeDESeq	R/Bioconductor package	Count-based (Tweedie distribution family)	More general than DESeq/edgeR, but new and not widely tested
Limma	R/Bioconductor package	Linear models on continuous data	Originally developed for microarray analysis, very thoroughly tested. Need to preprocess counts to continuous values
SAMSeq (samr)	R package	Nonparametric test	Adapted from the SAM microarray DE analysis approach. Works better with more replicates
NOISeq	R/Bioconductor package	Nonparametric test	
CuffDiff	Linux command line tool	Isoform deconvolution + count-based tests	Can give differentially expressed isoforms as well as genes (also differential usage of TSS, splice sites)
BitSeq	Linux command line tool and R package	Isoform deconvolution in a Bayesian framework	Can give differentially expressed isoforms. Also calculates (gene and isoform) expression estimates
ebSeq	R/BioConductor package	Isoform deconvolution in a Bayesian framework	Can give differentially expressed isoforms. Can be used in a pipeline preceded by RSEM expression estimation





Bioinformatics task 1



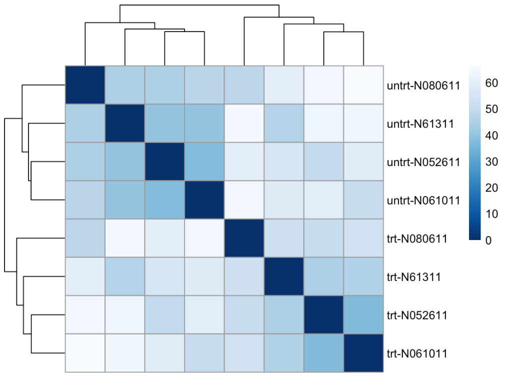
Expression analysis

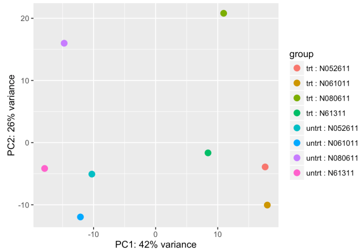
Experiment design

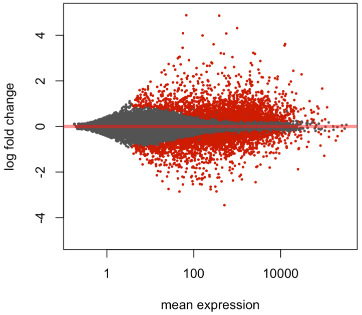
Biological question

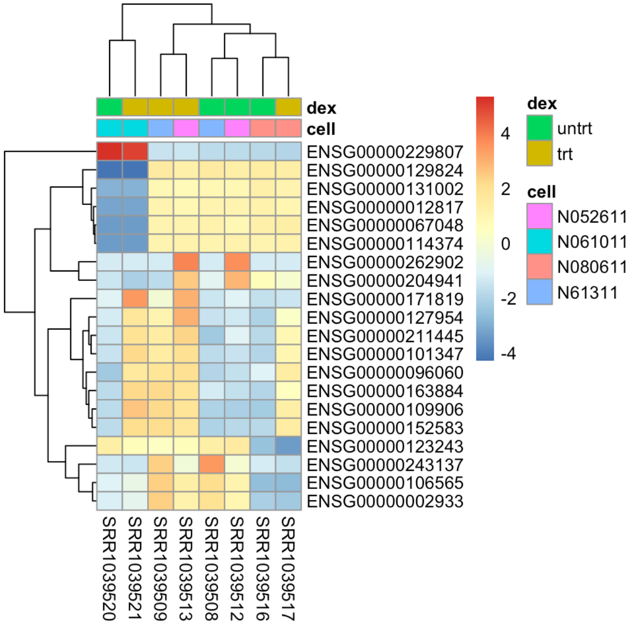
DESeq analysis as exercise

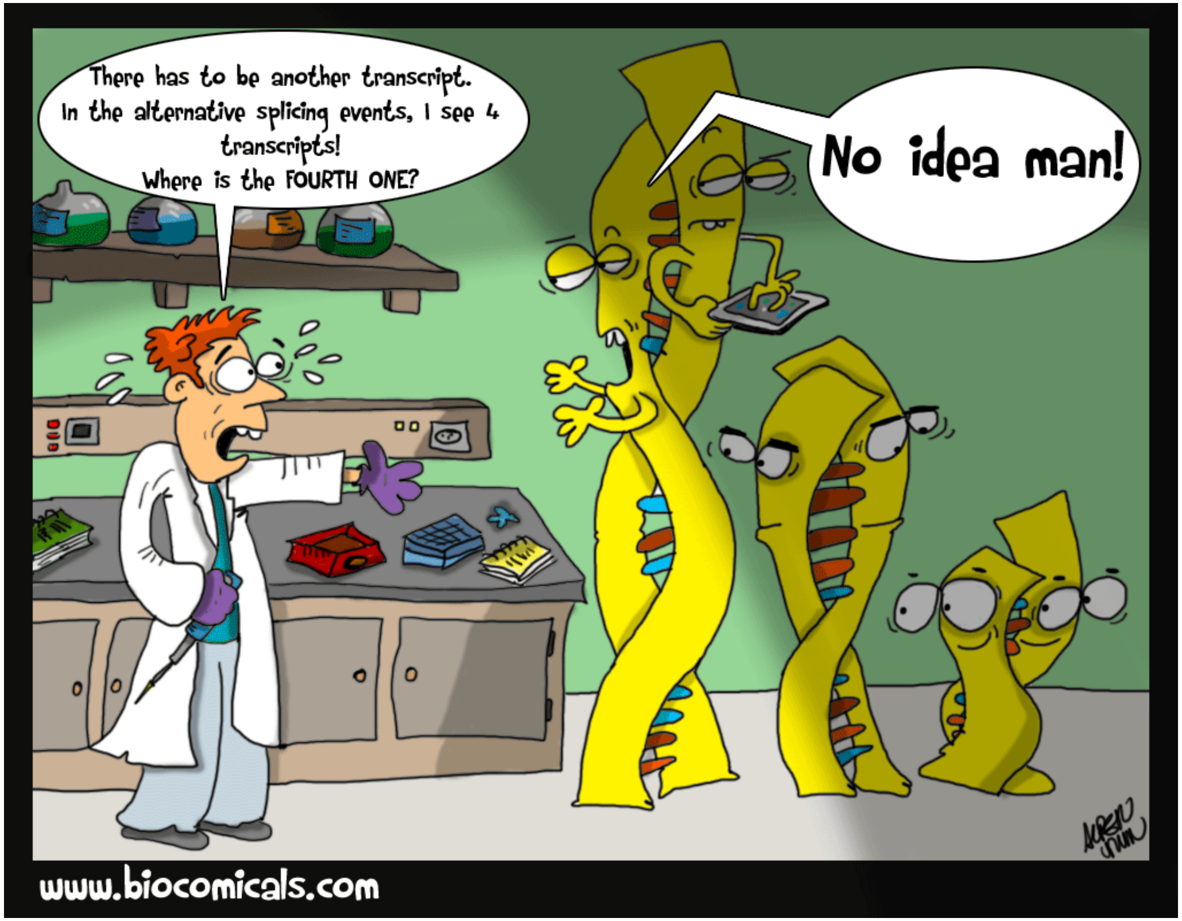
- Differentially expressed genes
- Complex design (more than one varying factor)
- Simple comparison of groups (less than 5 biological replicates per group)





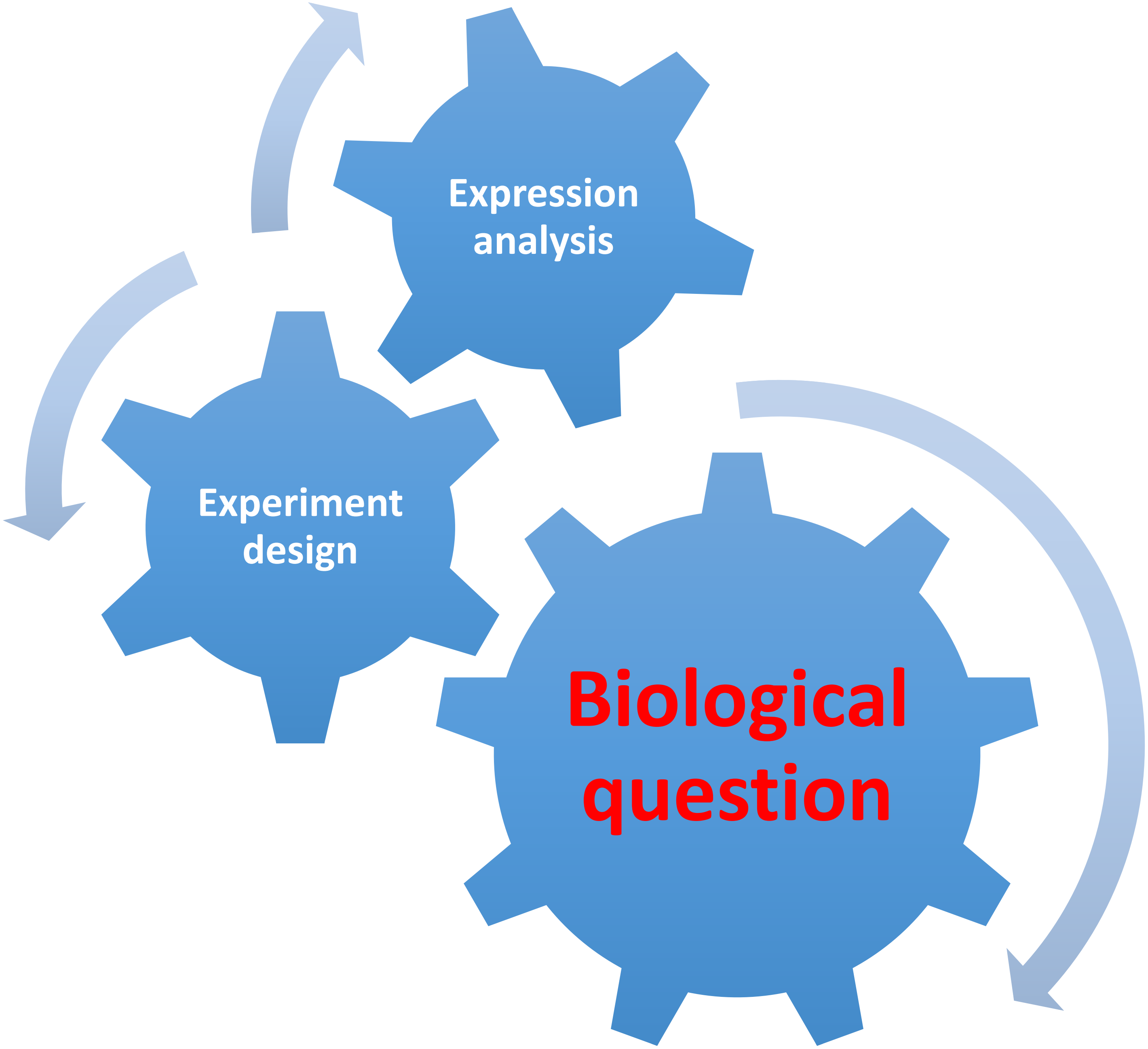





There has to be another transcript.
In the alternative splicing events, I see 4 transcripts!
Where is the FOURTH ONE?
No idea man!
www.biocomicals.com

How to spot biological functions embedded in a gene list?

Bioinformatics task 2



Expression analysis

Experiment design

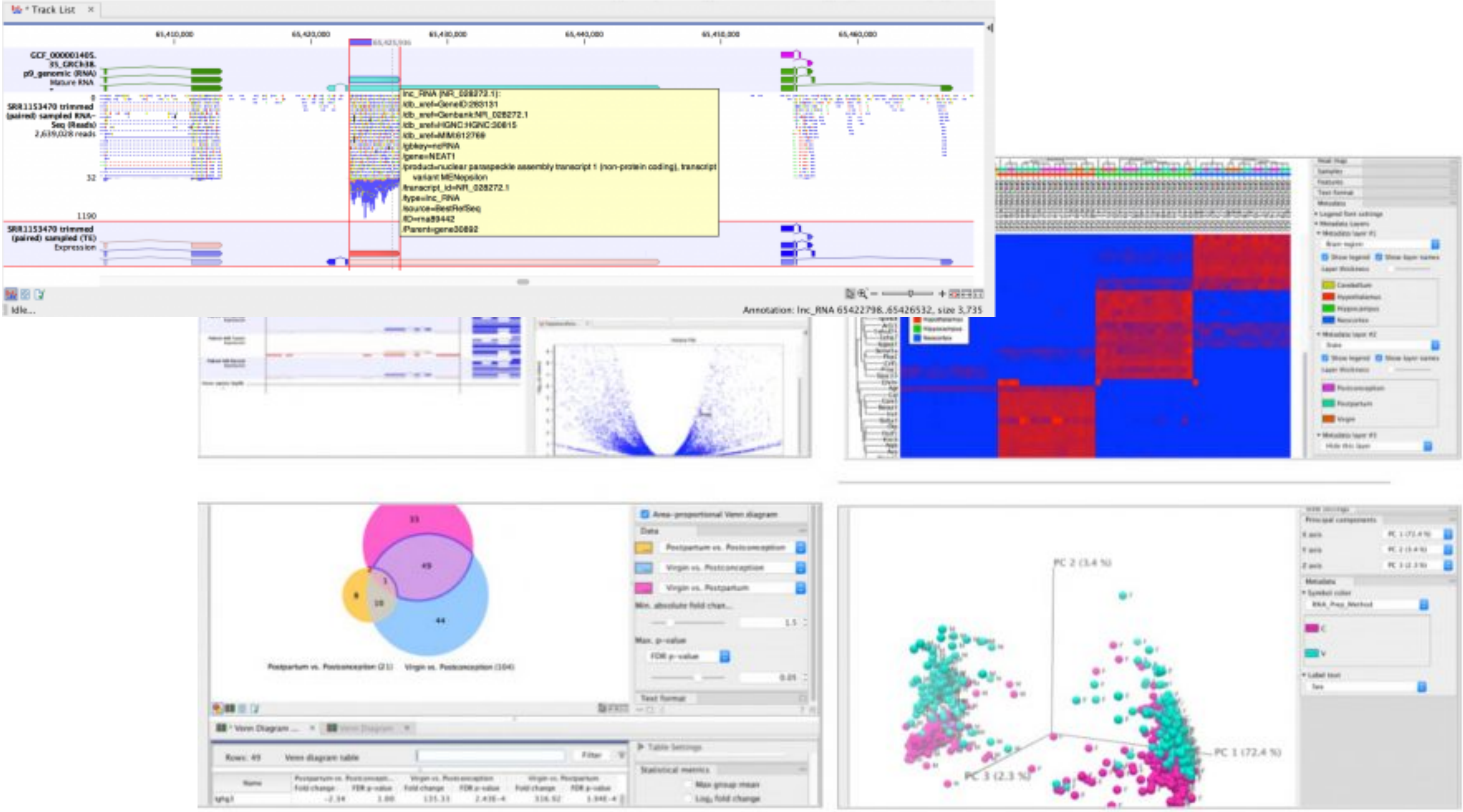
Biological question

Expression Analysis using RNA-Seq

- 10 individuals out of the 726 present in the original dataset [Lin et al., 2016],and have reduced each sample to 250,000 reads mapping to chromosome 2R.
  - The reads from 10 Drosophila samples.
  - An Excel spreadsheet that contains the metadata associated with each individual.
  - SRR_ID is an SRA identifier unique for each individual.
  - DGRP_Number describes the strain of the fly.
  - Sex stated as M for males and F for females.
  - Environment stated as 2 and 3 for different calendar times for collecting the flies.
  - RNA_Prep_Method using QIAGEN RNeasy kit in all cases but following either the
  - centrifuge or the vacuum based protocol.
  - Lane of the sequencer on which the sample was loaded.
  - A workflow to rapidly and efficiently proce.

Comparison of normalization and differential expression analyses using rna-seq data from 726 individual drosophila melanogaster. BMC genomics,17(1):1.

CLC Genomics Workbench 10



lnc_RNA (NR_028272.1):
/db_xref=GeneID:283131
/db_xref=Genbank:NR_028272.1
/db_xref=HGNC:HGNC:30815
/db_xref=MIM:612769
/gbkey=ncRNA
/gene=NEAT1
/product=nuclear paraspeckle assembly transcript 1 (non-protein coding), transcript
variant MENepsilon
/transcript_id=NR_028272.1
/type=lnc_RNA
/source=BestRefSeq
/ID=rna89442
/Parent=gene30892

Idle...

Annotation: lnc_RNA 65422798..65426532, size 3,735

Name	Postpartum vs. Postconception Fold change	Postpartum vs. Postconception FDR p-value	Virgin vs. Postconception Fold change	Virgin vs. Postconception FDR p-value	Virgin vs. Postpartum Fold change	Virgin vs. Postpartum FDR p-value
Igfbp3	-2.34	1.00	135.33	2.43E-4	316.92	1.34E-4

thanks for your attention