Oxford Genomics CENTRE

UNIVERSITY OF OXFORD

RNA-Seq Data Analysis
7-8th June, 2018

Organised and delivered by Bioinformatics Core at WHG:

Helen Lockstone M.Sc.

Ben Wright PhD

Eshita Sharma PhD

Santiago Revale, M.Sc.

wellcome centre human genetics

What we do when we do RNAseq?

Oxford Genomics CENTRE

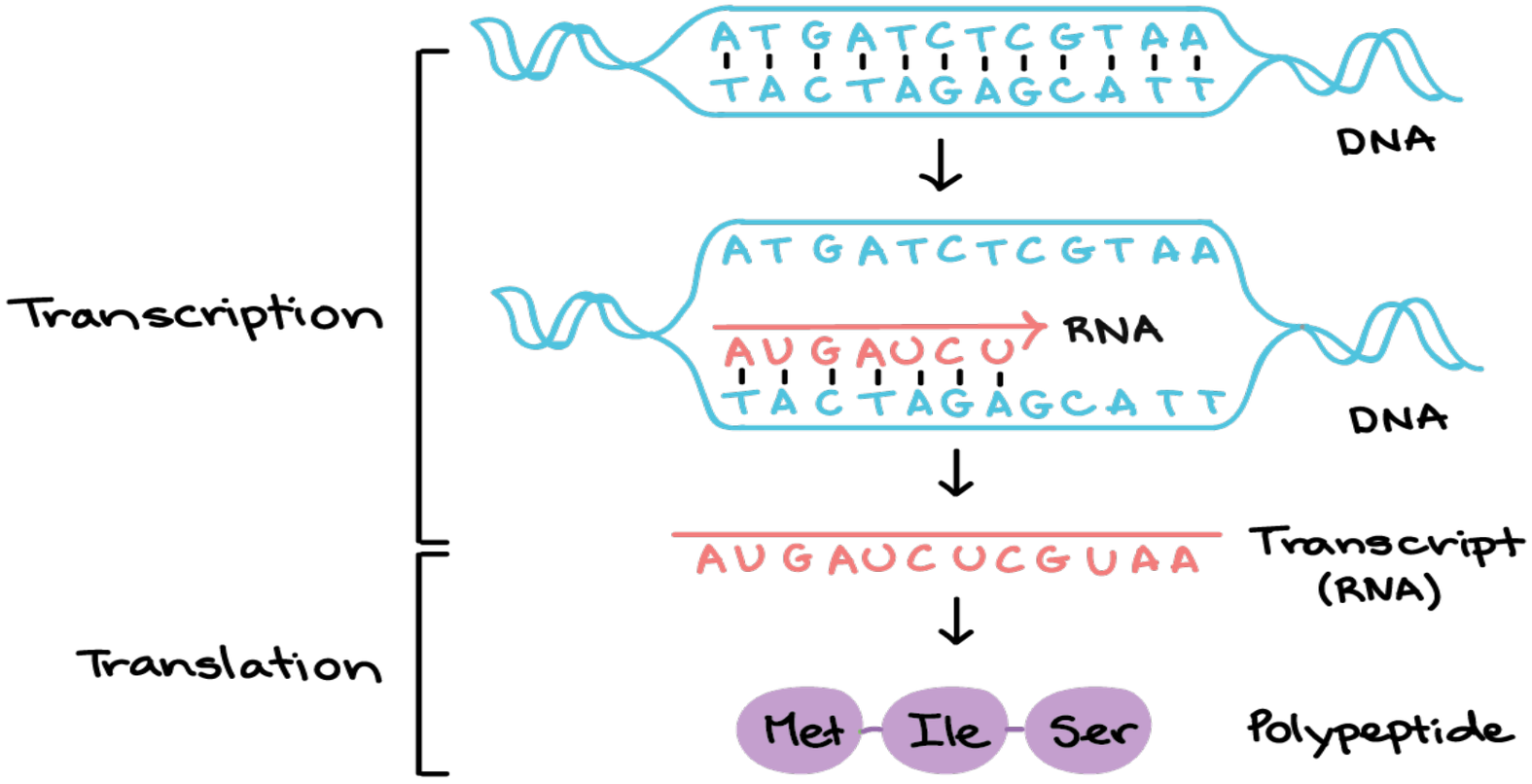
- **What it is?**
- **Scope of RNAseq**
- **Usual approaches for RNAseq library preparation?**
- **Considerations for RNAseq experiments**
- **General methods for RNAseq data analysis.**

Eshita Sharma,
eshita.sharma@well.ox.ac.uk
Research Associate in Funtional Genomics, Bioinformatics Core

wellcome centre human genetics

UNIVERSITY OF OXFORD

What it is?
RNA - Mid-point of the information cascade



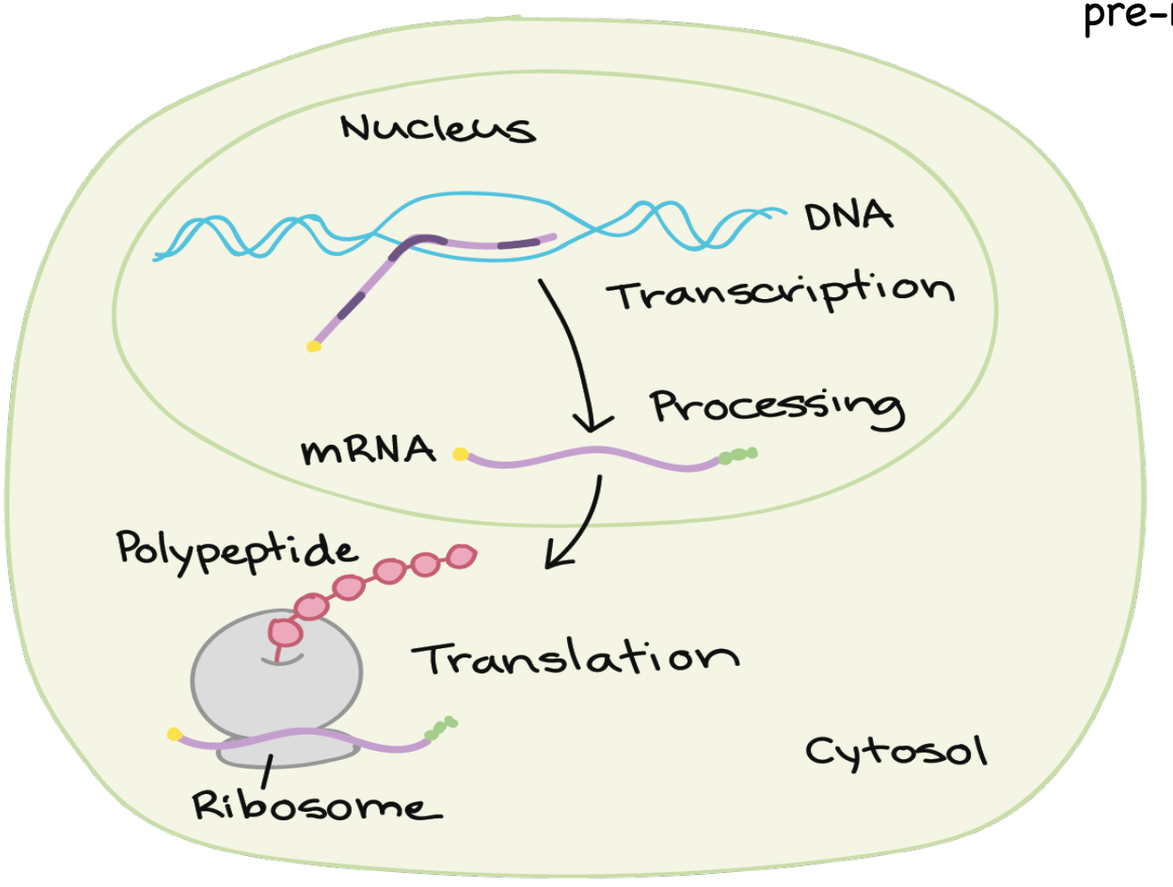
- Image credit Khan Academy Open Courses

We identify the mRNA molecule and extrapolate the knowledge to say something about the proteins and DNA

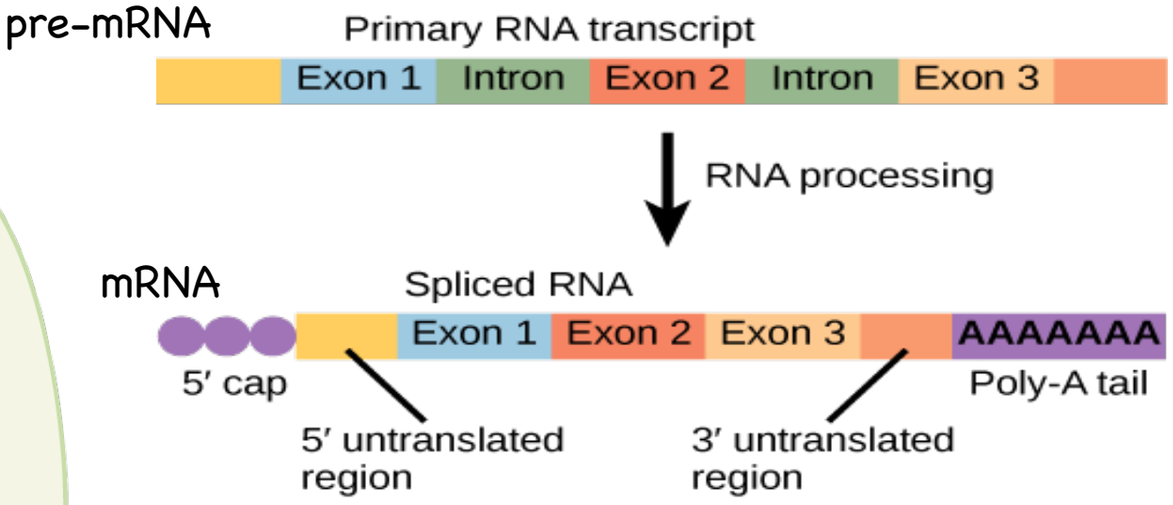
The RNA repertoire or Transcriptome

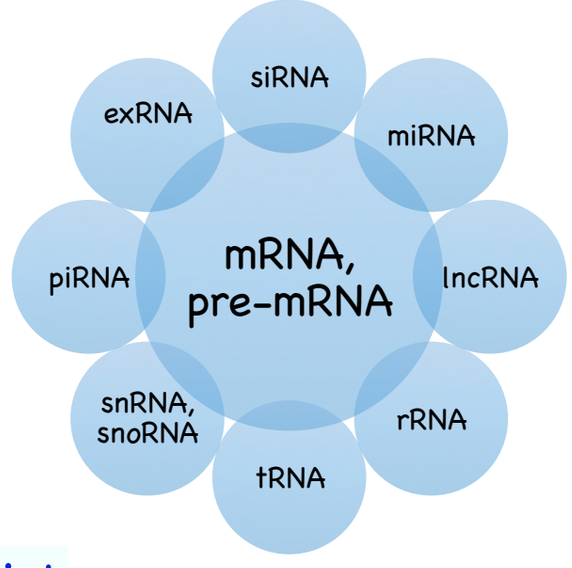
sum total of all RNA molecules expressed from the Genome

Oxford Genomics CENTRE



- Image credit Khan Academy Open Courses





RNA repertoire is dynamic!
It varies in time and space.

wellcome centre human genetics

UNIVERSITY OF OXFORD

RNAseq is a method for Transcriptome profiling

Image of the transcribed genome at any point of time!

Oxford Genomics CENTRE

- Single genes
- RNA transcripts
- Presence/Absence
- Quantification
- Length

Northern Blot

- Multiple genes
- cDNA molecules
- Presence/Absence
- Quantification

RT-PCR

How do we take this image?

Micro-arrays

- Multiple full transcriptomes
- cDNA molecules
- Presence/Absence
- Quantification

RNAseq

- Multiple full transcriptomes
- cDNA molecules (usually)
- Presence/Absence
- Quantification
- Length/Splicing
- Sequence

wellcome centre human genetics

UNIVERSITY OF OXFORD

Scope of 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.

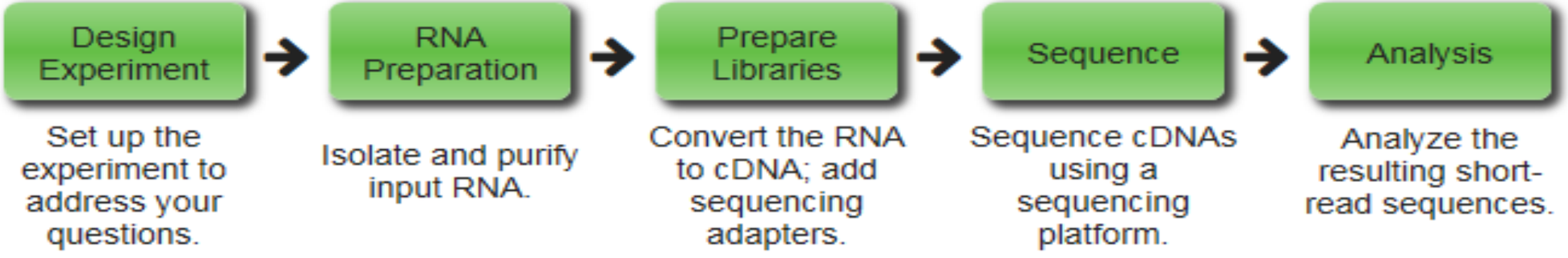
Common main goals

- 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.

Usual approaches for RNAseq library preparation

polyA enriched mRNAseq	Ribodepleted Total RNAseq	SMARTer/Smartseq2 mRNAseq	Small RNAseq <200nt ncRNAseq	3' mRNAseq mRNAseq
Oligo dT_{25} selection	Biotinylated ribosomal RNA probes	Oligo dT primer used for Reverse transcription	3' adapter ligation	Oligo dT_{25} primed RT (First strand synthesis)
Fragmentation of mRNA	Bind ribosomal RNA	Template switching by RT	5' adapter ligation	Removal of RNA template
1^{st} strand → 2^{nd} strand → cDNA synthesis of fragments	extracted with streptavidin beads	PCR pre-amplification of full-length cDNA	1^{st} strand cDNA synthesis	Random priming and second strand synthesis
Adapter ligation → PCR amplification	Fragmentation → cDNA synthesis → adapters → PCR	Tn5 transposase tagmentation & library prep.	PCR enrichment → Size selection	Bead purification of tagged cDNA library → PCR
Mature mRNA	Mature mRNA, nascent RNA,non-coding transcripts	Full-length mature mRNA	Focus on 21-25nt miRNA/ siRNA involved in gene regulation	200-300bp insert libraries of 3' ends of mRNA
Directionality	Works with low-quality RNA, e.g. FFPE samples	Pre-amplification of low-input RNA	Size specificity, Directional, Low-input and Low depth	Lower sequencing depth; poor-quality RNA works
Requires good quality total RNA	Requires high sequencing depth	Requires good quality total RNA; no directionality	Requires good quality total RNA	Mainly useful for expression quantification
100ng-1ug; 8-10 samples/ lane HiSeq4000	100ng to 1ug; 4-6 samples/lane HiSeq4000	10pg to 10ng; 10-30 samples/lane HiSeq4000	1ug total RNA; 20+ samples/lane HiSeq2500; 50bp SE	0.5 ng – 2 ug; 48 samples per lane HiSeq4000; 50-75bp SE

Typical RNAseq experiment



Key considerations for Experimental Design

How is an experiment designed?

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

Before we begin…

Oxford Genomics CENTRE

Define research question

State objectives and null hypothesis

Is there a reference?

Yes

Genome reference and gene annotations

Yes

Transcriptome reference

No

De novo or reference guided assembly required

Do we want to analyse alternative splicing? Identify novel isoforms?

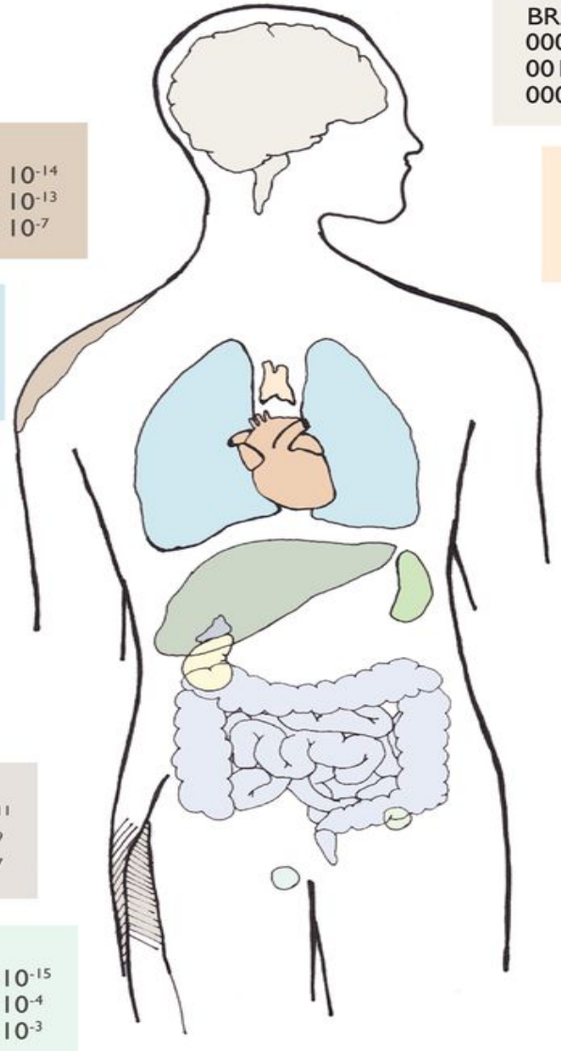
Type of sample/ Library/ Replicates/ Coverage/ Analysis Pipeline

wellcome centre human genetics

UNIVERSITY OF OXFORD

Sample selection

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}

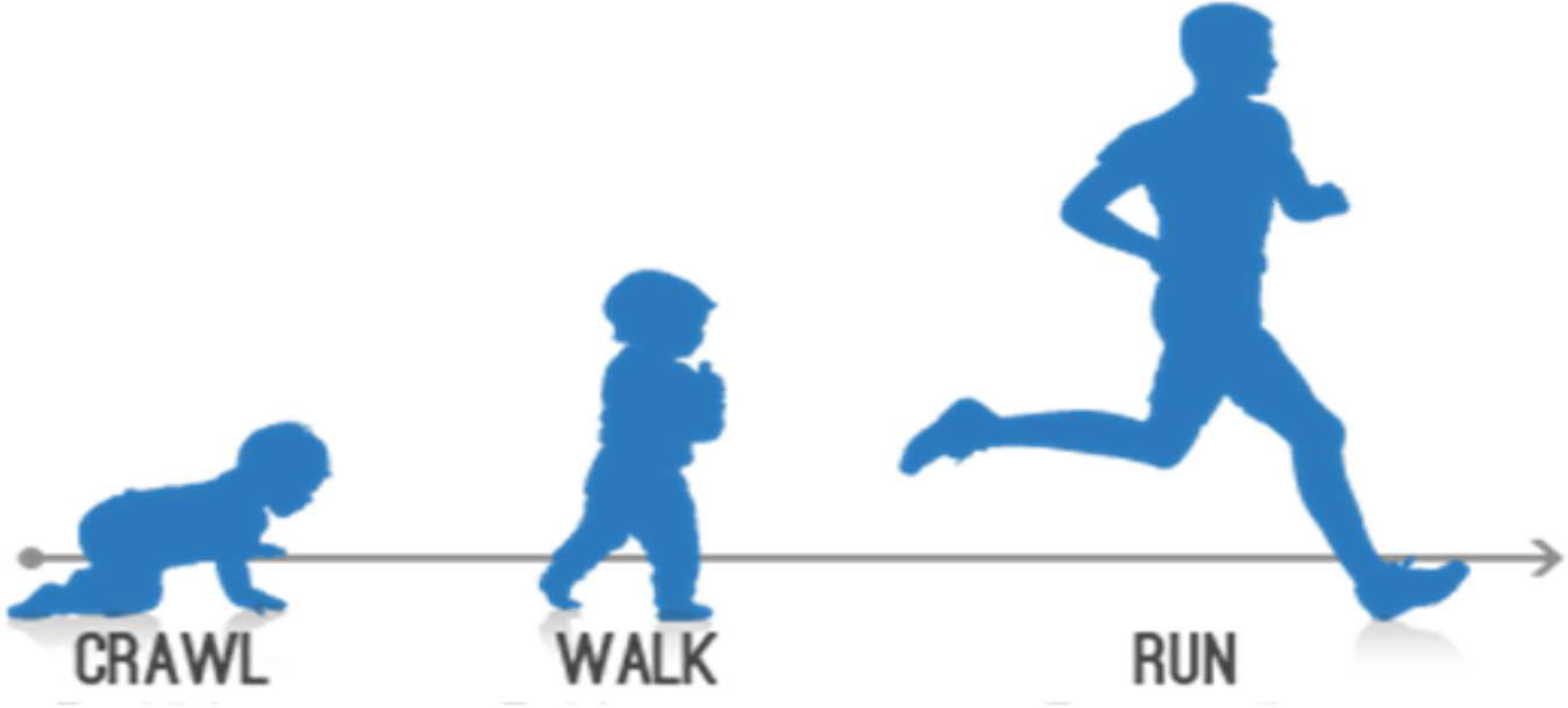
OVARY

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

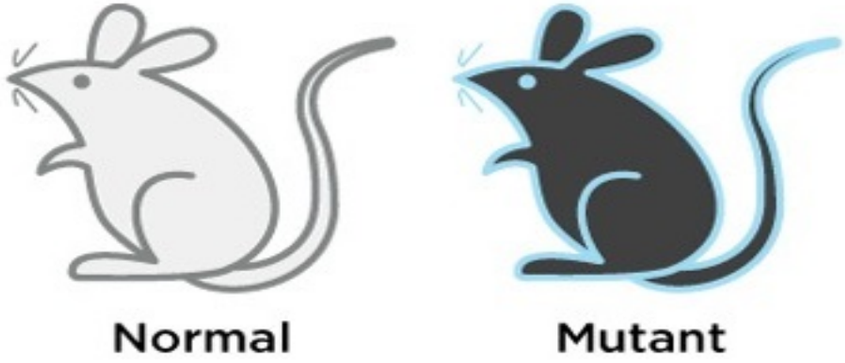
wellcome centre human genetics

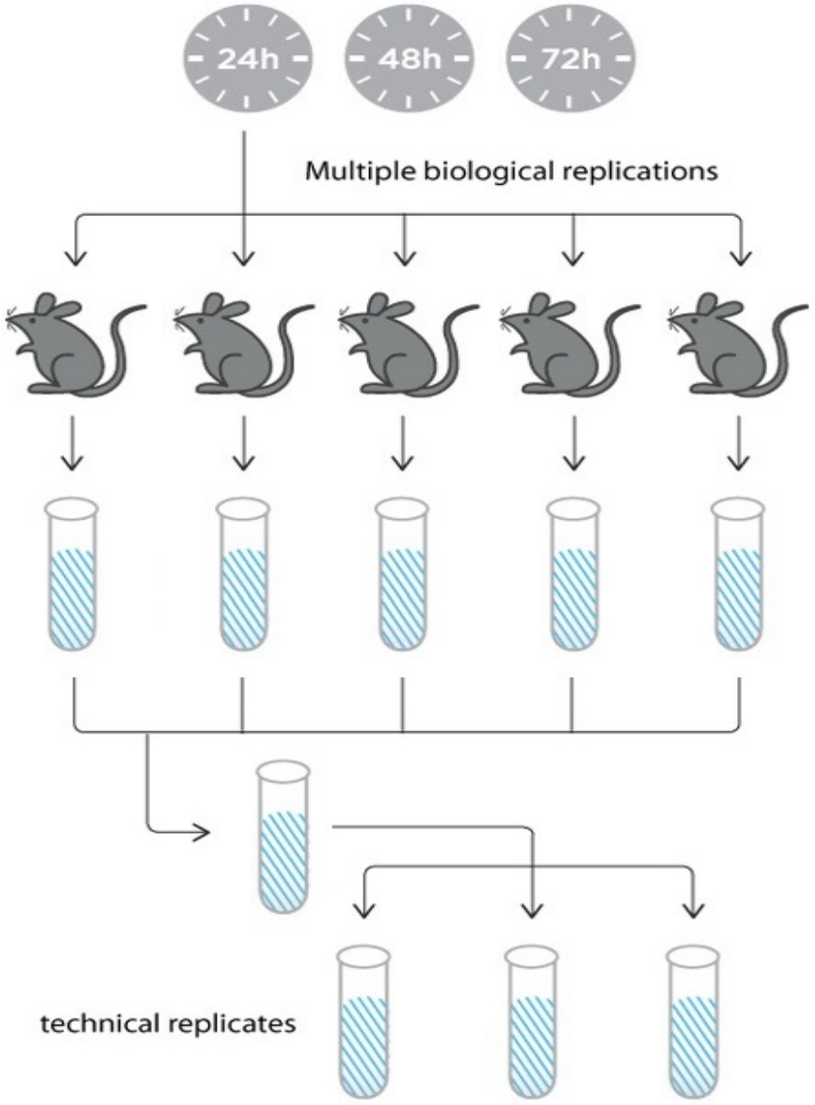
UNIVERSITY OF OXFORD

Time of sampling



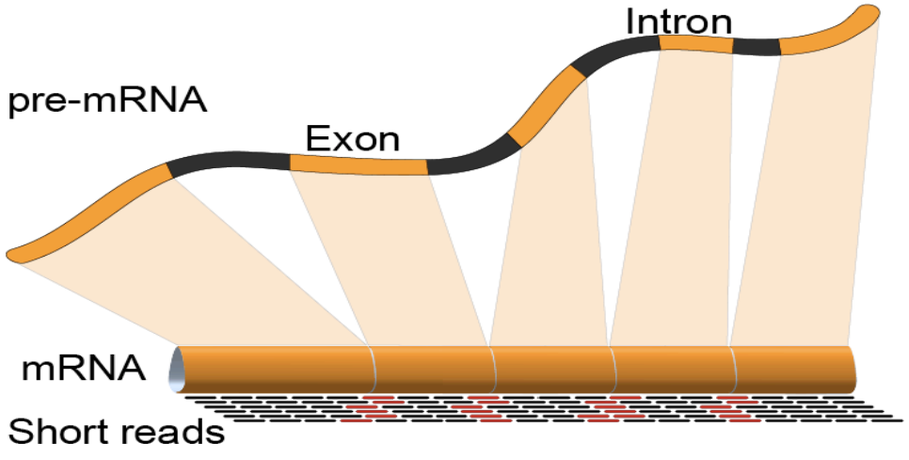
Replicates (technical / biological)



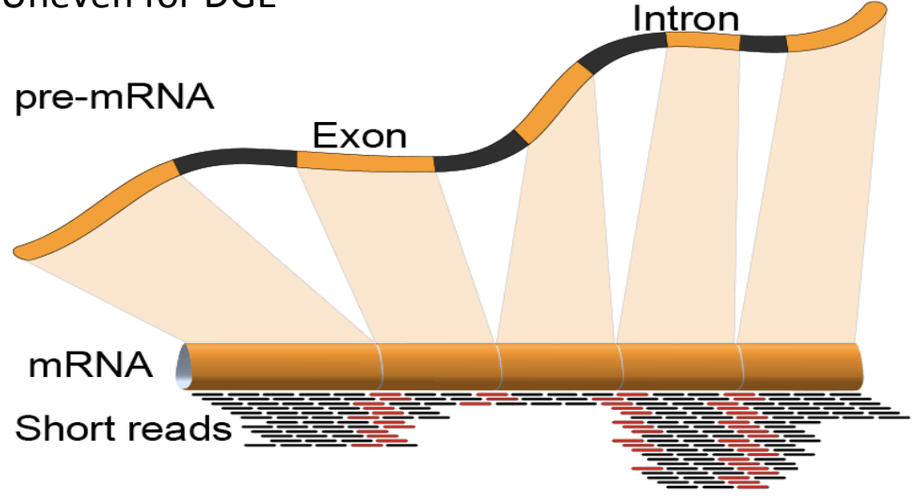


Coverage

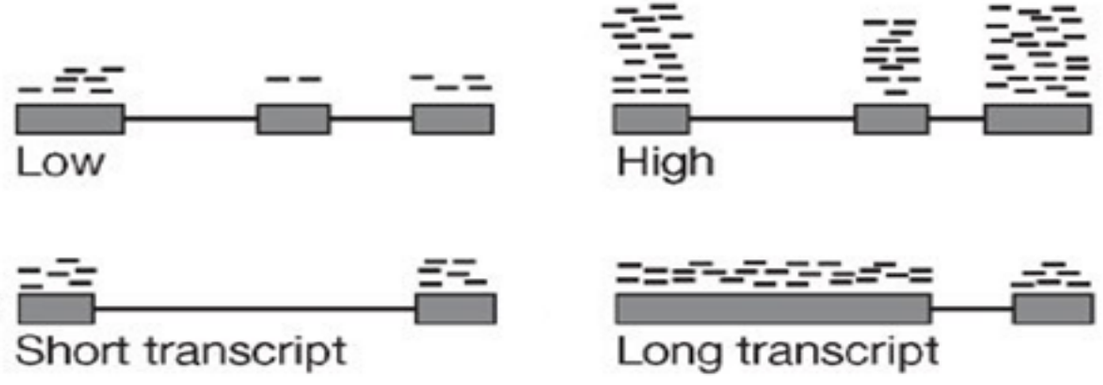
- Even for annotation



- Uneven for DGE

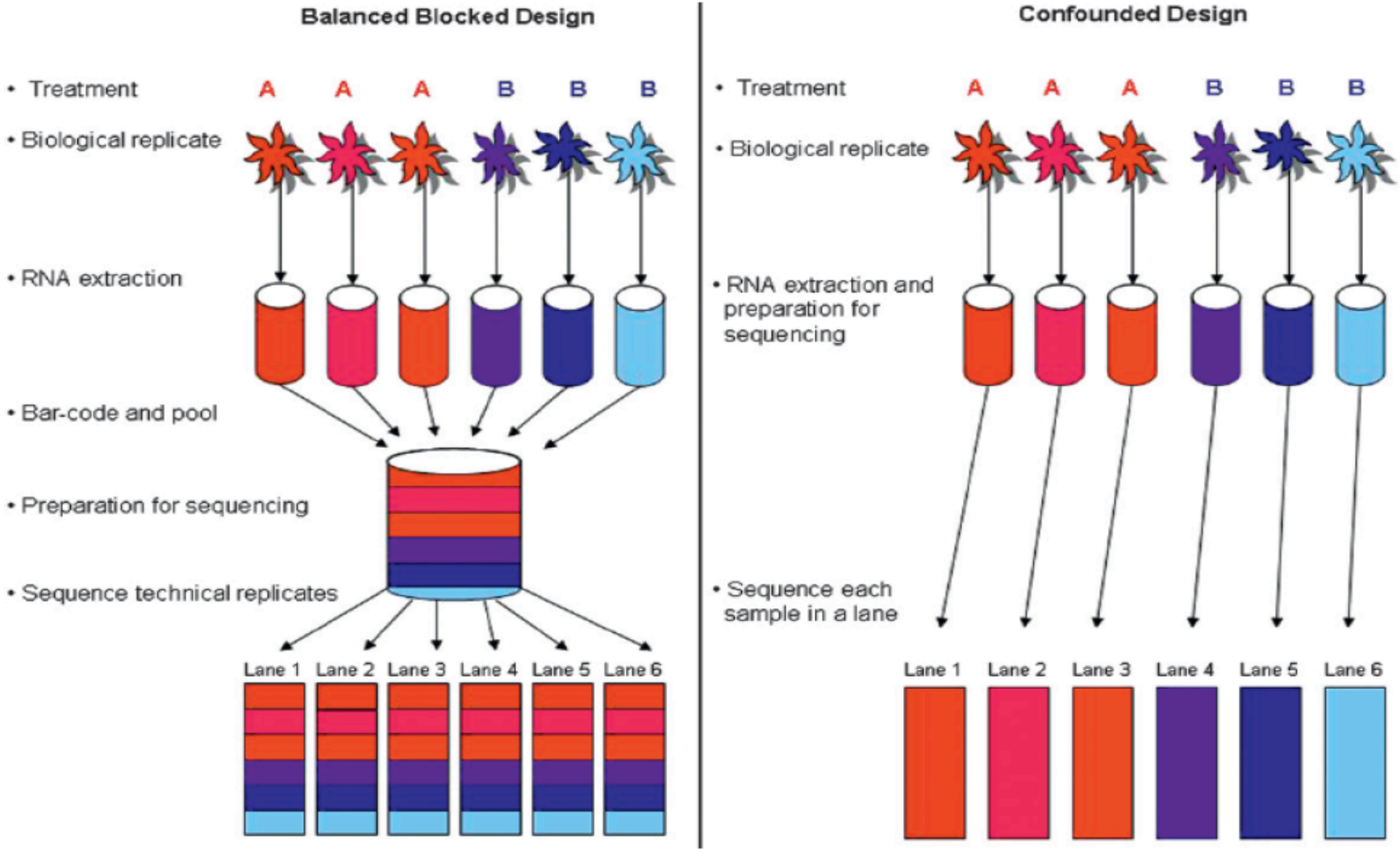


- Target transcript properties (low abundant vs high abundant transcripts)



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

Randomization and Blocking



Data management & Downstream analysis and 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

Subjectivity of the analysis

Oxford Genomics CENTRE

- ✓ 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.

Adapter trimming

Alignment

Assembly

Quantification

Count normalization

Alternative splicing/ Isoform level analysis

Differential expression analysis

wellcome centre human genetics

UNIVERSITY OF OXFORD

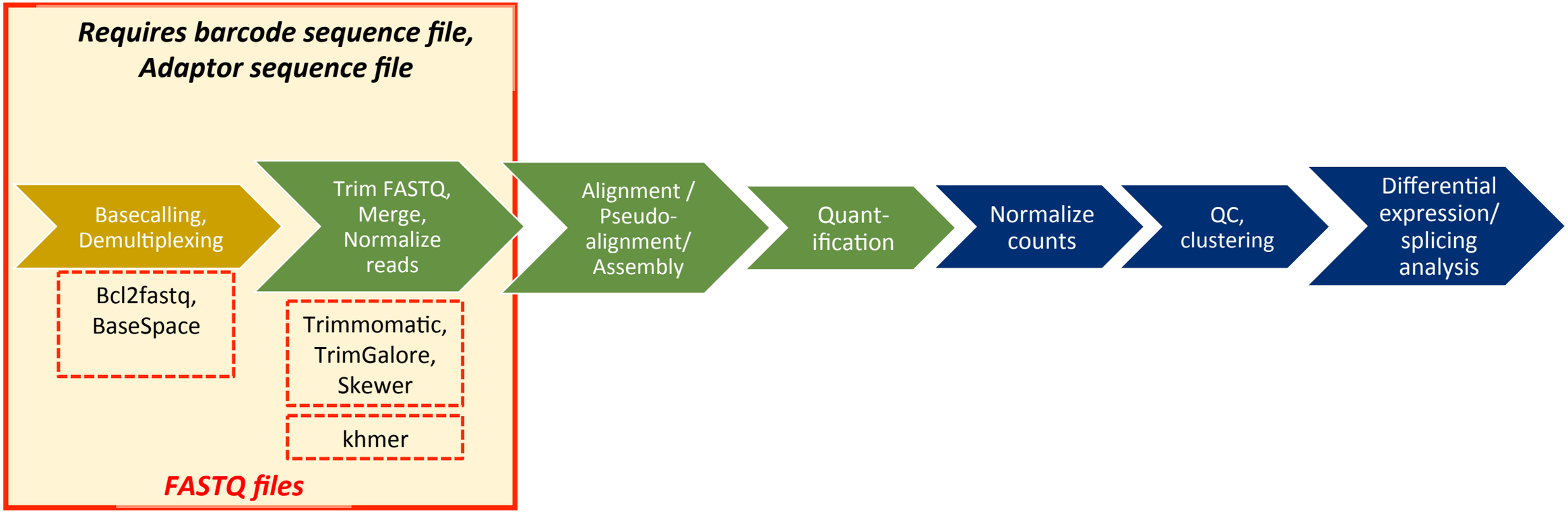
Data Analysis

Typical 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

Read processing



Read processing

Oxford Genomics CENTRE

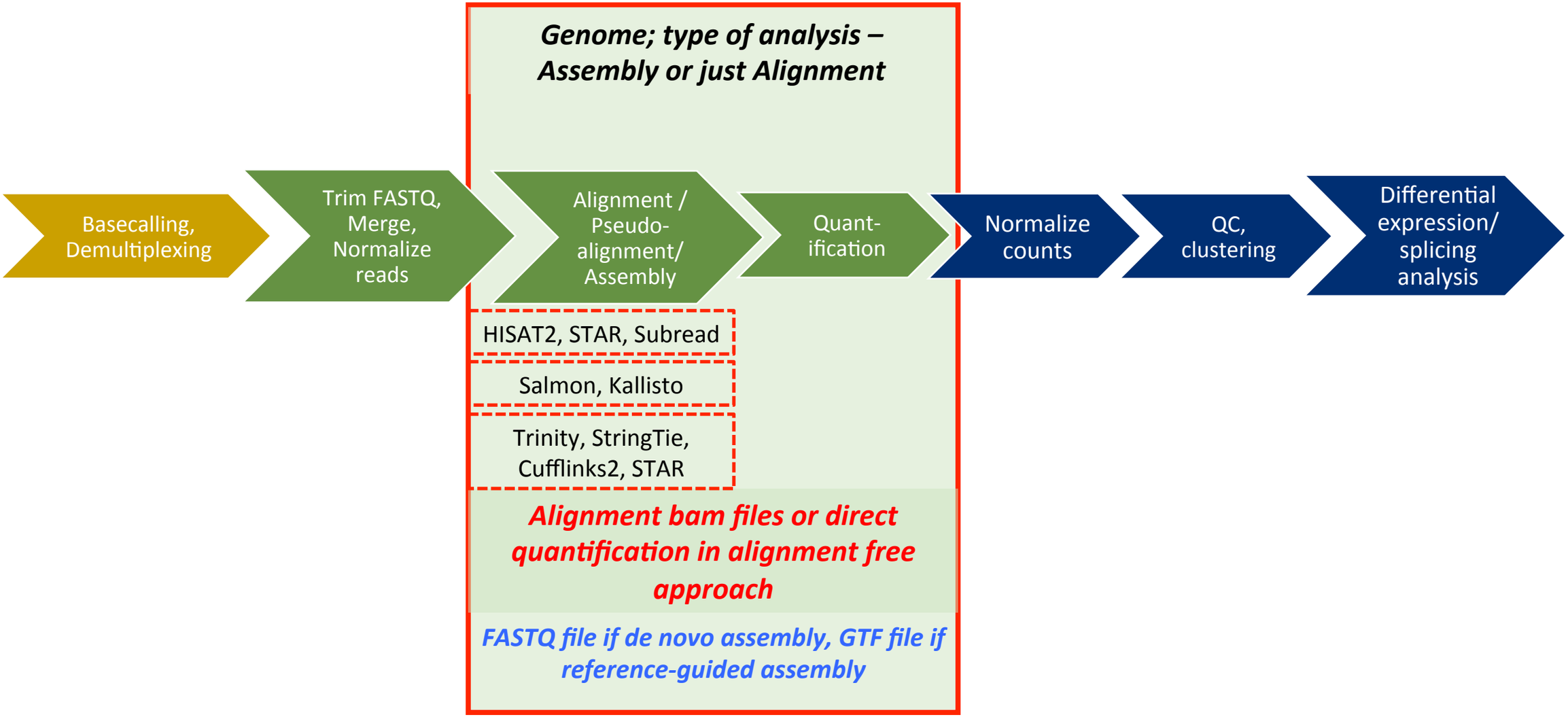
For *de novo* assembly

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

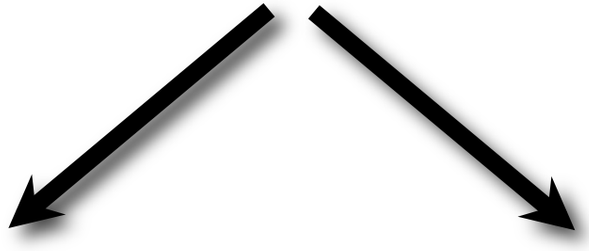
Read processing, alignment/assembly+alignment, quantification



de novo assembly or reference mapping?

Oxford Genomics CENTRE

When to use each?



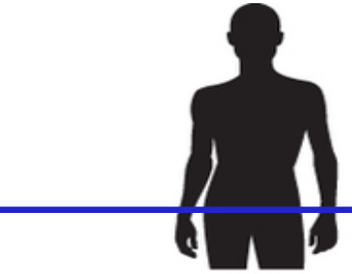
de novo
(do not know the transcriptome)
(main goal is to discover NOT to quantify)

reference
(do know the transcriptome)
(main goal is to quantify NOT to discover)

wellcome centre human genetics

UNIVERSITY OF OXFORD

Protocol for *de novo* RNAseq

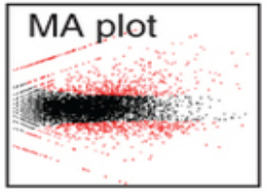


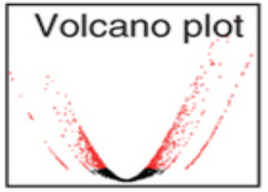
Reads
(per sample)

Abundance estimation

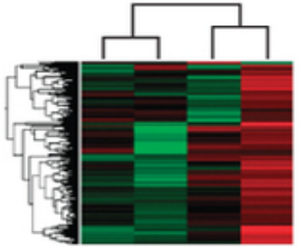
Assembled
transcripts
(all samples)

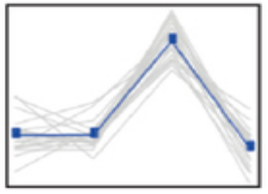
Identify differentially expressed transcripts





Expression patterns, transcript clusters





Combine reads

Normalization?

De novo assembly

Assembled
transcripts

Identify coding regions

Functional
annotation

Annotated transcriptome

NCBI BLAST

HMMER Pfam

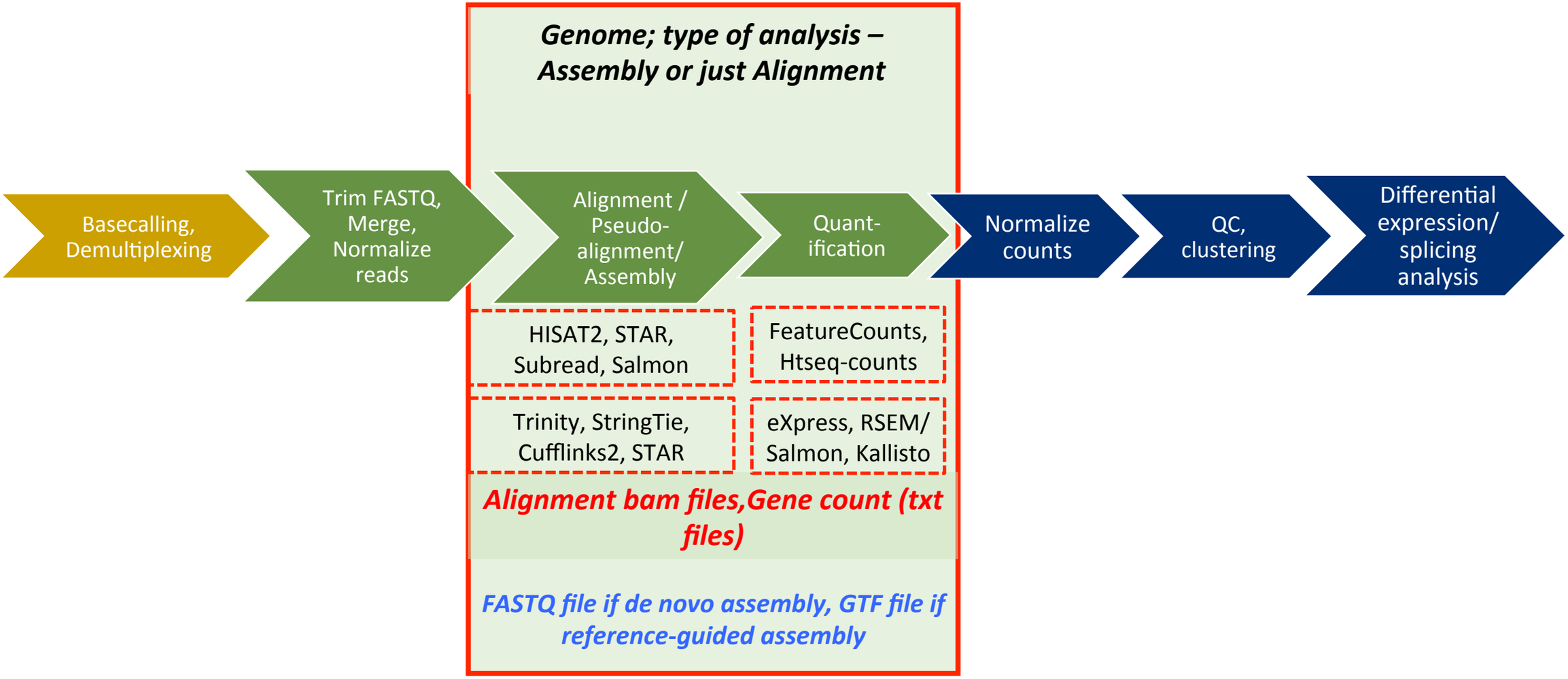
UniProt

eggNOG version 3.0

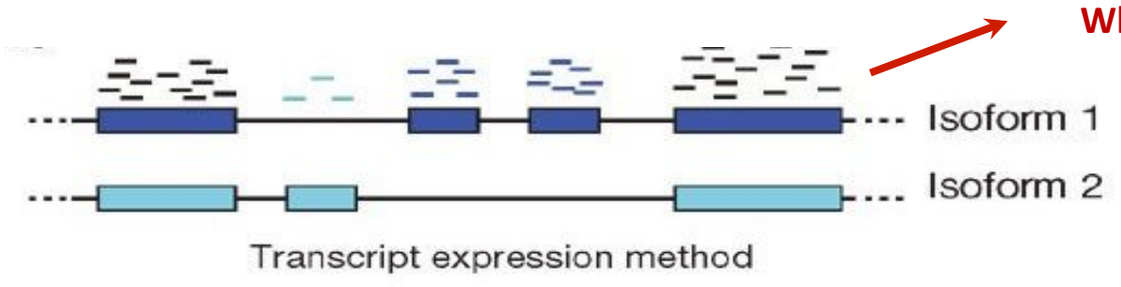
the Gene Ontology

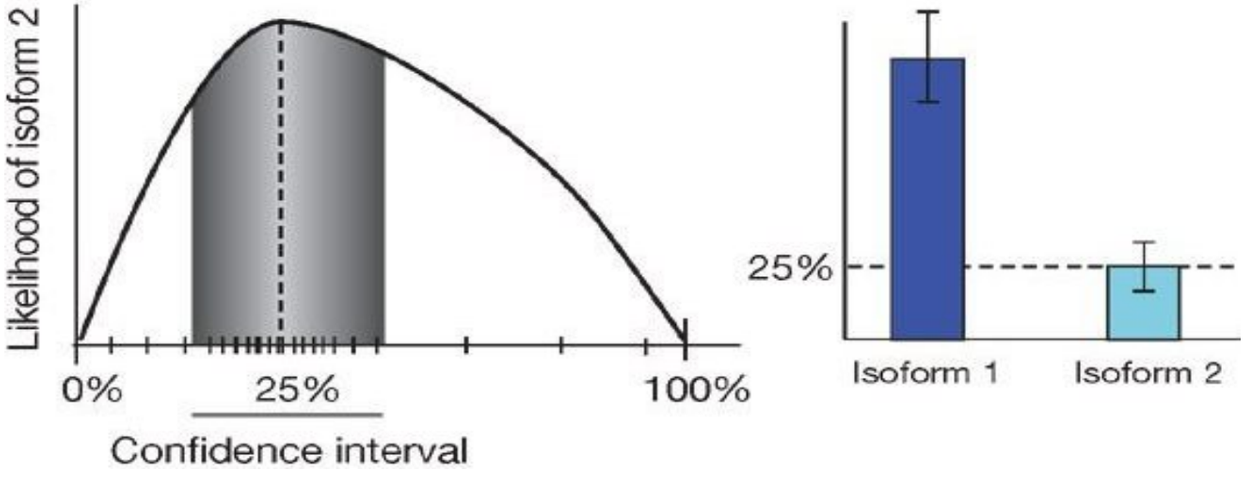
Protocol for reference-based RNAseq

Read processing, alignment/assembly+alignment, quantification



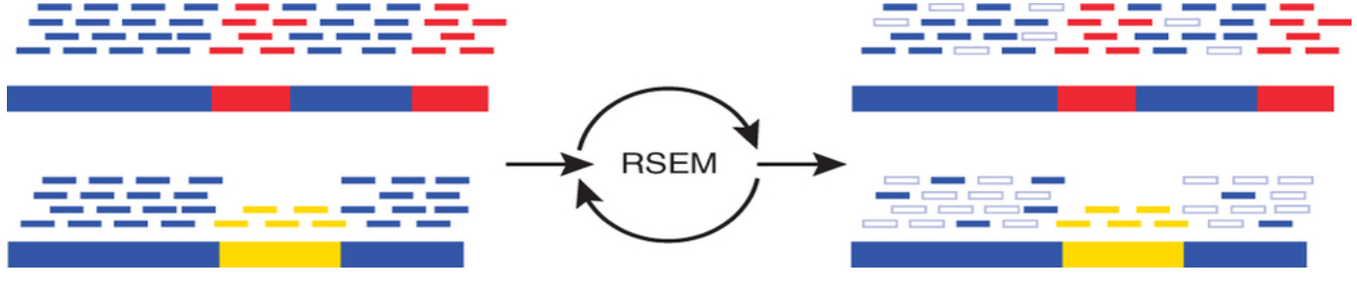
Expression quantification



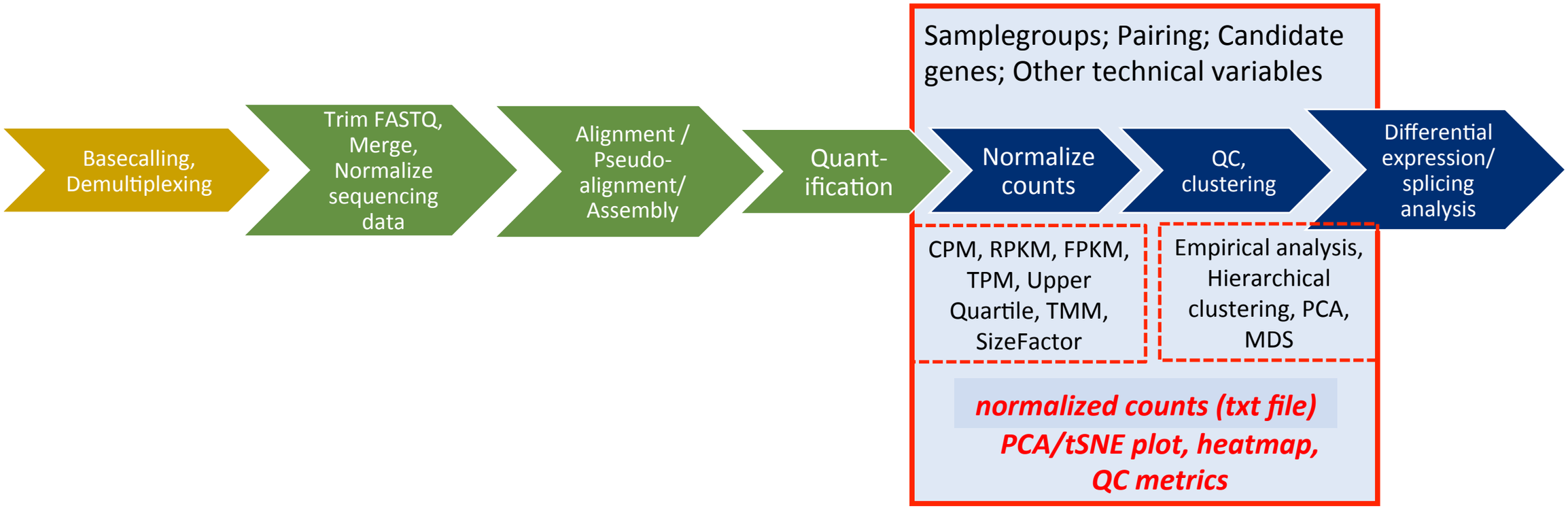


Genome guided: Cufflinks2, StringTie

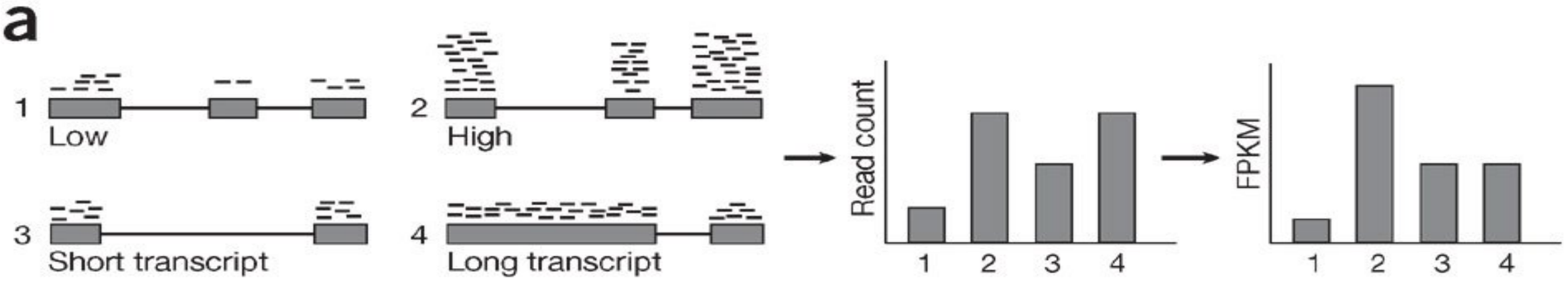
Trancriptome guided: RSEM, eXpress, Salmon, Kallisto



Normalization, Expression quantification



Count 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**.

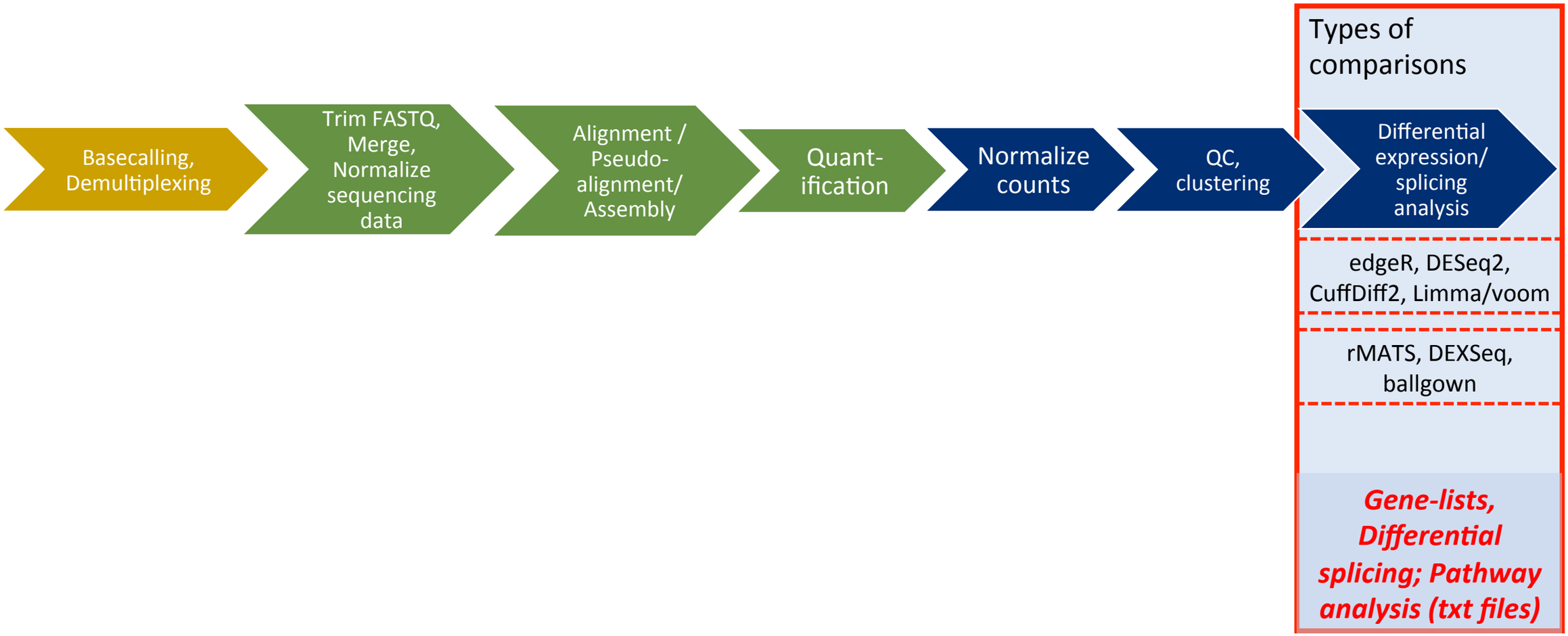
Statistical model should take into account "length" and "sequencing depth".

Count normalization

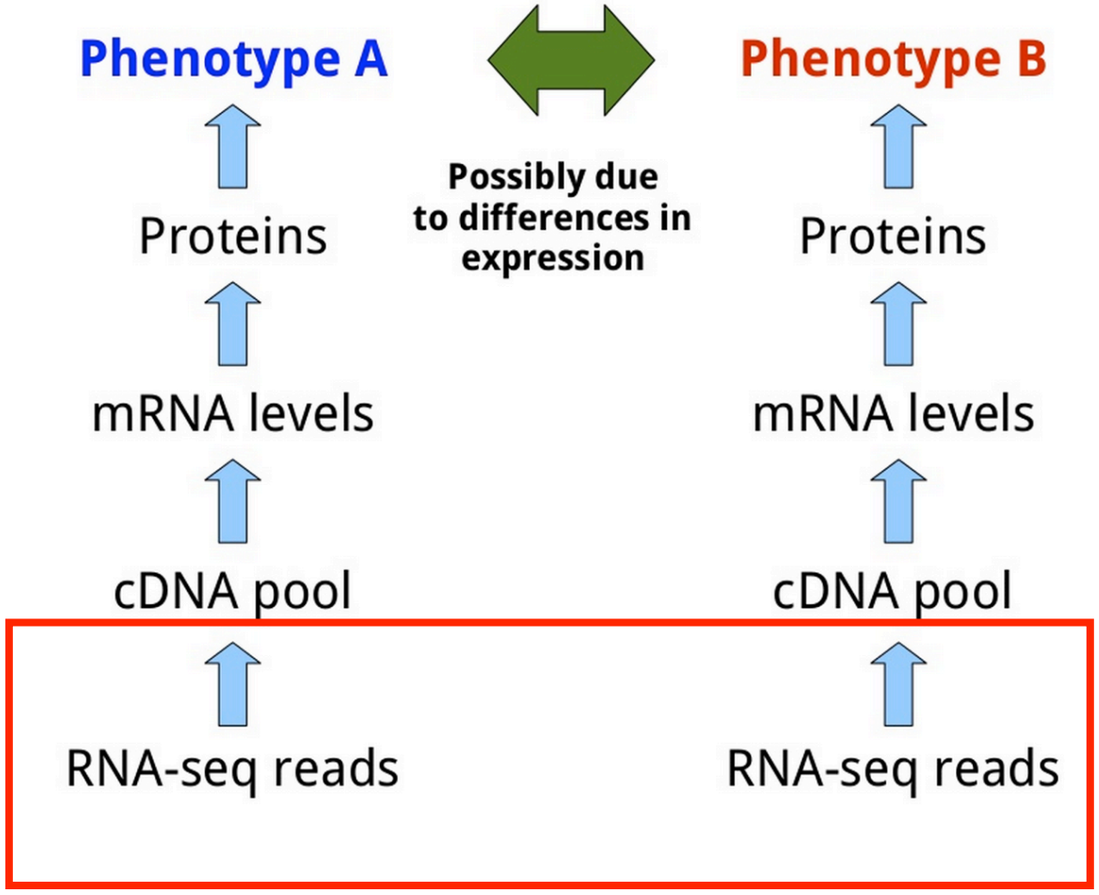
Oxford Genomics CENTRE

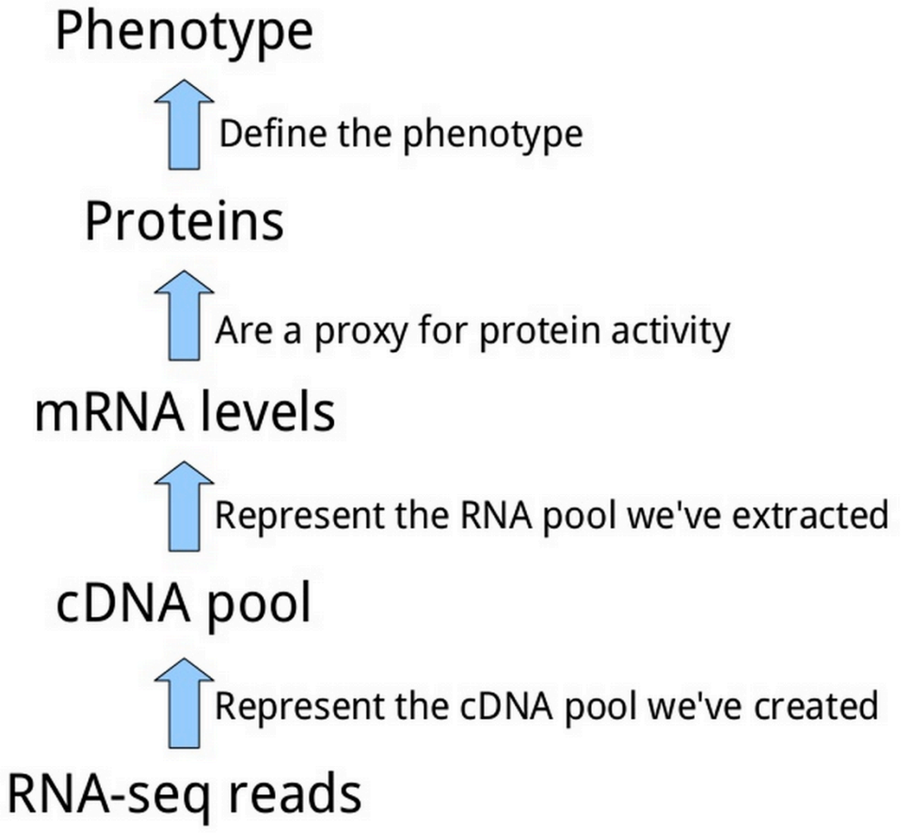
	Counts	CPM	RPKM/FPKM	TPM
Value	Integer	Fraction	Fraction	Fraction
Depth-bias	✗	✔	✔	✔
Length-bias	✗	✗	✔	✔
Compare same genes across samples	✗	✔	✔ (but may have bias)	✔
Compare different genes in sample	✗	✗	✔	✔
Compare different genes across samples and across experiments	✗	✗	✗	✔
Can be used for barplots/ boxplots of single genes	✗	✔	✔ (but may have bias)	✔
Can be used for heatmaps with multiple genes (log transformed)	✗	✔ (as long as we don't compare the colour of different genes)	✔ (but may have bias)	✔

Differential expression analysis



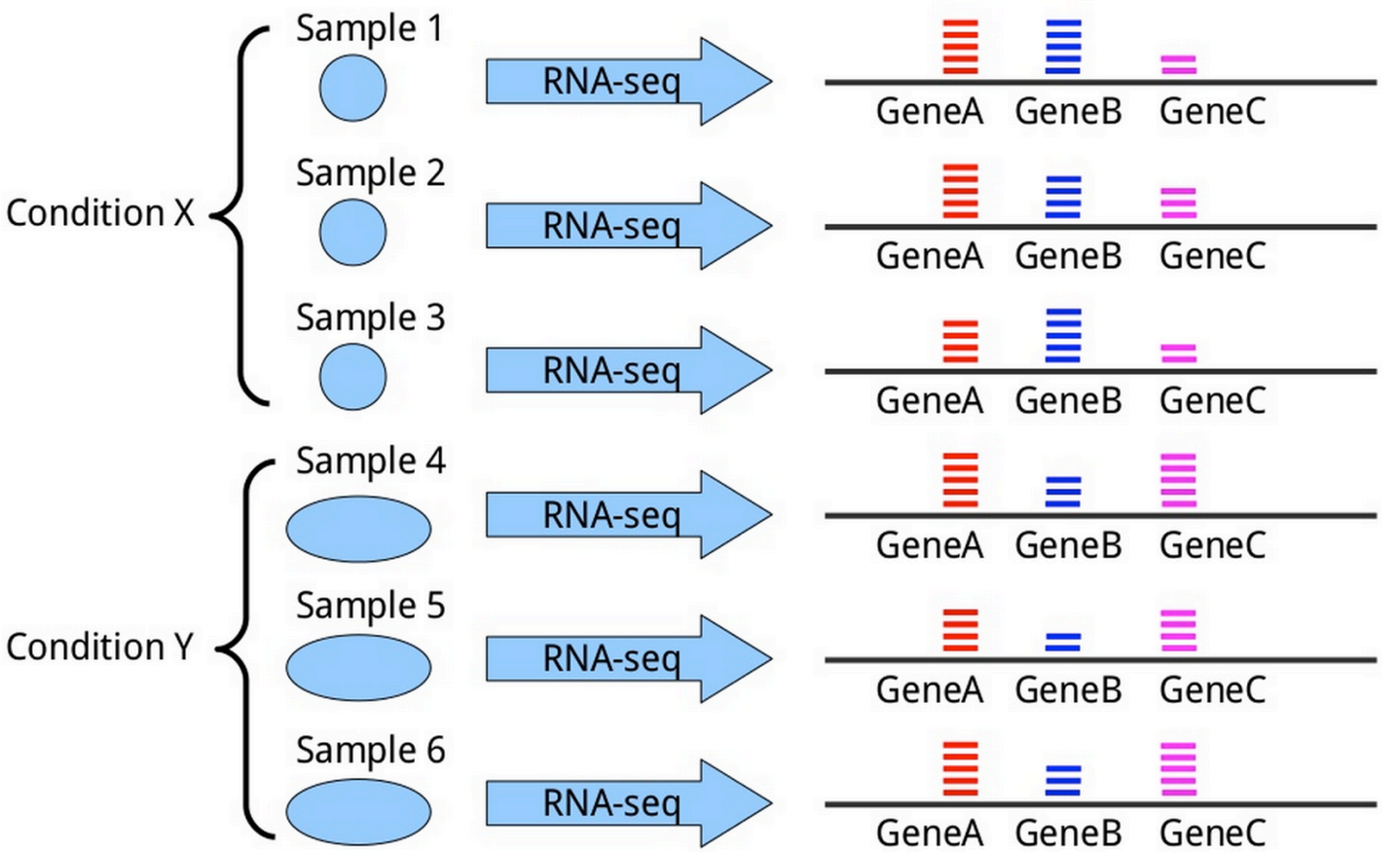
Our assumptions and comparison





http://www.slideshare.net/joachimjacob/1rna-seqpart1working-tothegoal?related=2

Statistical testing for Differential expression



http://www.slideshare.net/joachimjacob/1rna-seqpart1working-tothegoal?related=2

Read in raw count data

Remove genes with low-expression (<10 reads per sample across group)

Normalize subset count tables using size factors (e.g. TMM normalization in edgeR)

Unsupervised clustering to identify technical effects and biological effects

Create design matrix with comparison of interest and technical/biological variability

Fit normalized expression matrix to linear models to identify coefficients for each gene

Identify DE genes with pre-defined statistical criteria (~FDR < 0.05)

What we do when we do RNAseq?

Oxford Genomics CENTRE

- **What it is?**
- **Scope of RNAseq**
- **Usual approaches for RNAseq library preparation?**
- **Considerations for RNAseq experiments**
- **General methods for RNAseq data analysis.**

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
- https://f1000research.com/articles/5-1408/

Thank You!

Eshita.sharma@well.ox.ac.uk