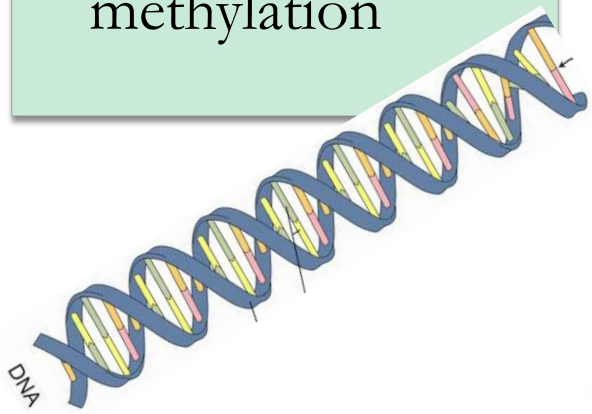
RNA-SEQ DATA ANALYSIS: TRANSCRIPTOME ASSEMBLY AND DIFFERENTIAL EXPRESSION ANALYSIS

Ashley Sawle (Ashley.Sawle@cruk.cam.ac.uk)
Guillermo Parada (guillermo.parada@sanger.ac.uk)

HTS Applications - Overview

DNA Sequencing

- Genome Assembly
- SNPs
- DNA methylation

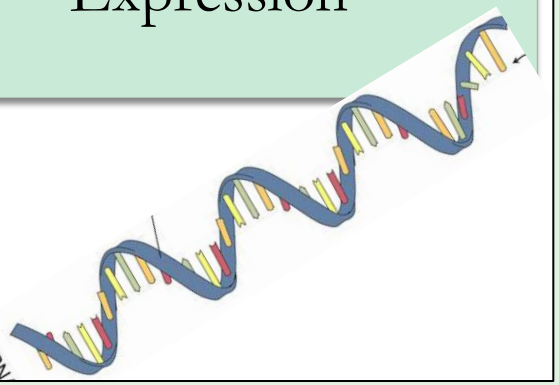


ChIP-sequencing

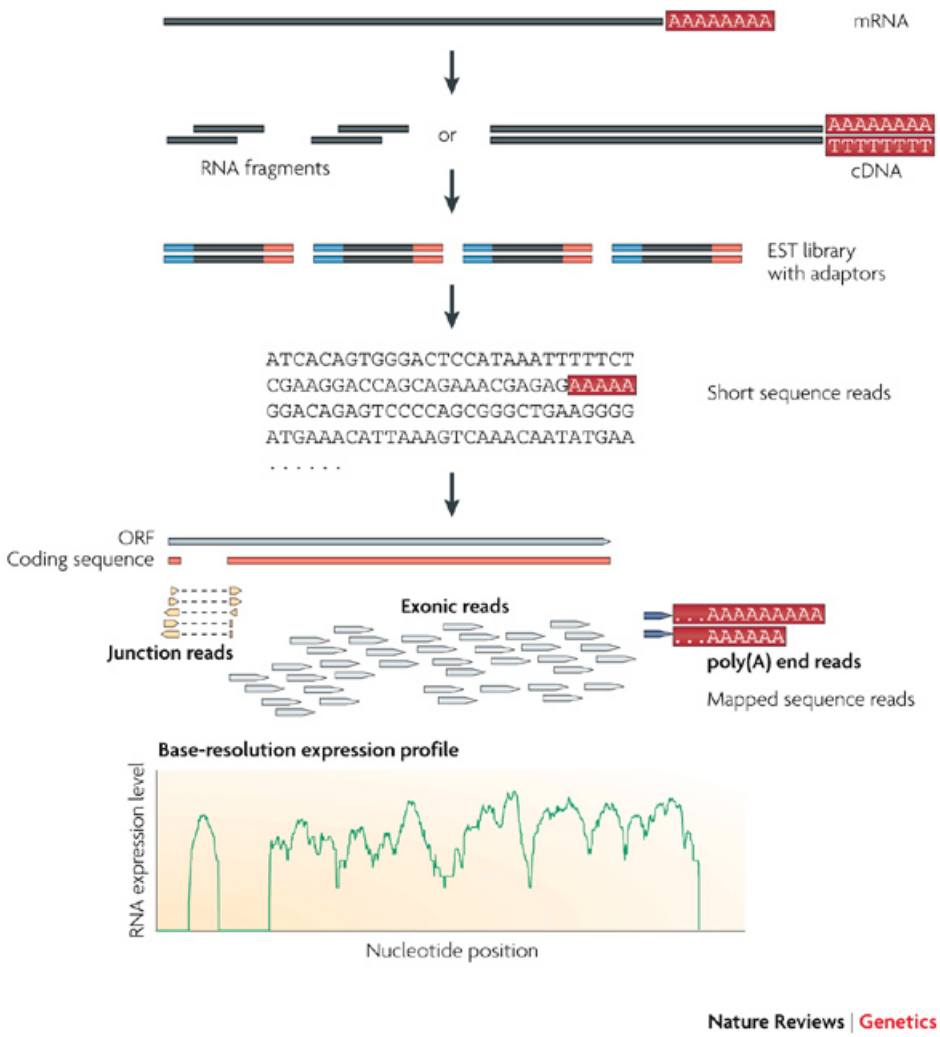
- Transcription Factor Binding Sites
- Chromatin Modification Regions

RNA-sequencing

- Transcriptome Assembly
- Gene Expression
- Differential Expression



RNA-seq workflow



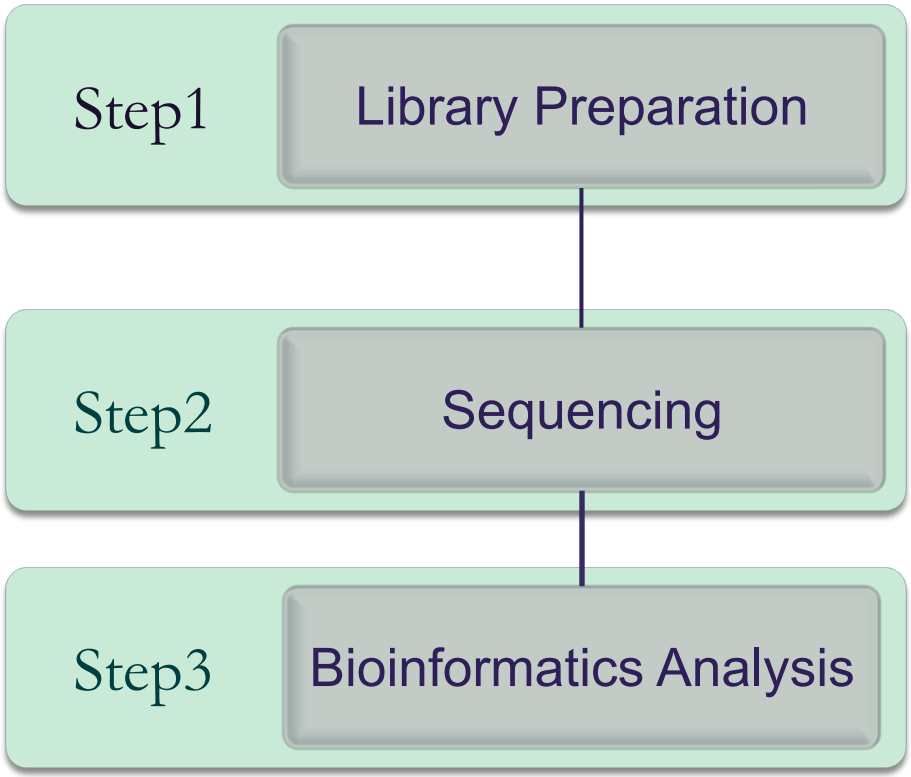
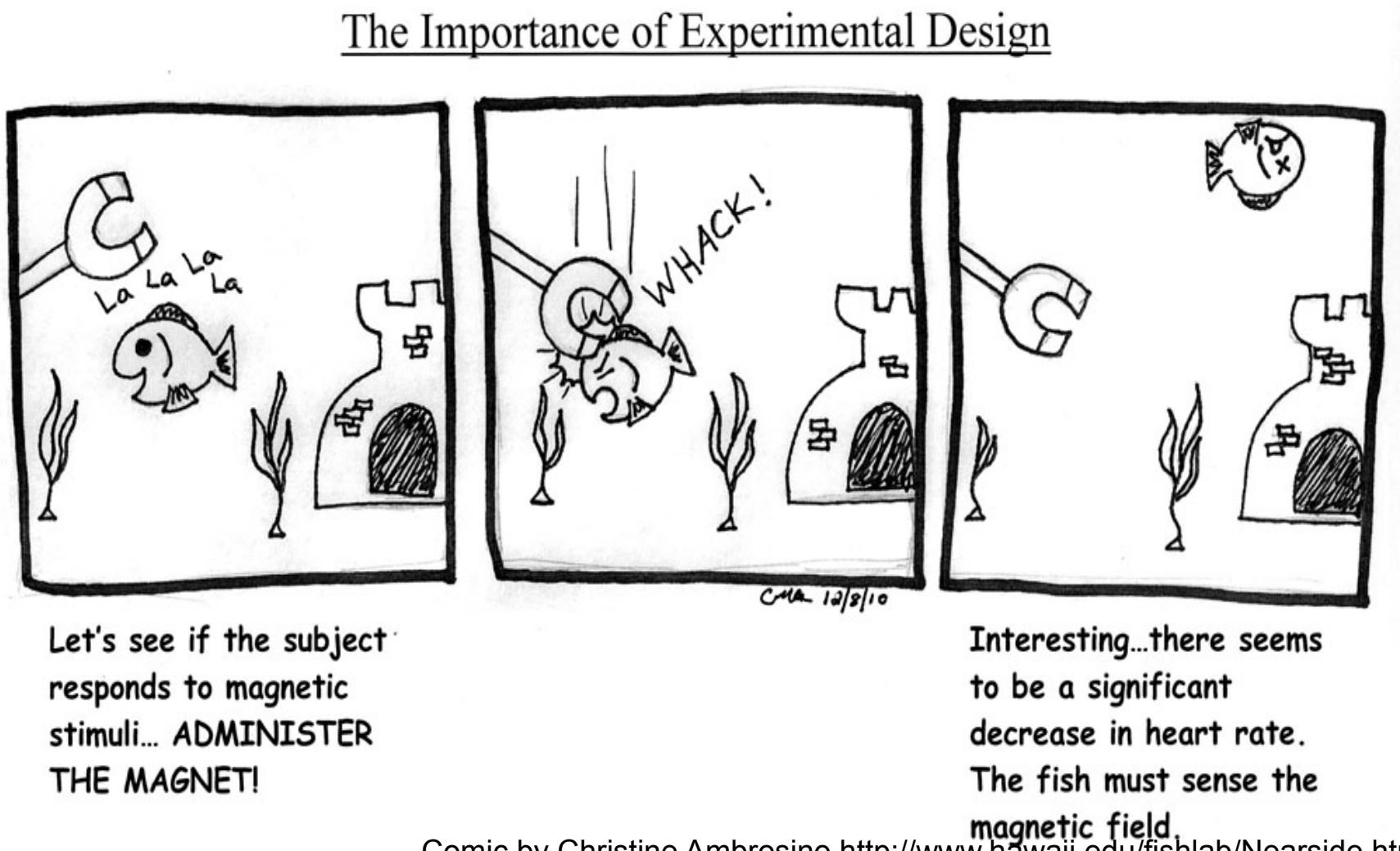


Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

Designing the right experiment



Comic by Christine Ambrosino http://www.hawaii.edu/fishlab/Nearside.htm

Designing the right experiment

- The design of the experiment is the first step and it is obviously determinant for all downstream analyses

- You have to evaluate all the eventualities and limitations of available technologies, designing the experiment according to your goals

Designing the right experiment

COVERAGE: How many reads do we need?

The coverage is defined as $C = (R_{length} \times R_{num}) / A_{length}$

R_{length} = length in nucleotides of the reads

R_{num} = number of sequenced reads

A_{length} = number of nucleotides of sequenced subject (genome, transcriptome, exome)

The amount of sequencing needed for a given sample is determined by the goals of the experiment and the nature of the RNA sample.

Designing the right experiment

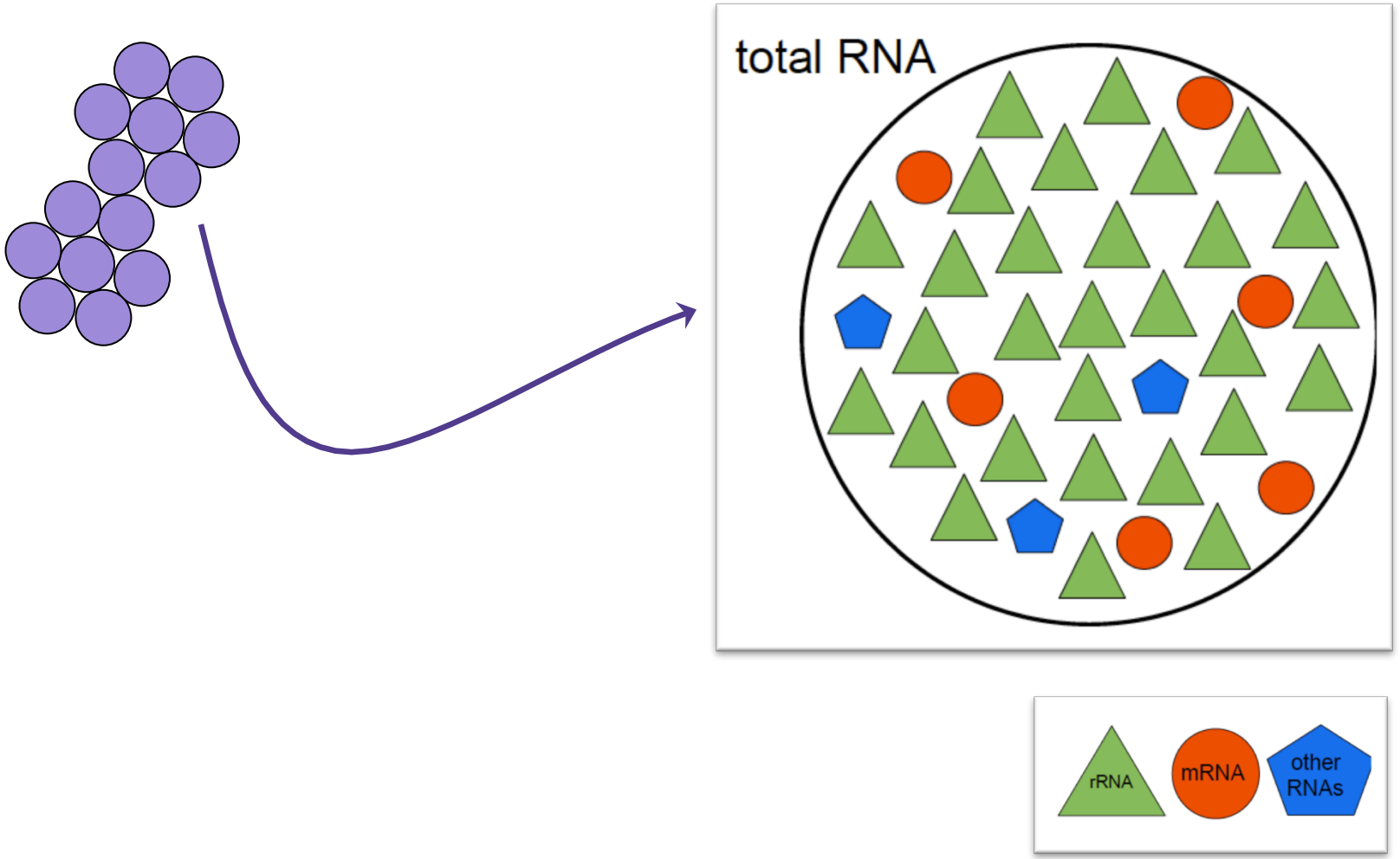
READ LENGTH: long or short reads?

The answer depends again on the experiment:

- GENOME RESEQUENCING
- De novo TRANSCRIPTOME
- TRANSCRIPTOME seq
- ChIP seq

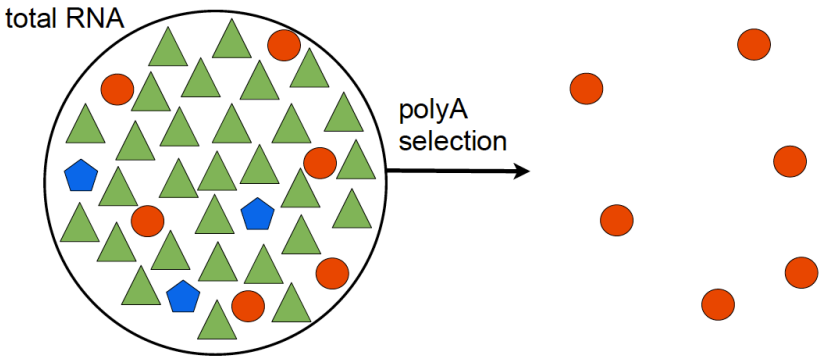
Read length is inversely proportional to the multi-mappability of a read, in a sample of 50 nt reads there is a small fraction (<0.01 %) that can be mapped to multiple positions of the human genome.

Step1 – Library Preparation



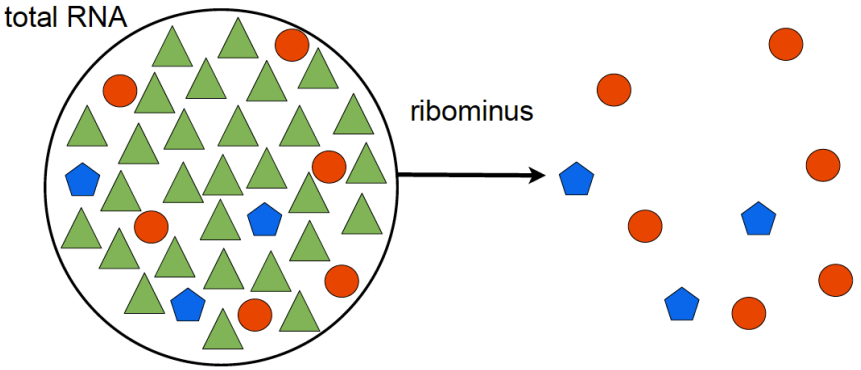
Step 1 – Library Preparation

polyA selection

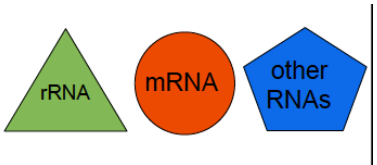


- poly(A+)-transcripts:
  - mRNAs
  - immature microRNAs
  - snoRNAs

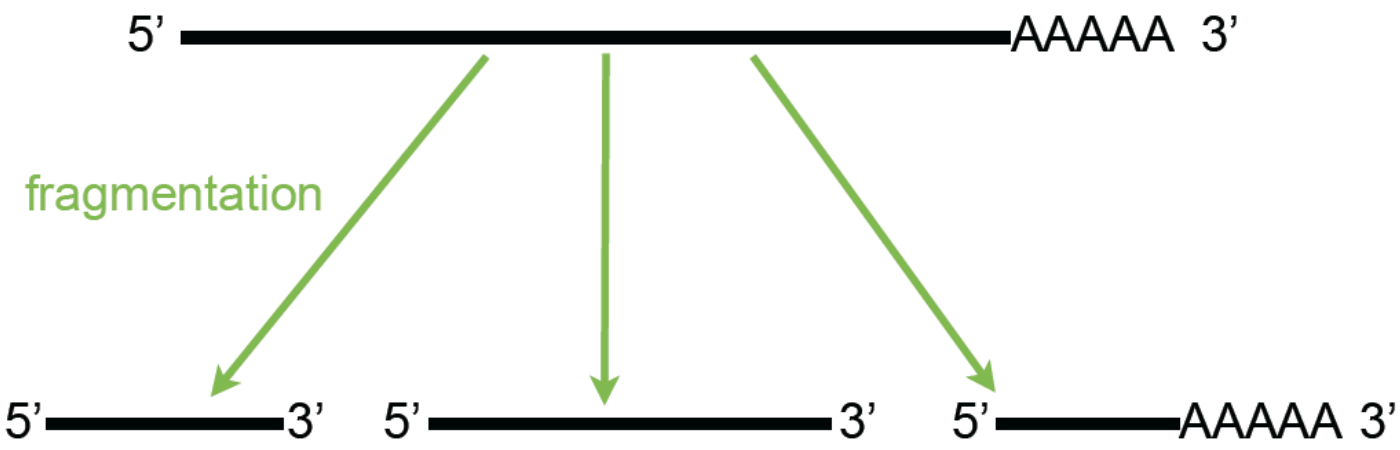
ribominus selection



- non poly(A+)-transcripts:
  - mRNAs
  - histone mRNAs
  - tRNAs
  - other small RNAs



Step 1 – Library Preparation



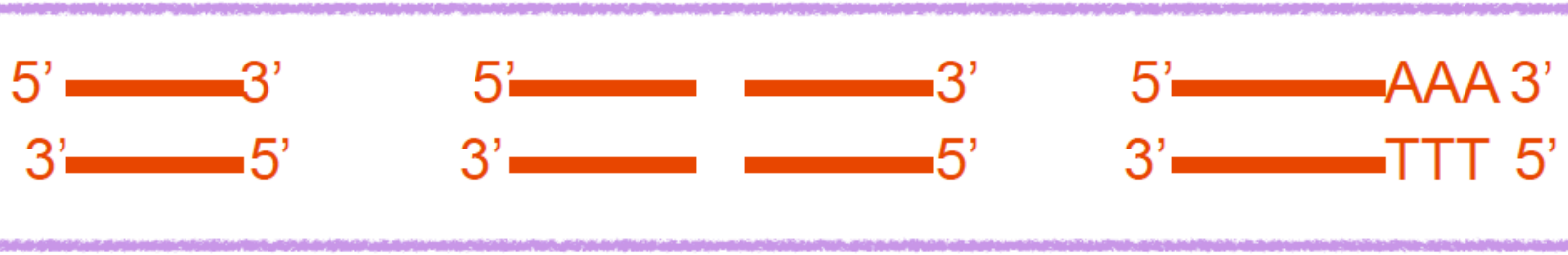
Step 1 – Library Preparation

1 strand cDNA synthesis

remove RNA strand

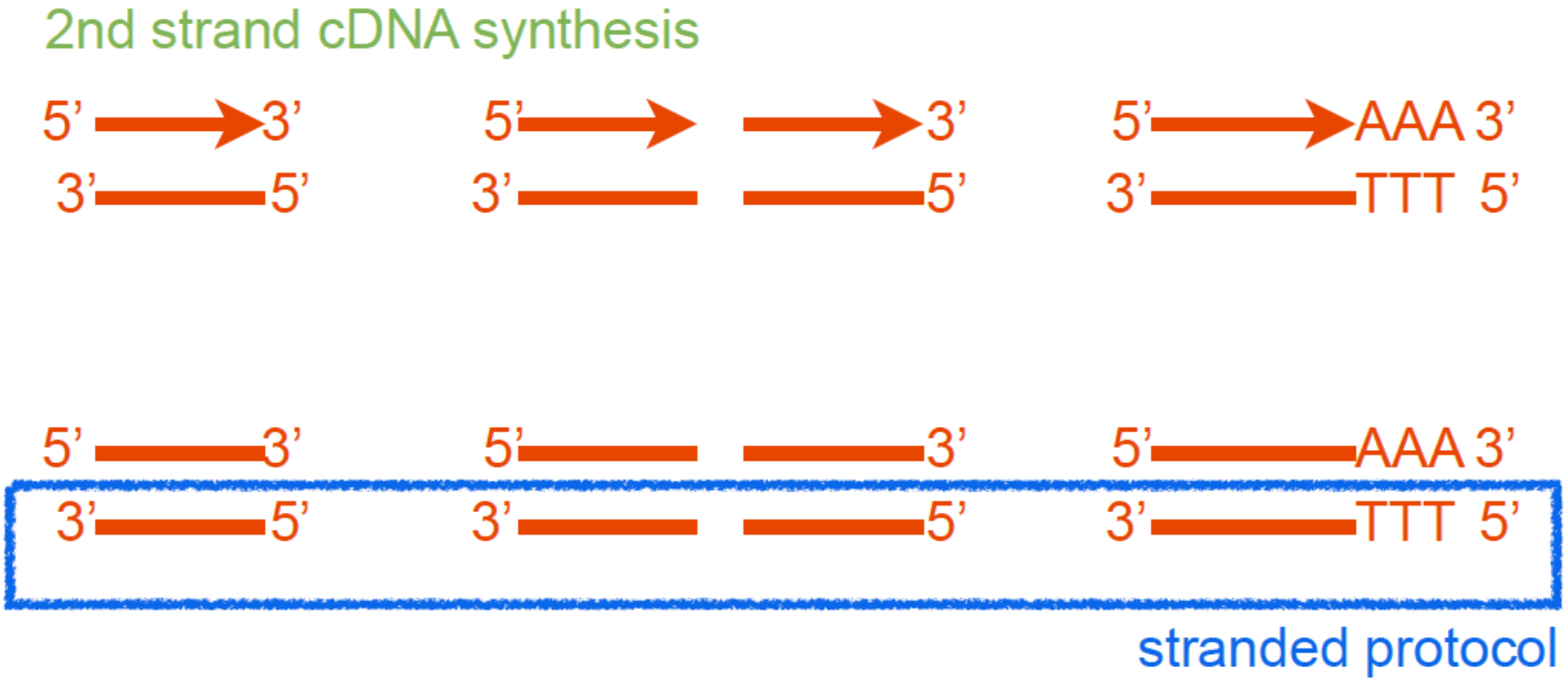
Step 1 – Library Preparation

2nd strand cDNA synthesis



Unstranded protocol

Step 1 – Library Preparation



Parkhomchuk, D., et al. (2009), Nucleic Acids Res, 37 (18), e123.

Step 1 – Library Preparation

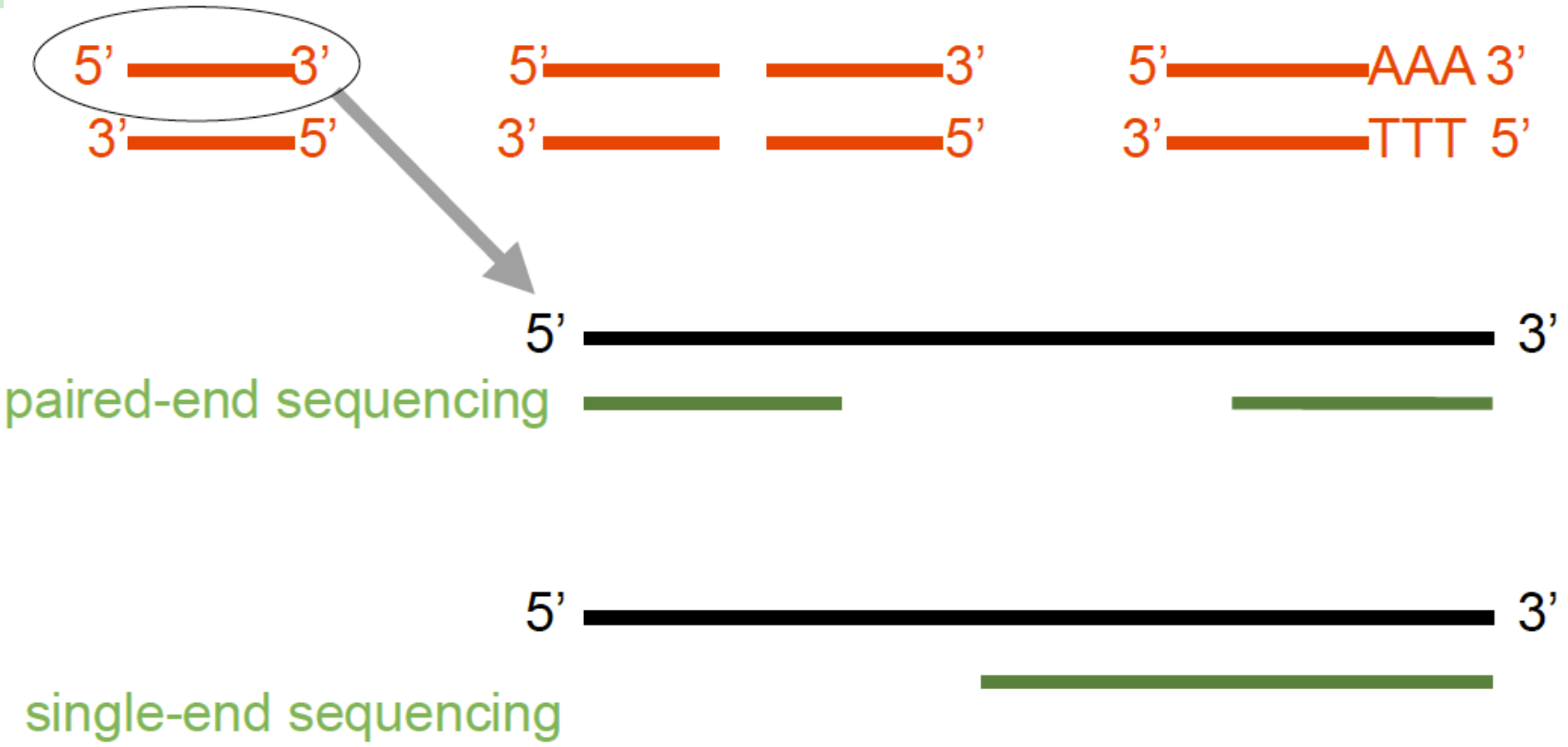
Fragmented cDNA

cDNA with adaptors

Size selection

PCR amplification

Step 2 – Sequencing



Single- vs paired-end sequencing

my_sequence.fastq

```
@HWI-BRUNOP16X_0001:1:1:1466:1018#0/1
AAGGAAGTGCTTGTCTGGCTAACACAGCNAGNCACGTGAC
+
aVfbe`^^^_TTTSSdffffdfffabbZbbfebafbbbbb
```

SE

my_sequence_1.fastq

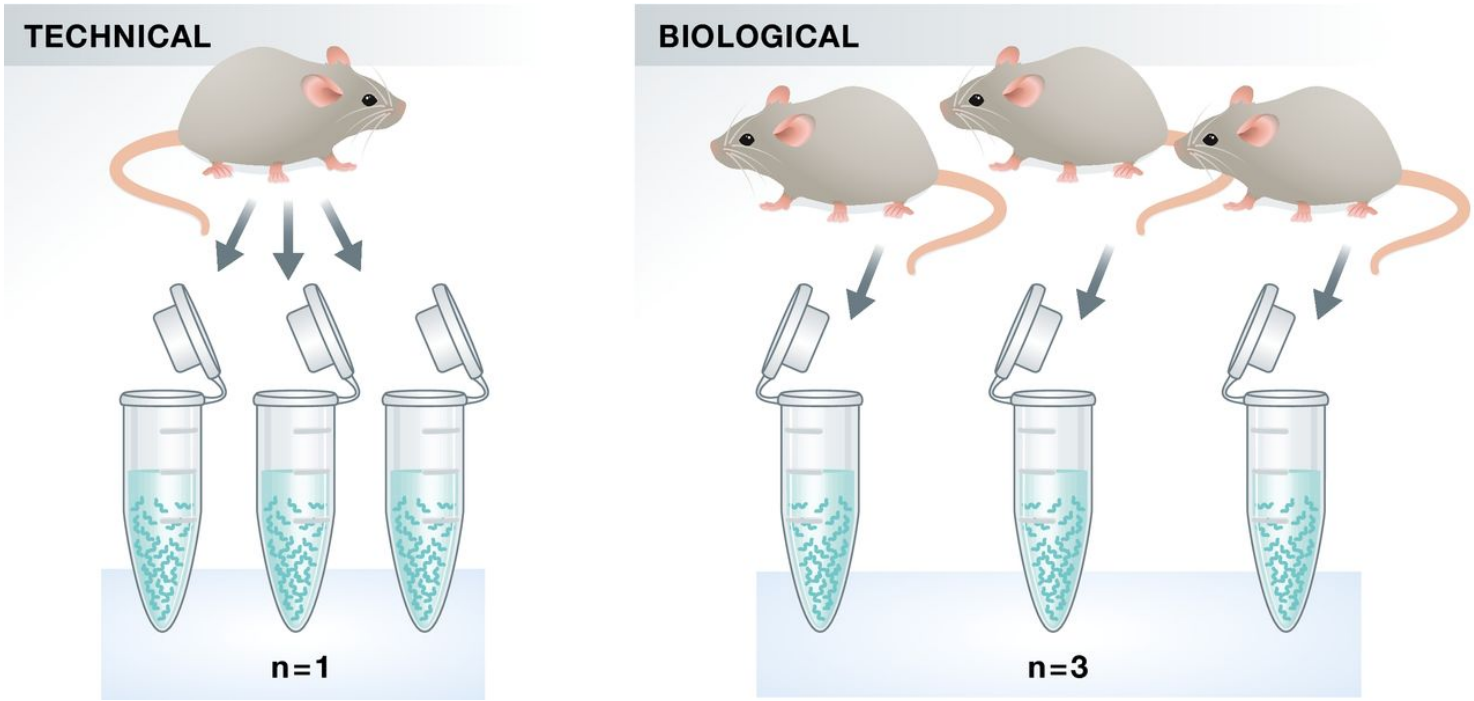
```
@HWI-BRUNOP16X_0001:1:1:1278:989#0/1
NAAATTTCGAATTTCTGTGAAGTAAGCATCTTCTTTGTCAT
+
BJJGGKIINN^^^^^QQNTUQOOTTTRTOTY^^Y^\\^^^\
```

my_sequence_2.fastq

```
@HWI-BRUNOP16X_0001:1:1:1278:989#0/2
AACCCACACAGGAGAGCAGCCTTACAGATGCAAATACTGTG
+
]K___fffffggghegggggggdggggfgggggeggggghh
```

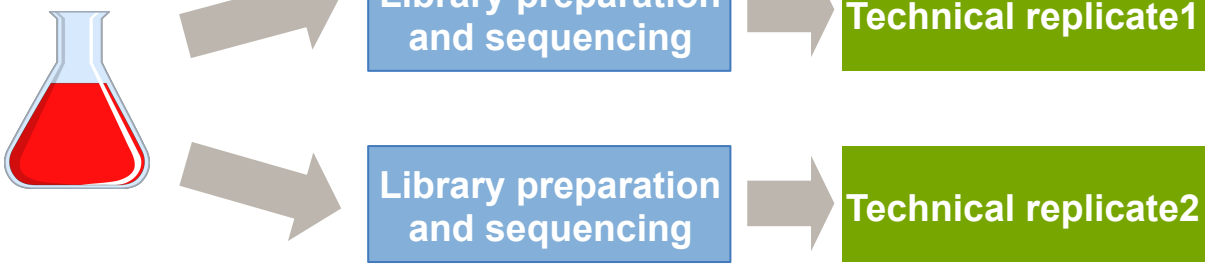
PE

Replicates – do I need them?



10.15252/embj.201592958 | Published online 21.09.2015 The EMBO Journal (2015) e201592958

Replicates – do I need them?

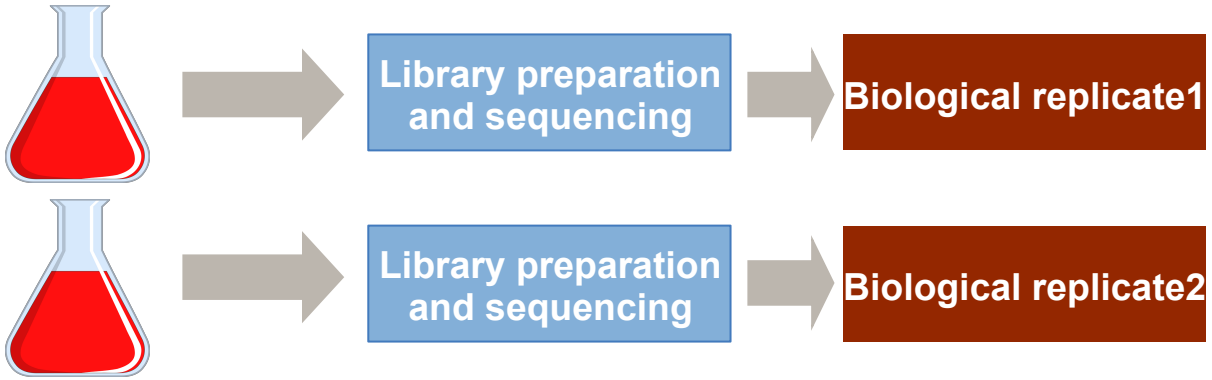


Technical

Same biological sample – same conditions

measure technical variation

Replicates – do I need them?



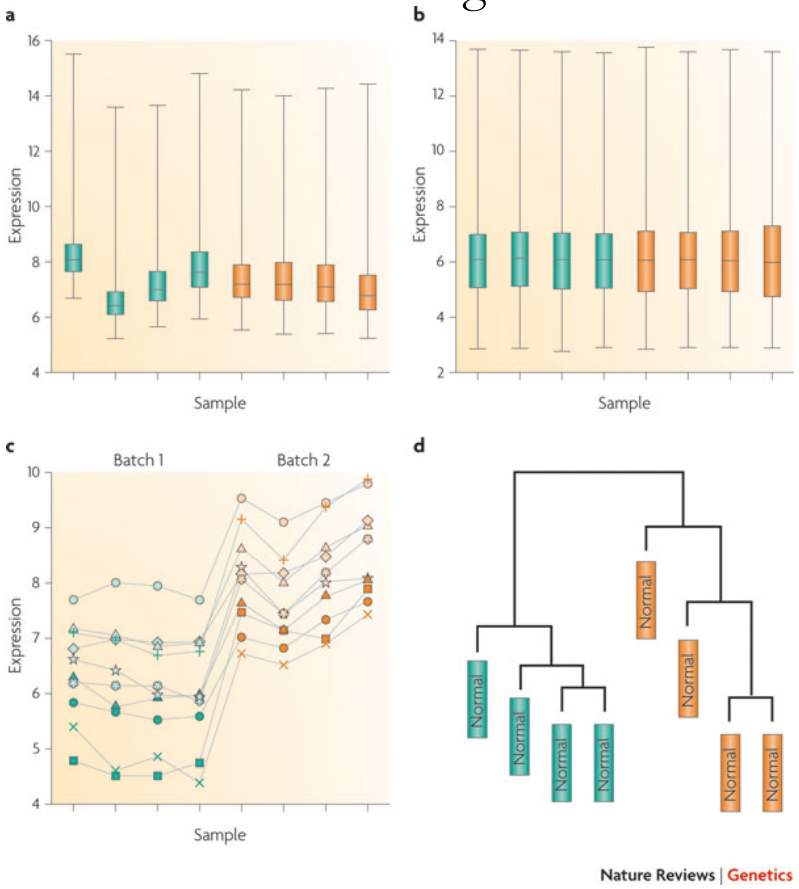
Biological

Different biological sample – same conditions

measure biological variation

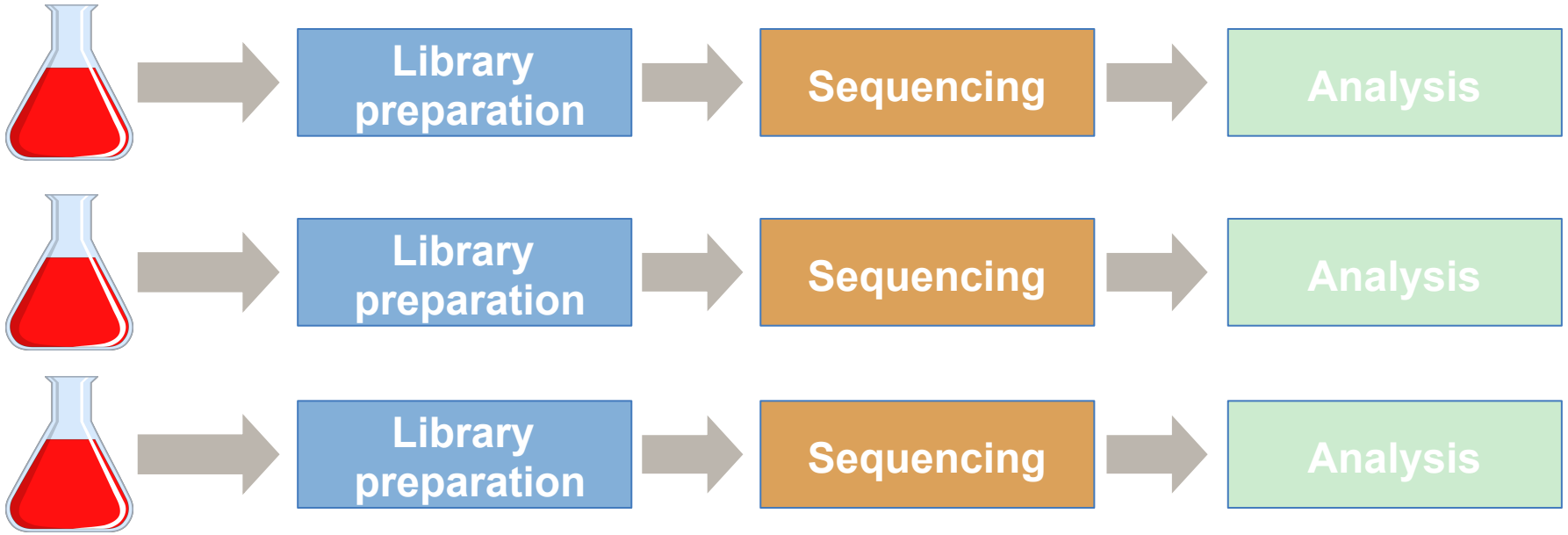
Controlling batch effects

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

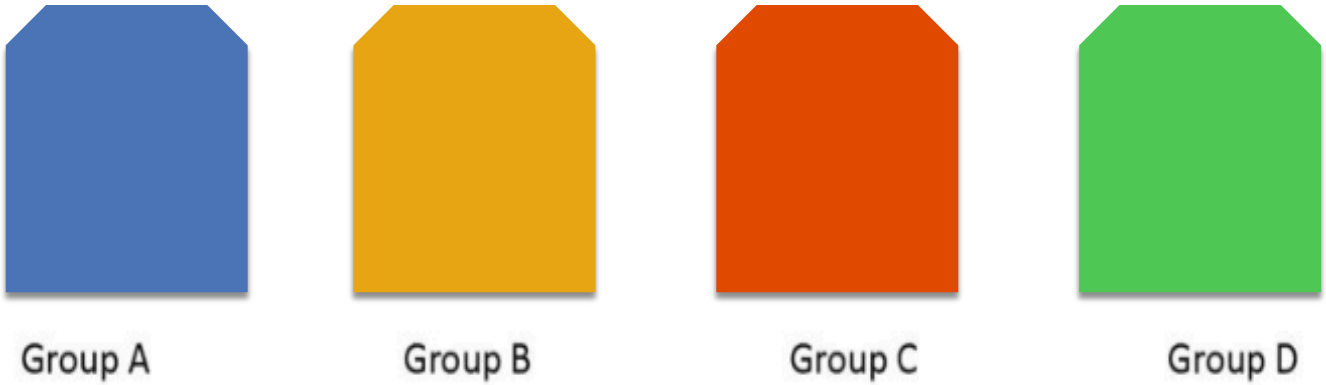


Leek et al. Nature Reviews Genetics 11, 733-739 (October 2010) | doi:10.1038/nrg2825

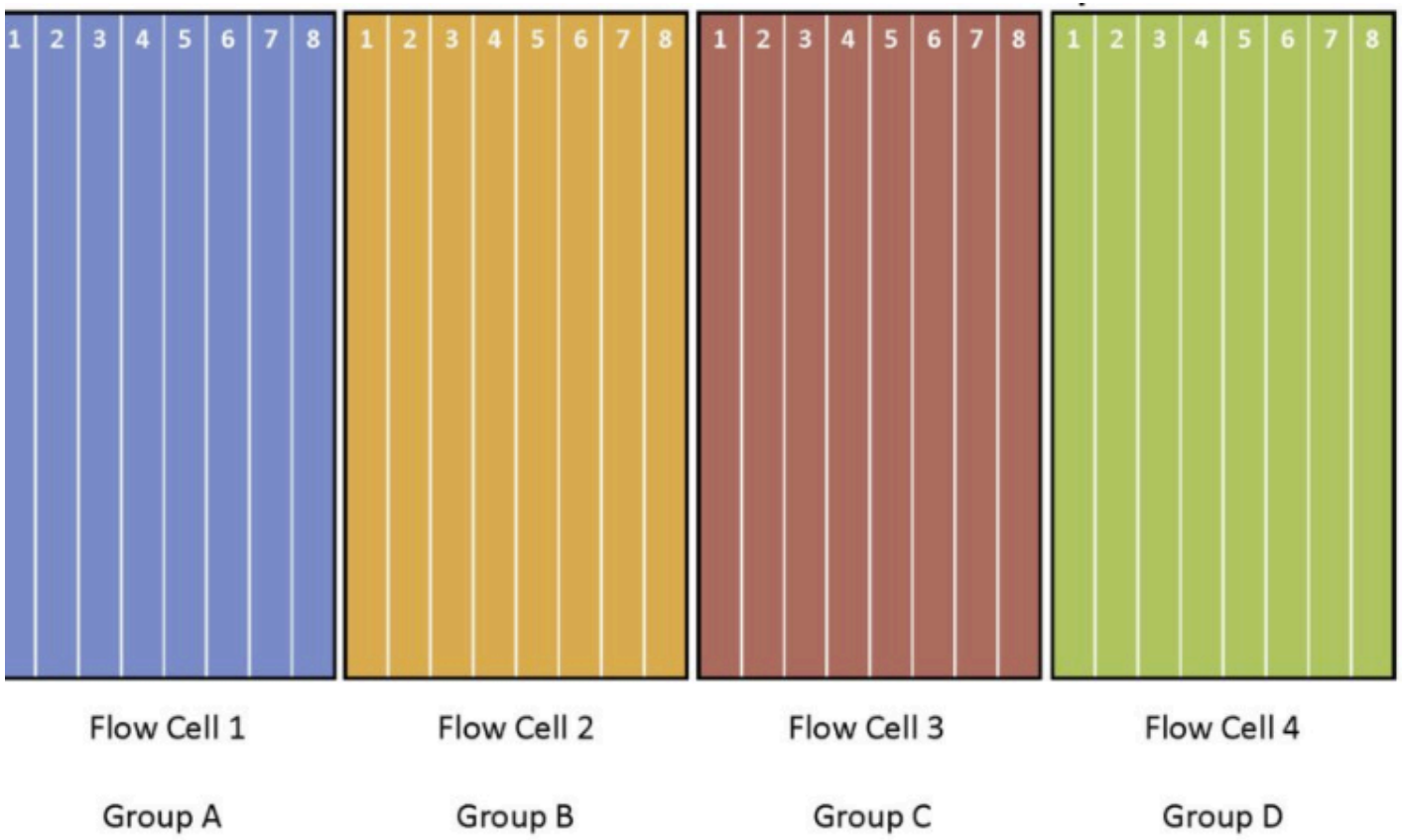
Controlling batch effects



Example of experimental design

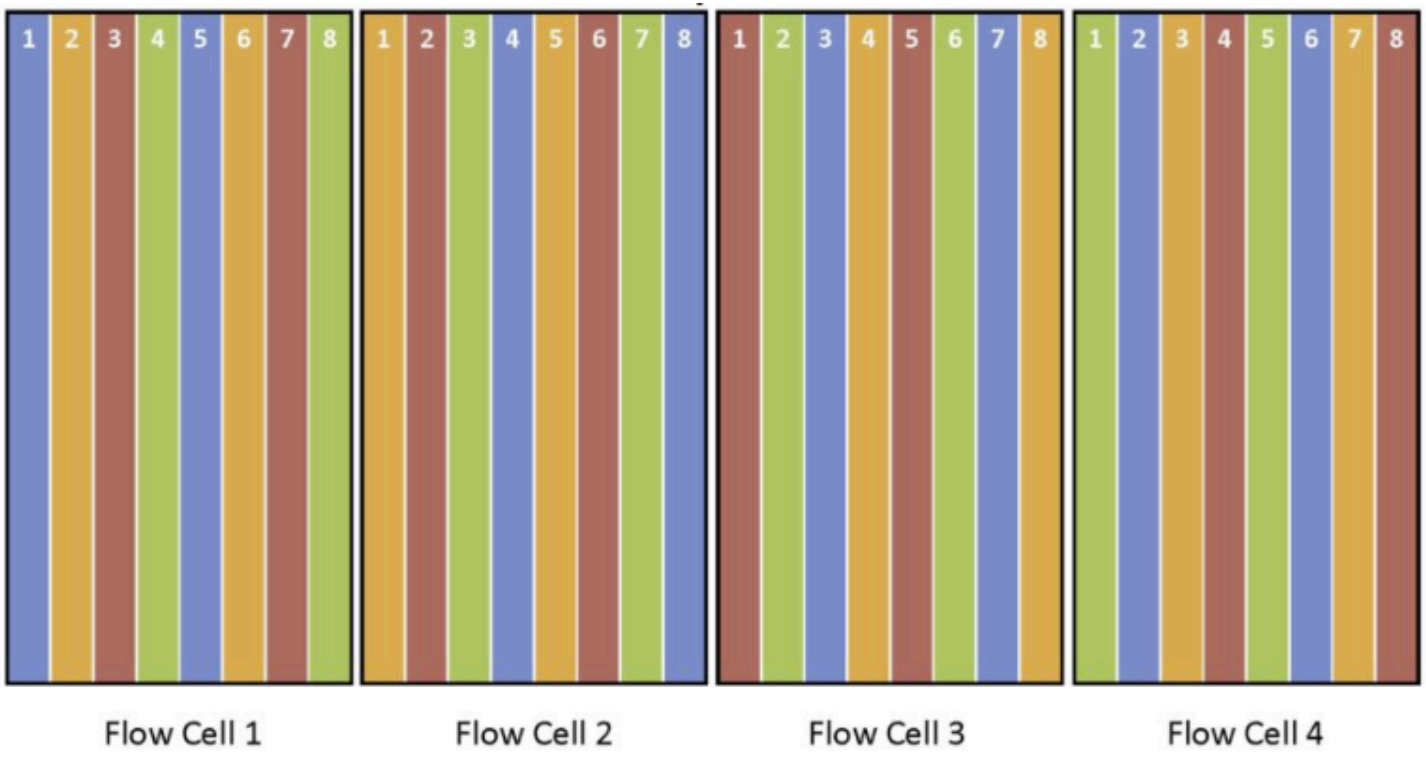


Example of experimental design

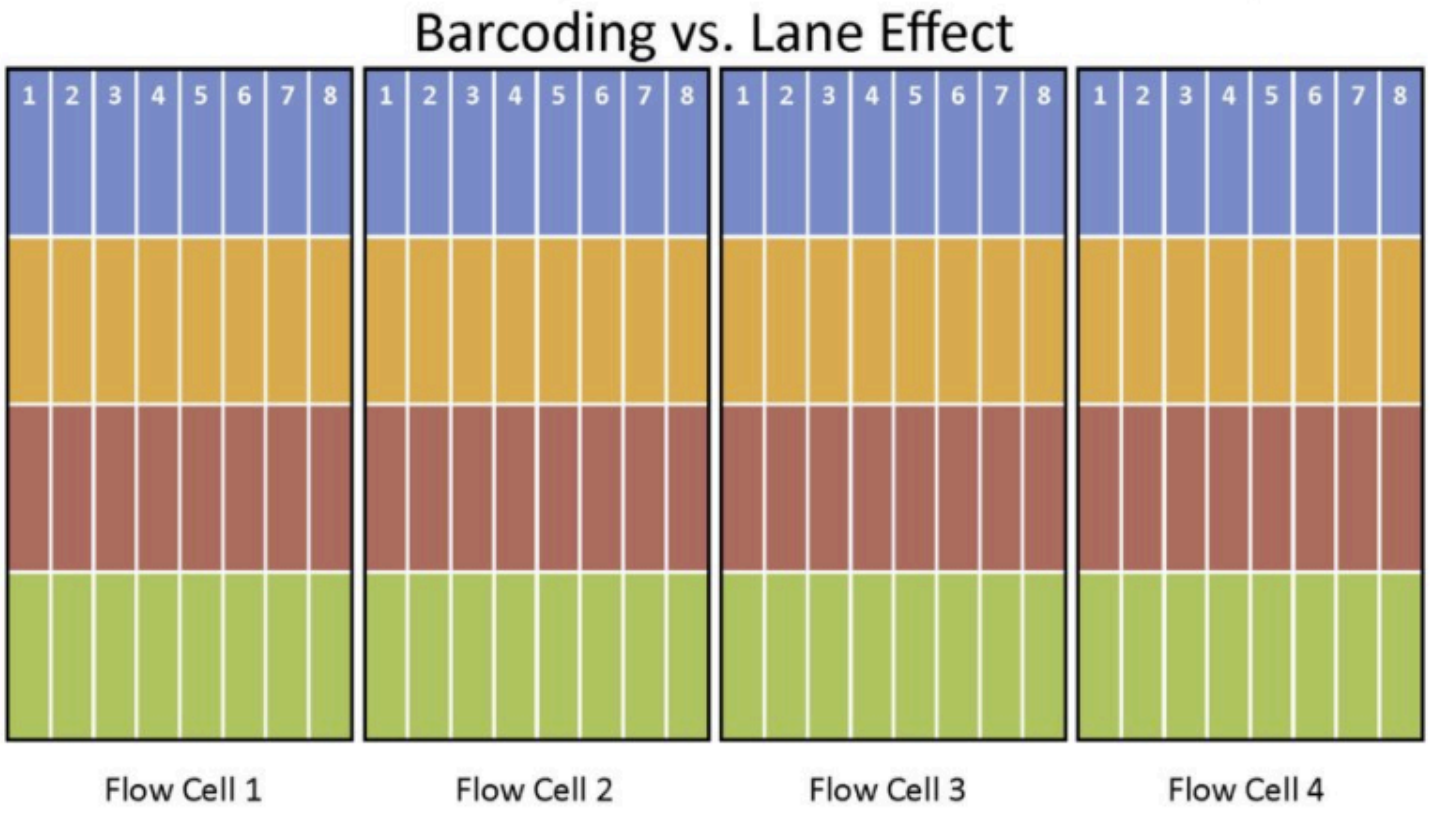


...better experimental design

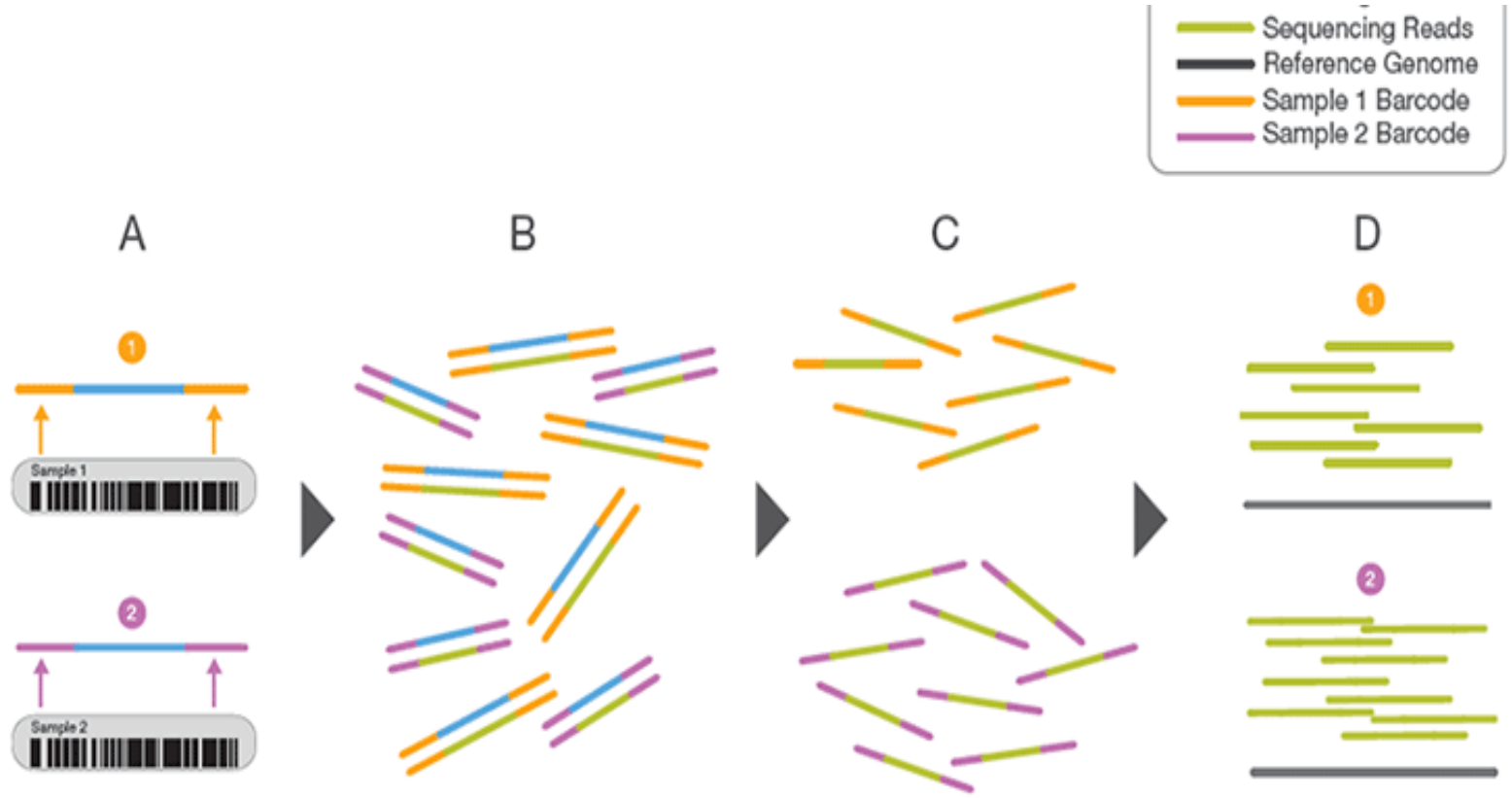
- Randomize samples with respect to the flow cell



...even better experimental design



Multiplexing to prevent batch effects



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.

B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.