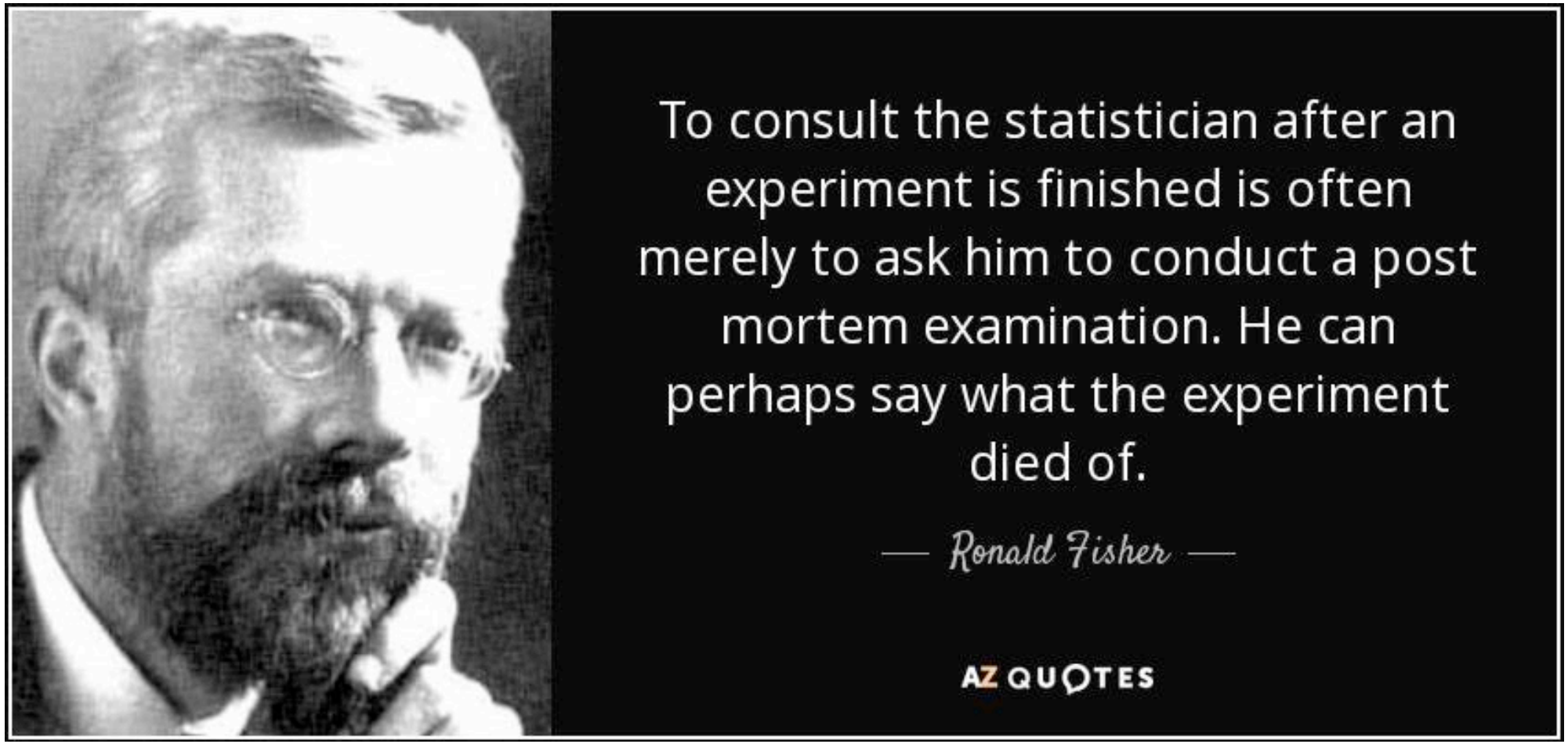
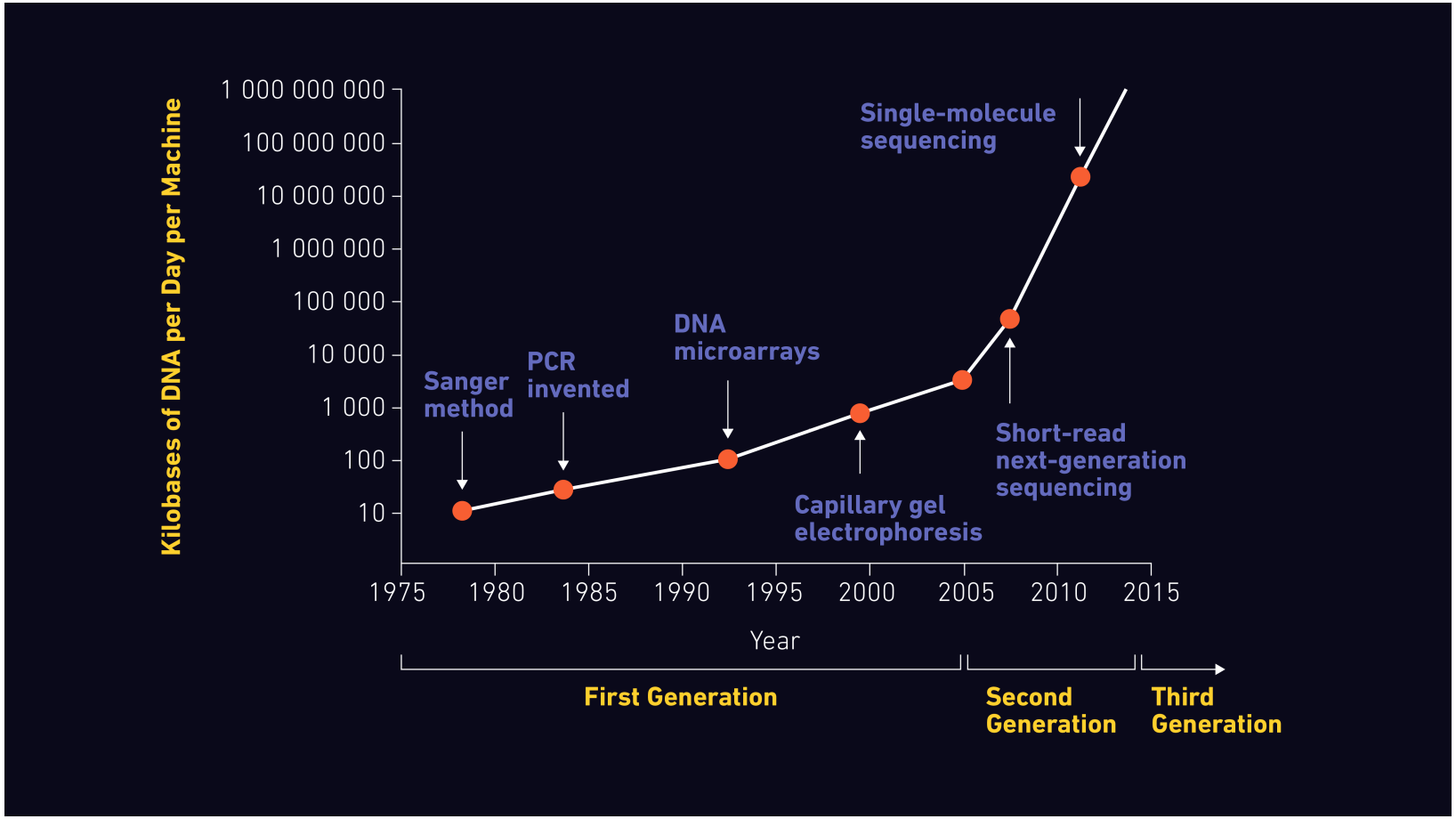
Experimental design and workflows for next generation sequencing

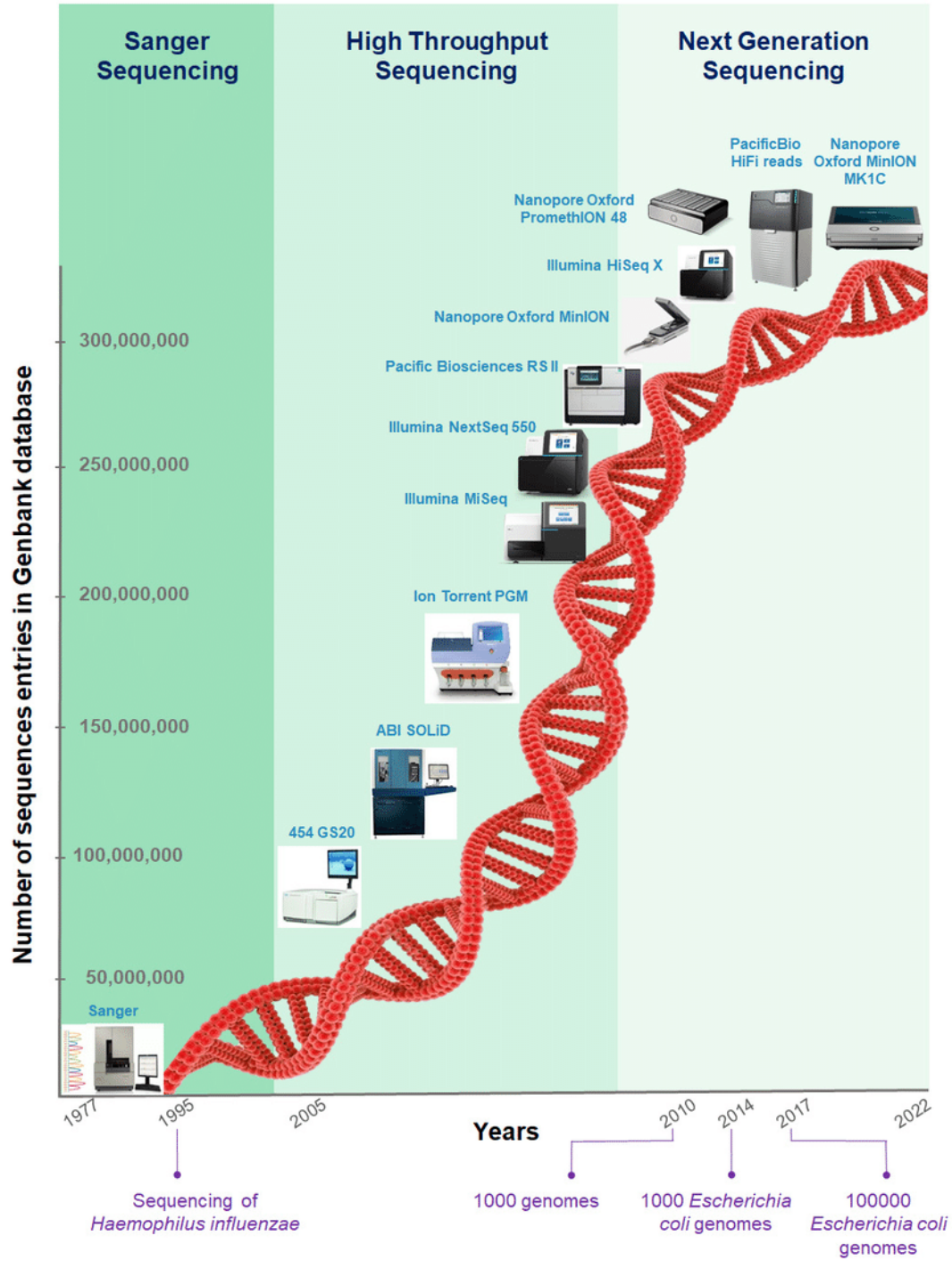
Mohamed Zahir Alimohamed


To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.
— Ronald Fisher —
AZ QUOTES

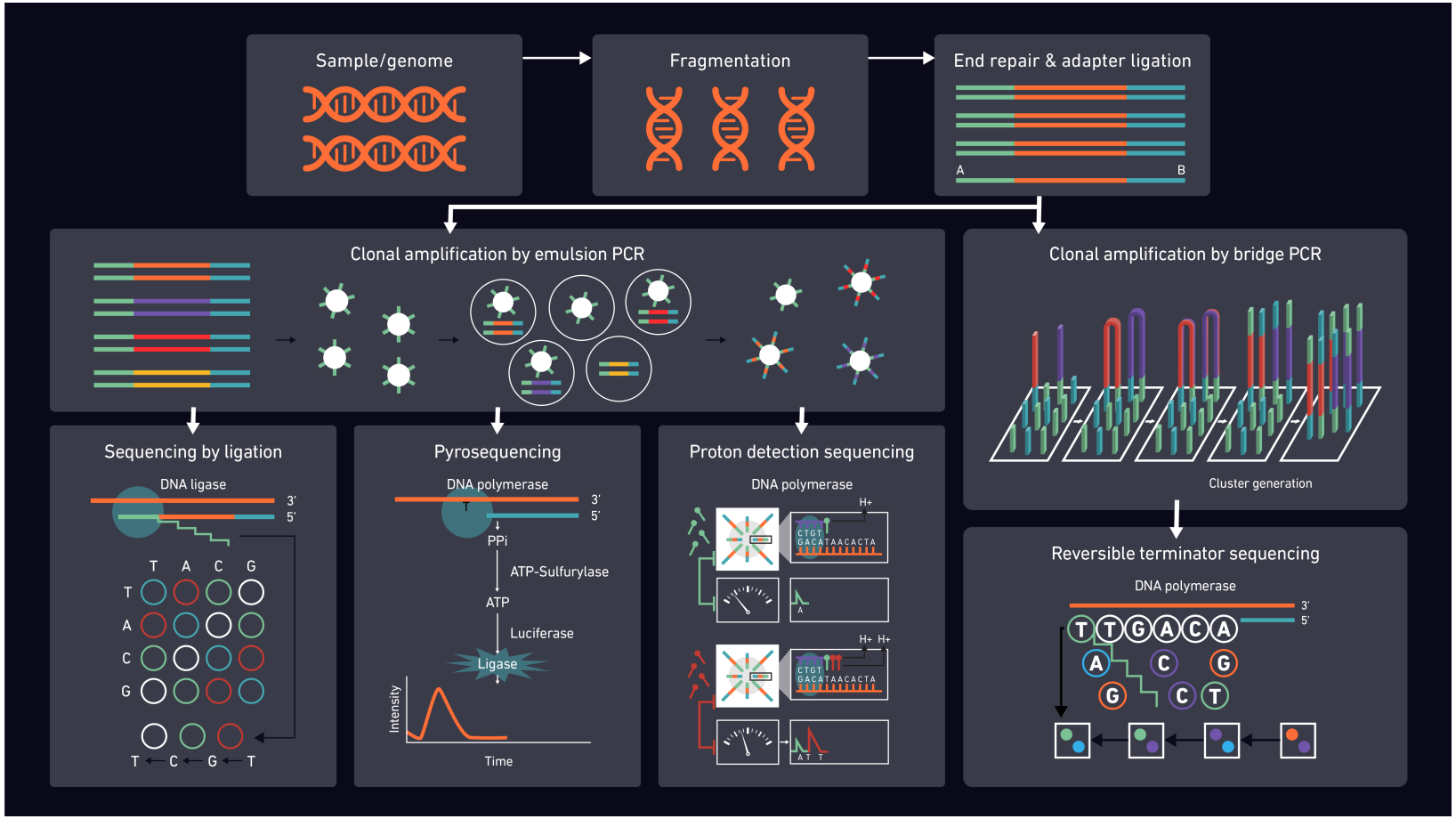
To carry out NGS without a sound design?

The evolution of sequencing methodologies




Sanger Sequencing
High Throughput Sequencing
Next Generation Sequencing
PacificBio HiFi reads
Nanopore Oxford MinION MK1C
Nanopore Oxford PromethION 48
Illumina HiSeq X
Nanopore Oxford MinION
Pacific Biosciences RS II
Illumina NextSeq 550
Illumina MiSeq
Ion Torrent PGM
ABI SOLiD
454 GS20
Sanger
Number of sequences entries in Genbank database
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
1977
1995
2005
2010
2014
2017
2022
Years
Sequencing of Haemophilus influenzae
1000 genomes
1000 Escherichia coli genomes
100000 Escherichia coli genomes

Sequencing platforms principles and chemistries



Selecting a sequencing strategy

- The following are some of the key considerations when deciding on the appropriate library preparation and sequencing platform:

(a) Research question being asked

(b) Sample type

(c) Short-read or long-read sequencing

(d) DNA or RNA sequencing – do you need to look at the genome or transcriptome?

(e) Is the whole genome required or only specific regions?

(f) Read depth (coverage) needed – experiment-specific

(g) Extraction method

(h) Sample concentration

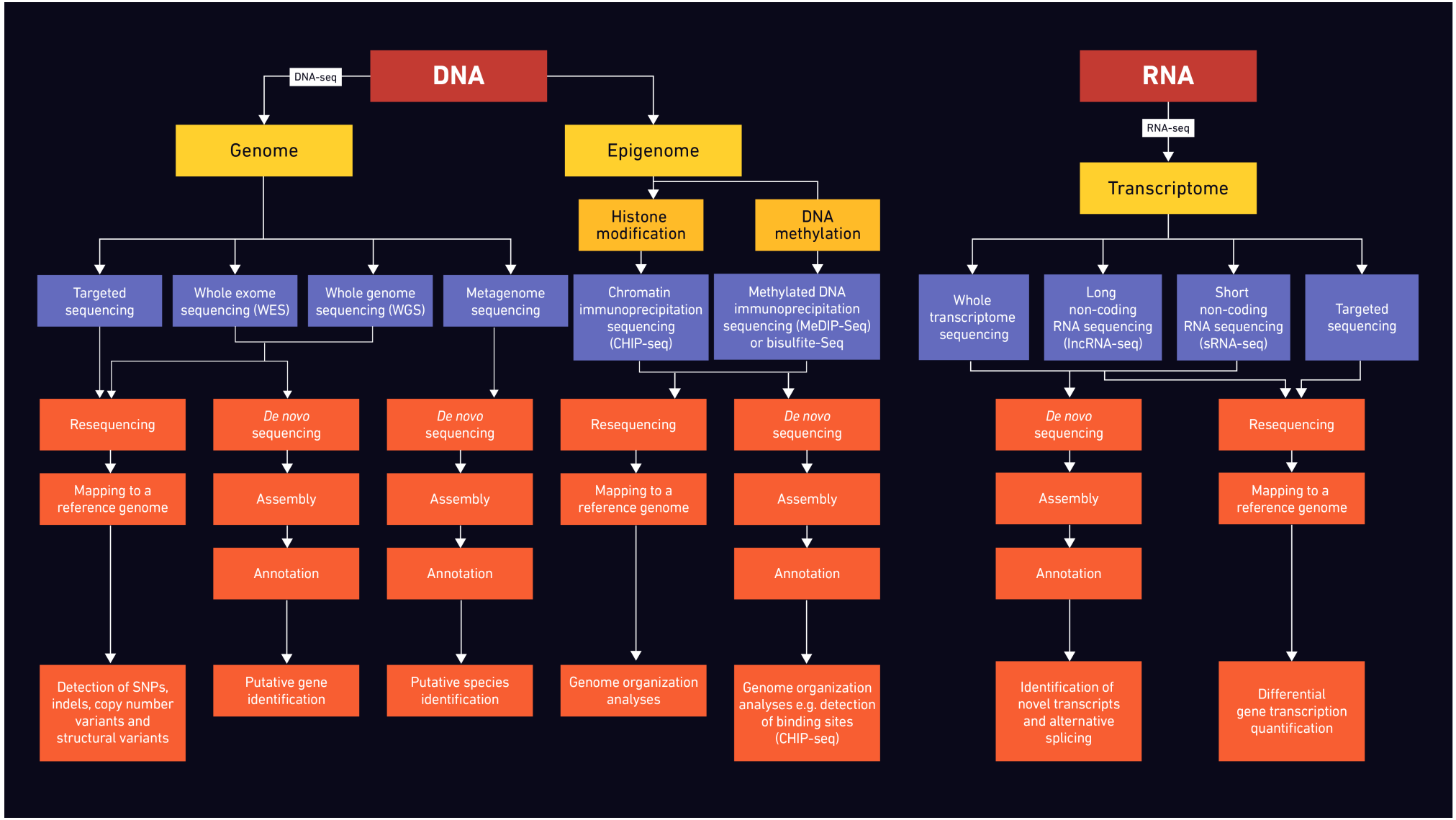
(i) Single end, paired end or mate pair reads

(j) Specific read length required

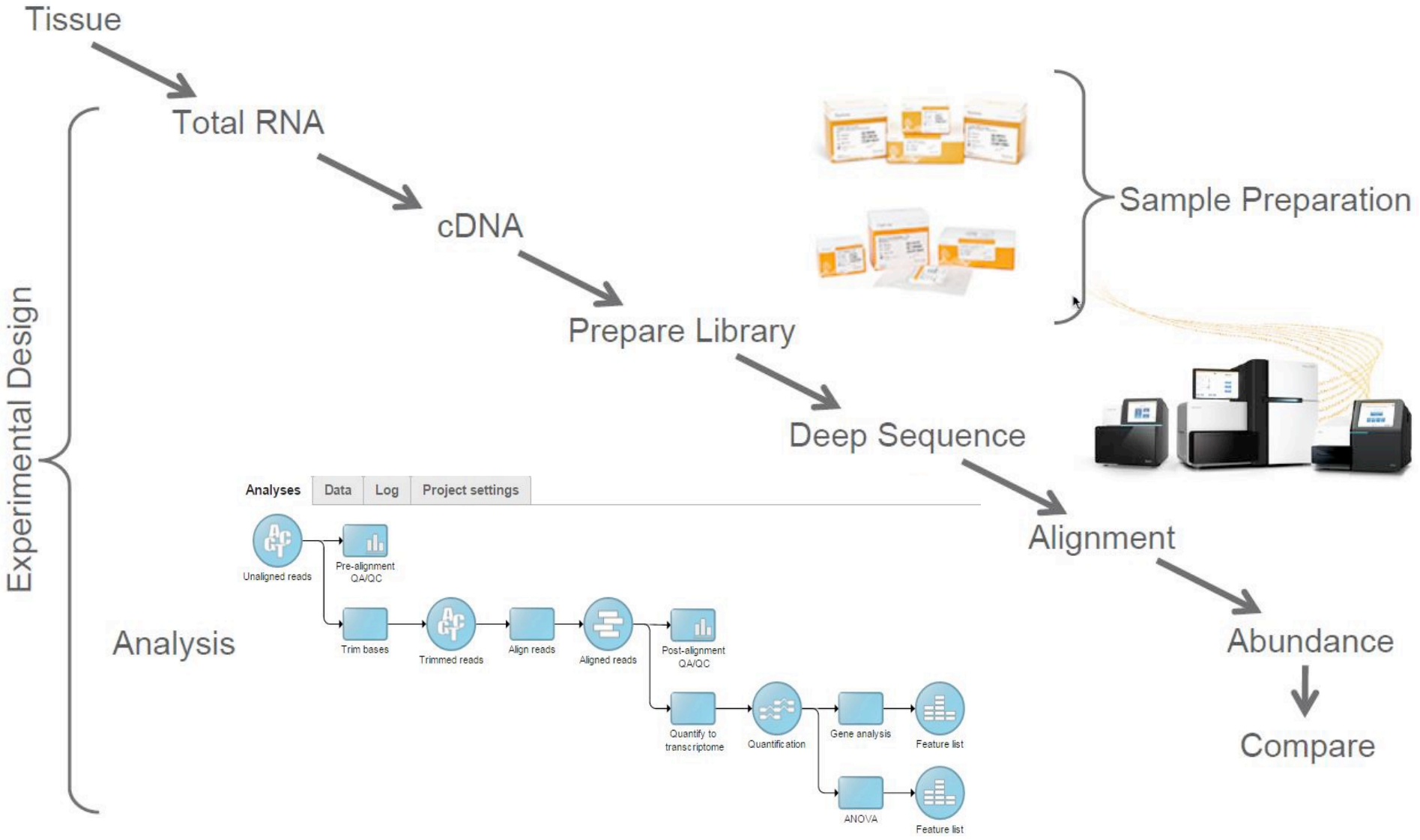
(K) Could samples be multiplexed?

(l) Bioinformatic tools – experiment dependent. Depending on the sample and the biological question, the entire process of sequence analysis can be adapted.

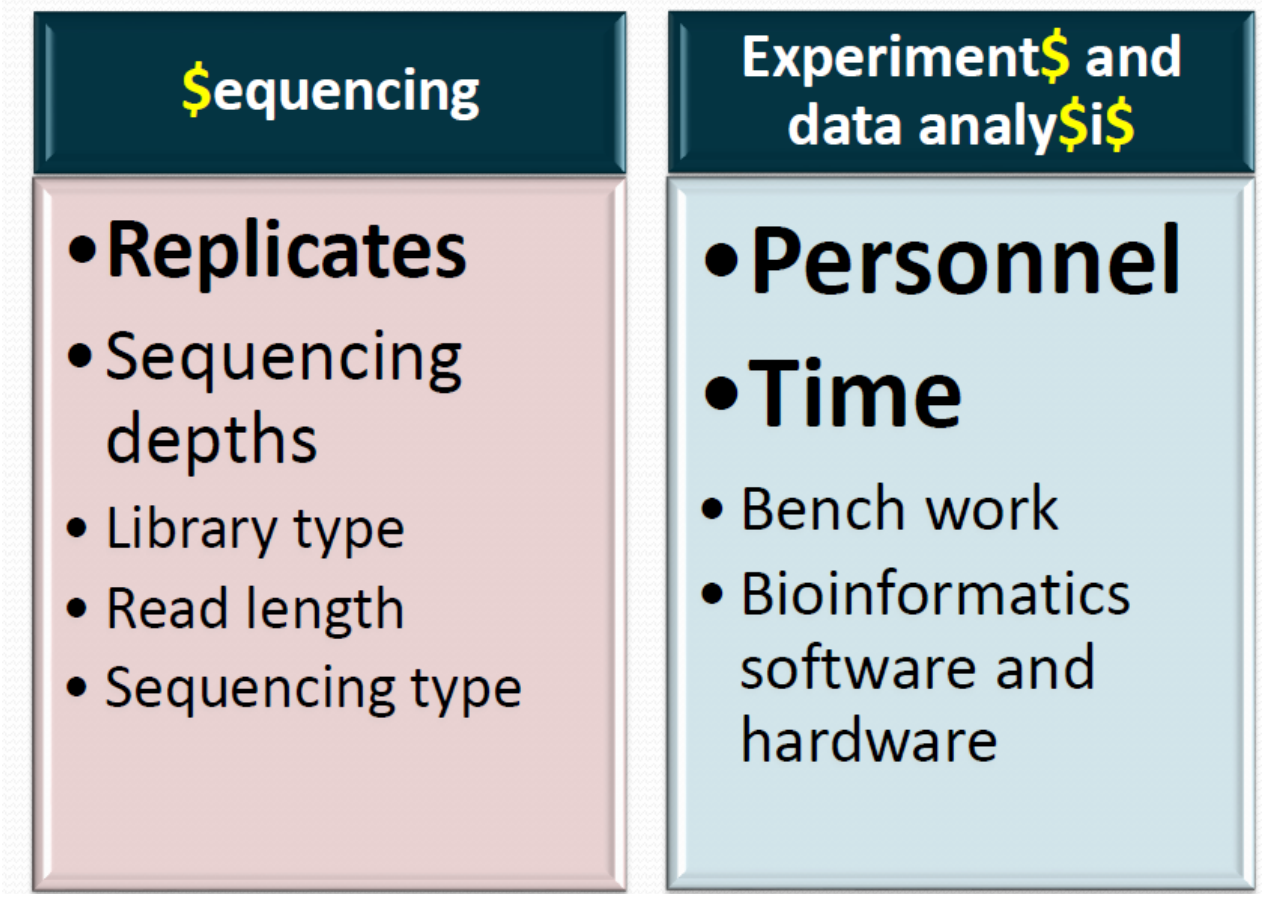
Flow diagram indicating possible sequencing strategies for different sample types.



RNA-seq experiments at a glance



NGS research cost – don’t pinch pennies



Why do you need replicates?

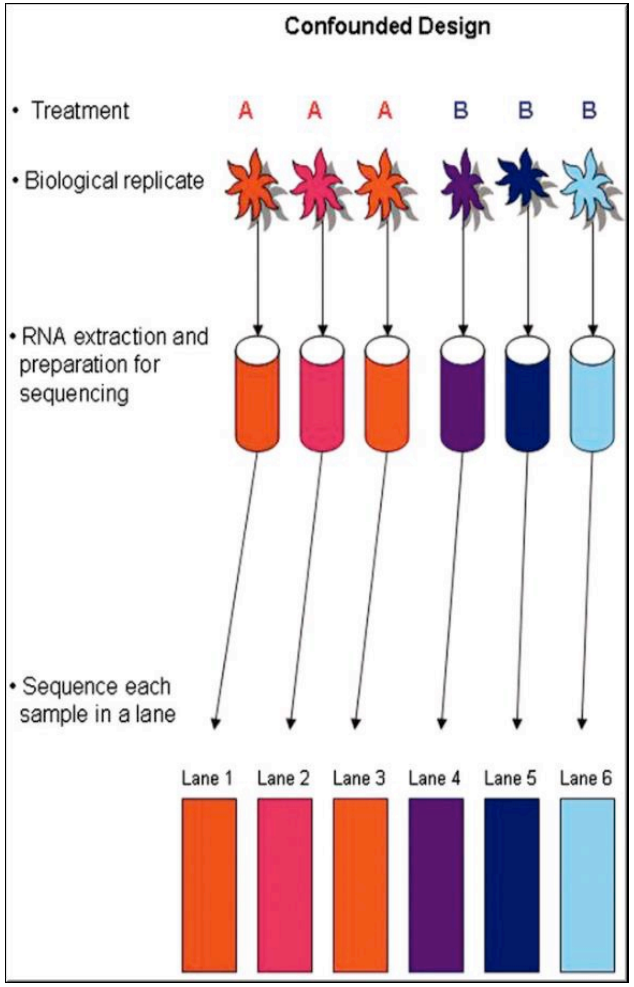
- »Access the variances of your measurements
- »Technical variances
- »Biological variances
- »Ensure the validity of your findings and allow generalization of your conclusions
- »Achieve the power of your studies

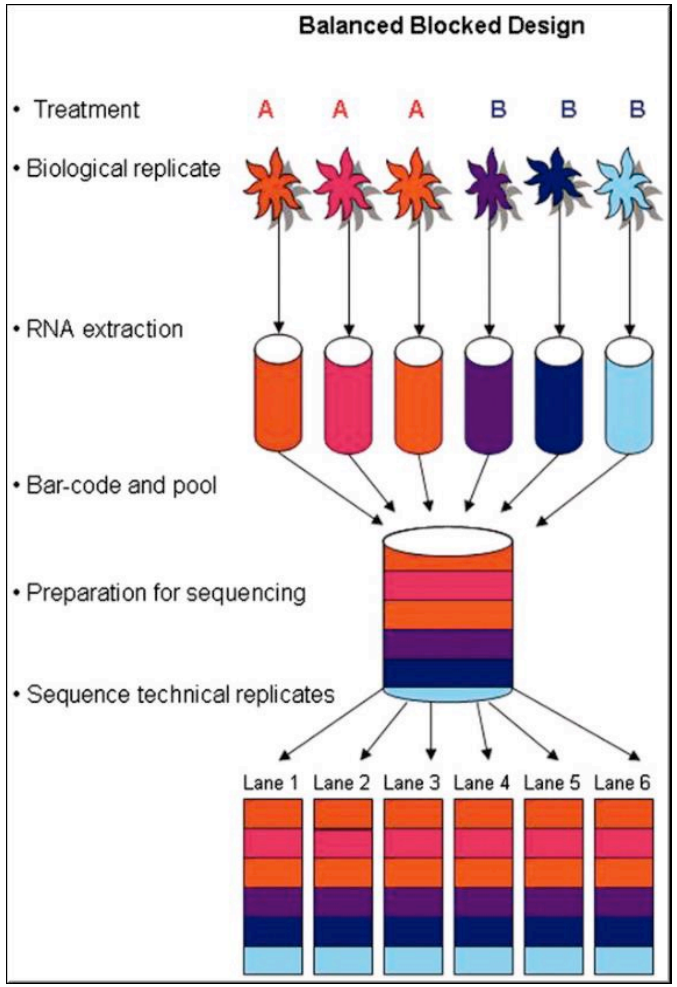
The variances in NGS studies

- »Technical variances
- »Arose from sample preparations, library construction and sequencing
- »Technical variances from sequencing are relatively small
- »More pronounced in low-expression genes
- »Biological variances
- »Often much larger than technical variances
- »Magnitude is depended on the studying system:
- »Individuals (such as human samples)>>inbred organisms>cell lines
- »Also affected by the nature of treatments

Reduce the technical variances and avoid confounding it with biological variances

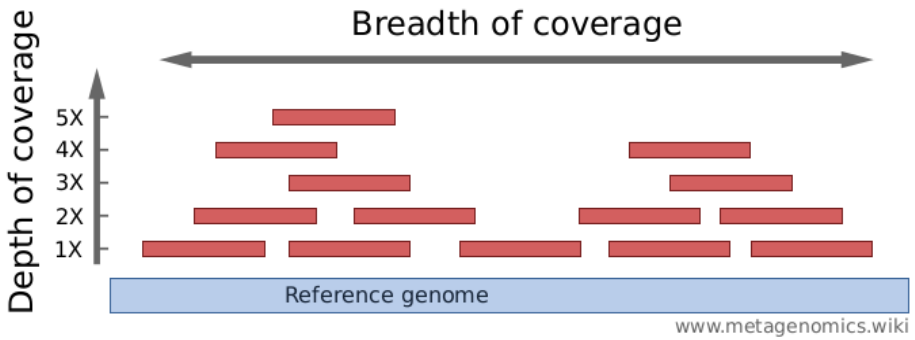
»Systematically reduce technical variances throughout your study
»Thoughtful design
»Block what you can and randomize what you cannot
»Avoid sequencing in batches if possible
»Increase sequencing depth





Sequencing depth vs read per sample

$$\text{Sequencing depth (average coverage)} = \frac{\text{Number of reads x read length}}{\text{Total length of the targeted sequence}}$$



Species	Number of genes	Transcriptome size (Mbp)
Homo sapiens	29230	70.1
Mus musculus	24080	61.4
*Gallus gallus***	4906	11.1
Drosophila melanogaster	18436	30.1
Caenorhabditis elegans	23933	28.0
Arabidopsis thaliana	27278	51.1
Saccharomyces cerevisiae	6692	8.9
*Escherichia coli****	4290	0.6

More replicates or deeper sequencing

- »More replicates:
- »Higher power of detection (more DEGs)
- »Better accuracy of differential expression results
- »More reads per sample:
- »Better detection of low-expressed genes (more DEGs)
- »Better accuracy of quantification (read counts)
- »Replicates over sequencing depth!
- »Validity rules!

Quick hits on sequencing depths for DNA-seq variants detection

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
Whole genome sequencing	Homozygous SNVs	15x	Bentley et al., 2008
	Heterozygous SNVs	33x	Bentley et al., 2008
	INDELs	60x	Feng et al., 2014
	Genotype calls	35x	Ajay et al., 2011
	CNV	1-8x	Xie et al., 2009; Medvedev at al., 2010
Whole exome sequencing	Homozygous SNVs	100x (3x local depth)	Clark et al., 2011; Meynert et al., 2013
	Heterozygous SNVs	100x (13x local depth)	Clark et al., 2011; Meynert et al., 2013
	INDELs	not recommended	Feng et al., 2014

*

Somatic mutations should be sequenced much deeper than germline mutations!

Other DNA based sequencing experiments

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
DNA Target-Based Sequencing	ChIP-Seq	10-14M (sharp peaks); 20-40M (broad marks)	Rozowsky et al., 2009; ENCODE 2011 Genome; Landt et al., 2012
	Hi-C	100M	Belton, J.M et al., 2012
	4C (Circularized Chromosome Confirmation Capture)	1-5M	van de Weken, H.J.G. et al., 2012
	5C (Chromosome Carbon Capture Carbon Copy)	15-25M	Sanyal A. et al., 2012
	ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing)	15-20M	Zhang, J. et al., 2012
	FAIRE-Seq	25-55M	ENCODE 2011 Genome; Landt et al., 2012
	DNAse 1-Seq	25-55M	Landt et al., 2012
DNA Methylation Sequencing	CAP-Seq	>20M	Long, H.K. et al., 2013
	MeDIP-Seq	60M	Taiwo, O. et al., 2012
	RRBS (Reduced Representation Bisulfite Sequencing)	10X	ENCODE 2011 Genome

How long should your reads be?

- »Shorter reads (50-75 bp)
- »Typical differential expression analysis
- »Longer reads (>=100 bp)
- »Required for *de novo* sequencing or gene model refining
- »Improves mapping specificity and resolving isoforms

How many reads do you need??


F1000Research F1000Research 2017, 6:997 Last updated: 16 NOV 2017

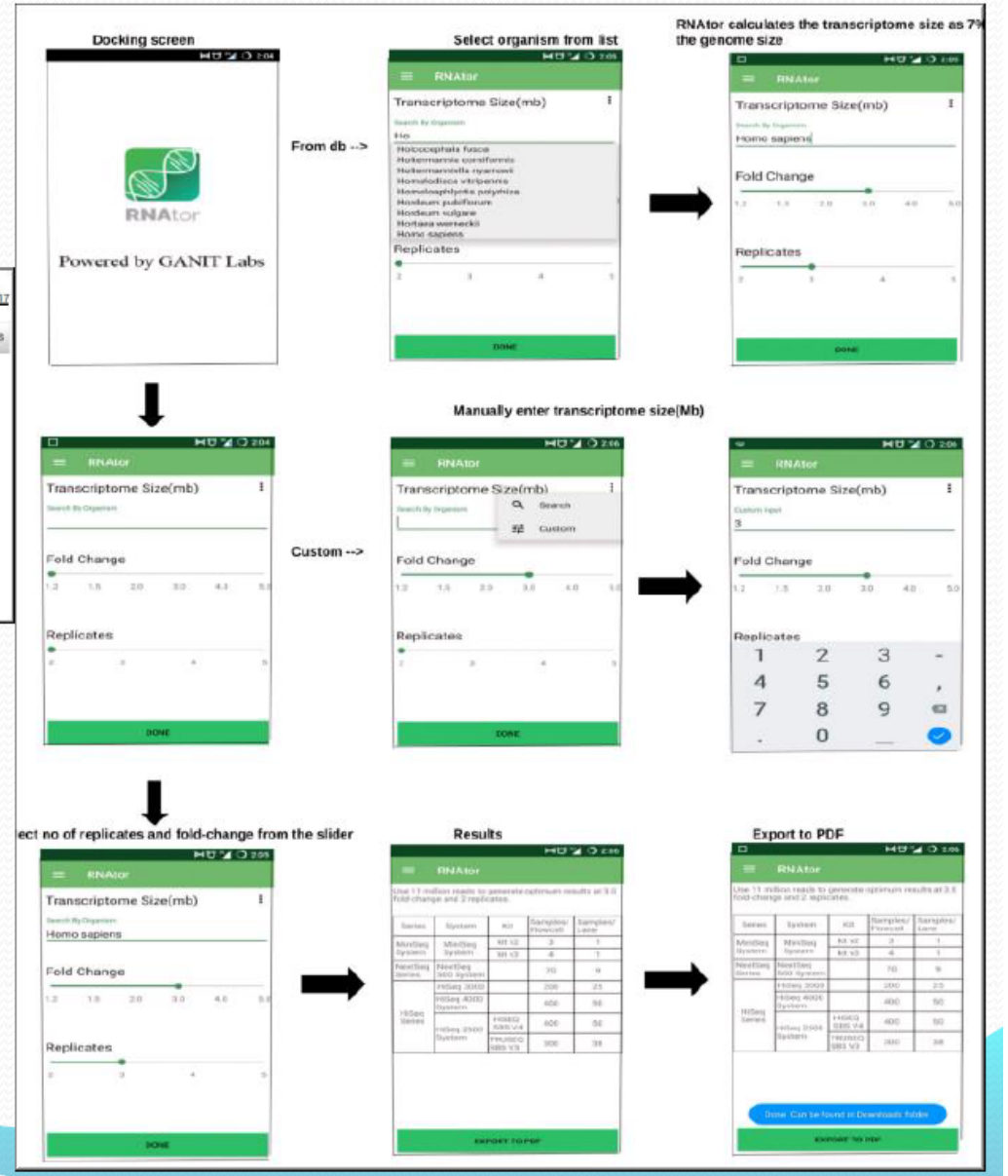
Check for updates

SOFTWARE TOOL ARTICLE

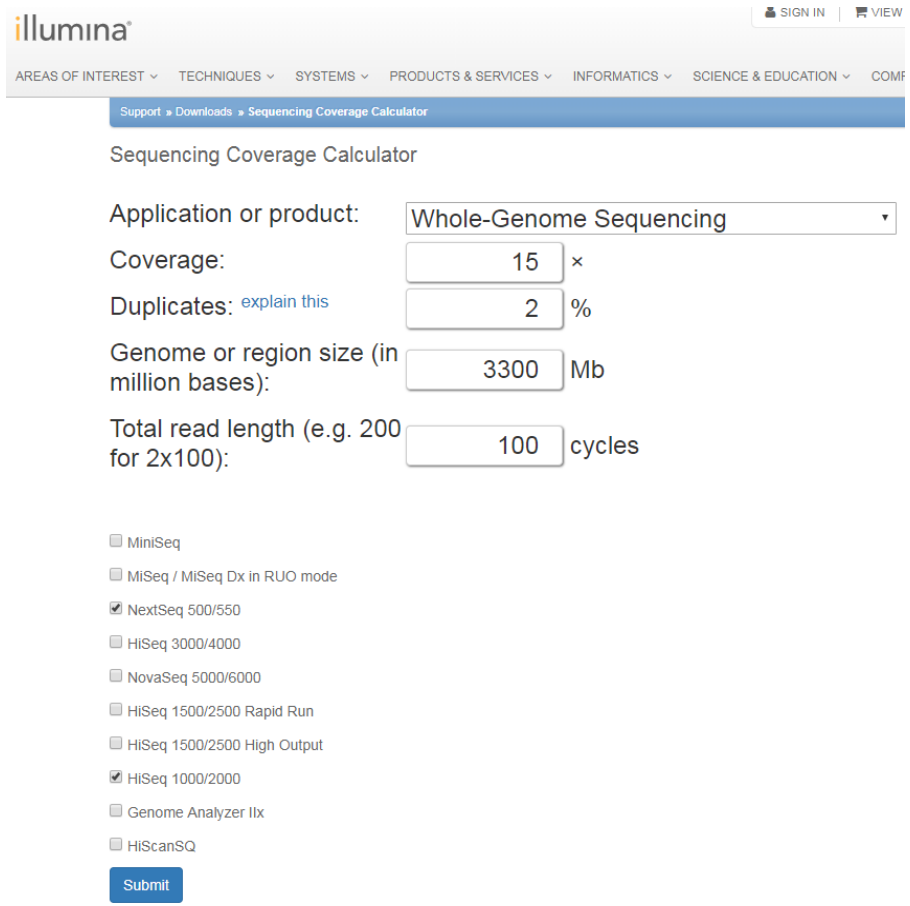
REVISED ***RNAtor*: an Android-based application for biologists to plan RNA sequencing experiments [version 2; referees: 1 approved with reservations, 1 not approved]**

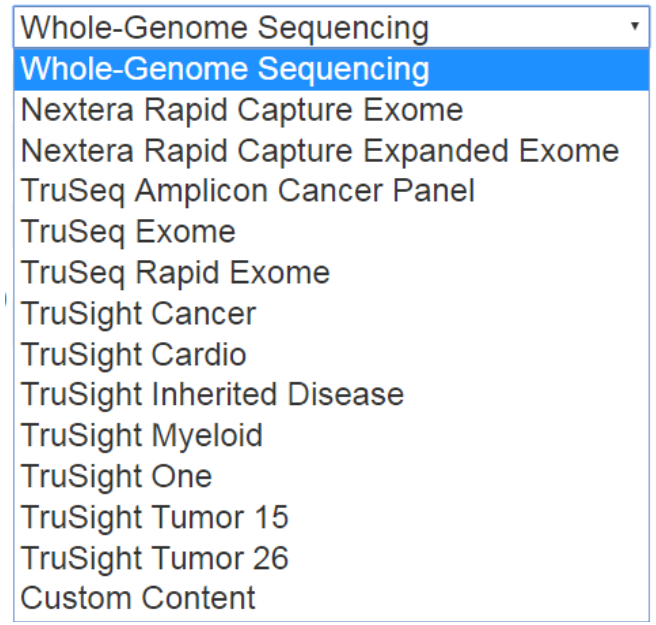
Shruti Kane[1], Himanshu Garg[1], Neeraja M. Krishnan[1], Aditya Singh[1], Binay Panda[1,2]

[1]Ganit Labs, Bio-IT Centre, Institute of Bioinformatics and Applied Biotechnology, Bangalore, India
[2]Strand Life Sciences, Bangalore, India



https://support.illumina.com/downloads/sequencing_coverage_calculator.html





DNA sequencing coverage

- How to estimate and achieve the desired NGS Coverage for DNA sequencing will depends on the application used and best practice as recommended by the scientific community.
- Whole genome recommendation is 10X to 30X, while CHIP-Seq is 100X.The Lander/Waterman equation is a method for computing coverage: C = LN / G. Thus, the total number of reads needed N =CxG/L, where C is the coverage, G size of haploid genome and L is the read length (e.g.100 base-long reads).

RNA sequencing

- RNA-Seq experiments should be performed with at least three or more biological replicates.

- The first step in any successful sequencing experiment is the preparation of the RNA to be sequenced. The number of RNA samples that can be analyzed on the core's sequencers will depend on RNA quality, depth of sequencing needed (Goal) and Output of the sequencing kit.

- NextSeq2000 P3 generates 1.2 to 1.4 Billion reads per run, while the P2 generates from 400-500 million reads.

- Mi-Seq V2 and V3 sequencing kits can generate respectively 17 and 25 million reads per run. We also have MiSeq QC runs affording 1 M reads (nano kit) and 4M reads (micro kit) per run.

Goal #1 : I want to focus on the coding transcriptome and I want to quantify gene expression at the gene level, with one abundance value generated per gene.

Method: Gene expression Profiling – mRNA-seq
≥ 25 Million reads per sample, 1 x 50 bp
Library prep: mRNA stranded
Next Seq P2: 16 sample pool
Next Seq P3: 40 sample pool

Goal #2: I want to focus on the abundance values of both coding and multiple forms of noncoding RNA and identify novel transcript isoforms, SNVs, gene fusions, and/or identify allele-specific expression.

Method: Total RNA Sequencing – rRNA depleted
≥ 50 Million reads for QC samples, 2 x 50 bp
≥ 100 Million for degraded samples, 2 x 100 bp
Library prep: stranded total RNA ribo-depletion
Next Seq P2: 8 sample pool
Next Seq P3: 20 sample pool