Oxford Genomics CENTRE

UNIVERSITY OF OXFORD

RNA-Seq Data Analysis
27-28 November, 2017

Taught module for DPhil programme in
Genomic Medicine and Statistics

Organised and delivered by Bioinformatics Core at WHG:

Helen Lockstone M.Sc.

Ben Wright PhD

Santiago Revale, M.Sc.

wellcome centre human genetics

Oxford Genomics Centre

Oxford Genomics Centre
The Wellcome Trust Centre for Human Genetics

Sequencing overview and RNA-Seq applications

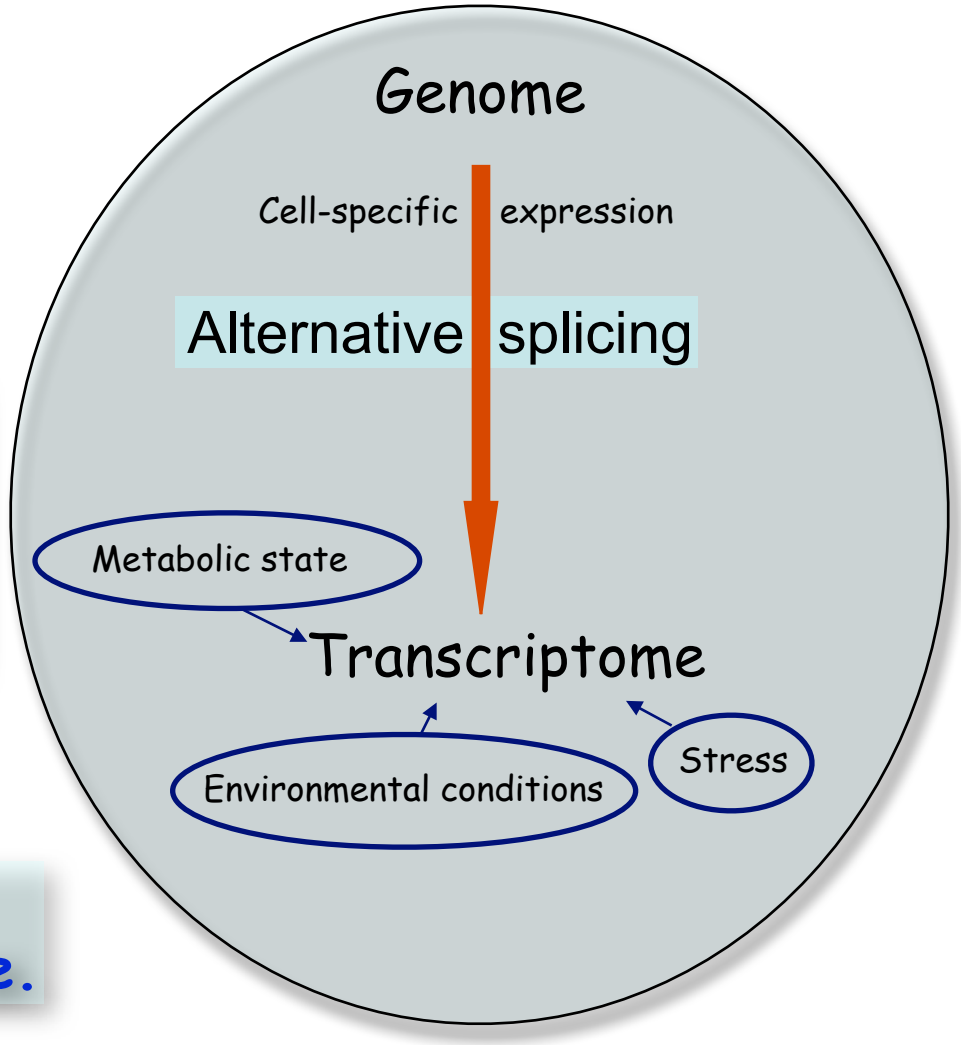
Santiago Revale, M.Sc.
Senior Bioinformatics Analyst, Bioinformatics Core Group

whg

UNIVERSITY OF OXFORD

Transcriptome

Set and quantity of RNAs in a cell, tissue or organism, for a specific developmental stage or physiological condition.



Is dynamic!
It varies in time and space.

RNAseq

Oxford Genomics CENTRE

It's a technology that uses the NGS capabilities to reveal a snapshot of the presence and quantity of RNA in a sample at a given moment in time.

wellcome centre human genetics

UNIVERSITY OF OXFORD

RNAseq: it's always about the goals!

Oxford Genomics CENTRE

At RNA transcript level, it provides the ability to:

- ✓ look at alternative gene spliced transcripts,
- ✓ post-transcriptional modifications,
- ✓ gene fusion,
- ✓ mutations/SNPs,
- ✓ changes in gene expression.

Can look at different populations of RNA to include:

- ✓ total RNA,
- ✓ mRNA,
- ✓ small RNA (miRNA, tRNA, ribosomal profiling, etc.)

Can be used to:

- ✓ determine exon/intron boundaries,
- ✓ verify or amend previously annotated 5' and 3' gene boundaries.

Common main goals

Oxford Genomics CENTRE

- Catalog all species of transcripts, e.g. messengers, non-coding, small, etc.
- Determine the transcriptional structure of genes, in terms of their starting sites, 5’ and 3’ ends, splicing patterns and other post-transcriptional modifications.
- Quantify the changes in the expression levels of each transcript during development and/or in different conditions.

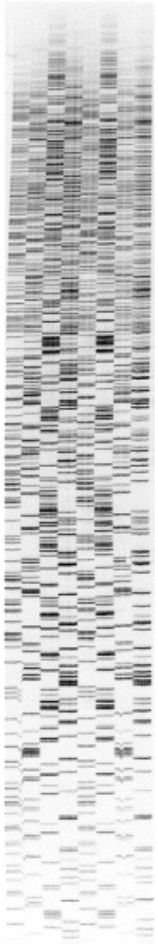
Practical approaches to RNAseq

Oxford Genomics CENTRE

- **Enriched for polyadenylated transcripts (polyA)**
  PolyA enrichment is good for mature protein coding transcripts – requires good quality total RNA.
- **Depleted for ribosomal RNA transcripts (ribodepleted)**
  Ribodepletion is good for including all non-polyadenylated RNA – also good for degraded material.
- **Amplified from low input material (SMARTer)**
  SMARTer is good for low input samples, down to a single cell – requires good quality total RNA or fresh cell lysate.
- **Small RNA species**
  Small RNA require a separate preparation but can be run in parallel to polyA.
- **3' mRNA**
  3'mRNA is good to minimise costs/maximise sequencing lane capacity – some tolerance to degraded material.

Genomics Platforms

Oxford Genomics CENTRE



wellcome centre human genetics

UNIVERSITY OF OXFORD

Sequencing Outputs

Oxford Genomics CENTRE

Illumina HiSeq 4000

Output range	105 – 1500 Gb	
Reads per run	2.1 – 5 billion	
Max. read length	2 x 150 bp	
Run time	< 1 – 3.5 days	
Samples sequenced per:	Flowcell	Lane
polyA	80	10
Ribodepleted	40	5
3’ mRNA	384	48
CHIPseq	80	10

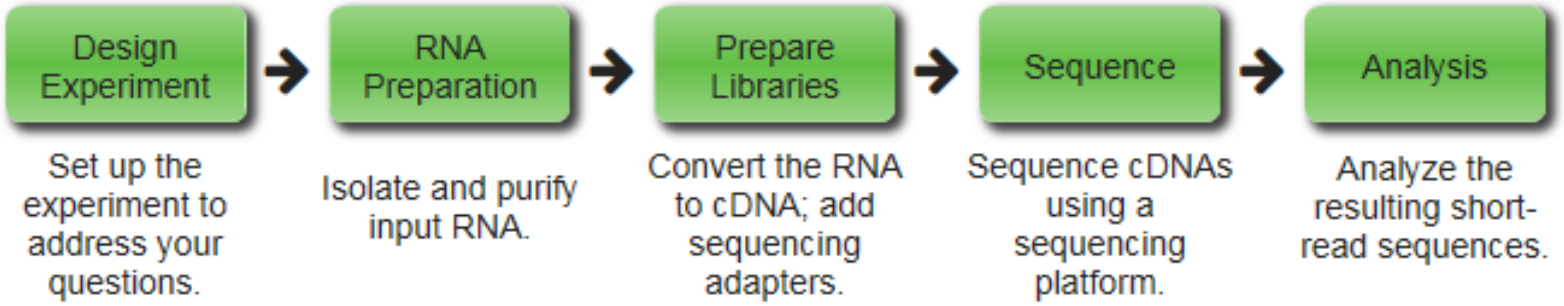
Illumina HiSeq 2500

Output range	9 – 1000 Gb	
Reads per run	0.3 – 4 billion	
Max. read length	2 x 250 bp	
Run time	< 1 – 6 days	
Samples sequenced per:	Run	Lane
Small RNA	168	21

wellcome centre human genetics

UNIVERSITY OF OXFORD

Typical RNAseq experiment



Experimental Design

How is an experiment designed?

Oxford Genomics CENTRE

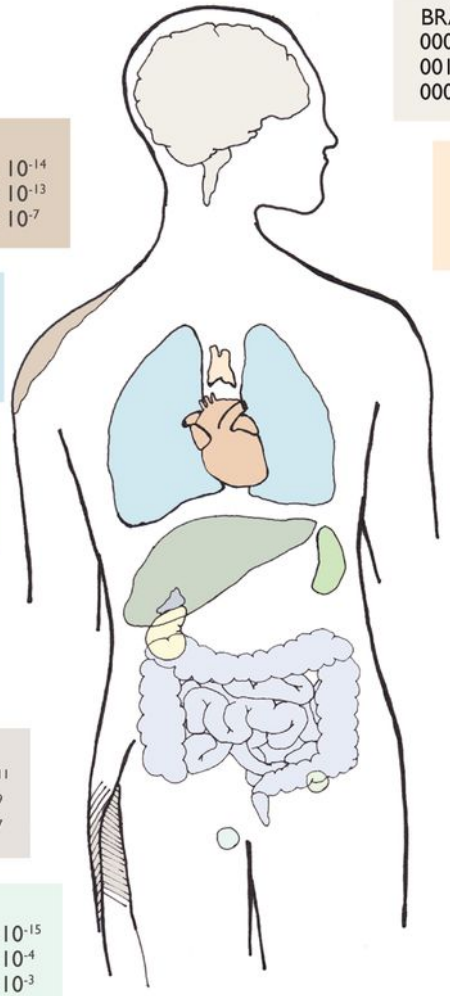
While a good design does not guarantee a successful experiment, a suitably bad design guarantees failure.

wellcome centre human genetics

UNIVERSITY OF OXFORD

Tissue specificity

Oxford Genomics CENTRE



BRAIN

0007268	synaptic transmission	8.9×10^{-41}
0016358	dendrite morphogenesis	1.2×10^{-10}
0007611	learning or memory	7.9×10^{-6}

SKIN

0031424	keratinization	2.9×10^{-14}
0006955	immune response	3.5×10^{-13}
0031069	hair follicle morphogenesis	4.1×10^{-7}

THYMUS

0019882	antigen presentation	7.1×10^{-21}
0045059	positive thymic T cell selection	9.8×10^{-8}
0045060	negative thymic T cell selection	2.6×10^{-7}

LUNG

0030324	lung development	6.2×10^{-16}
0006954	inflammatory response	2.1×10^{-15}
0043330	response to exogenous dsRNA	6.2×10^{-6}

HEART

0006099	tricarboxylic acid cycle	2.5×10^{-15}
0045214	sarcomere organization	7.5×10^{-12}
0008016	regulation of heart contraction rate	8.3×10^{-7}

ADRENAL

0006700	C21-steroid hormone biosynthesis	4.6×10^{-8}
0017157	regulation of exocytosis	4.2×10^{-4}
0006584	catecholamine metabolism	1.4×10^{-3}

LIVER

0008203	cholesterol metabolism	2.6×10^{-8}
0007596	blood coagulation	2.0×10^{-7}
0000050	urea cycle	5.0×10^{-5}

KIDNEY

0001822	kidney development	1.4×10^{-6}
0007588	excretion	1.3×10^{-3}
0001736	establishment of planar polarity	2.9×10^{-3}

SPLEEN

0050766	positive regulation of phagocytosis	4.5×10^{-9}
0030183	B cell differentiation	1.5×10^{-7}
0030217	T cell differentiation	2.6×10^{-7}

MUSCLE

0006941	striated muscle contraction	7.7×10^{-11}
0005977	glycogen metabolism	1.8×10^{-9}
0045445	myoblast differentiation	8.0×10^{-7}

INTESTINE

0006955	immune response	7.0×10^{-13}
0007586	digestion	9.3×10^{-5}
0050892	intestinal absorption	4.6×10^{-4}

TESTIS

0007059	chromosome segregation	9.1×10^{-15}
0007276	gametogenesis	8.1×10^{-4}
0006349	imprinting	1.5×10^{-3}

OVARY

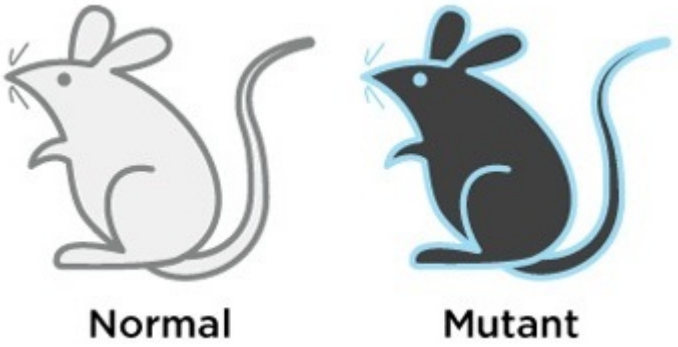
0007059	chromosome segregation	1.0×10^{-12}
0007276	gametogenesis	8.6×10^{-8}
0006349	imprinting	3.5×10^{-5}

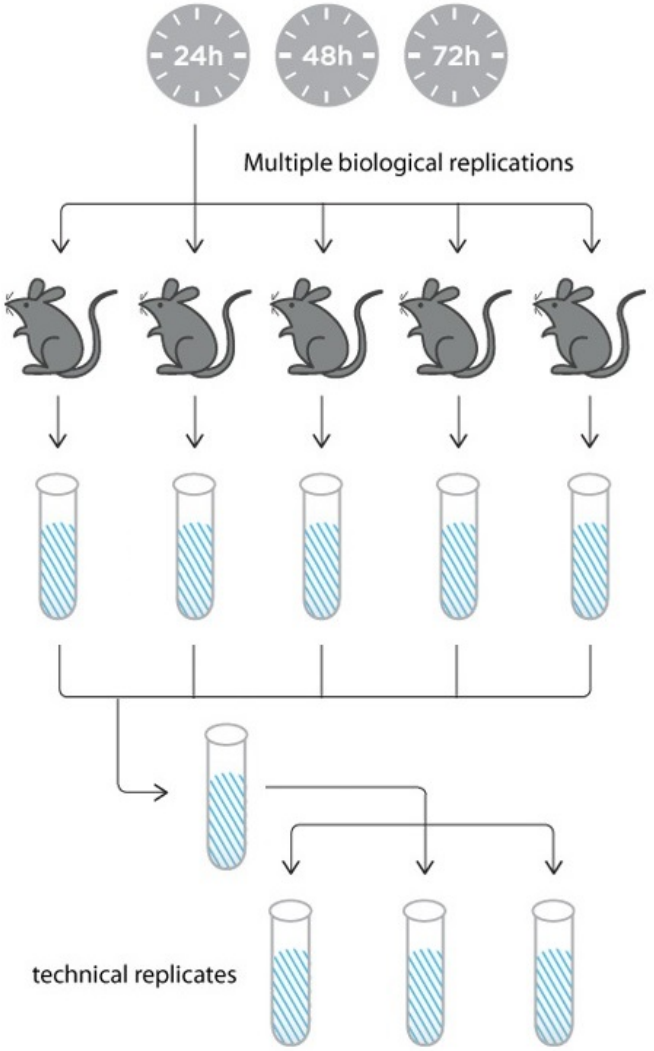
wellcome centre human genetics

UNIVERSITY OF OXFORD

Time dependent

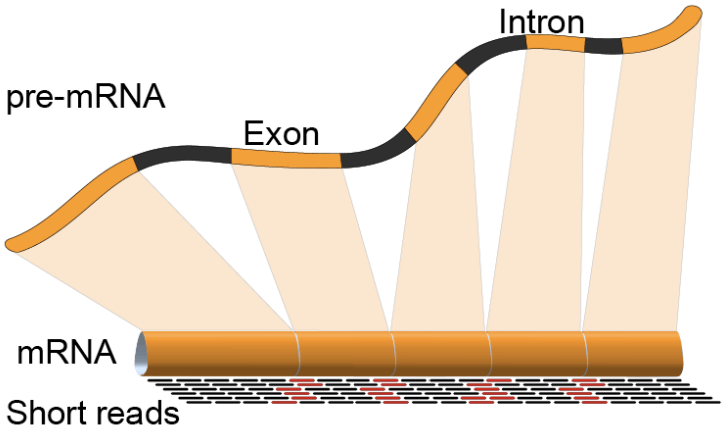
Replicates (technical / biological)



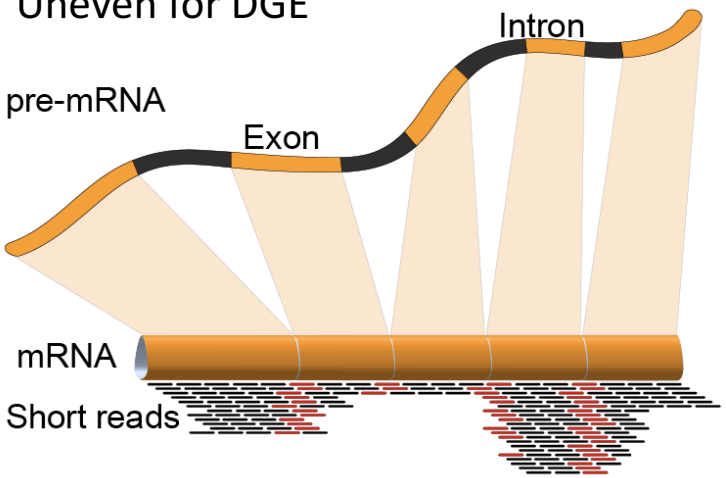


Coverage

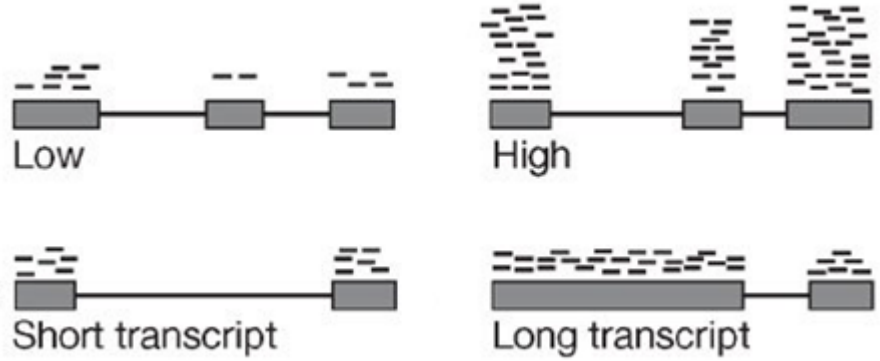
- Even for annotation



- Uneven for DGE



- Target transcript properties (low abundant vs high abundant transcripts)



- Allele might not be detected (not in the genome/not being expressed)
- Estimate expression of each allele

Subjectivity of the analysis

- ✓ Multitude of algorithms and pipelines available.
- ✓ Most approaches correct, but have to be tailored to the needs of the investigators in order to better capture the desired effect.

Data management & Downstream interpretation of the data

Oxford Genomics CENTRE

- ✓ Several Gigabytes (70-75 Gb Avg)
- ✓ Different layers of interpretations have to be considered (e.g. biological, clinical, regulatory functions, etc.)

wellcome centre human genetics

UNIVERSITY OF OXFORD

Recommendations based upon experimental objectives.

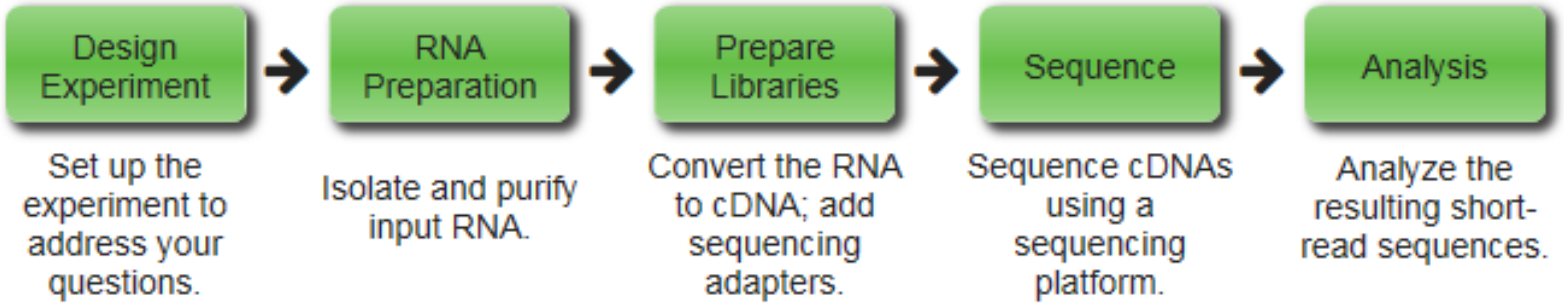
Oxford Genomics CENTRE

Criteria	Annotation	Differential Gene Expression
Biological replicates	Not necessary but can be useful	Essential
Coverage across the transcript	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not as important; however the only reads that can be used are those that are uniquely mappable.
Depth of sequencing	High enough to maximize coverage of rare transcripts and transcriptional isoforms	High enough to infer accurrate statistics
Role of sequencing depth	Obtain reads that overlap along the length of the transcript	Get enough counts of each transcript such that statistical inferences can be made
DSN	Useful for removing abundant transcripts so that more reads come from rarer transcripts	Not recommended since it can skew counts
Stranded library prep	Important for de Novo transcript assembly and identifying true anti-sense trancripts	Not generally required especially if there is a reference genome
Long reads (>80 bp)	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not generally required especially if there is a reference genome
Paired-end reads	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not important

wellcome centre human genetics

UNIVERSITY OF OXFORD

Typical RNAseq experiment



Typical RNAseq experiment

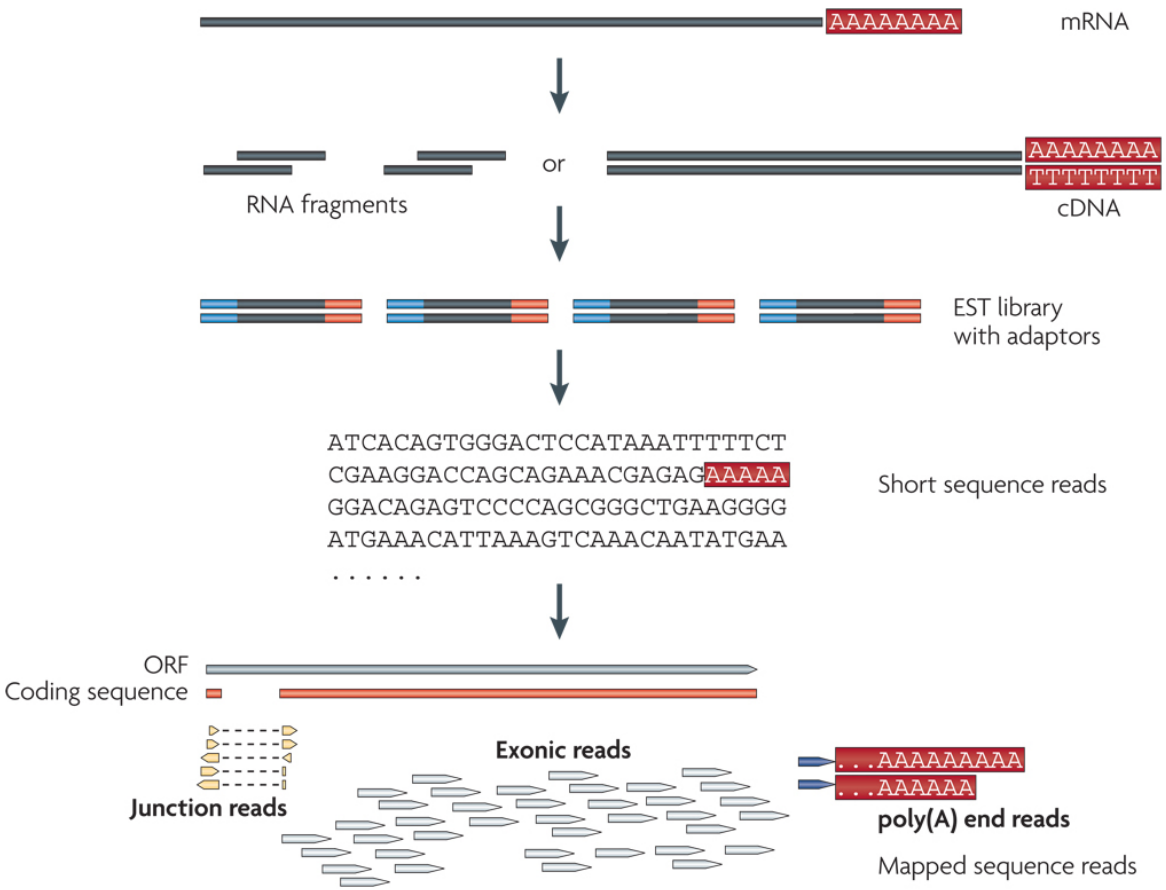
Oxford Genomics CENTRE

RNA Preparation

Isolate and purify input RNA.

Prepare Libraries

Convert the RNA to cDNA; add sequencing adapters.

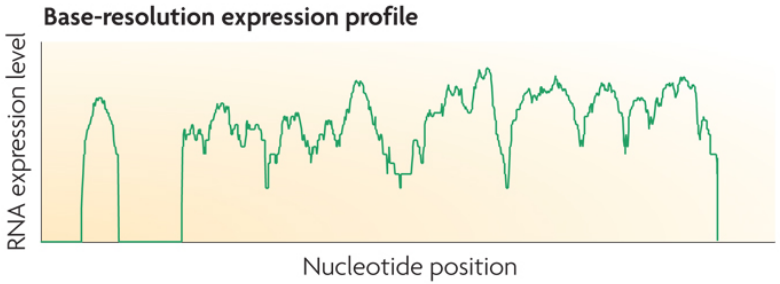


Sequence

Sequence cDNAs using a sequencing platform.

Analysis

Analyze the resulting short-read sequences.



wellcome c
human gene

UNIVERSITY OF OXFORD

Data Analysis

Stereotypical RNA-seq analysis pipeline

Oxford Genomics CENTRE

1. Demultiplex, filter, and trim sequencing reads.
2. Normalize sequencing reads (if performing *de novo* assembly)
3. *de novo* assembly of transcripts (if ref genome is not available)
4. Map sequencing reads to reference genome or transcriptome
5. Annotate transcripts assembled or to which reads have been mapped
6. "Count" mapped reads to estimate transcript abundance
7. Perform statistical analysis to identify differential expression (or differential splicing) among samples or treatments
8. Perform multivariate statistical analysis/visualization to assess transcriptome-wide differences among samples

wellcome centre human genetics

UNIVERSITY OF OXFORD

1. Read processing

Oxford Genomics CENTRE

Table 5.1 Read Processing Software

For *de novo* assembly

Software	De-mulitplexing	Adaptor Trimming	Quality Filtering/ Trimming	K-mer Filtering	K-mer Normalization
FASTX-Toolkit	✔	✔	✔		
Goby	✔	✔			
khmer				✔	✔
NGS_backbone		✔	✔		
Stacks	✔	✔	✔	✔	✔
trimmomatic		✔	✔		
biopieces	✔	✔	✔		

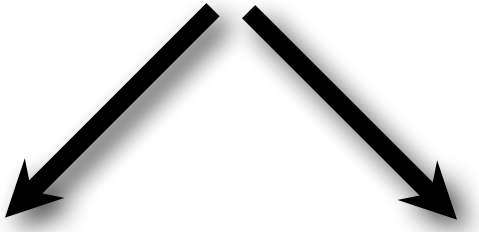
wellcome centre human genetics

UNIVERSITY OF OXFORD

de novo assembly or reference mapping?

Oxford Genomics CENTRE

When to use each?

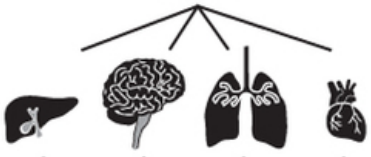


de novo	**reference**
(do not know the transcriptome) (main goal is to discover NOT to quantify)	(do know the transcriptome) (main goal is to quantify NOT to discover)

wellcome centre human genetics

UNIVERSITY OF OXFORD

Protocol for *de novo* RNAseq



Reads
(per sample)

Combine reads

Normalization?

Assembled
transcripts
(all samples)

Abundance estimation

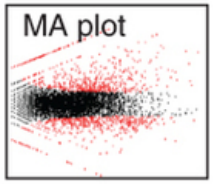
De novo assembly

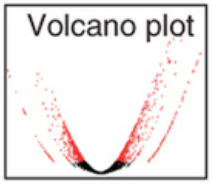
Assembled
transcripts

Identify differentially expressed transcripts

Identify coding regions

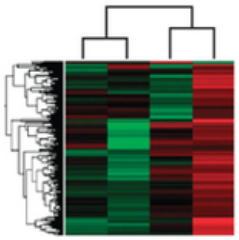
Functional
annotation

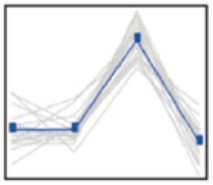




Annotated transcriptome

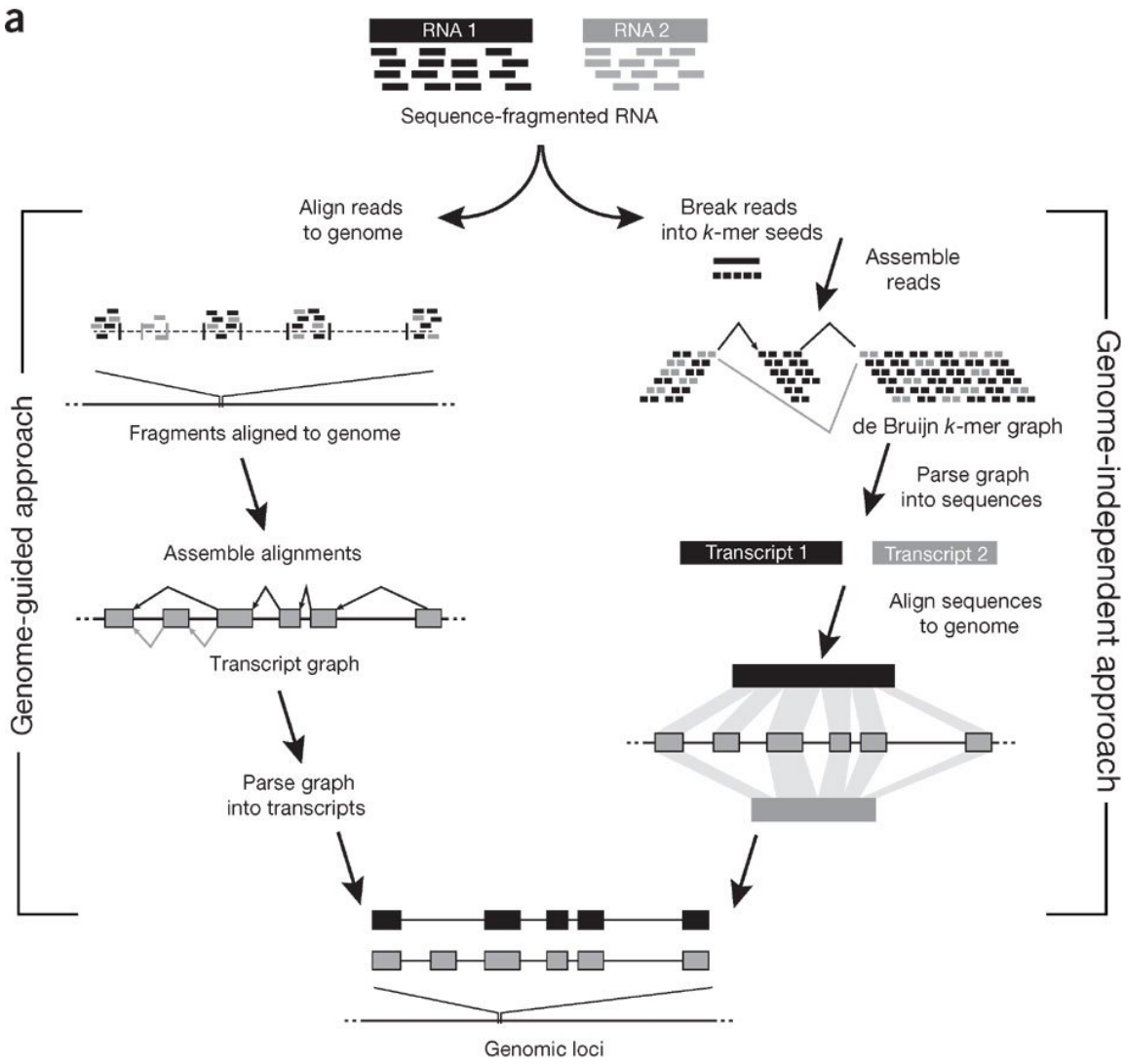
Expression patterns, transcript clusters





Protocol for reference-based RNAseq

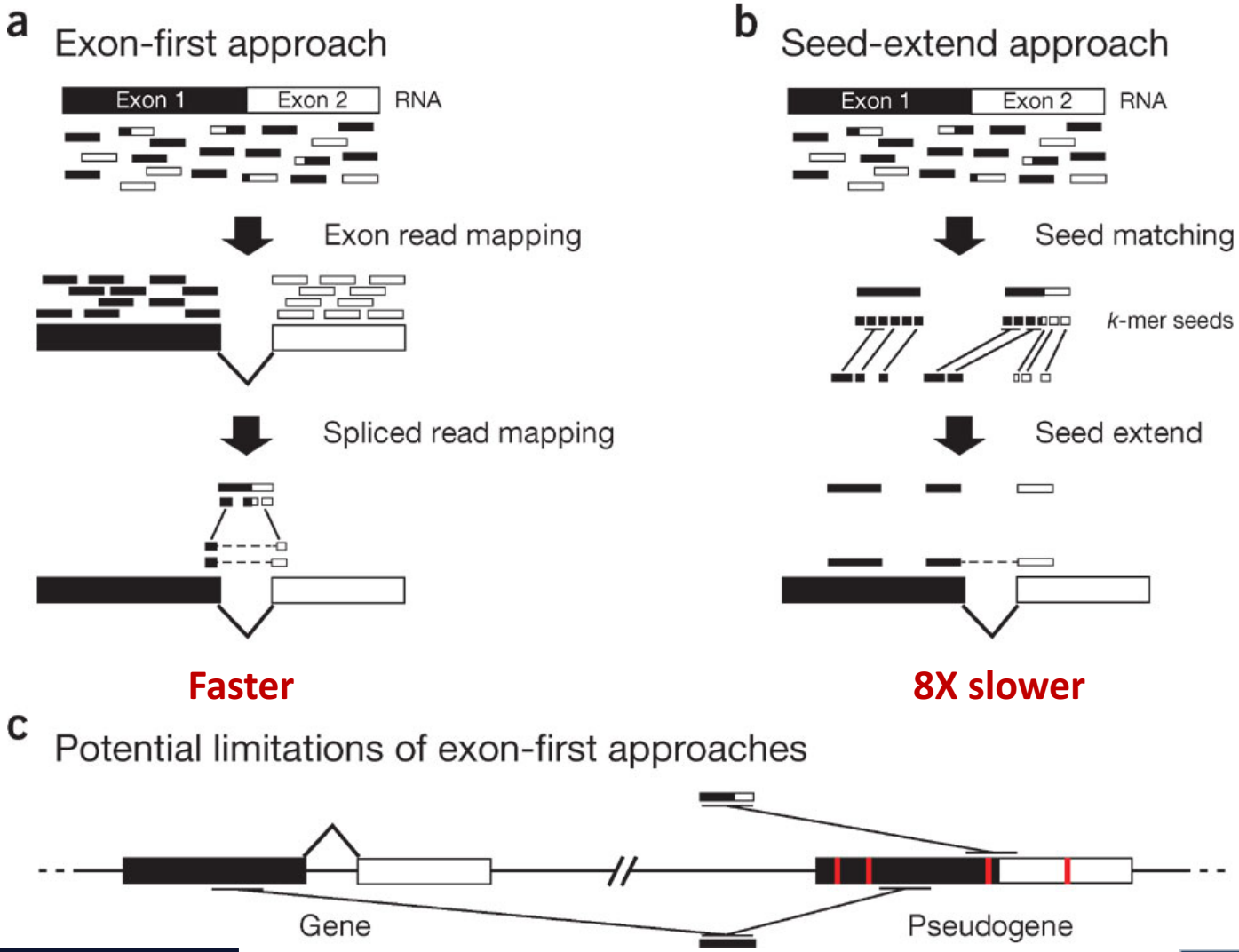
3. *de novo* transcriptome assembly



4. Map RNA data against genome

Oxford Genomics CENTRE

e.g. TopHat



e.g. GSNAP

Faster

8X slower

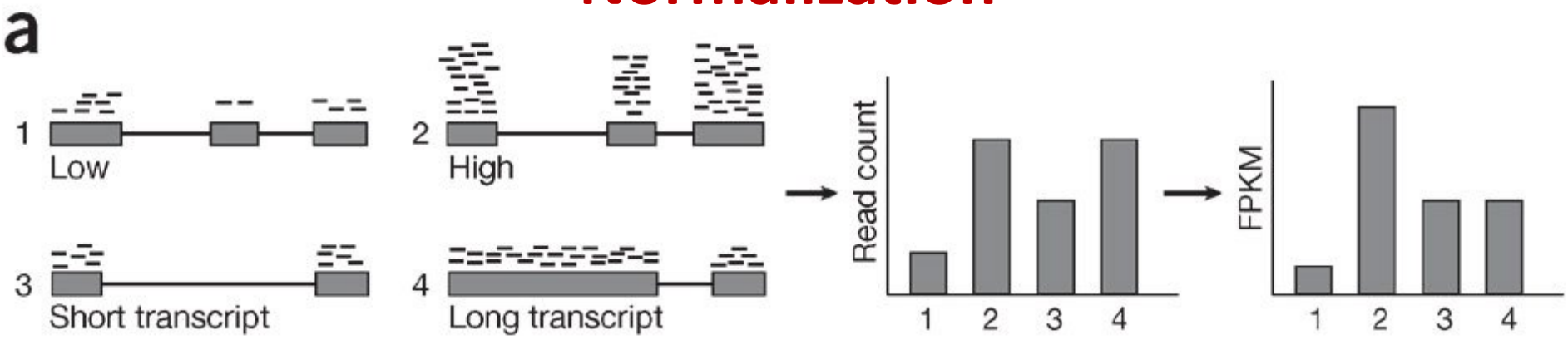
wellcome centre human genetics

UNIVERSITY OF OXFORD

6. Expression quantification

Oxford Genomics CENTRE

Normalization



Influence of length: Counts are proportional to the transcript length times the mRNA expression level.

Influence of sequencing depth: The higher sequencing depth, the higher counts.

"Gene counts" should be corrected in order to minimize these biases: **normalization**.

Statistical model should take into account "length" and "sequencing depth".

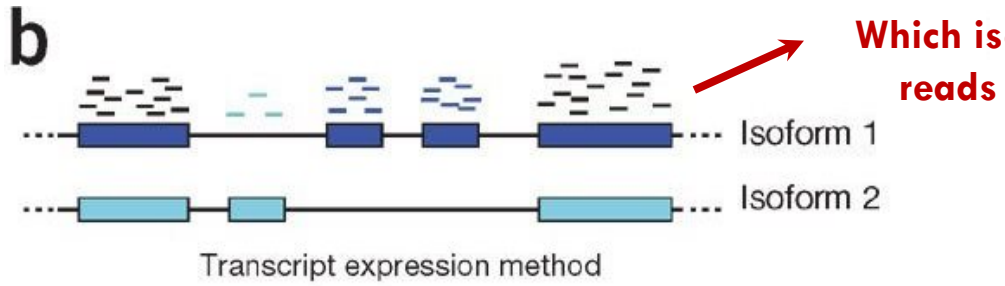
Some normalization methods

- **RPKM** (Mortazavi *et al.*, 2008): Counts are divided by the transcript length (kb) times the total number of millions of mapped reads.
- **Upper-quartile** (Bullard *et al.*, 2010): Counts are divided by upper-quartile of counts for transcripts with at least one read times the square root of transcript length.
- **Quantiles**, as in microarray normalization (Irizarry *et al.*, 2003).
- **FPKM** (Trapnell *et al.*, 2010): Instead of counts, Cufflinks software generates FPKM values (Fragments Per Kilobase of exon per Million fragments mapped) to estimate gene expression, which are analogous to RPKM.

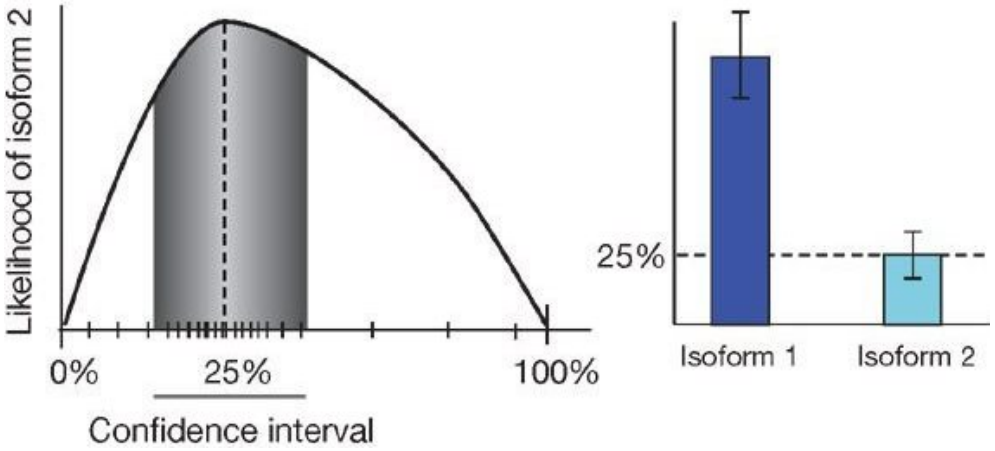
6. Expression quantification

Oxford Genomics CENTRE

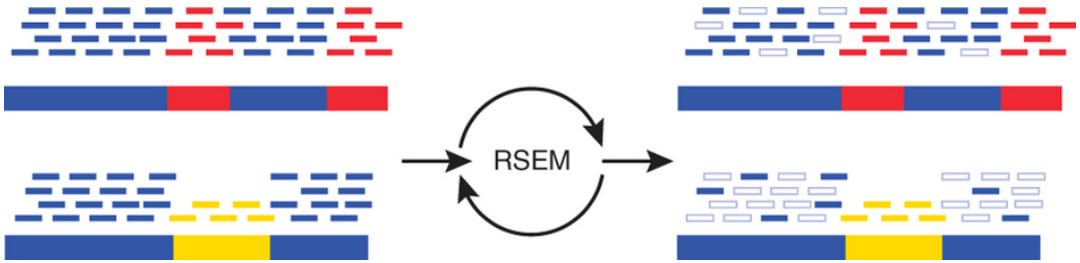
Isoforms



Which isoform do this reads belong to?



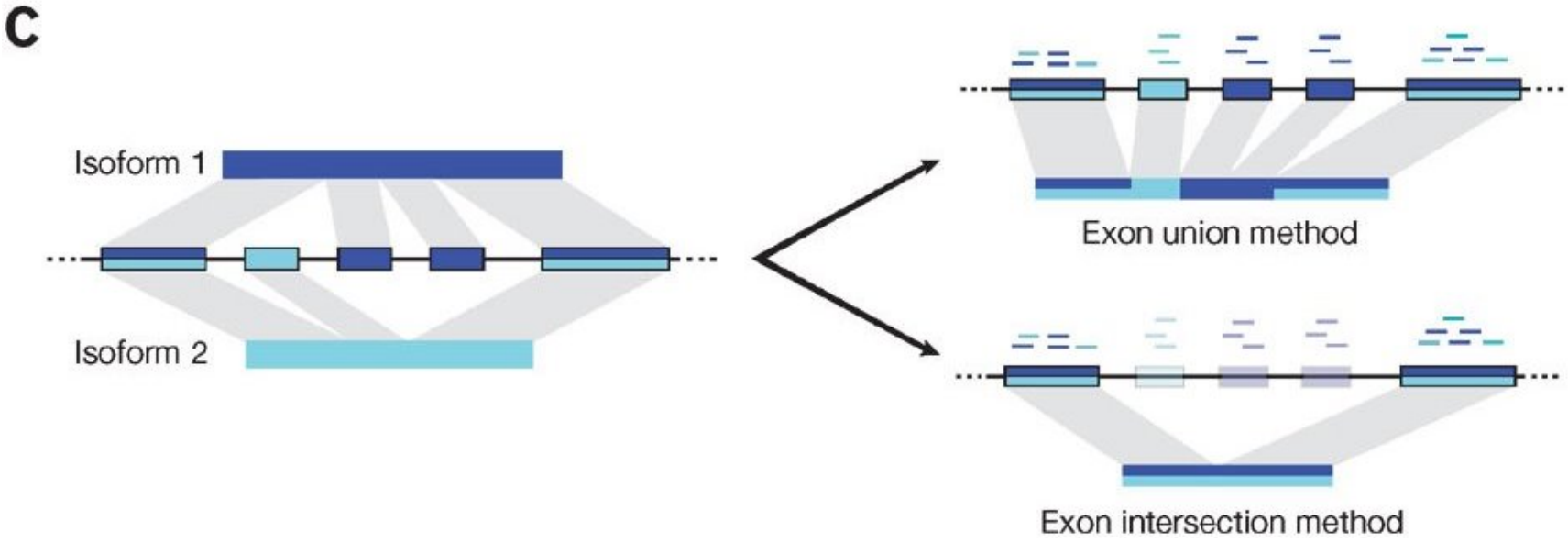
e.g. Cufflinks, RSEM



6. Expression quantification

Oxford Genomics CENTRE

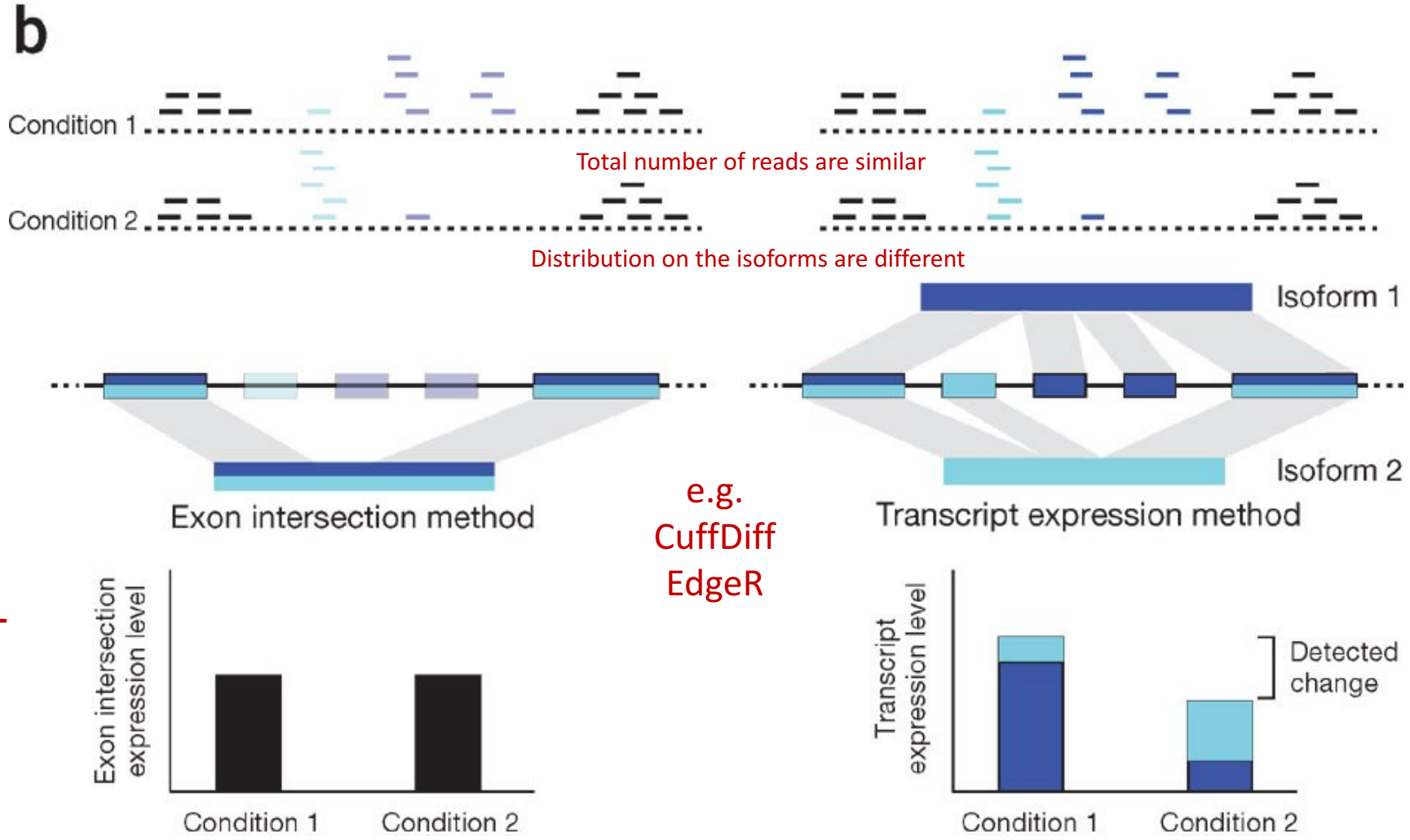
Gene expression



wellcome centre human genetics

UNIVERSITY OF OXFORD

7. Differential expression analysis



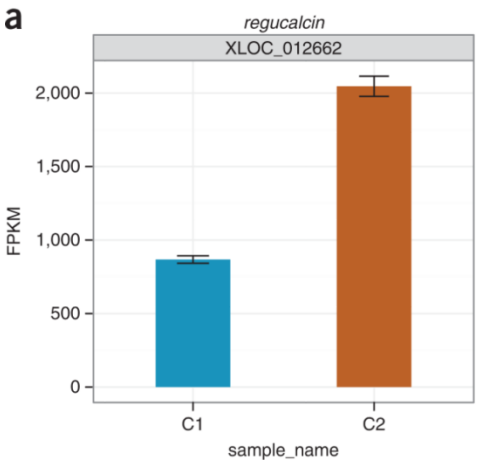
Gene expression level

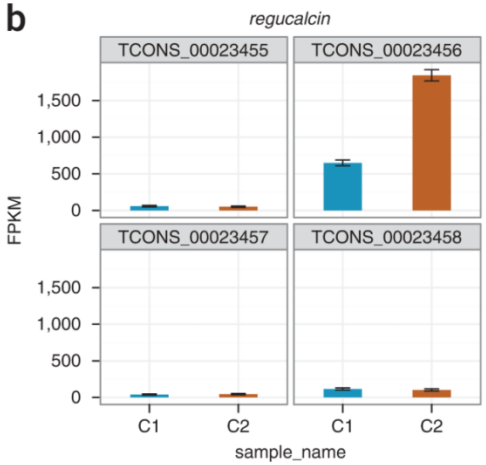
e.g.
CuffDiff
EdgeR

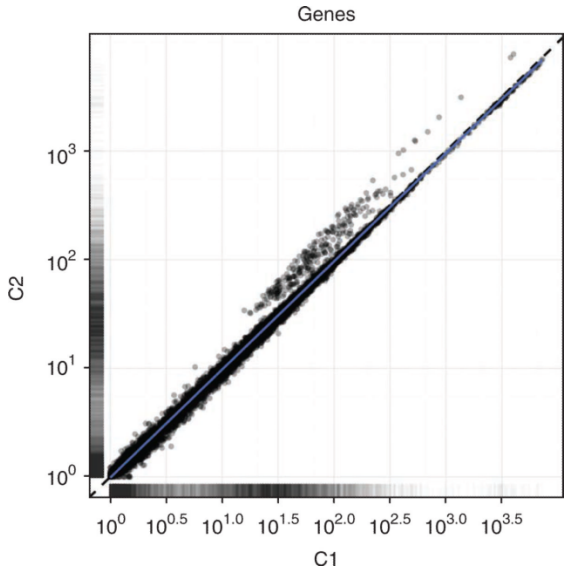
Transcript expression level

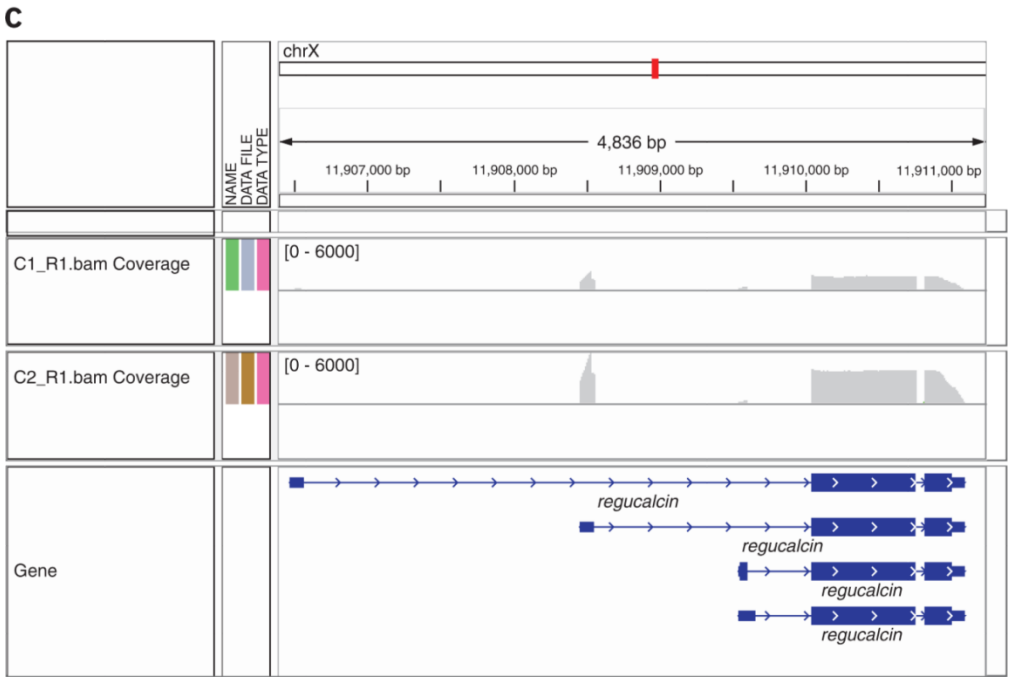
8. Data visualization

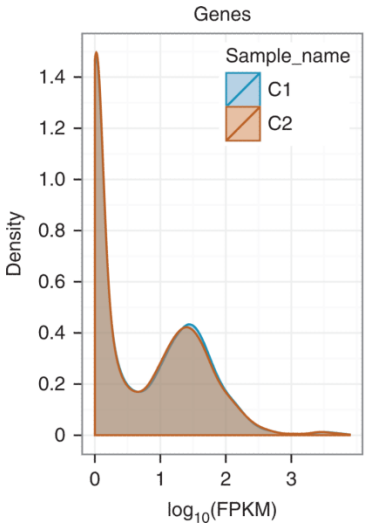
Oxford Genomics CENTRE

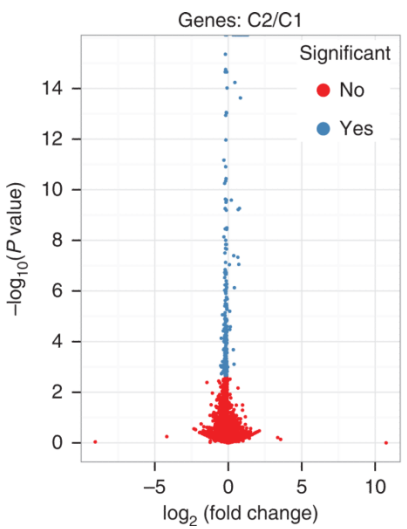






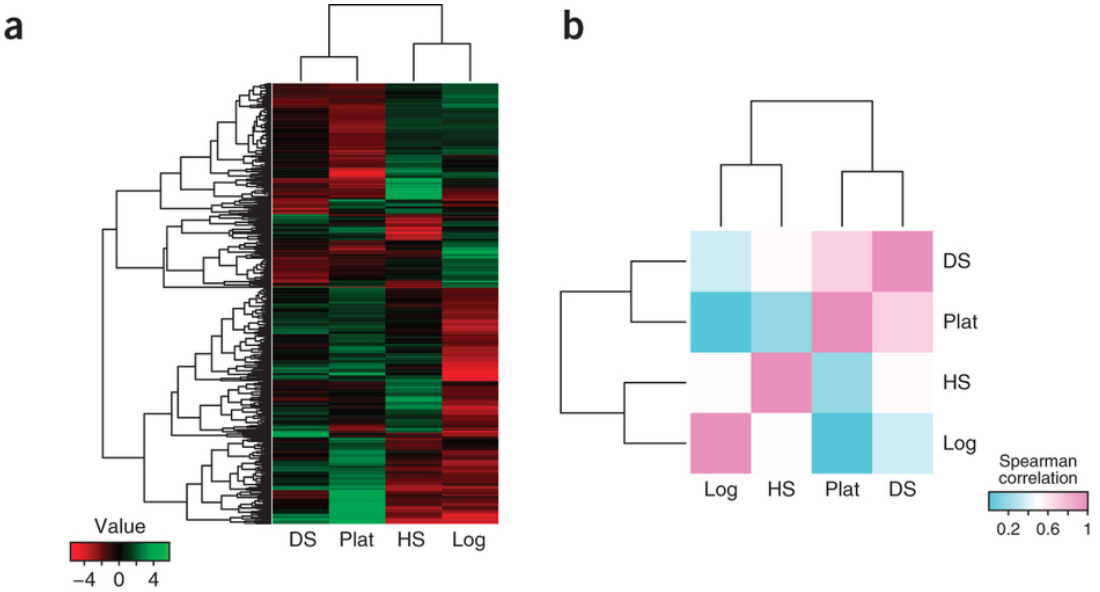






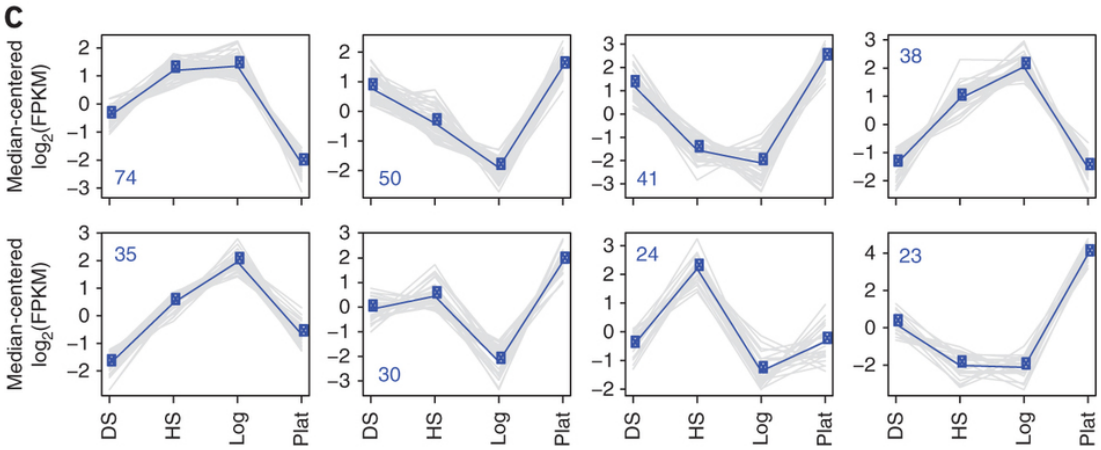
8. Data visualization

Hierarchical clustering of transcripts and samples.



Spearman correlation matrix

Transcript clusters extracted from the hierarchical clustering.

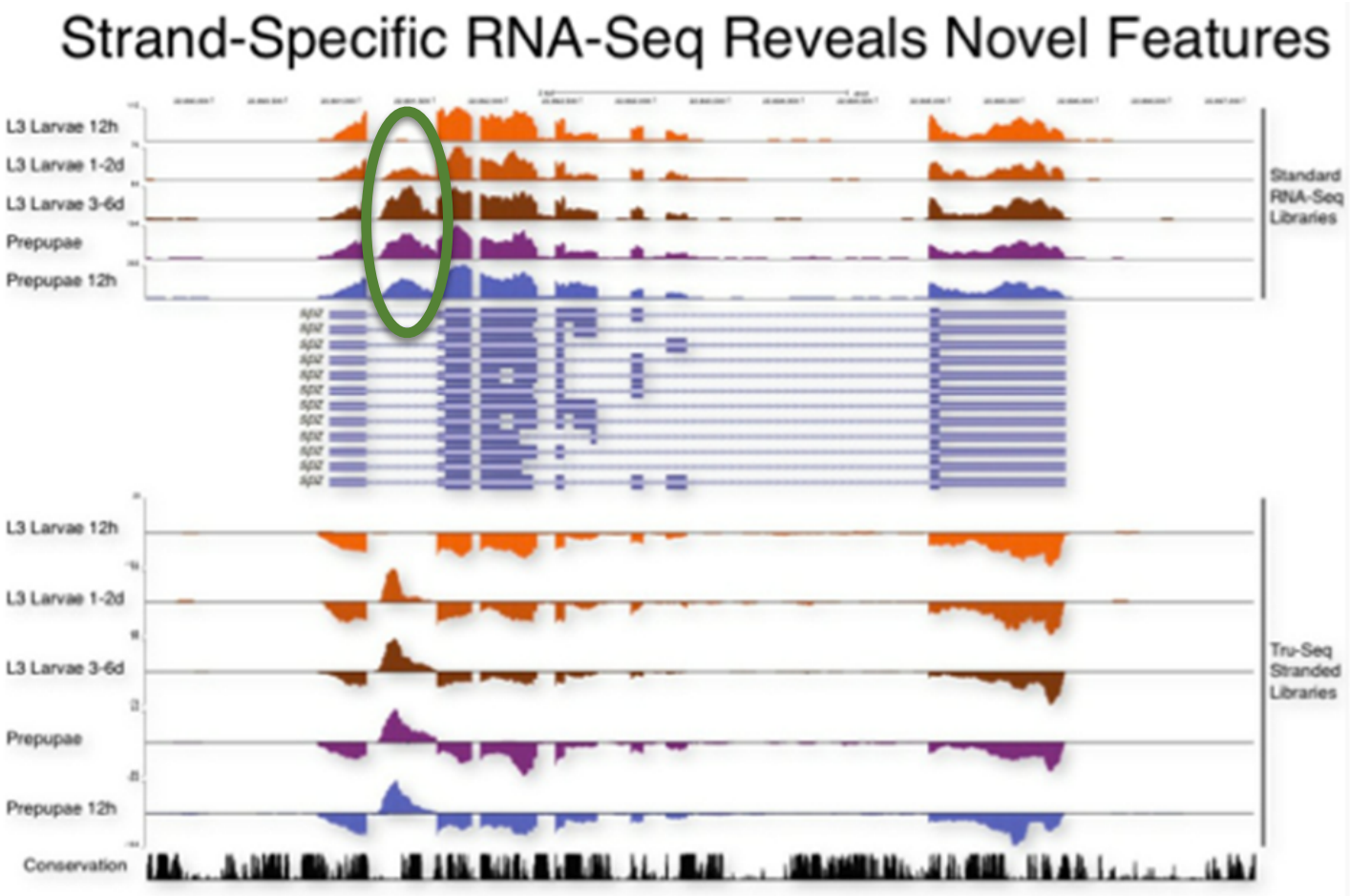


wellcome centre human genetics

Oxford Genomics CENTRE

UNIVERSITY OF OXFORD

8. Data visualization



spz gene Drosophila

transcripts first intron
missinterpreted as new exon

spz gene Drosophila

transcripts first intron
are from opposite strand
is a new protein

Some applications:

1. Refine Transcriptomes
2. Focus on Non-Coding RNAs

References

- Cresko Lab, University of Oregon. RNA-seqlopedia. http://rnaseq.uoregon.edu/
- Garber M , Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Met. 2011; 8: 469–477.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011; 29: 644–652.
- Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013; 8: 1494–1512.
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan; 10(1): 57–63.
- List of RNAseq bioinformatic tools. http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools