General overview of RNAseq

Irina Mohorianu (CSCI)

RNAseq. Overview

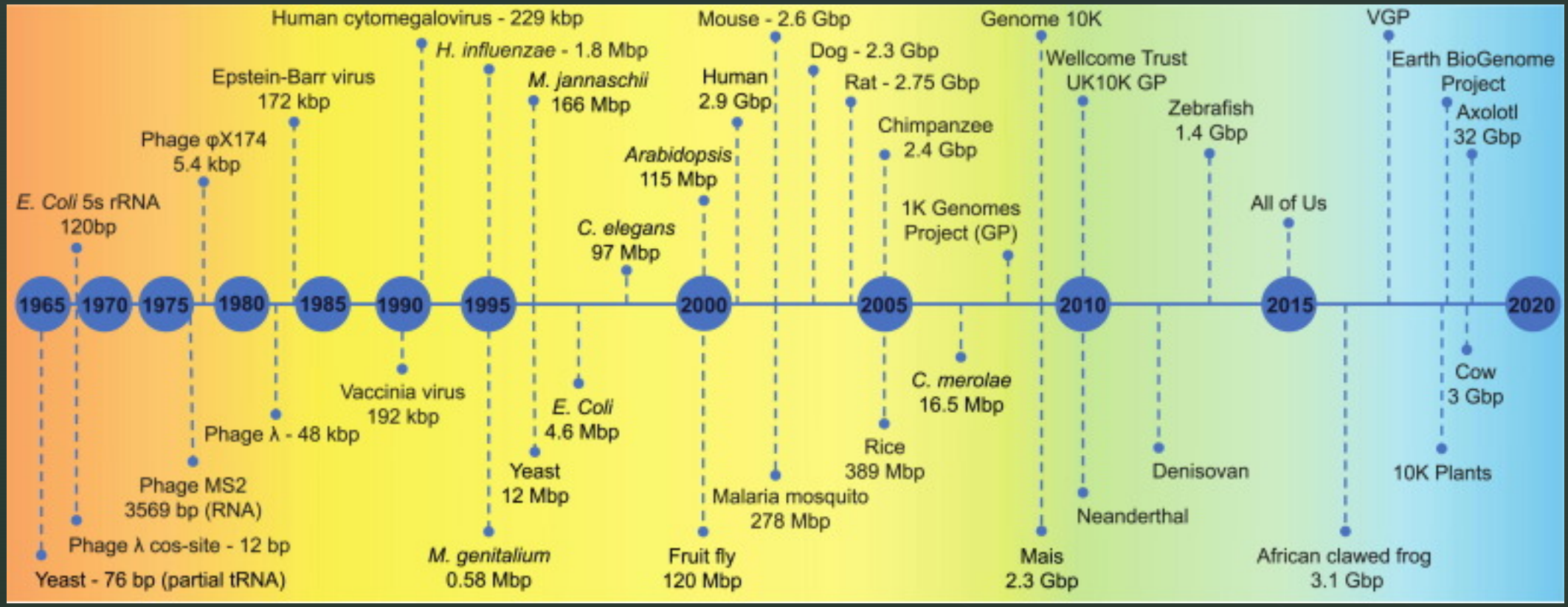
RNA-seq produces millions of sequences from complex RNA samples.

[1] Measure gene expression. Assess the distribution of signal across transcripts
[2] Discover and annotate complete/ new transcripts.
[3] Characterize alternative splicing and polyadenylation.
[4] infer regulatory interactions and build Gene Regulatory Networks

In this lecture we'll focus on a general overview of RNAseq.
We'll discuss characteristics or sequencing details that can have an effect on the quantification of gene expression.

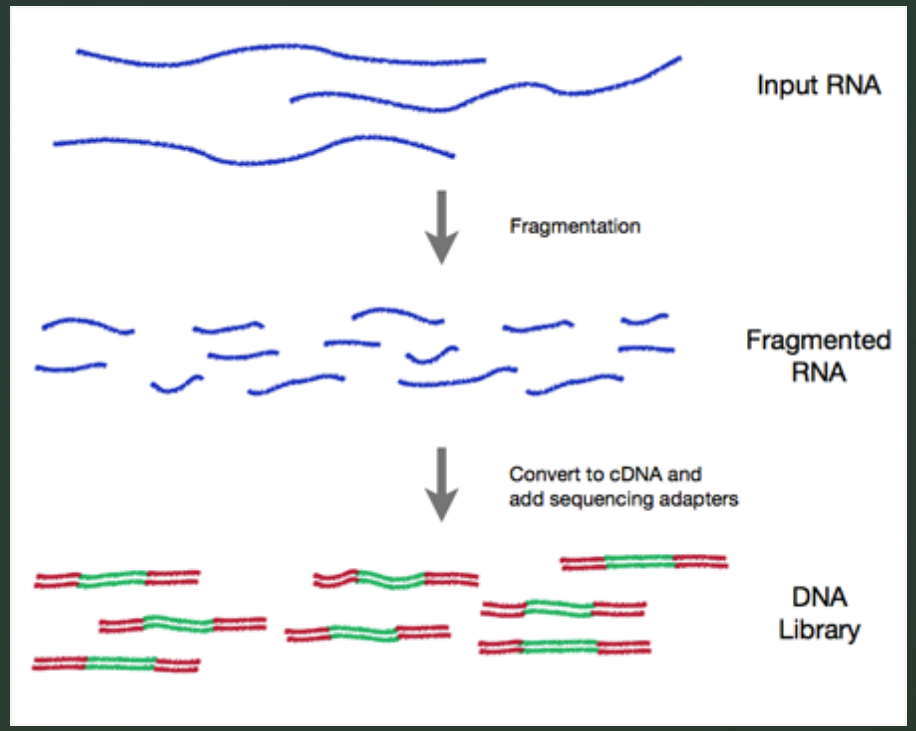
For creating this material three main sources were used:
[a] **RNA sequencing: the teenage years**; Rory Stark, Marta Grzelak & James Hadfield
Nature Reviews Genetics volume 20, pages631–656(2019)
[b] **Long walk to genomics: History and current approaches to genome sequencing and assembly**; Alice M Giania, Guido R Gallob, Luca Gianfrancheschi, G Formenti
Computational and Structural Biotechnology Journal Volume 18, 2020, Pages 9-19
[c] **RNA-seqlopedia** https://rnaseq.uoregon.edu/

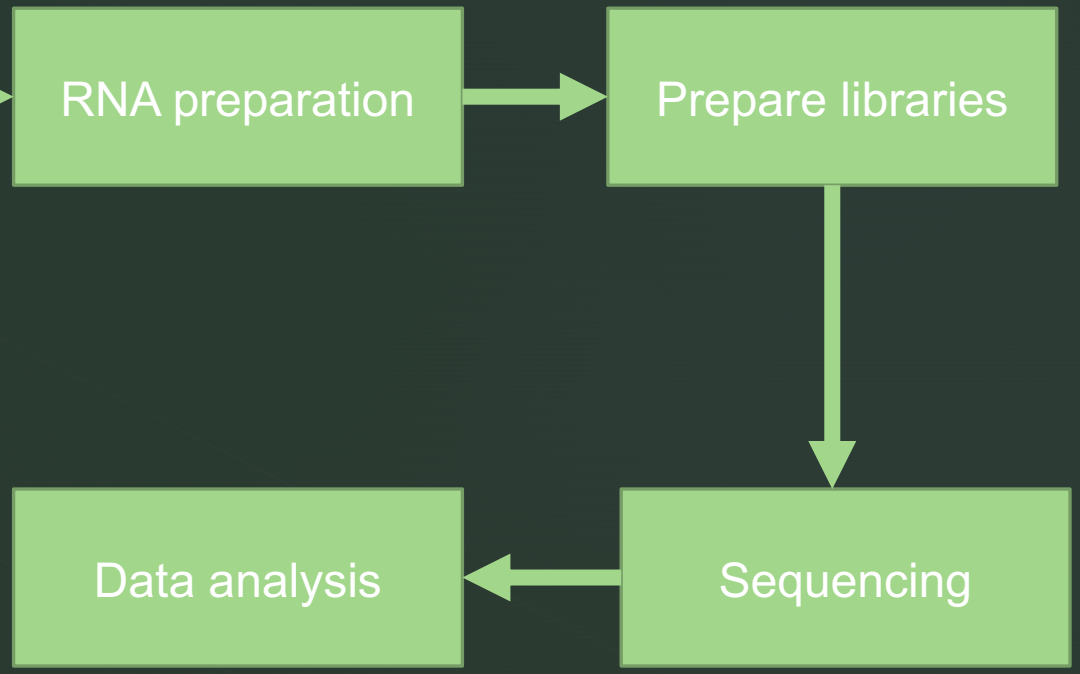
RNAseq. The history



Milestones in genome assembly. Timeline illustrating many of the major genome assembly achievements ranging from the beginning of the sequencing era to the large-scale genome projects currently ongoing. Each genome or genome project (GP) is placed under a color-coded background according to the sequencing approach adopted. Light red: early sequencing methods, Yellow: Sanger-based shotgun sequencing, Green: NGS, Light blue: TGS. *Ref: Gallob et al 2020*

RNAseq.
From the cells to the numbers





Information flow for an RNAseq experiment.
RNA-seqlopedia

RNAseq. Experimental design

Experimental objectives:

[1] qualitative

[2] quantitative

Qualitative data includes identifying expressed transcripts, and identifying exon/intron boundaries, transcriptional start sites (TSS), and poly-A sites. We refer to this type of information as "annotation".

Quantitative data includes measuring differences in expression, alternative splicing, alternative TSS, and alternative polyadenylation between two or more treatments or groups. We focus specifically on experiments to measure differential gene expression (DGE).

RNAseq. Experimental design

Criteria	Annotation	Differential gene expression
Biological replicates	Not necessary but useful	Essential
Transcript Coverage	Important for de Novo assembly identifying transcriptional isoforms	Not as important; useful reads = uniquely mapped
Depth of sequencing	High to maximize coverage of rare transcripts and transcriptional isoforms	High for accurate statistics
Role of seq depth.	reads that overlap on the transcript	sufficient counts per transcript
Stranded library prep	essential Novo transcript assembly identifying true anti-sense transcripts	not required if ref genome exists
Long reads (>80 bp)	essential Novo transcript assembly identifying transcriptional isoforms	not required if ref genome exists
Paired-end reads	essential Novo transcript assembly identifying transcriptional isoforms	not required if ref genome exists

RNAseq. Experimental design. Annotation

Goal: identification of genes and genic architecture (based on expressed RNA).
what is present rather than how much of it is present
Essential parameter: coverage of the transcript [ends are important]

Coverage depends on the method used to prepare the library.
[1] **oligo-dT priming** for first-strand synthesis can be used to accurately annotate 3′-ends but often fails to evenly cover the 5′-end.
[2] **Random priming** suffers from uneven coverage due to sequence/structure priming biases, and can result in poor coverage of either of the ends.

Regardless of the method used to prepare the libraries, we observe uneven coverage of individual transcripts. To obtain reads that span the entire transcript more reads are required (i.e. deeper sequencing).

Challenges:
[a] reads are much shorter than the biological transcripts
[b] in many cases a reference genome is unavailable. Both of these issues create problems for mapping and de novo transcript assembly.

RNAseq. Experimental design. Differential gene expression

Goal: [1] quantitatively measure differences in transcripts abns expression between groups.
[2] to appropriately interpret differences in read counts

Sources of variances associated with the counts.:

Sampling variance: the millions of reads represent only a small fraction of the nucleic acid that is actually present in the library.; also a sample = snapshot in time

Technical variance: Library preparation and sequencing procedures involve a series of complex chemical reactions which all contribute to between-sample variance.

Biological variance: the aim is to measure differences between individuals.
Biological systems are inherently complex and very sensitive to perturbations.
Bio-variance = the nascent variance that is present within a treatment or control group.

RNAseq. Experimental design. Differential gene expression. Replicates

Decisions about the number and types of replicates are driven by extrinsic and intrinsic factors.

Extrinsic factors : cost, availability of samples, and feasibility of experiments.

Intrinsic factors are more difficult to grasp without prior information about the system and more ambiguous from a decision standpoint.

These include [1] the degree of transcriptional variability among samples,
[2] whether certain genes or transcripts are of special interest,
[3] whether these genes of interest are expressed at low levels
[4] how many experimental factors are of interest
(i.e. the complexity of an experimental design).

Because the extrinsic considerations are often "hard and fast", it is useful to start from those constraints and subsequently build the study design around them.

RNAseq. Experimental design. Differential gene expression. Replicates

Technical replication : sequencing multiple libraries derived from the same biological sample.

Non-biological variation in estimated transcript abundance across these samples can arise from unintended differences during library preparation or sequencing.

Biological replication: sequencing multiple libraries derived from synonymous biological samples. Assess variation among different individuals or tissues, usually with respect to experimental treatments. **Usually biological variation is large relative to technical variation.**

Unless you are genuinely interested in comparing technical aspects of RNA-seq, or you expect technical variation to be especially great for a large majority of the target transcripts, we recommend greater resource allocation to biological replication.

RNAseq. Experimental design. Differential gene expression. Replicates

How many replicates should be sequenced?

Statistical hypothesis tests are prone to two types of error: type I errors, type II errors

Failure to reject the null hypothesis of no difference when there actually is a difference (a "**false negative**") is known as type II error [β = probability of occurrence].

The number of replicates per group in an experiment directly affects type II error, and therefore "statistical power" (which is 1-β).
Power depends on the magnitude of the effect of one condition relative to another on the variable of interest, which is in part determined by the degree of variation among individuals.

Power also depends on the acceptable maximum probability of type I error (the event in which the null hypothesis is rejected in favor of the alternative when the null hypothesis is actually true, a "**false positive**").

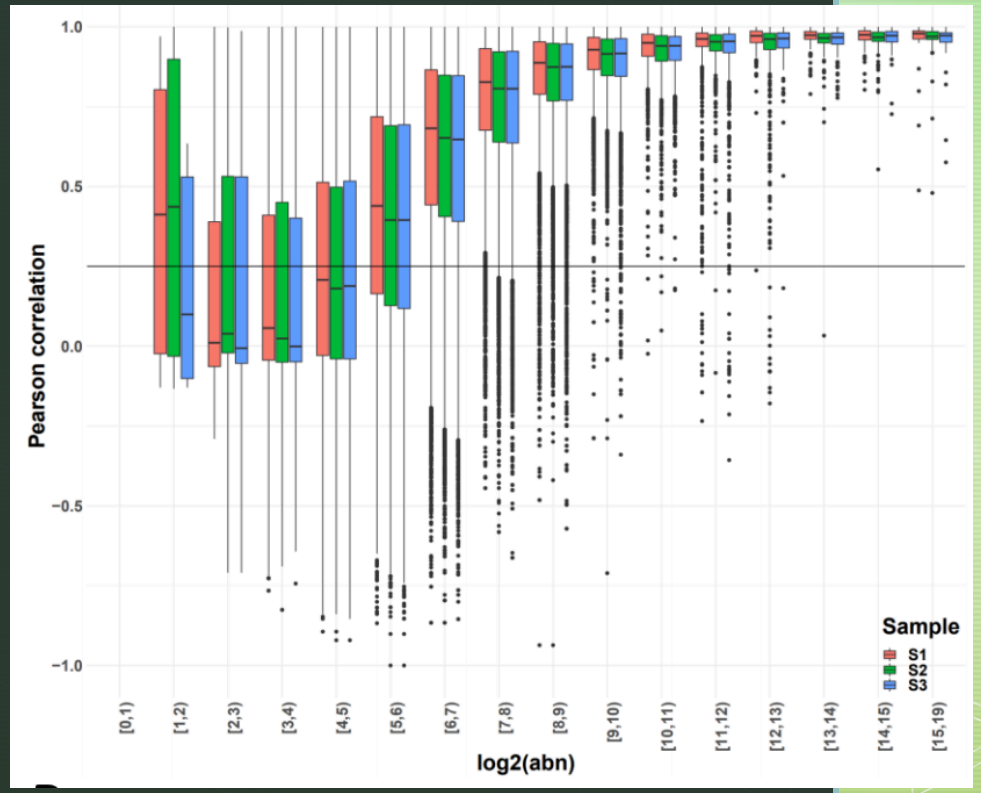
RNAseq. Experimental design. Differential gene expression. Replicates

To estimate the #replicates for a given hypothesis test to achieve a desired level of power, you need to have some understanding of the treatment effect size, which depends on variance in read counts across individuals within each treatment.

Effect sizes are easier to reliably estimate for genes with at least moderate sequencing coverage in one treatment than for genes with sparse coverage across treatments.

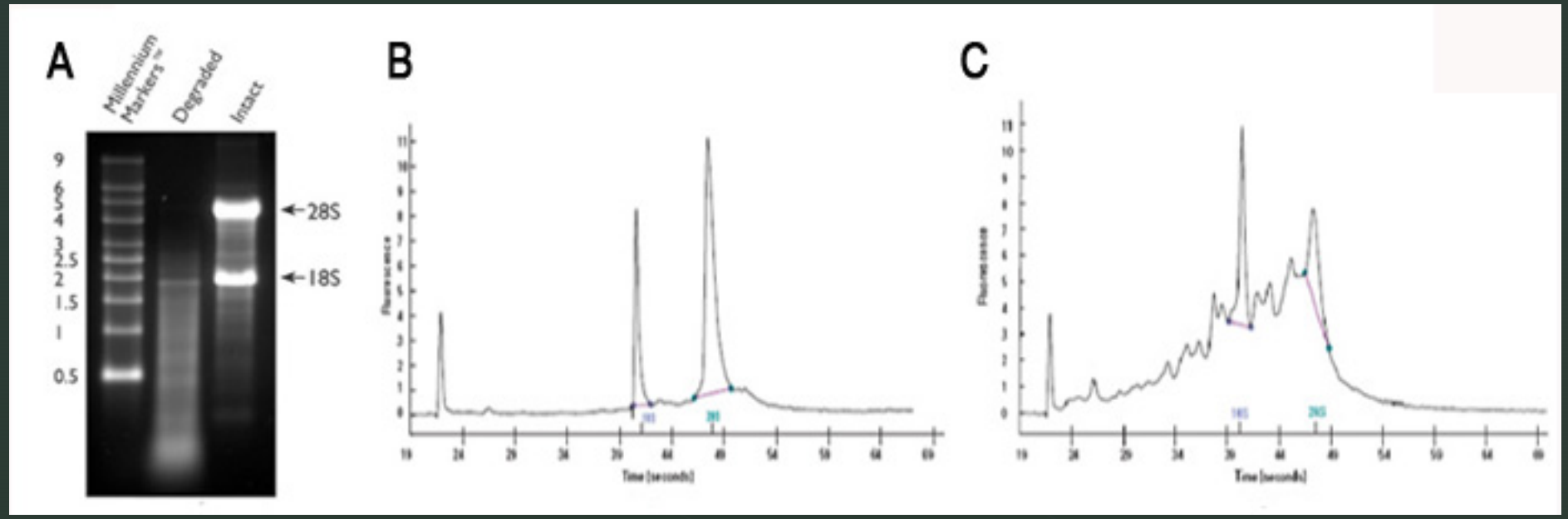
What is the optimal sequencing depth?

Variation due to the sampling process largely contributes to the total variance among individuals for transcripts represented by few reads. This means that identifying a treatment effect on genes with shallow coverage is not likely amidst the high sampling noise



noisyR: Enhancing biological signal in sequencing datasets by characterizing random technical noise, Moutsopoulos et al 2021 https://doi.org/10.1101/2021.01.17.427026

RNA preparation. RNA quality



(A) Two µg of degraded total RNA and intact total RNA were run on a 1.5% denaturing agarose gel. The 18S and 28S ribosomal RNA bands are clearly visible in the intact RNA sample. The degraded RNA appears as a lower molecular weight smear. Image from www.ambion.com

(B) There are two well-defined peaks corresponding to the 18S and 28S ribosomal subunits and the ratio between the 28S and 18S peaks is approximately 2:1.

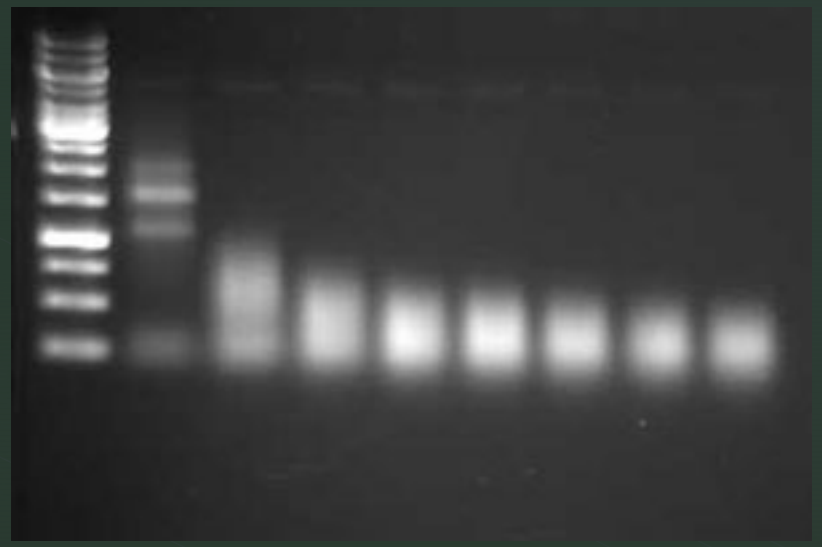
(C) is an example of partially degraded RNA. The 2:1 ratio between the ribosomal peaks is absent and there is a high presence of degraded products.

RNA fragmentation

Most current sequencing platforms are capable of providing only relatively short sequence reads (~40-400bp depending upon the platform). Most protocols incorporate a fragmentation step to improve sequence coverage over the transcriptome.

Protocols differ on the timing of the fragmentation.

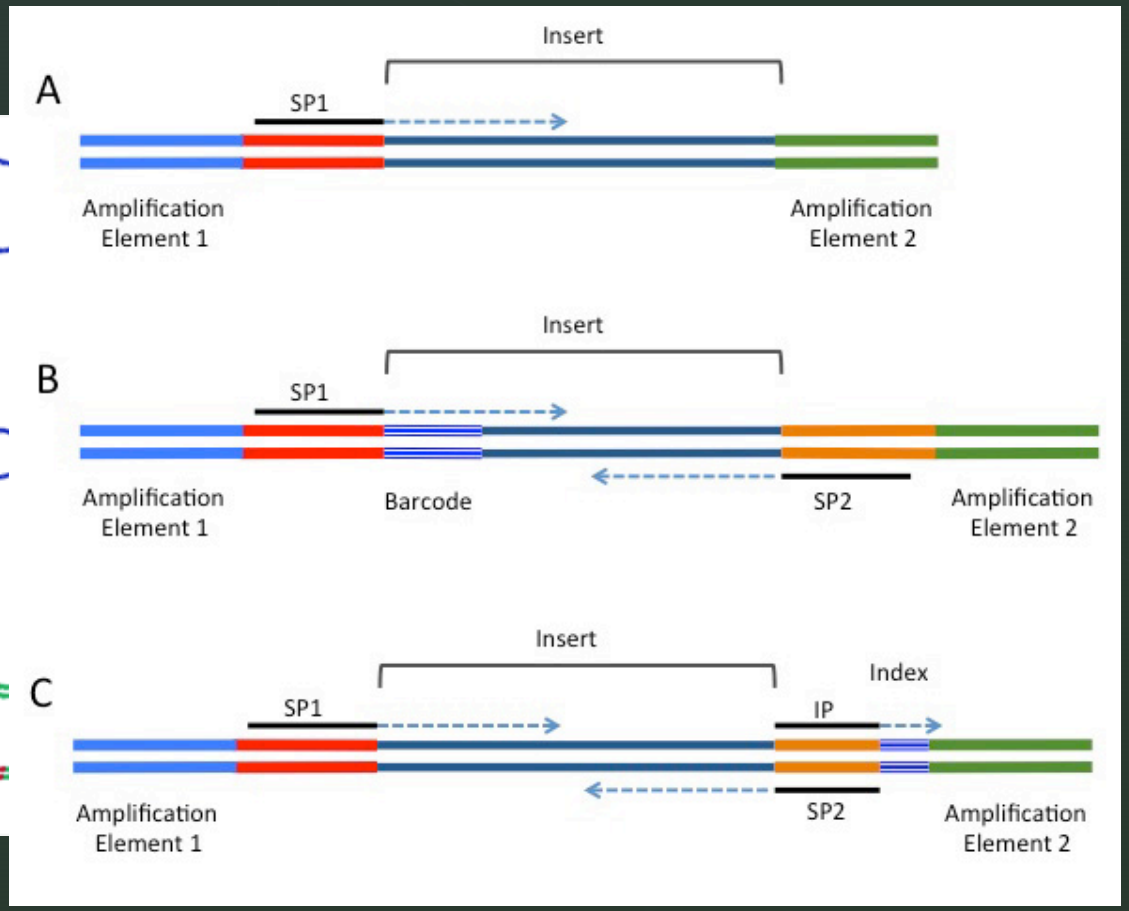
Most of the original protocols fragmented cDNA; however, fragmentation of the RNA (before converting it to cDNA) is used. Four different methods are commonly used to fragment RNA: enzymatic, metal ion, heat, and sonication.



Example of metal induced fragmentation. (Courtesy of UT FGRS.)

Library preparation

To sequence the cDNAs, specific adapter sequences must be present at the ends of the fragments. The roles and composition of the adapter sequences vary depending upon the sequencing platform. Adapters contain several different, essential functional elements.



The elements shown here are specific to the Illumina platform.

The sequencing primers are referred to as SP1 (primary sequencing primer) and SP2 (paired-end primer).

IP refers to the sequencing primer for the index read.

(A) Minimal adapter components.
(B) "In-line" barcode configuration supporting a paired-end read.
(C) Index configuration supporting a paired-end read.

Functional elements contained in sequencing adapters

Adapter element	Requirement	Location	Function
Amplification element	Required	5′ and 3′ terminus	Clonal amplification of the construct
Primary sequencing priming site	Required	next to the insert	Initiating the primary sequencing
Barcode/Index	Optional	5′-end of the insert	unique label of different samples.
Paired-end sequencing priming site	Optional	Adjacent to insert	Sequencing the insert 3’5’
Index sequencing priming site	Optional		Sequencing of the index

Library preparation. Multiplexing samples

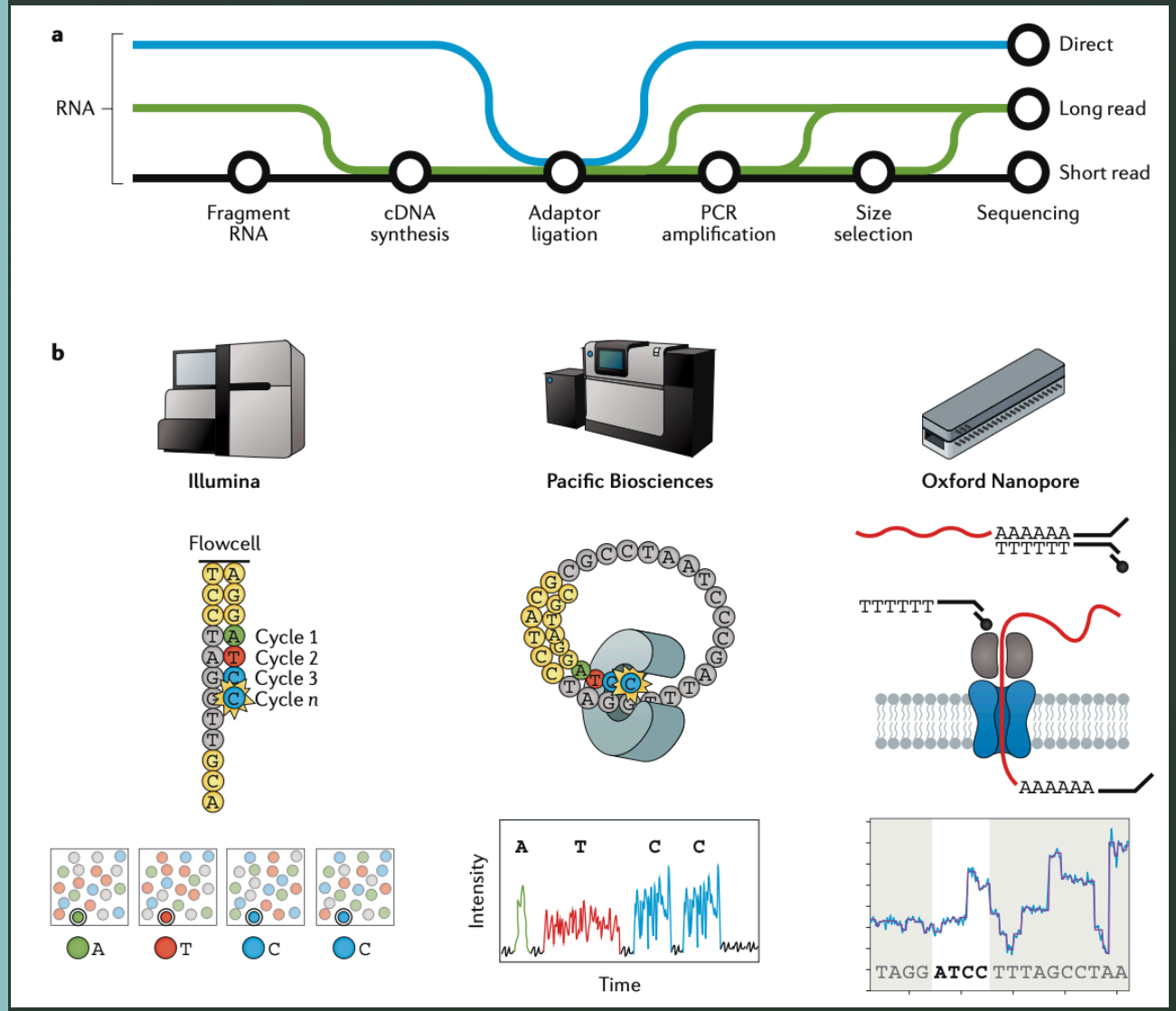
Most high-throughput sequencers produce millions of reads in a single reaction.

The number of reads that can be obtained can exceed the needs of the experiment.

It is often desirable to pool ("multiplex") libraries from multiple experiments into a single sequencing reaction. To identify which experiment/ sample a given sequence comes from, each library is prepared using adapters containing different tags (commonly known as an index or barcode).

The tags are typically short (~8nt) sequences. The sequence of the insert and tag can be associated, allowing one to identify which sample the insert came from.

RNAseq. Sequencing



The Illumina short-read sequencing technology was used to generate more than 95% of the published RNA-seq data available on the Short Read Archive (SRA)

However, long-read cDNA sequencing and, most recently, dRNA-seq methods may soon present a challenge to its dominance, as users seek out methods that can deliver improved isoform-level data

Figure from Stark et al

RNAseq. sequencing

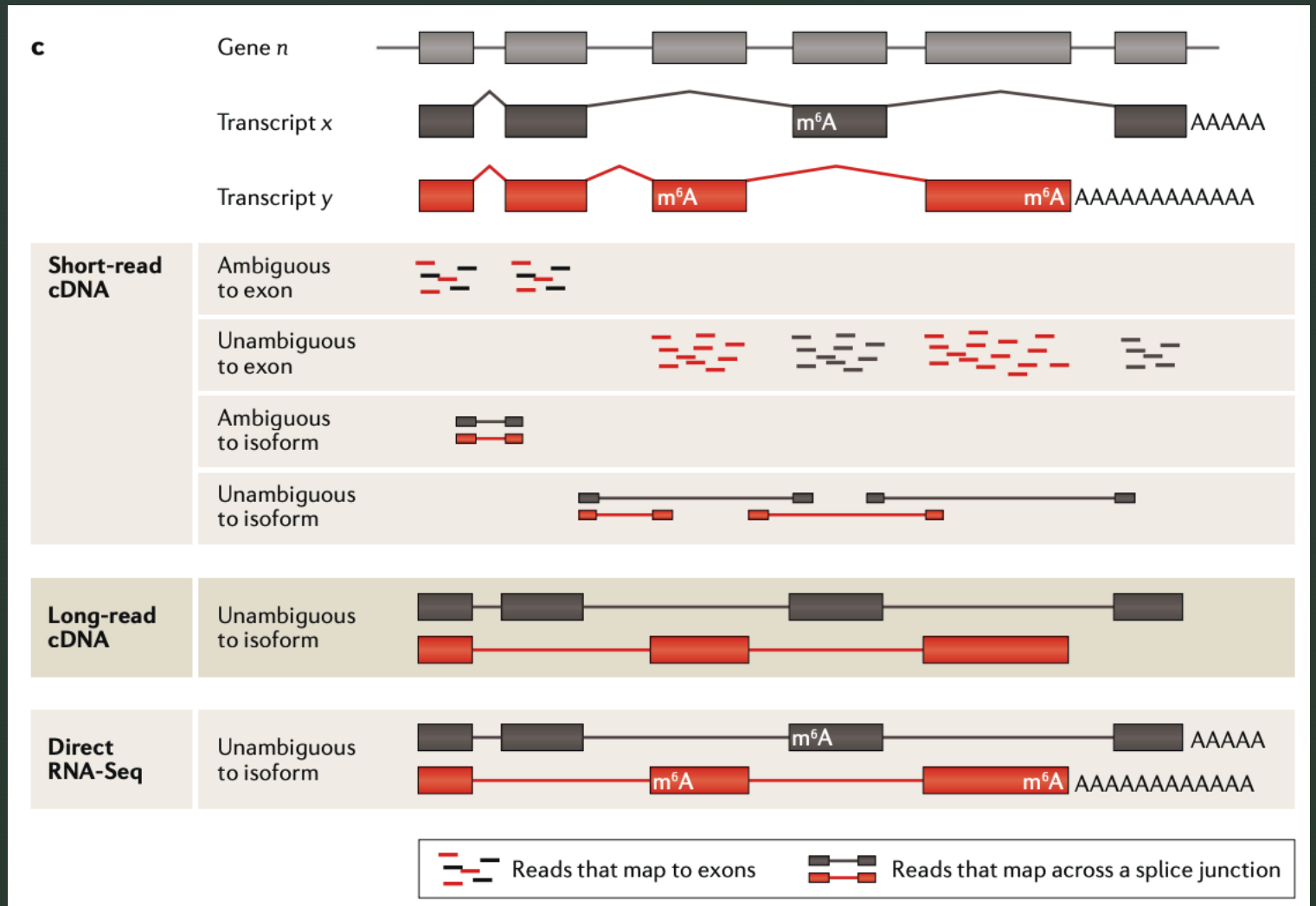


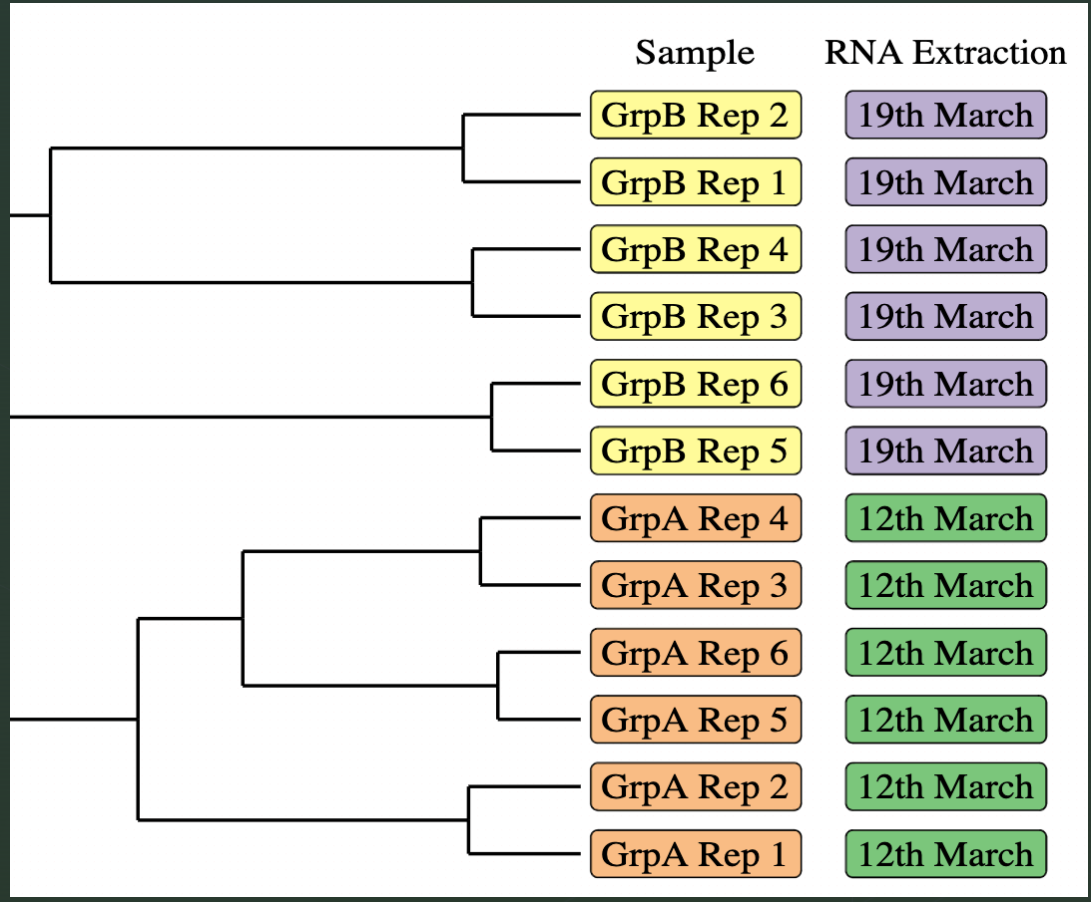
Figure from Stark et al

RNAseq. Sequencing Batch effects

Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study.

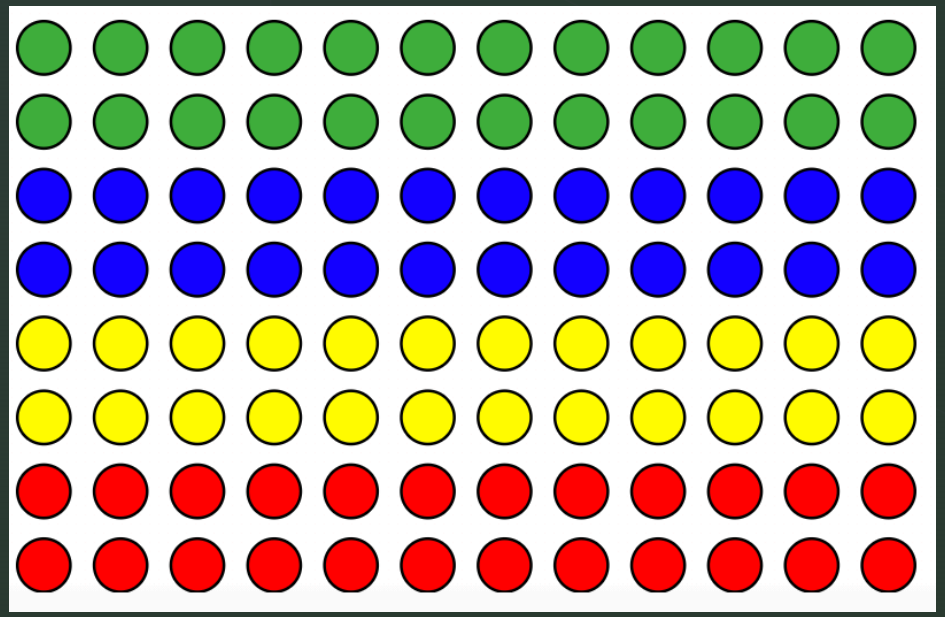
Batch effects are problematic if they are confounded with the experimental variable.

Batch effects that are randomly distributed across experimental variables can be controlled for.

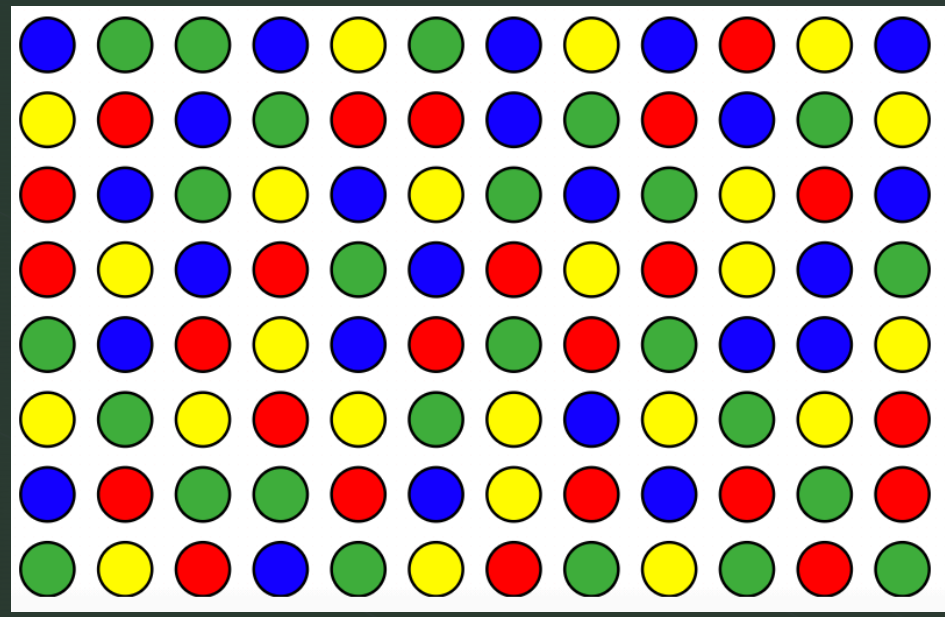


RNAseq. Sequencing Batch effects

Solution to avoid batch effects: randomisation



Structured distribution of samples.



Randomized distribution of samples

Data analysis. Gene abundances

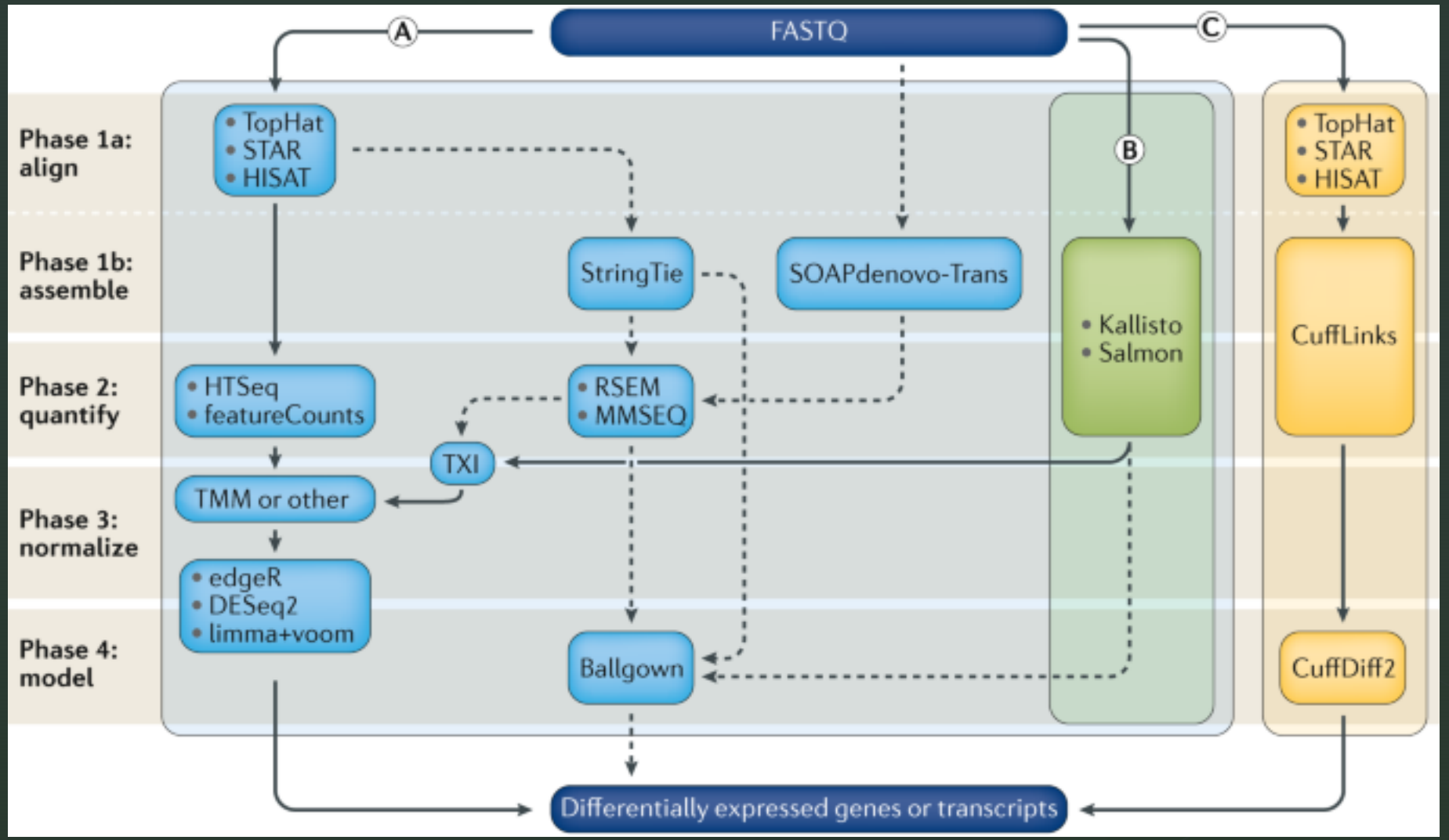


Figure from Stark et al