

*IN-BIOS[9,5]000 2018*

---

# Experimental Design - from a HTS perspective

---

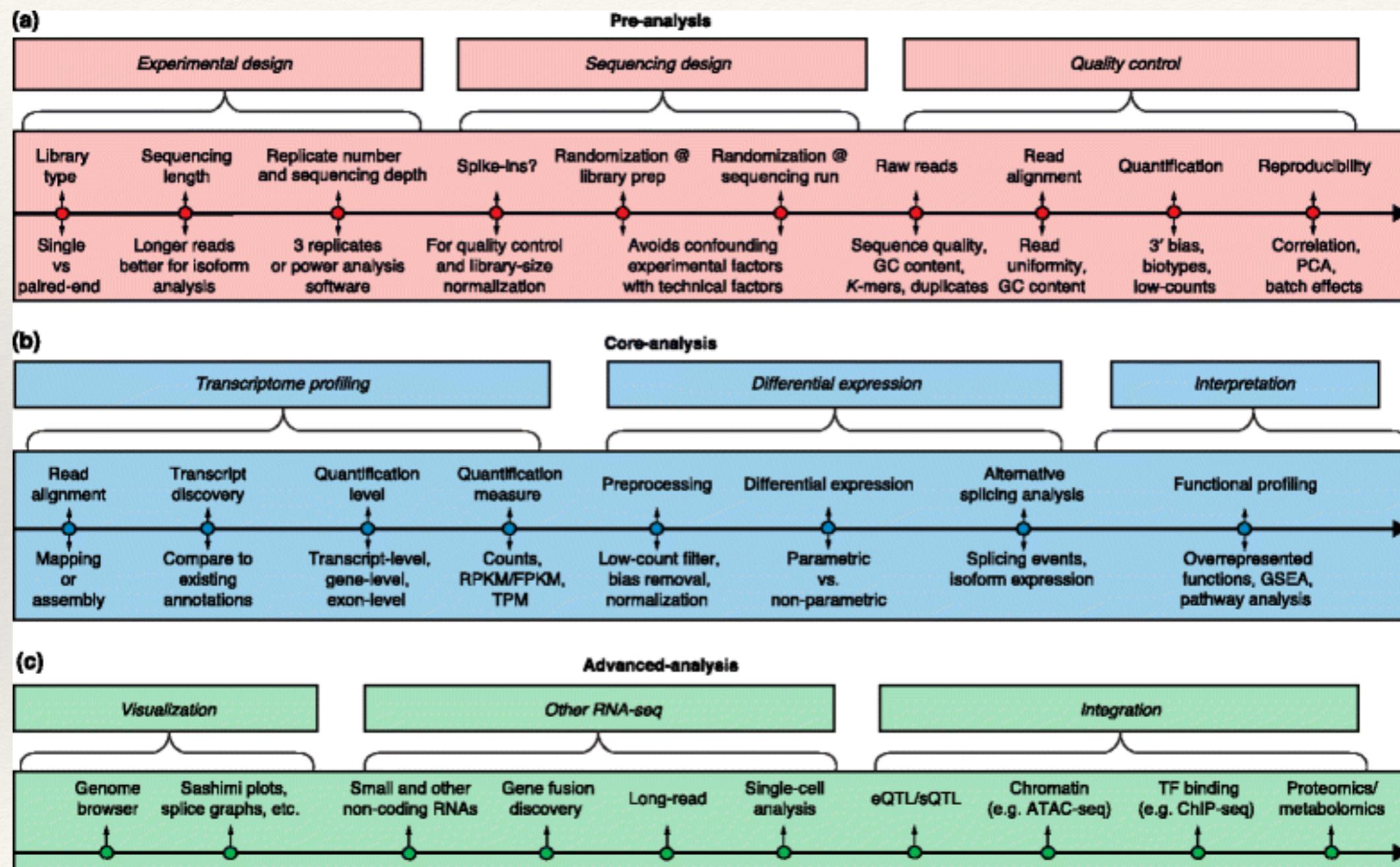
Arvind Sundaram  
Oct 23, 2018

---

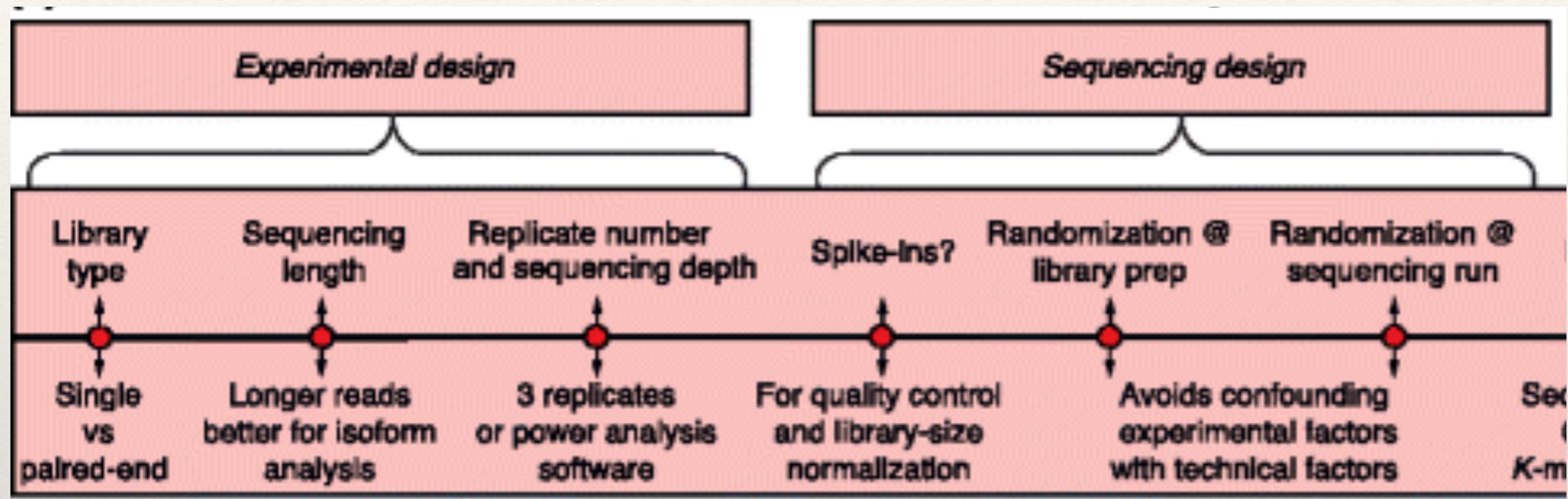
Norwegian Sequencing Centre  
OUS, Ullevål, Oslo

---

# RNA seq analysis pipeline



# Experimental design



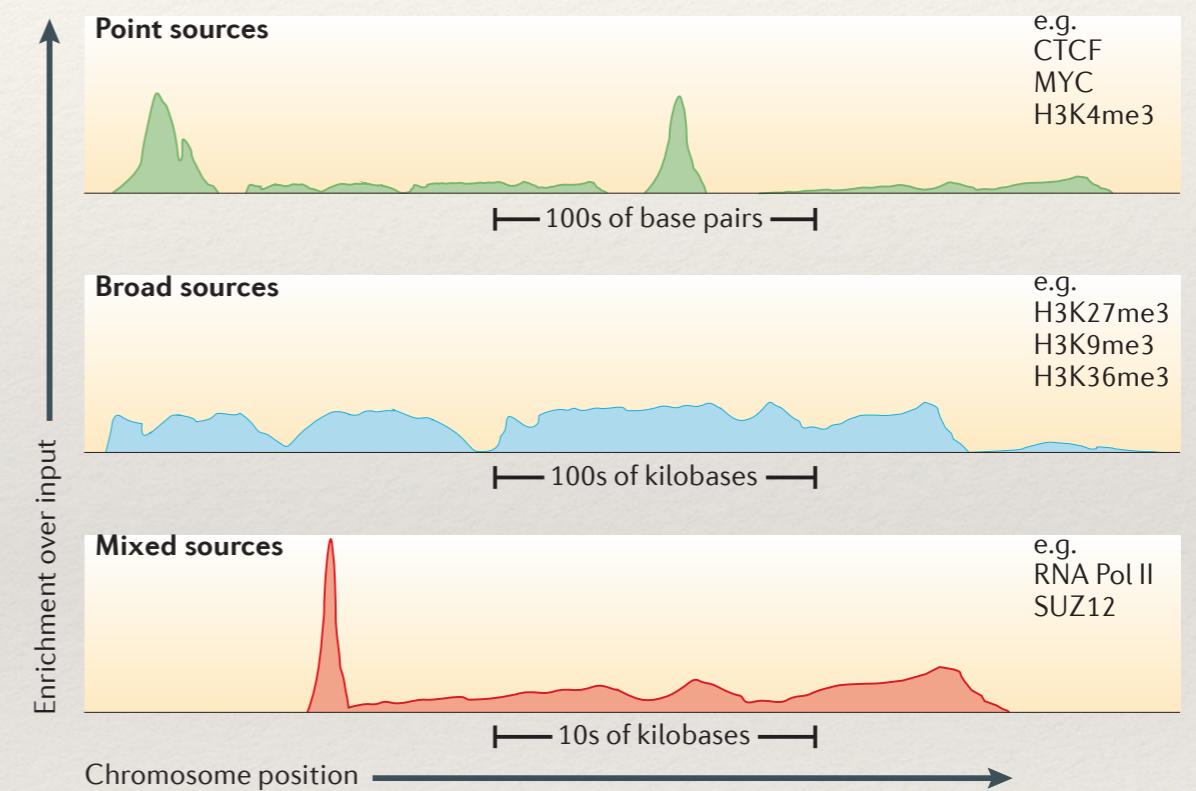
Design of the experiment and sequencing plan have a direct effect on downstream analyses and interpretation of data

# Experimental design

- ❖ Biological question
- ❖ Platform choice
- ❖ Technology variation
  - ❖ Technical bias
    - ❖ Run/Lane bias
    - ❖ Index/barcode bias
    - ❖ Duplicates
    - ❖ Error rates
  - ❖ Sample variation
    - ❖ PCR amplification?
  - ❖ Sequencing depth
  - ❖ Data analysis
  - ❖ Species-specific information
    - ❖ Is there a genome sequence available??
    - ❖ Genome size (c-value)
    - ❖ [genomesize.com](http://genomesize.com)

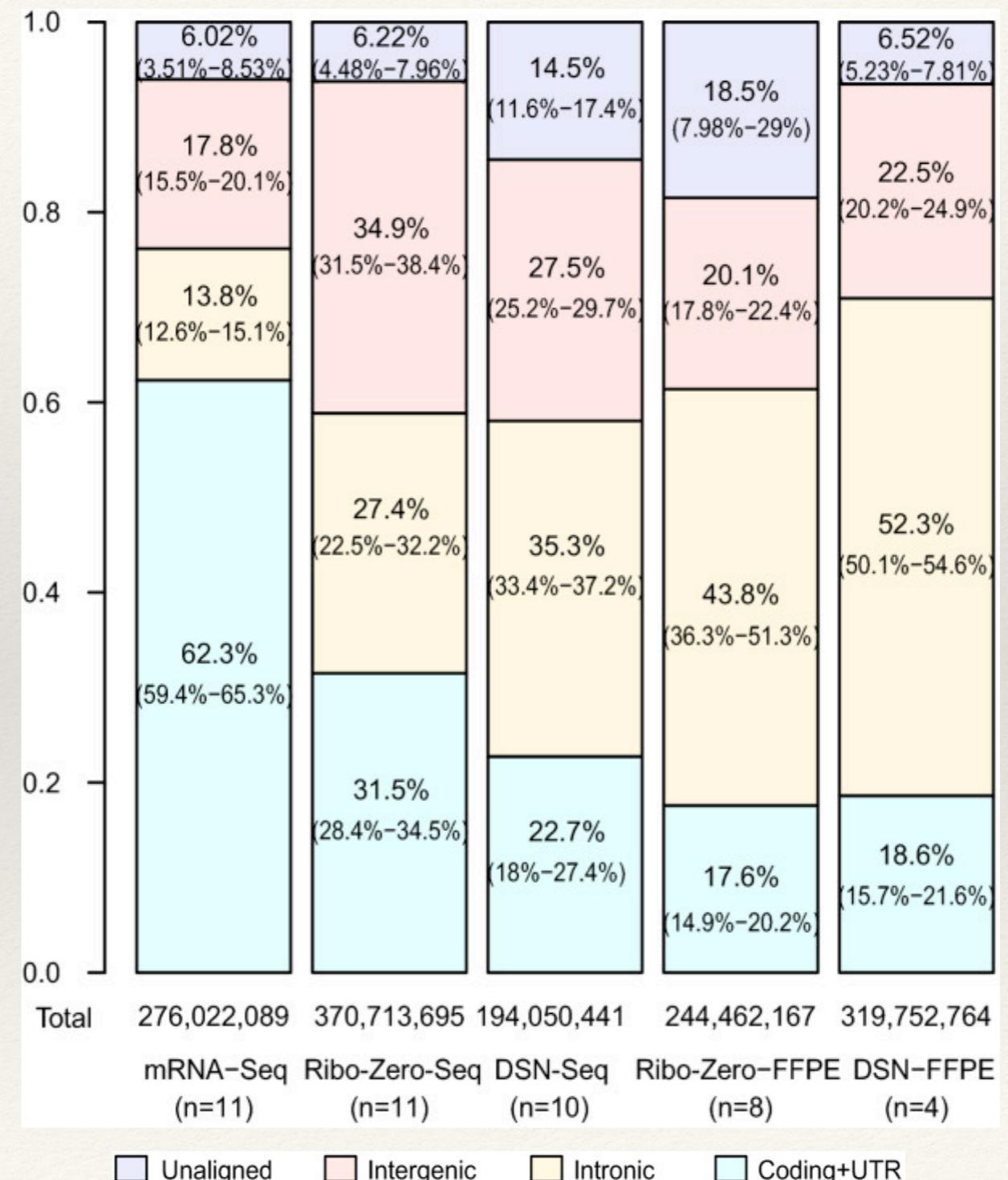
# Biological question

- ❖ Know your targets
  - ❖ Whole genome
  - ❖ Targeted (re)seq
    - ❖ Exome
    - ❖ ChIP-seq
  - ❖ .....



# Biological question

- ❖ Know your targets
  - ❖ Whole genome
  - ❖ Targeted (re)seq
    - ❖ Exome
    - ❖ ChIP-seq
  - ❖ RNA-seq
    - ❖ rRNA depleted?
    - ❖ polyA enriched?
  - ❖ microRNA



# Experimental design

- ❖ Short or long fragments
- ❖ Short or long reads
- ❖ Single or paired end
- ❖ Multiplexing
  - ❖ single or dual index
  - ❖ more barcodes?
- ❖ Library prep method
  - ❖ Depth required
  - ❖ Coverage required
- ❖ Replicates
  - ❖ biological
  - ❖ technical

# Platform choice: Read length



MiniSeq  
MiSeq  
NextSeq  
HiSeq 2500  
HiSeq 3/4000  
HiSeq X  
NovoSeq



Roche 454  
SOLiD  
Ion Torrent



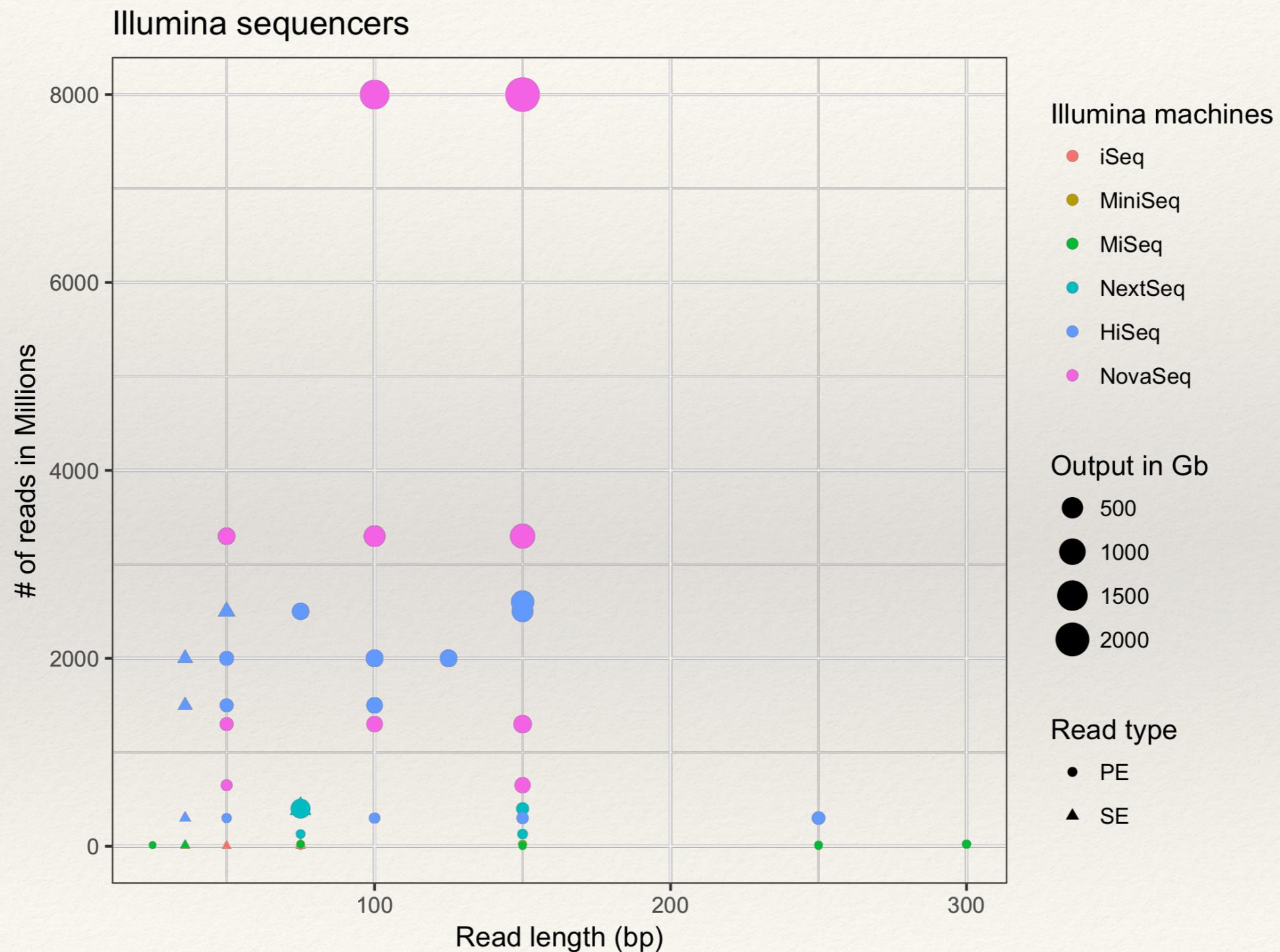
PACBIO®

RS II  
Sequel

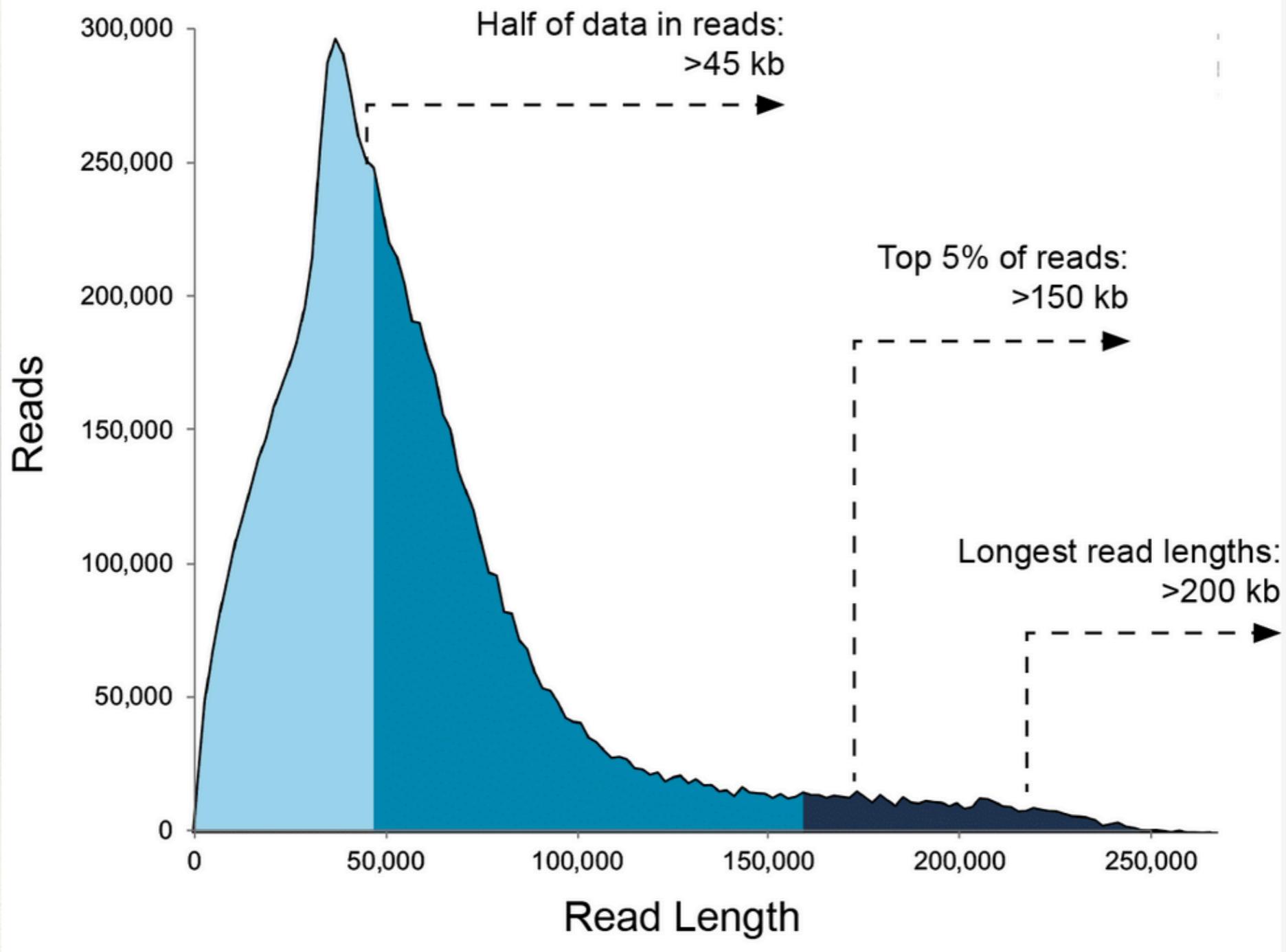


MinION  
Flongle  
GridION  
PromethION

# Illumina data output

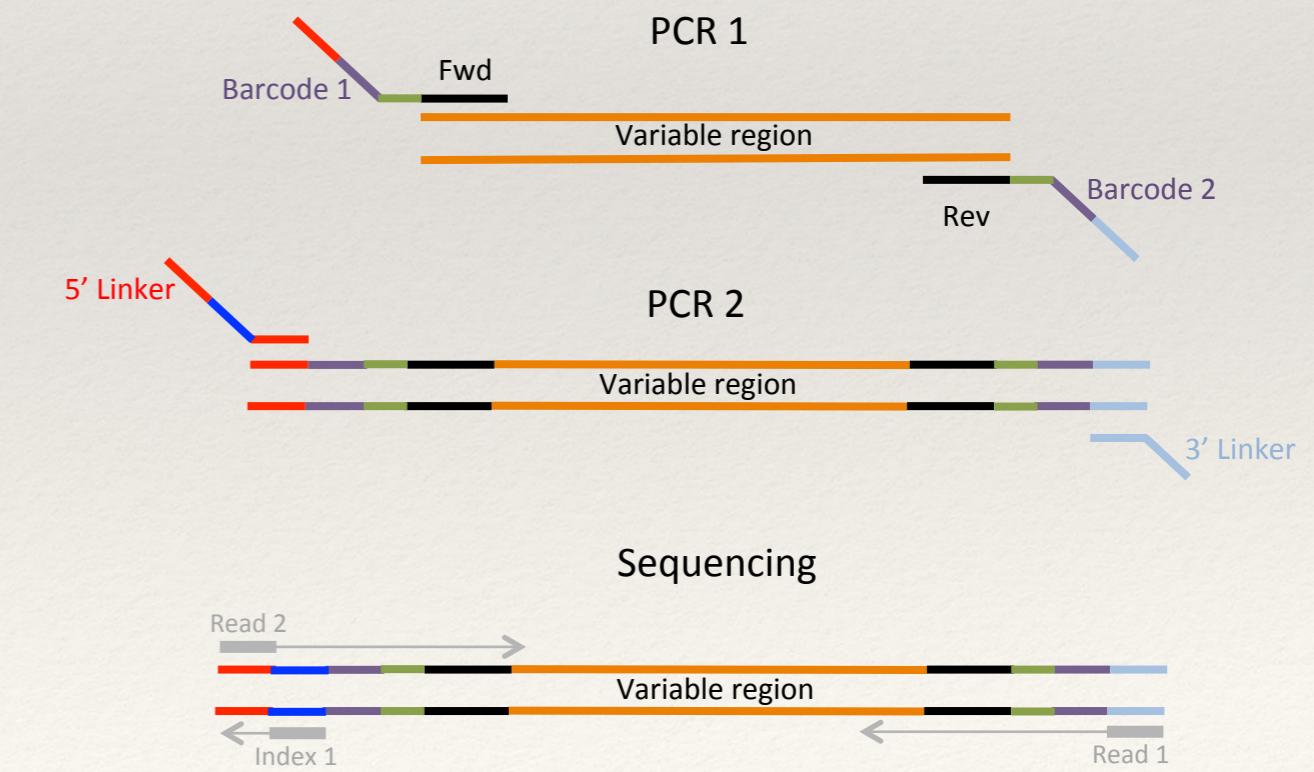
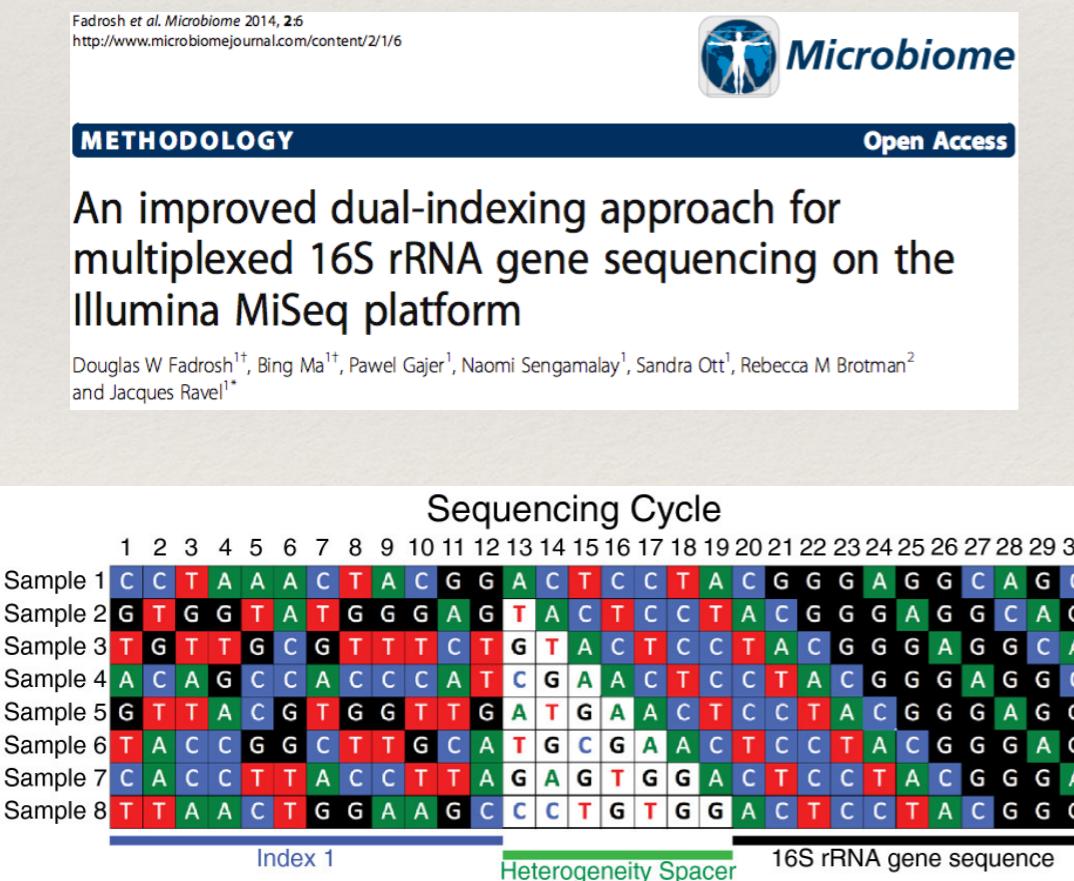


# Pacbio/Nanopore data output



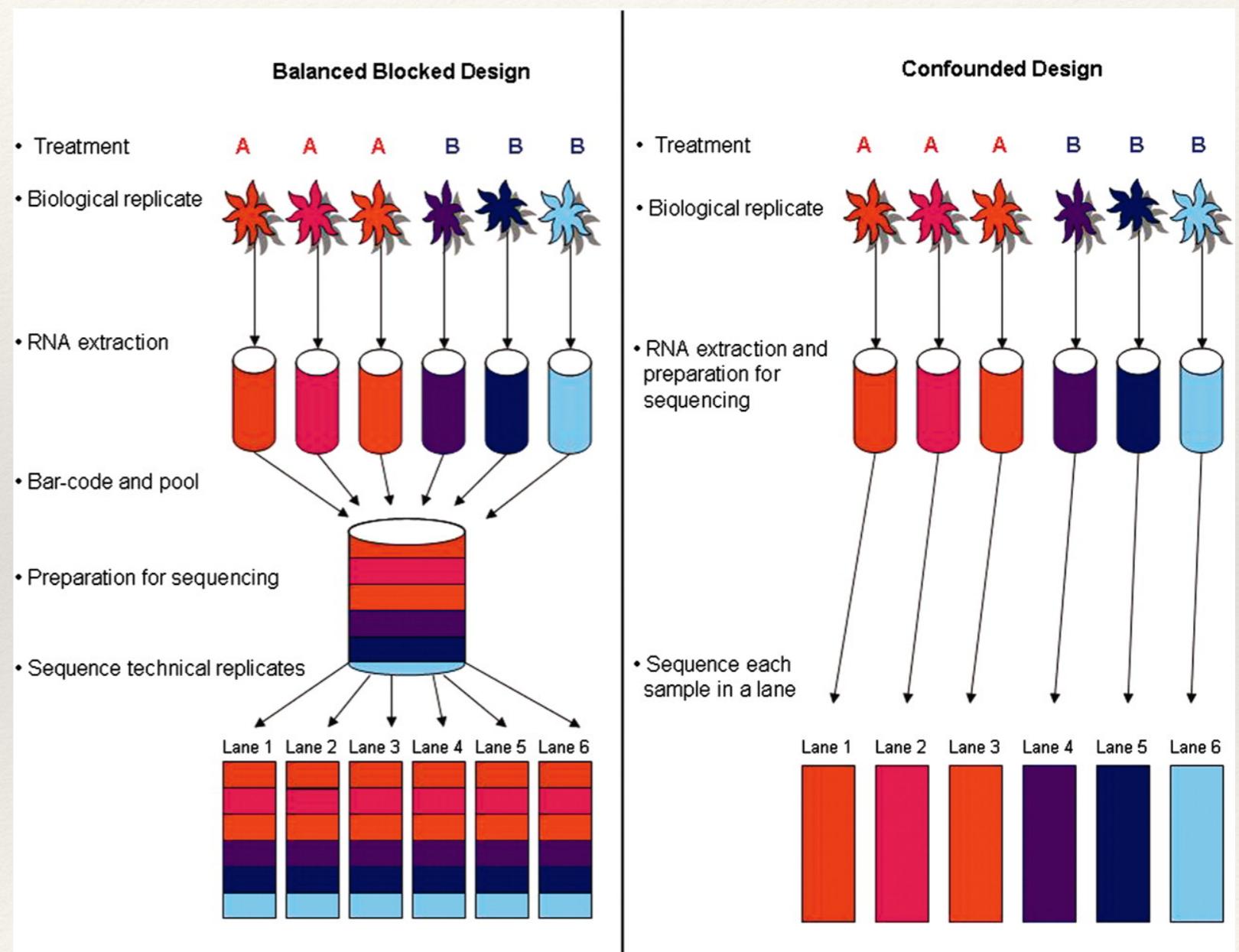
# Indexing

- ❖ Dual index possible
- ❖ Dual internal barcodes possible
  - ❖ multiplex up to 4000 samples.



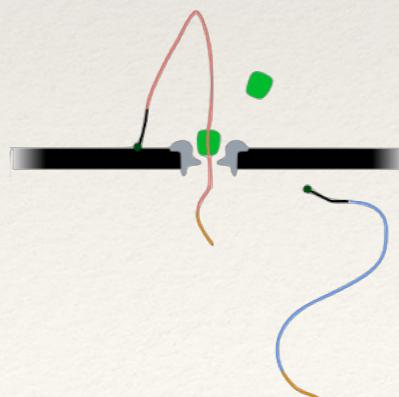
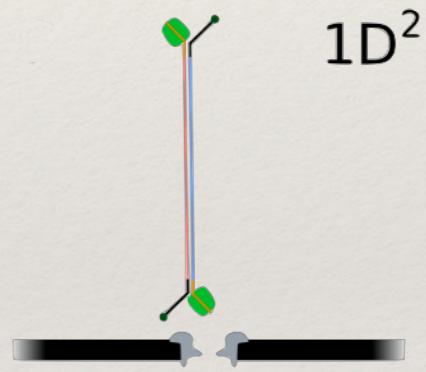
# Technical bias

- ❖ Lane / flowcell bias
- ❖ Index / barcode bias
- ❖ Batch effect
- ❖ Randomisation is key

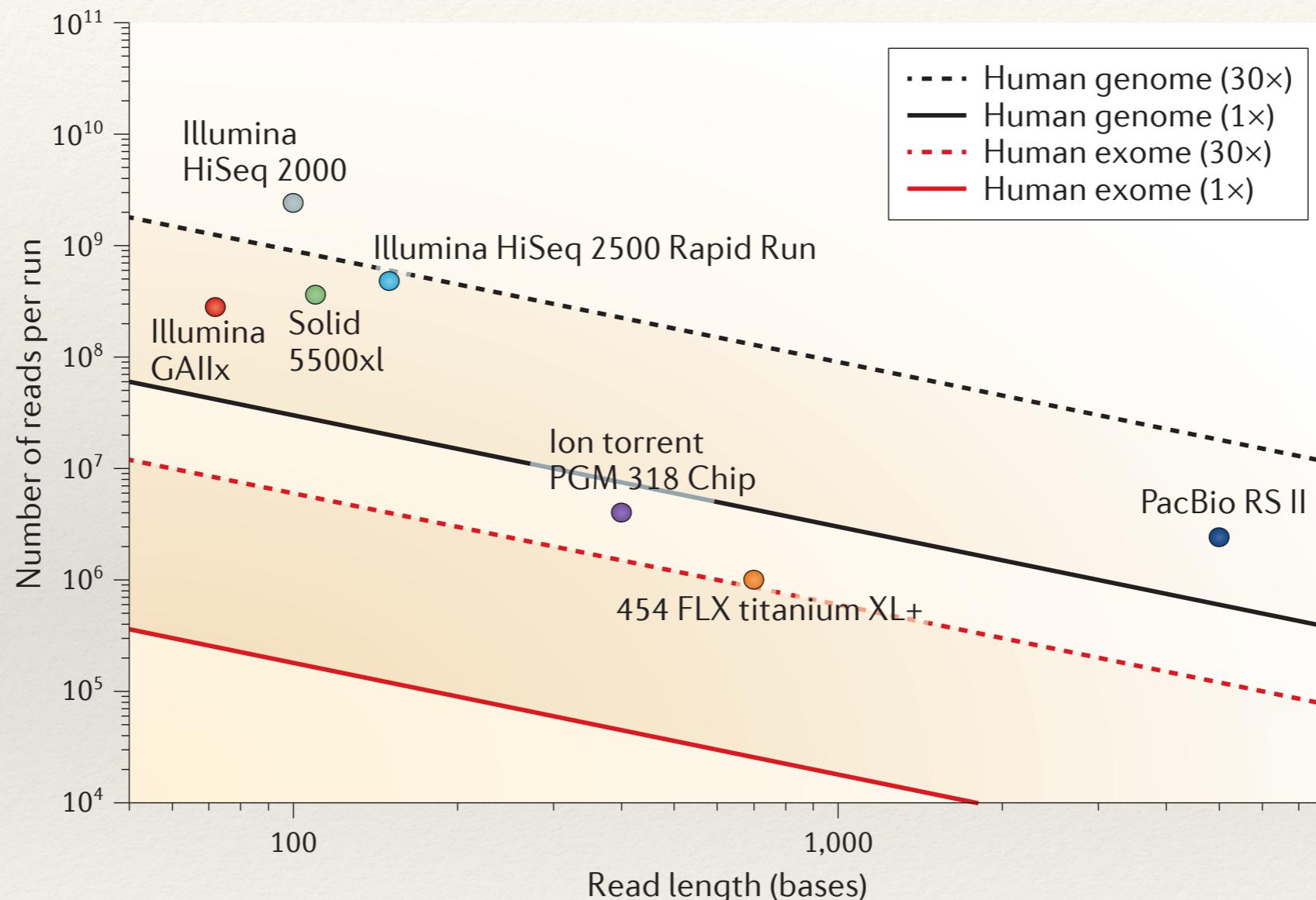


# Error rates

- ❖ Illumina has low error rates
- ❖ Pacbio and Oxford Nanopore have relatively high error rates
  - ❖ Cyclic sequencing can reduce the error rate in Pacbio
  - ❖ 1D<sup>2</sup> sequencing can reduce the error rate in Oxford Nanopore
- ❖ Deep sequencing is used to correct for errors

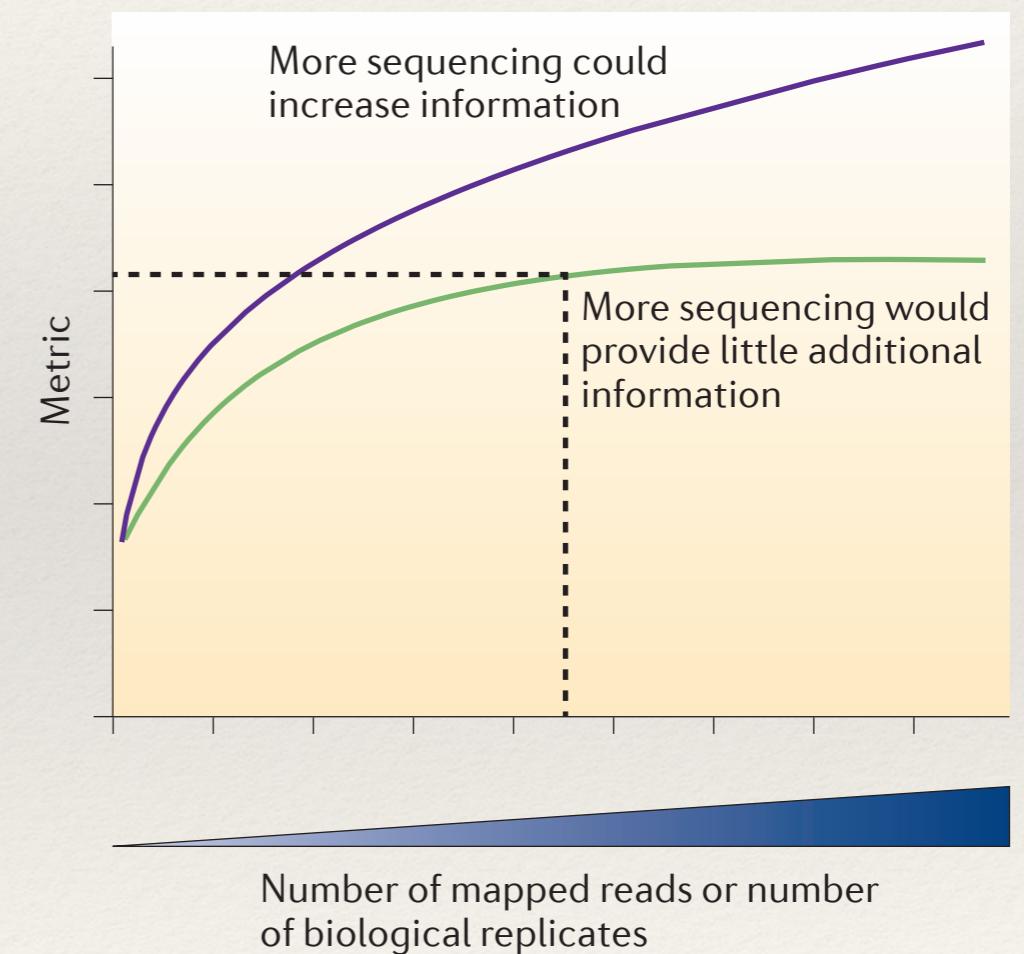
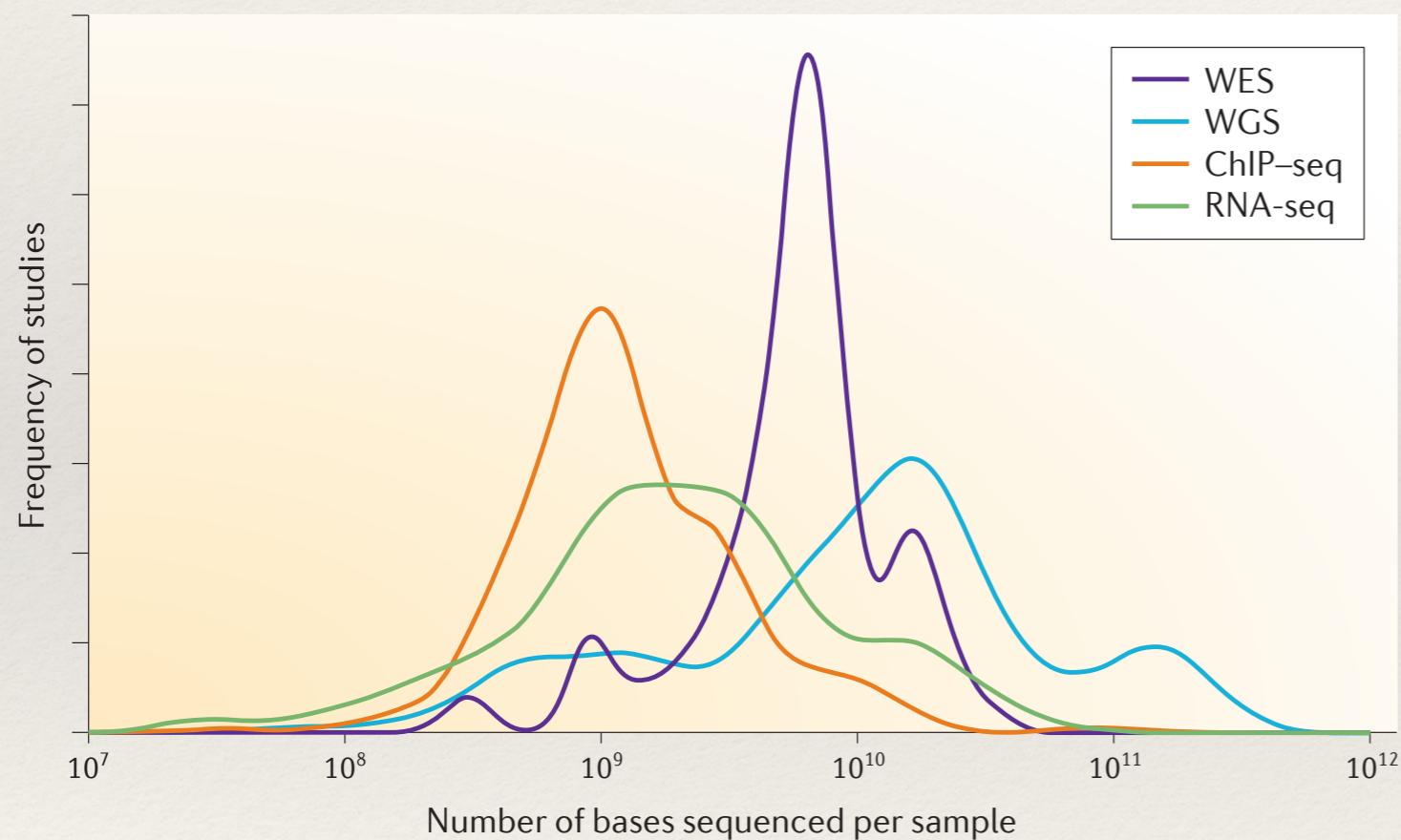


# Sequencing depth and coverage

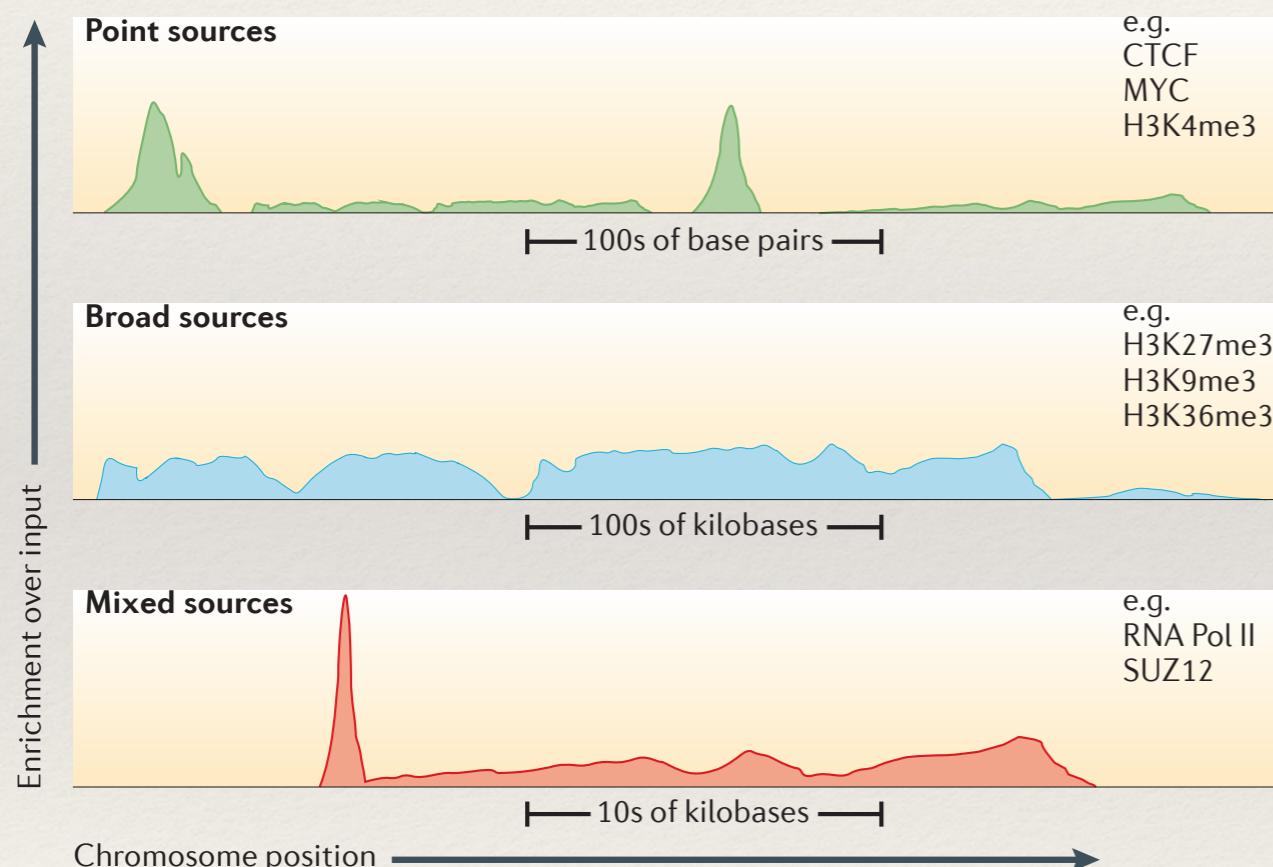


GAIix, Genome Analyzer Iix; PacBio, Pacific Biosciences; PGM, personal genome machine.

# Sequencing depth and coverage



# Sequencing depth and coverage



Techniques	Read counts in representative studies
DNasel-seq and FAIRE-seq	20–50 million
CLIP-seq	7.5 million; 36 million
iCLIP and PAR-CLIP	8 million; 14 million
ChIP and CHART	26 million
4C	1–2 million
ChIA-PET	20 million
5C	25 million
Hi-C	>100 million
MeDIP-seq	60 million
CAP-seq	>20 million
ChIP-seq	>10 million per sample (point source); >20 million per sample (broad source)

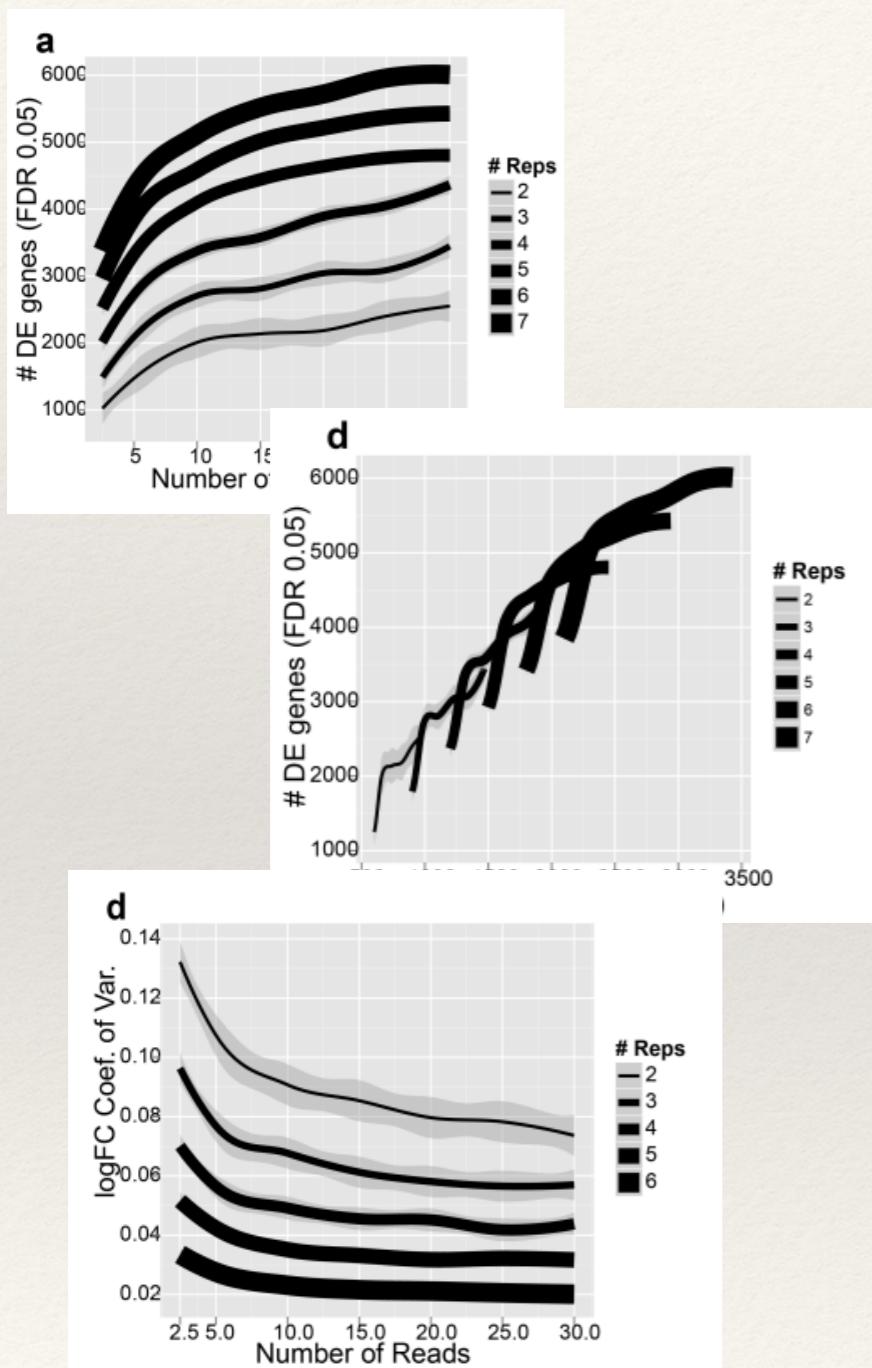
# Replicates and Depth

- ❖ Sound experimental design
- ❖ Number of replicates
  - ❖ Biological variation
  - ❖ Technical replicates - not so important
- ❖ Sequencing depth

**Table 1** Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

Effect size (fold change)	Replicates per group		
	3	5	10
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

# Replicates vs Depth

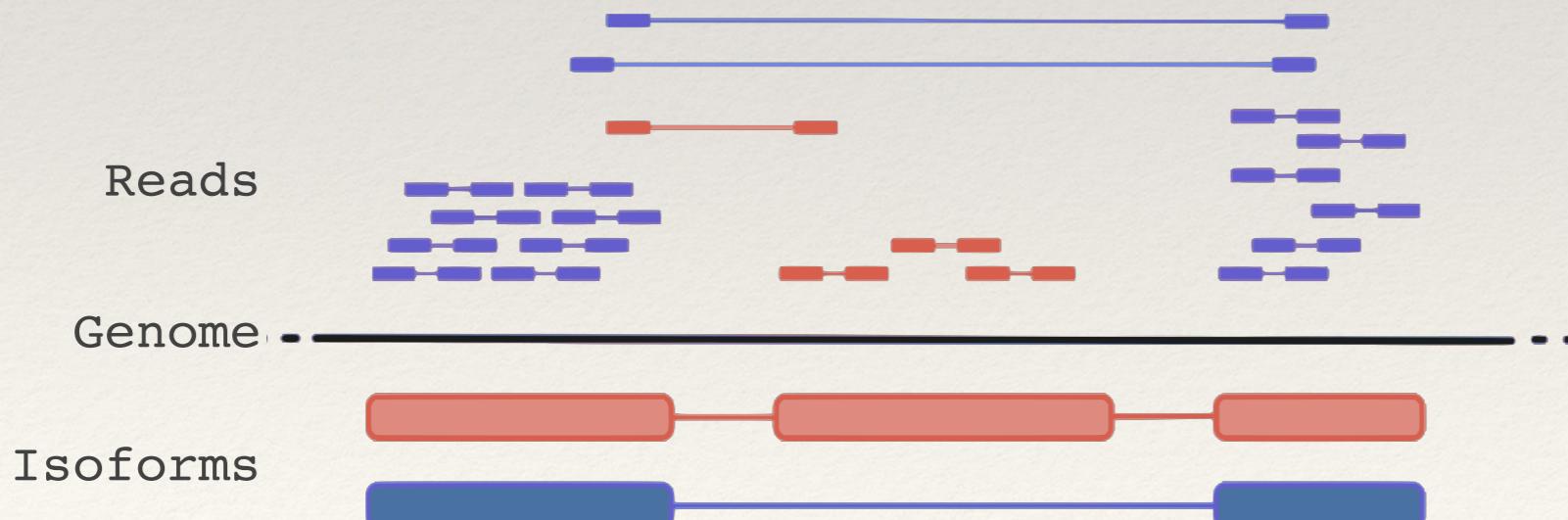


**Table 1.** Cost efficiency for power to detect DE genes (cost per 1% power given each experimental design where the variables are). Assumptions made during calculations are described in Methods. \* indicates lowest cost per 1% power in each replication level. Units are in dollars.

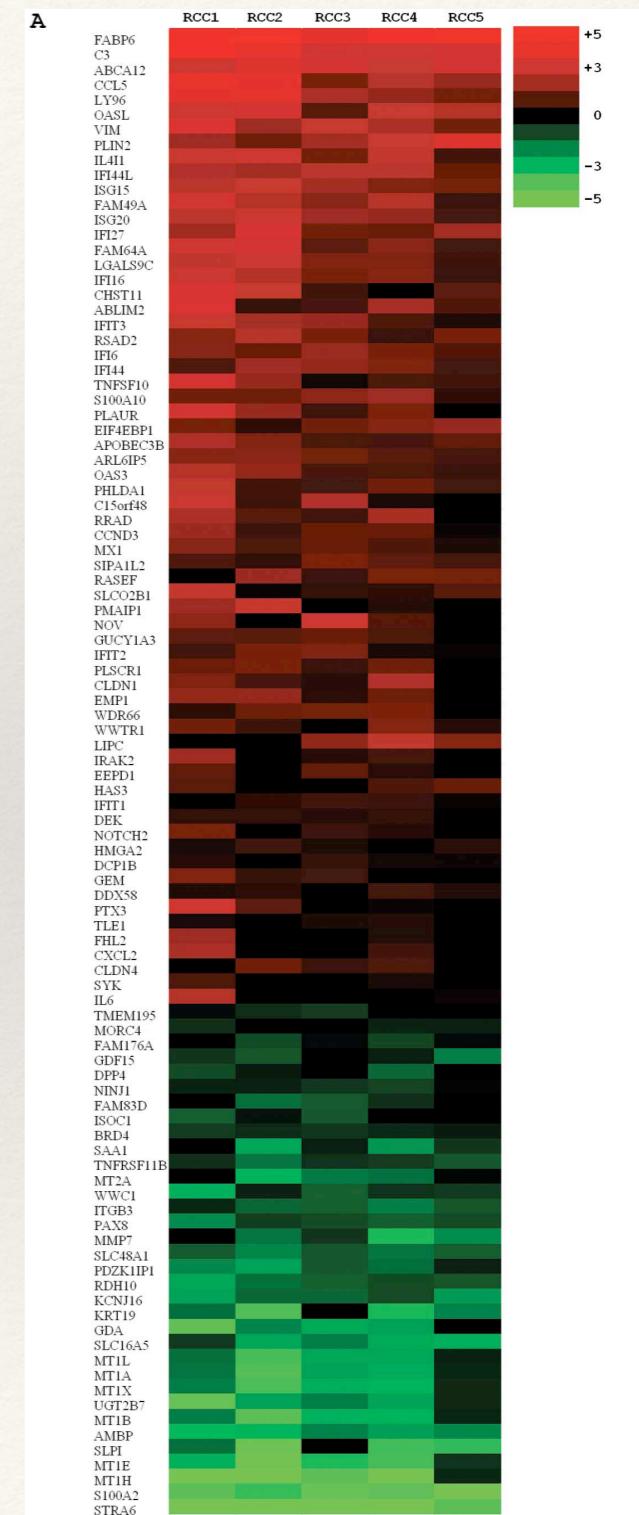
Relative Cost	2.5M	5M	10M	15M	20M	25M	30M
2 replicates	24.2	17.2	14.4*	15.8	16.7	17.0	17.8
3 replicates	23.4	17.2	15.3*	16.3	17.1	18.5	19.4
4 replicates	23.1	17.7	16.5*	17.5	18.6	19.8	21.2
5 replicates	23.8	19.0	18.1*	19.4	21.0	22.8	24.9
6 replicates	25.0	20.7	20.6*	22.4	24.6	27.0	29.4
7 replicates	26.8	23.0*	23.5	26.0	28.7	31.5	34.3

# Depth: example

- ❖ RNA sequencing
  - ❖ Highly expressed known transcripts
  - ❖ Novel isoforms
  - ❖ Low expressed / rare transcripts



More  
depth



---

# Design prior to sequencing

---

- ❖ Sources of variation
  - ❖ Dynamic range - Not all samples get sequenced the same way
  - ❖ Technical variation - biases inherent to the technology
  - ❖ Biological variation
- ❖ Controlling for variation
  - ❖ Randomisation
  - ❖ Blocking: Pool and sequence across several lanes
  - ❖ Replication

---

# Pre-processing

---

- ❖ Remove sequencing adapters
  - ❖ Trim / remove low quality reads
  - ❖ Remove sequencing spike-ins (PhiX for Illumina), if any
- Make sure paired end data is always paired and in correct order!

---

# Simple truth

---

To consult the statistician after an experiment is finished is often merely to ask him (her) to conduct a post mortem examination. He (she) can perhaps say what the experiment died of.

- Ronald Fischer