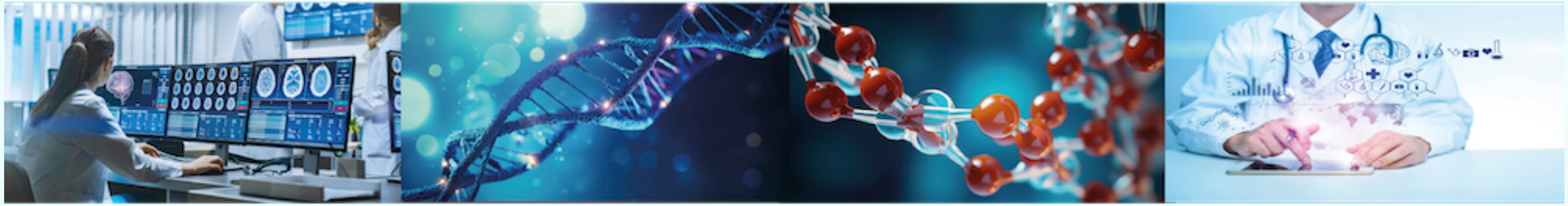


Programming for Biomedical Informatics

Lecture 9 “Measuring Gene Expression”

https://github.com/tisimpson/pbi

Ian Simpson
ian.simpson@ed.ac.uk

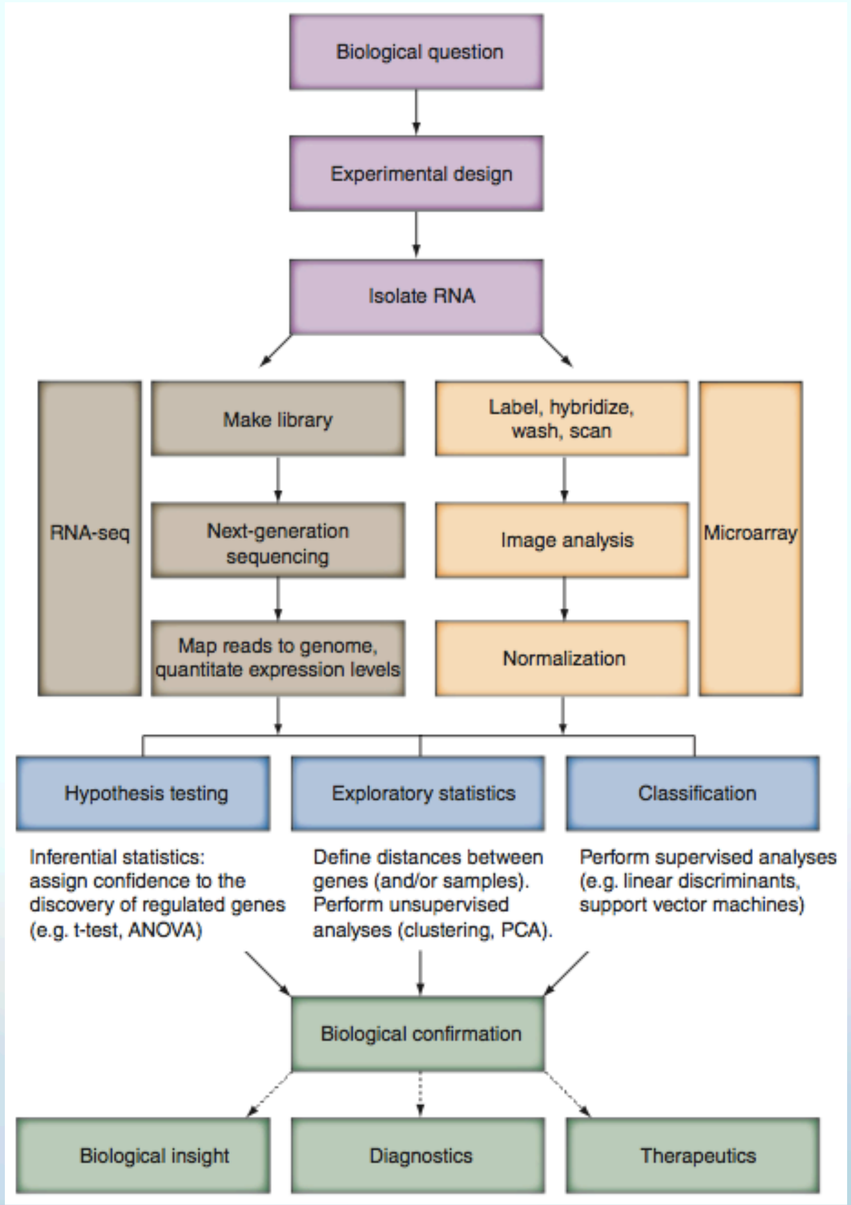
Background

THE UNIVERSITY of EDINBURGH
informatics

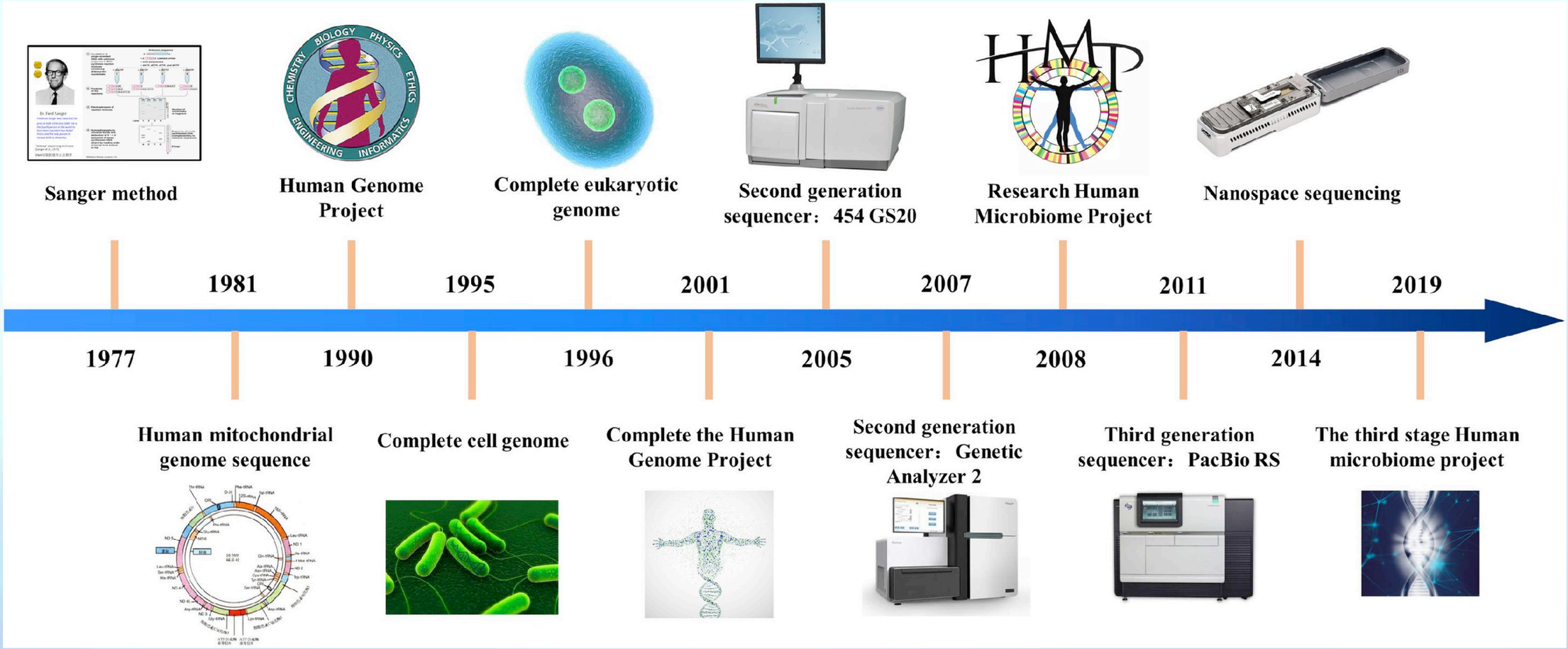
Why Measure Gene Expression?

- **Understanding Biological Processes**: Gene expression analysis helps us understand how genes and the regulatory networks they form control biological processes such as cell division, metabolism, aging, and response to environmental changes.
- **Disease Diagnosis and Prognosis**: Changes in gene expression patterns are often associated with diseases. Measuring gene expression can help in diagnosis, predicting disease progression, and assessing risk.
- **Developmental Biology**: Gene expression studies play a critical role in unraveling the complexities of development in organisms, showing how genes are dynamically regulated as organisms develop from embryos to adults.
- **Gene Function Discovery**: By measuring when and where genes are expressed, we can infer the function of unknown genes in comparison to genes of known function.
- **Drug Discovery and Development**: Gene expression analysis can identify molecular targets for new drugs and help us understand the potential side effects of drugs. It is an essential component of preclinical research.
- **Personalised Medicine**: Individual differences in gene expression can influence how patients respond to drugs (pharmacogenomics). Medical treatments can be more closely tailored to individual patients, thereby increasing the efficacy and safety of therapies.
- **Response to Therapy**: Measuring gene expression can help determine how well a patient is responding to a treatment and whether adjustments are necessary.
- **Microbial Pathogenesis**: In infectious diseases, understanding gene expression changes in pathogens and host organisms can reveal details about the mechanism of infection and host defence, leading to better therapeutic strategies.
- **Evolutionary Studies**: Gene expression analysis contributes to our understanding of evolutionary processes by revealing how gene regulation changes among species over time and contributes to phenotypic diversity and adaptation.

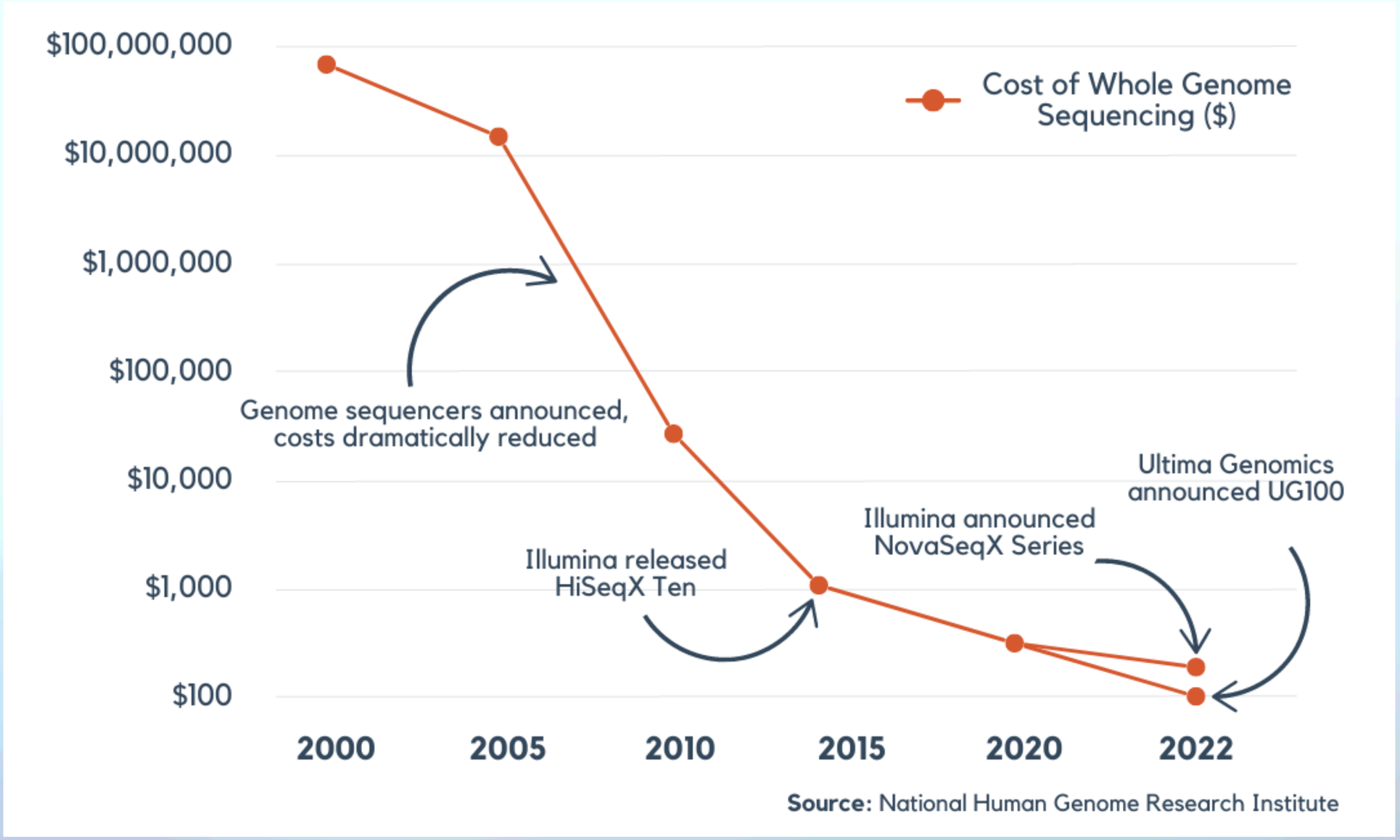
Workflow for Assessing Gene Expression



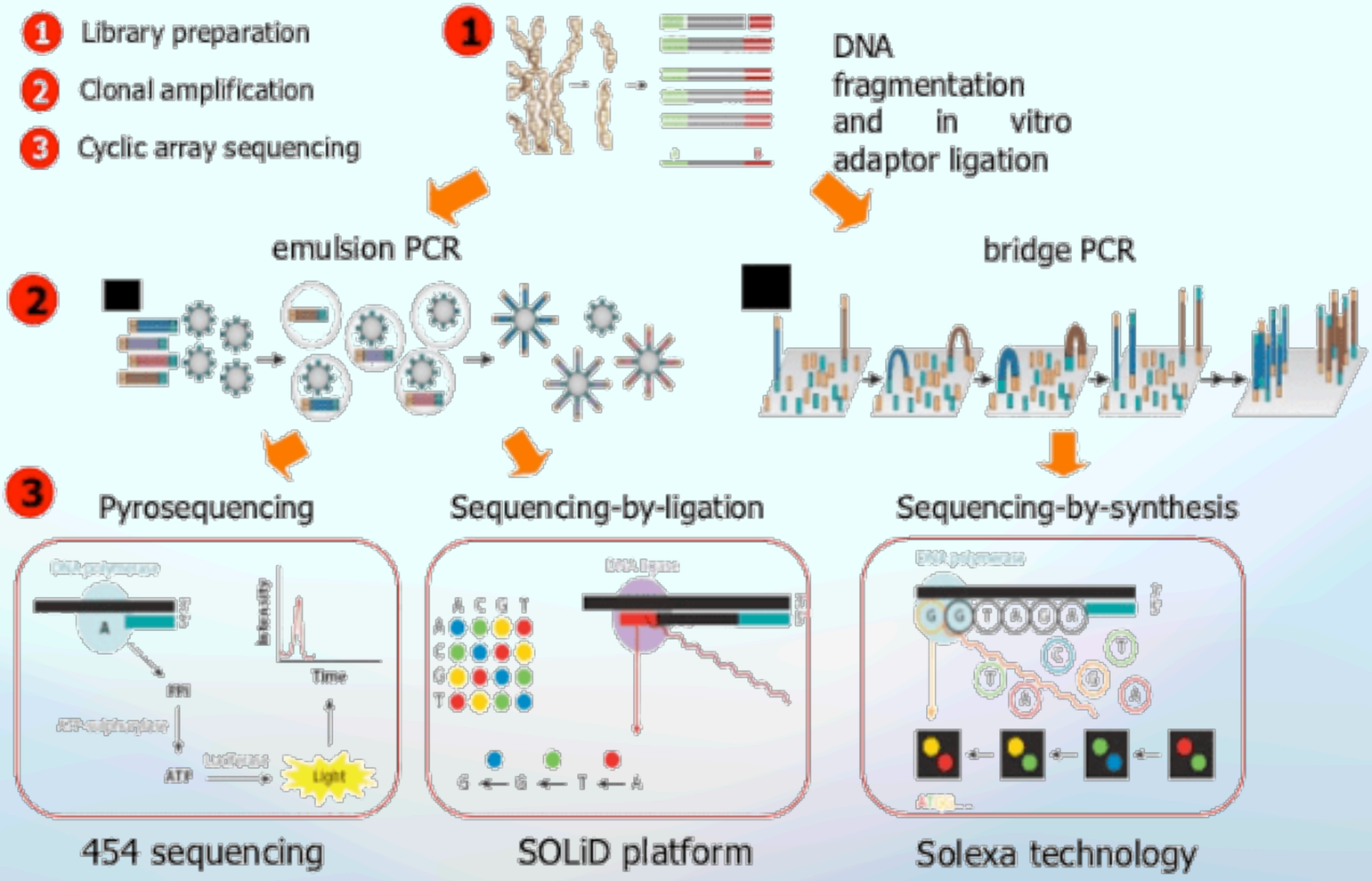
Nucleotide Sequencing Technologies, Past & Present



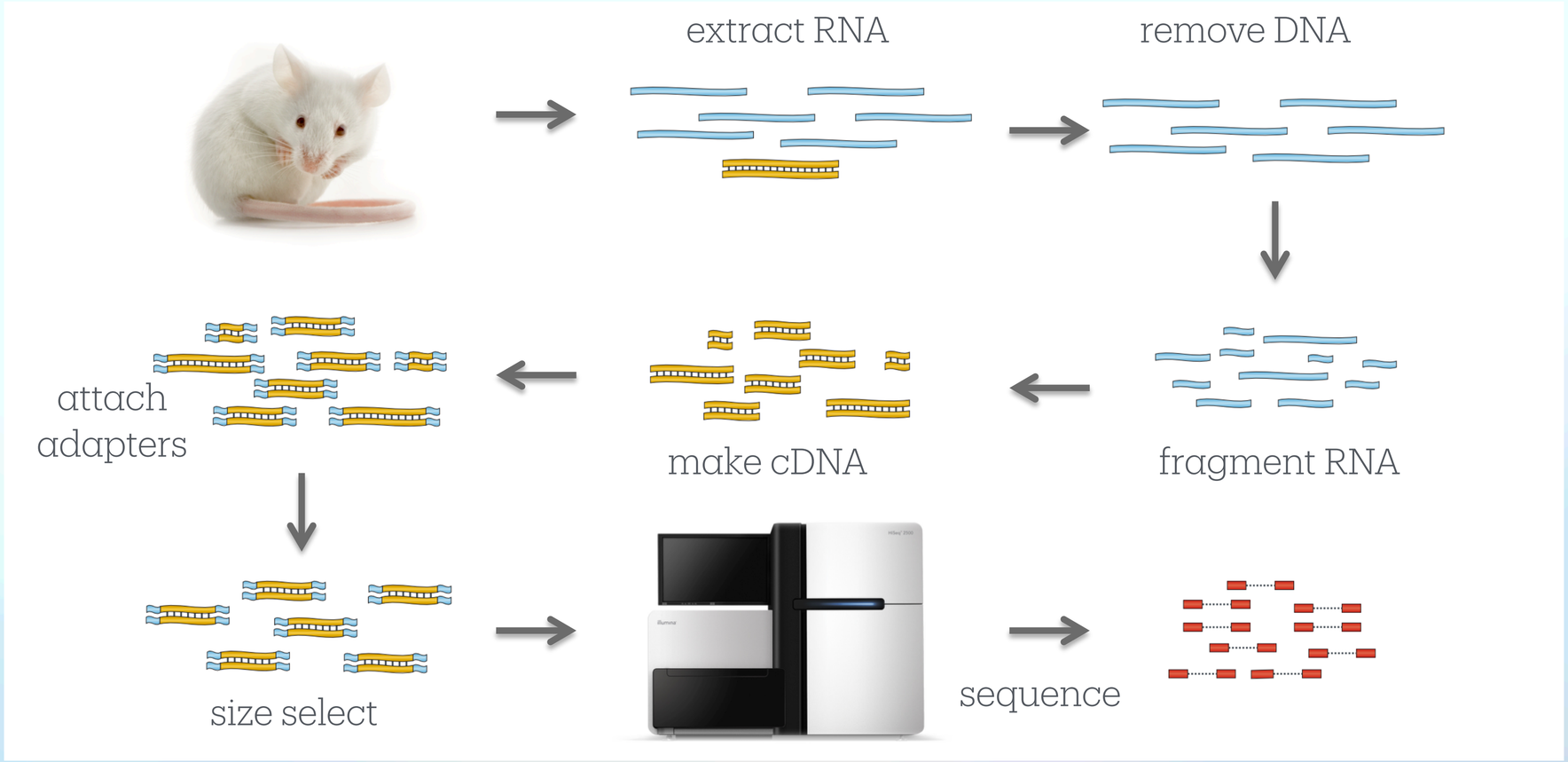
Sequencing Costs



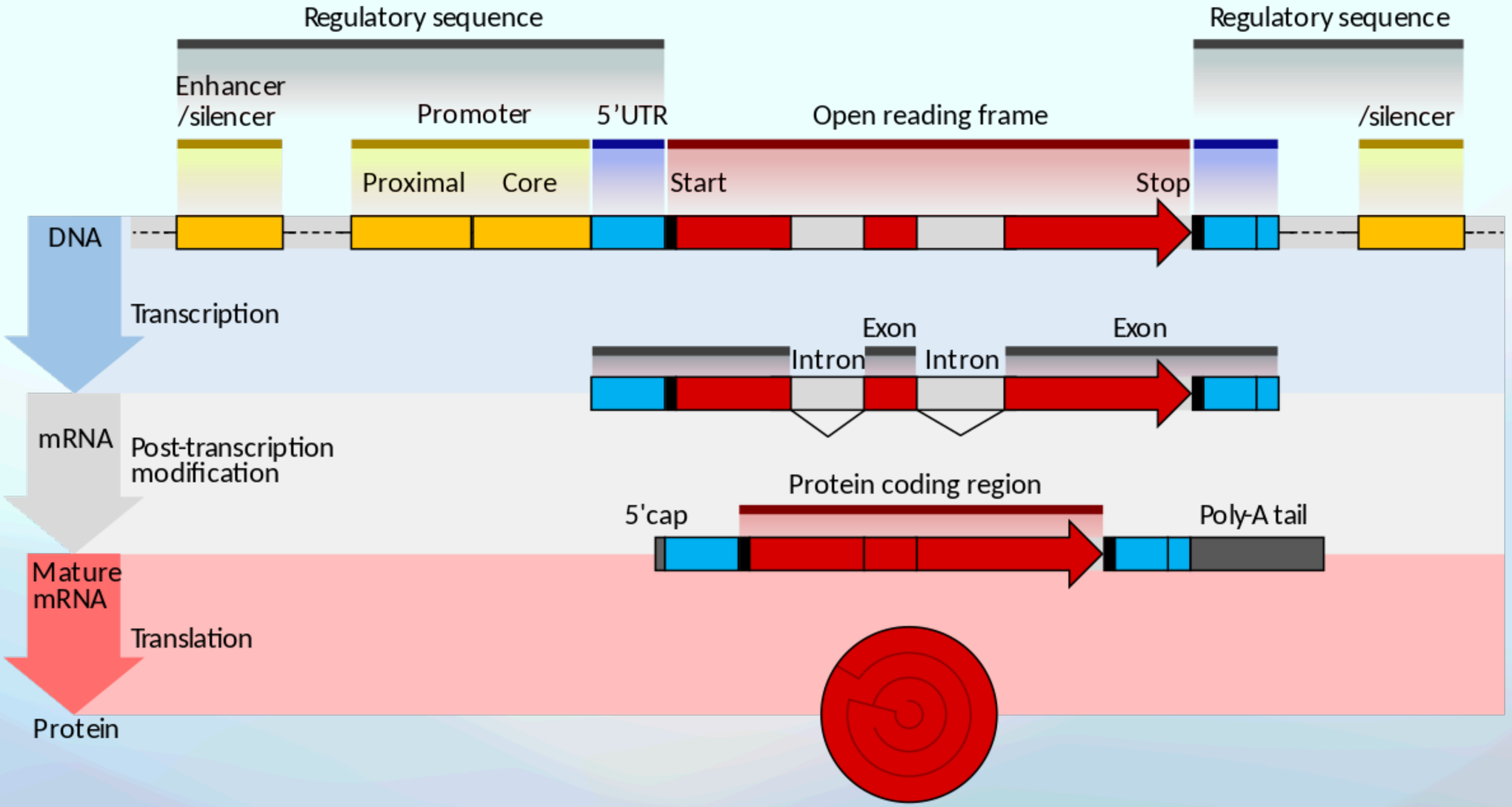
Next Generation Sequencing



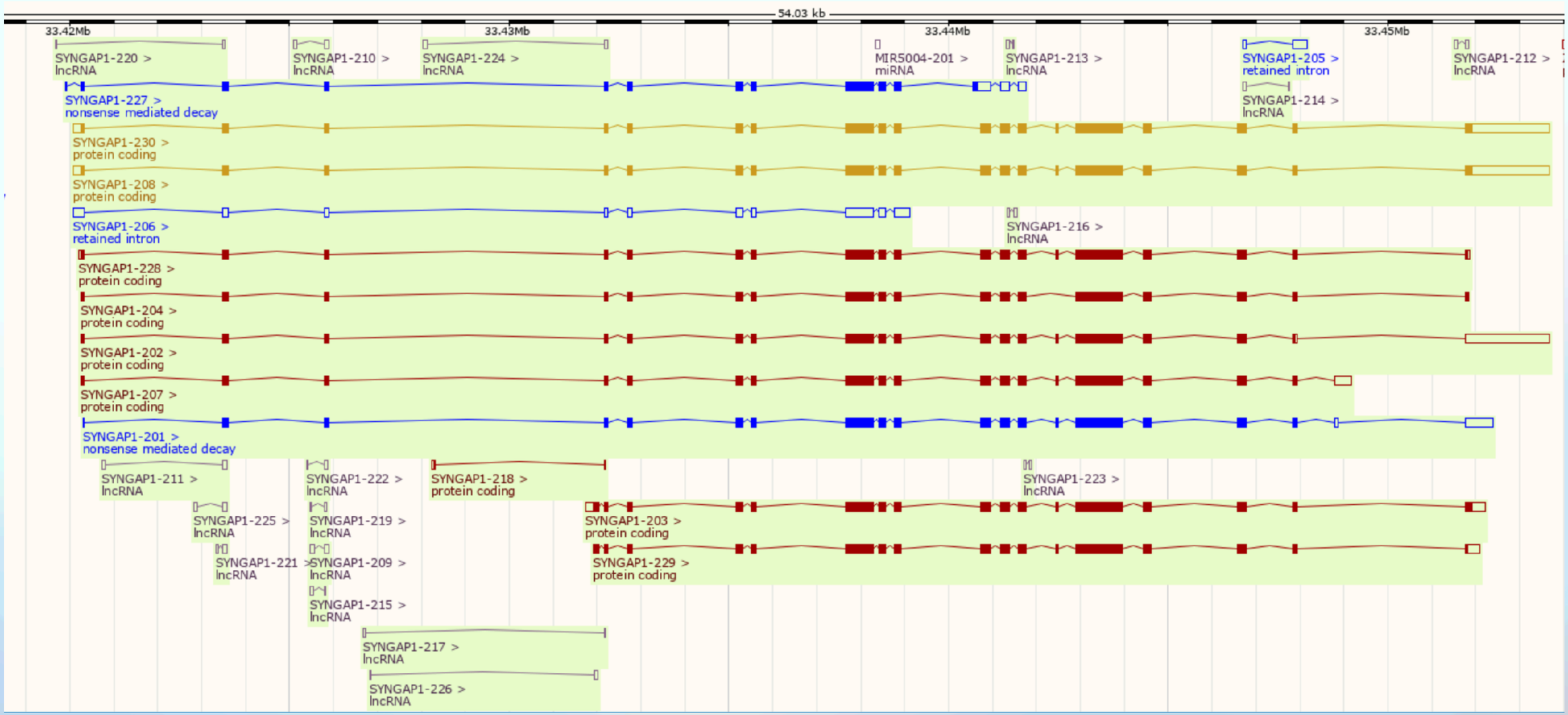
RNA-seq Data Generation



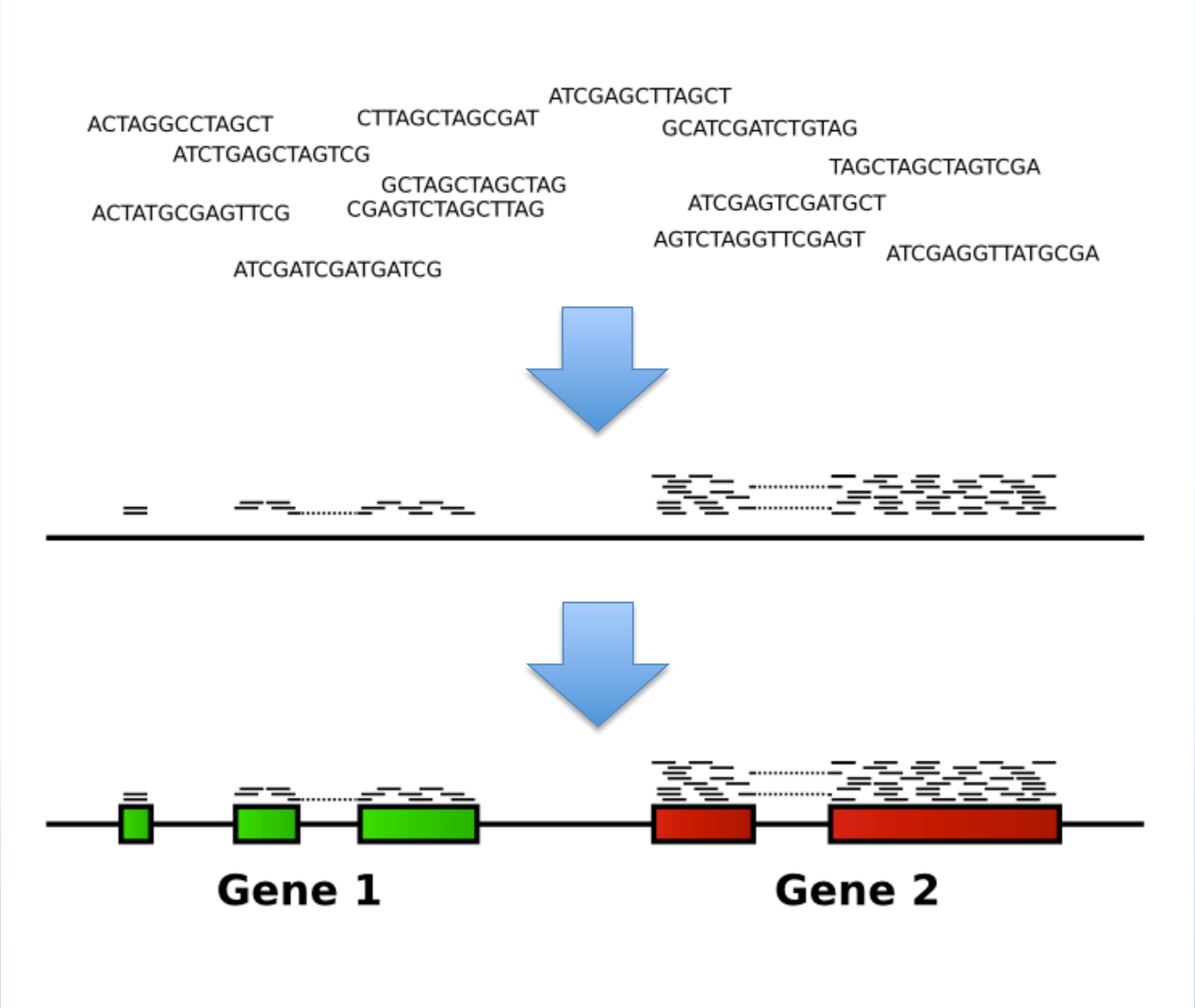
Gene Structure



Gene Models


54.03 kb
33.42Mb
33.43Mb
33.44Mb
33.45Mb
SYNGAP1-220 >
lncRNA
SYNGAP1-210 >
lncRNA
SYNGAP1-224 >
lncRNA
MIR5004-201 >
miRNA
SYNGAP1-213 >
lncRNA
SYNGAP1-205 >
retained intron
SYNGAP1-212 >
lncRNA
SYNGAP1-227 >
nonsense mediated decay
SYNGAP1-214 >
lncRNA
SYNGAP1-230 >
protein coding
SYNGAP1-208 >
protein coding
SYNGAP1-206 >
retained intron
SYNGAP1-216 >
lncRNA
SYNGAP1-228 >
protein coding
SYNGAP1-204 >
protein coding
SYNGAP1-202 >
protein coding
SYNGAP1-207 >
protein coding
SYNGAP1-201 >
nonsense mediated decay
SYNGAP1-211 >
lncRNA
SYNGAP1-222 >
lncRNA
SYNGAP1-218 >
protein coding
SYNGAP1-223 >
lncRNA
SYNGAP1-225 >
lncRNA
SYNGAP1-219 >
lncRNA
SYNGAP1-203 >
protein coding
SYNGAP1-221 >
lncRNA
SYNGAP1-209 >
lncRNA
SYNGAP1-229 >
protein coding
SYNGAP1-215 >
lncRNA
SYNGAP1-217 >
lncRNA
SYNGAP1-226 >
lncRNA

Transcriptome Quantification



Read Mapping - Aligning Reads to References

- Find the genomic location of millions of sequence reads
- Sequence alignment problem for each read
- Why is this hard?

The FASTQ Sequence Format

The FASTQ format stores DNA sequence data as well as associated Phred quality scores of each base.

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

DNA read

Base quality score

FASTQ quality scores

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

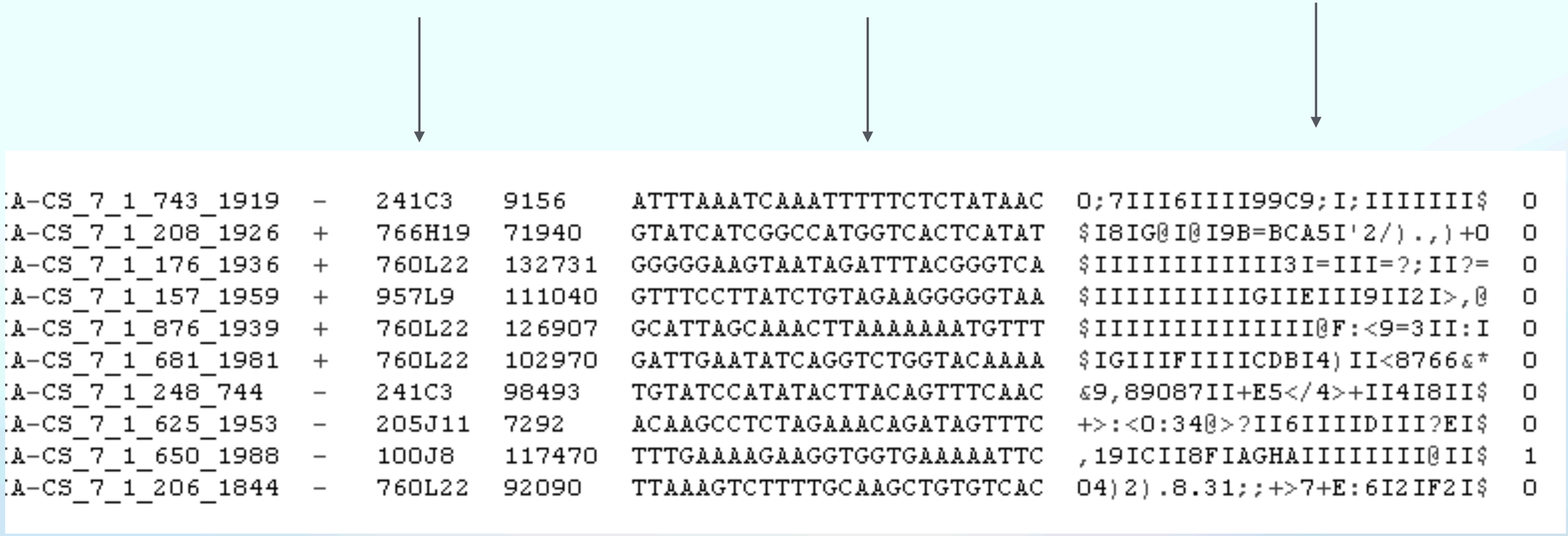
$P=10^{-Q/10}$ $Q=-10\log_{10}(P)$

Read Mapping - Aligning Reads to References

References to which reads match

reads

quality scores



```
:A-CS_7_1_743_1919   -   241C3    9156    ATTTAAATCAAATTTTTCTCTATAAC   O;7III6IIII99C9;I;IIIIIII$   0
:A-CS_7_1_208_1926   +   766H19   71940   GTATCATCGGCCATGGTCACTCATAT   $I8IG@I@I9B=BCA5I'2/).,)+O   0
:A-CS_7_1_176_1936   +   760L22   132731  GGGGGAAGTAATAGATTTACGGGTCA   $IIIIIIIIIIII3I=III=?;II?=   0
:A-CS_7_1_157_1959   +   957L9    111040  GTTTCCTTATCTGTAGAAGGGGGTAA   $IIIIIIIIIGIIEIII9II2I>,@   0
:A-CS_7_1_876_1939   +   760L22   126907  GCATTAGCAAACTTAAAAAAATGTTT   $IIIIIIIIIIIII@F:<9=3II:I   0
:A-CS_7_1_681_1981   +   760L22   102970  GATTGAATATCAGGTCTGGTACAAAA   $IGIIIFIIIICDBI4)II<8766&*   0
:A-CS_7_1_248_744    -   241C3    98493   TGTATCCATATACTTACAGTTTCAAC   &9,89087II+E5</4>+II4I8II$   0
:A-CS_7_1_625_1953   -   205J11   7292    ACAAGCCTCTAGAAACAGATAGTTTC   +>:<O:34@>?II6IIIIDIII?EI$   0
:A-CS_7_1_650_1988   -   100J8    117470  TTTGAAAAGAAGGTGGTGAAAAATTC   ,19ICII8FIAGHAIIIIIIII@II$   1
:A-CS_7_1_206_1844   -   760L22   92090   TTAAAGTCTTTTGCAAGCTGTGTCAC   O4)2).8.31;;+>7+E:6I2IF2I$   0
```

Alignment File Standards

(1) The query name of the read is given (M01121...)

(2) The flag value is 163. This is a decimal generated from a binary string of flag states for the alignment

(3) The reference sequence name, chrM, refers to the mitochondrial genome

(4) Position 480 is the left-most coordinate position of this read

(5) The Phred-scaled mapping quality is 60 (an error rate of 1 in 10^6)

(6) The CIGAR string (148M2S) shows 148 matches and 2 soft-clipped (unaligned) bases

```
home/bioinformatics$ samtools view 030c_S7.bam | less
M01121:5:000000000-A2DTN:1:2111:20172:15571     163     chrM
480     60      148M2S  =       524     195     AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACC
AACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATGTTTAGACGGG  BBBBBFFB5@FFGGGFGEGGGEGAAACGHFHFEGGAGFFH
AEFDGG?E?EGGGFGHFGHF?FFCHFH00E@EGFGGEEE1FFEEEHBGEFFFGGGG@</0
1BG212222>F21@F11FGFG1@1?GC<G11?1?FGDGGF=GHFFFHC.-
RG:Z:Sample7    XC:i:148        XT:A:U  NM:i:3  SM:i:37
AM:i:37 X0:i:1  X1:i:0  XM:i:3  XO:i:0  XG:i:0  MD:Z:19C109C0A17
```

(7) An = sign shows that the mate reference matches the reference name

(8) The 1-based left position is 524

(9) The insert size is 195 bases

(10) The sequence begins AATCT and ends ACGGG (its length is 150 bases)

(11) Each base is assigned a quality score (from BBBBB ending FHC.-)

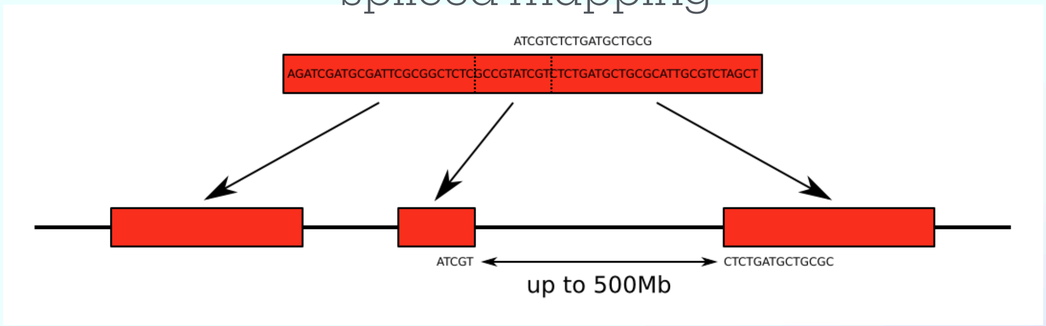
(12) This read has additional, optional fields that accompany the MiSeq analysis

- A standard format has been introduced for aligned sequences called Sequence Alignment/Map (SAM). Its binary version (which is compressed) is called BAM
- Aligned BAM files are available at repositories (Sequence Read Archive at NCBI, ENA at Ensembl)
- SAMTools is a software package commonly used to analyse SAM/BAM files (http://samtools.sourceforge.net/)

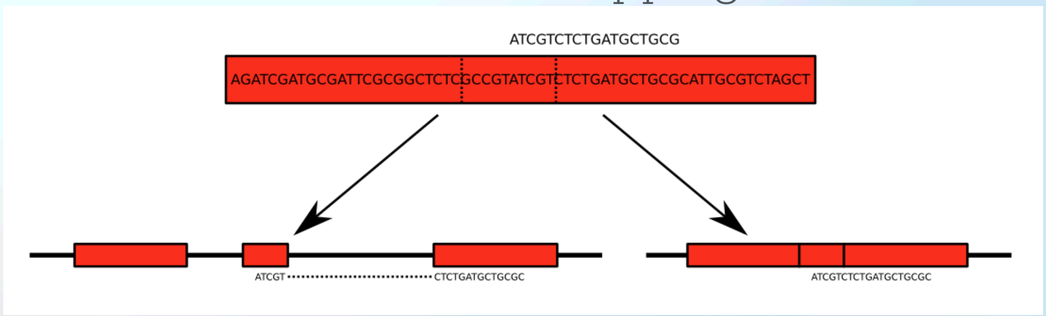
Difficulties Mapping RNA-seq Reads

- Errors in the reads
- Mismatches
- Insertions & Deletions
- Errors in the Reference Sequence
- Genetic Variation
- Spliced Mapping
- Multi-mapping
- Efficiency vs. Accuracy

spliced mapping



multi-mapping



STAR: Ultrafast Universal RNA-seq Aligner

- Builds genome index (uncompressed suffix array) and for each read
- finds “maximal mappable” seeds
- clusters and stitches seeds together via local alignment
- Mapping speed is traded against RAM
- ~30Gb RAM for aligning to the human genome
- 550 million paired-end reads/hour

Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR.
Curr Protoc Bioinformatics. 2015 Sep 3;51:11.14.1-11.14.19. doi: 10.1002/0471250953.bi1114s51.

Dobin A, Gingeras TR. Optimizing RNA-Seq Mapping with STAR.
Methods Mol Biol. 2016;1415:245-62. doi: 10.1007/978-1-4939-3572-7_13.

From Mapped Reads to Quantification

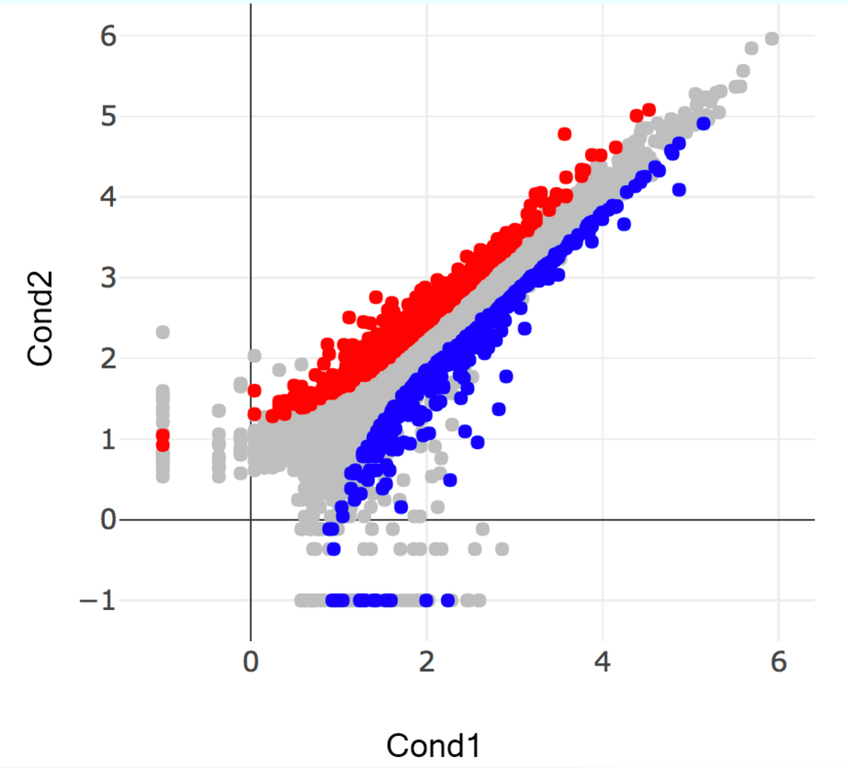
- Mapped reads need to be counted and summarised for some meaningful biological entity
- This could be at the gene or transcript level or even sub-sets of exons
- There are a huge number of tools for this:
  - *Salmon, Kallisto*
  - *Sailfish, RSEM, eXpress, Cufflinks*
  - *MISO, Scripture, FluxCapacitor*
  - *SLIDE, PSGInfer, IsoLasso, SpliceTrap*
  - *ERANGE, IsoEM, iReckon, DRUT, rQuant*
- When working with pre-processed gene expression data you will often download data that has already been mapped, normalised, and summarised ready for use
- It is still very important that you understand what has been done to the raw data!

http://seqanswers.com

Differential Expression Analysis (DEA)

- Identify genes (or isoforms) expressed in significantly different quantities between conditions
- Treatment versus control, Diseased versus healthy, Different tissues, developmental stages….
- More complicated designs might include experimental factors with multiple levels
- How can we tell if a gene or isoform is differentially expressed?
- Need to understand how abundance varies within sample groups
- Need biological replicates
- RNA-seq replicates are expensive
- Need statistical methods with enough detection power to perform well with small sample sizes
- Makes assumptions about the distribution of data

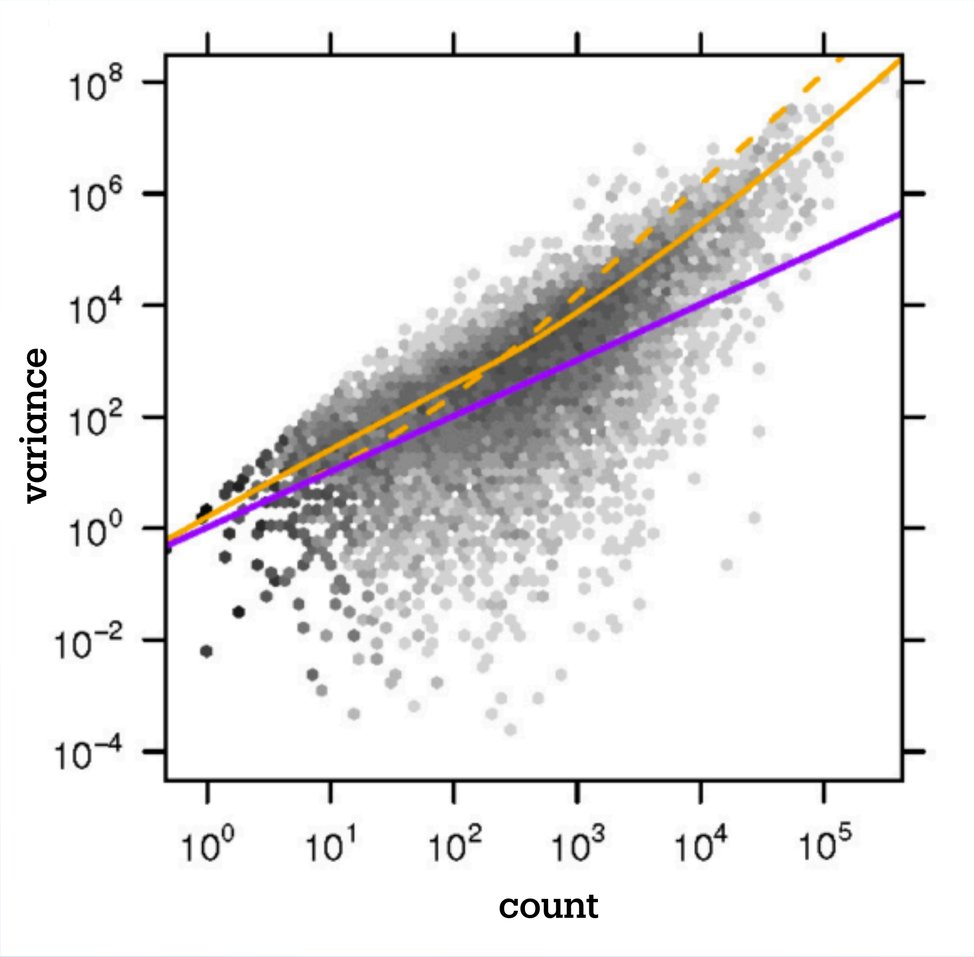
Visualising Expression Using a Simple Scatterplot



Scatter plot used to visualise differences between two samples, plotting the expression counts.

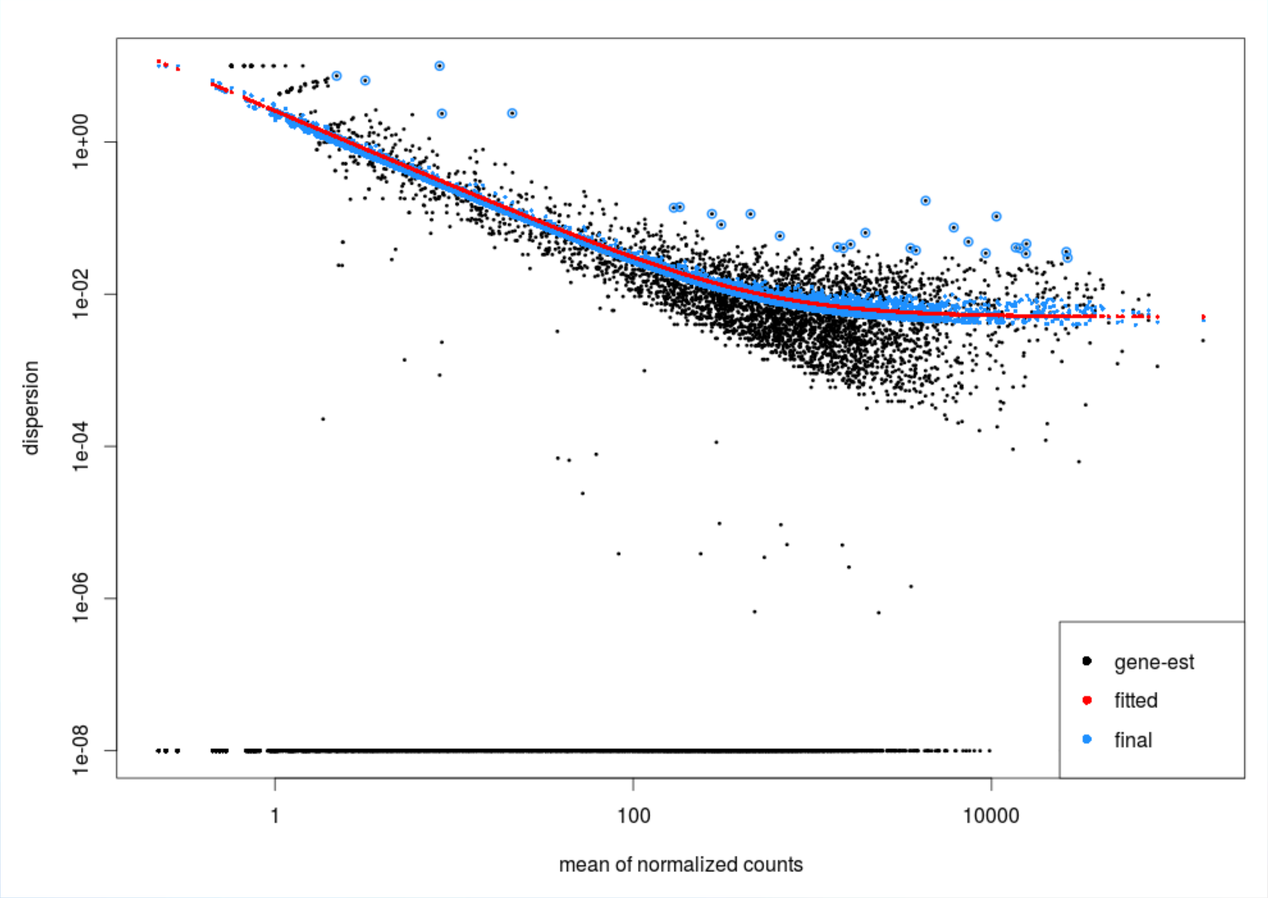
- We can make such a plot before any normalisation has taken place. This example shows a very stereotypical normalised plot. What would a plot that hadn't been normalised look like?
- Note that most genes are essentially expressed identically between samples - this is an assumption fundamental to all differential gene expression analyses.
- Looking at some real RNA-seq count distributions reveals some of the key features of this kind of data that need to be taken into account when developing statistical methods to quantify differences between samples.

Negative Binomial Distribution



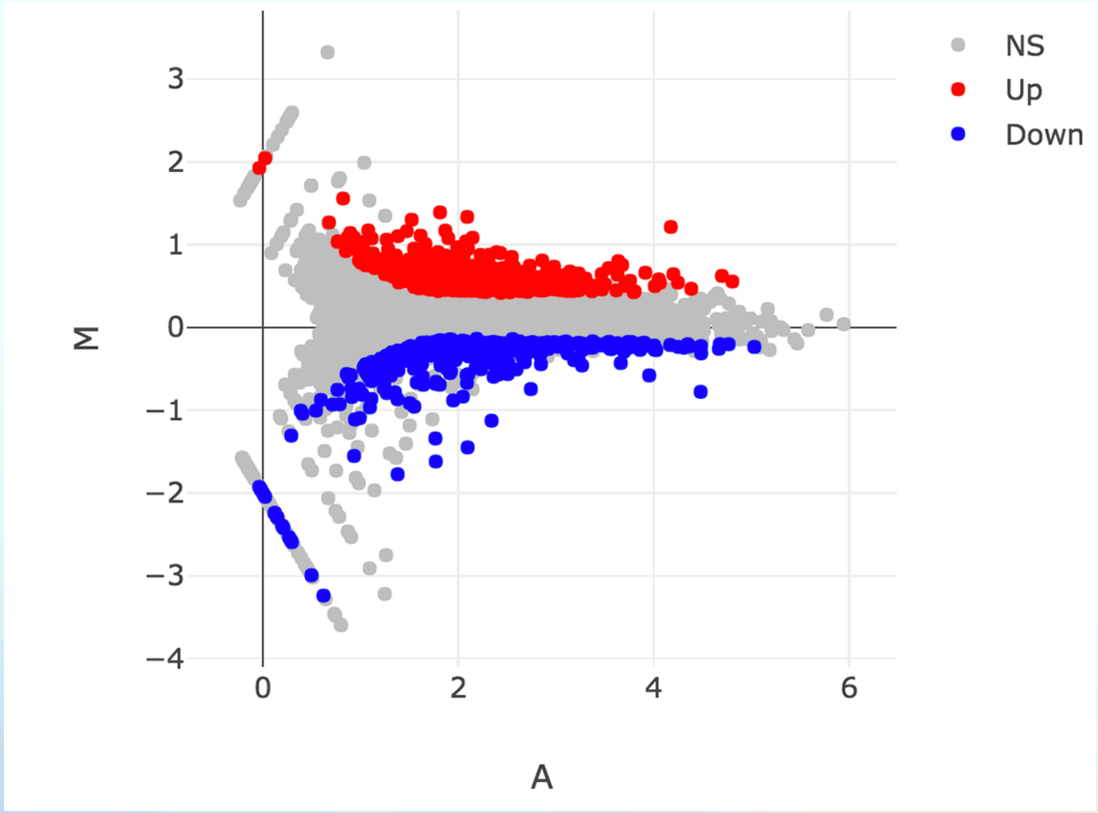
- The negative binomial distribution is commonly used to model RNA-seq data due to the variability and dispersion observed in count data.
- RNA-seq data typically exhibits over-dispersion relative to what would be expected under a Poisson distribution. This occurs when the variance in the data is greater than the mean which is common due to biological variability between samples and technical variability in measurements.
- A Poisson distribution assumes that the mean and variance are equal so this cannot be used (it is a commonly used distribution for count based data). The negative binomial distribution has an extra parameter which allows the variance to exceed the mean.
- RNA-seq datasets often exhibit a large number of zeros (genes with no reads mapped), which may represent true biological zeroes (genes not expressed) as well as zeroes resulting from technical limitations (genes expressed at levels too low to be detected). The negative binomial framework can be extended to include zero-inflation models that better handle this.
- Using a negative binomial distribution for RNA-seq data leads to more robust inference of differential expression. Tools and software for analysing RNA-seq data that use the negative binomial distribution (like DESeq2, EdgeR) can more accurately detect genes that are significantly differentially expressed between conditions.

Measuring Gene Expression Dispersion



- Dispersion is derived from the fitted negative binomial model and is a measure of the level of variance as a proportion of the mean count (essentially a normalised variance). This dispersion model fit is a fundamental part of the significance estimation procedure.
- In RNA-seq data, the variance of the gene expression counts tends to increase with the mean. This heteroscedasticity means that genes with higher expression levels often show higher absolute variability in their counts.
- This must be taken into account when assessing differential expression or significantly different genes will be biased by genes that have lower expression levels
- Differences in library size and sequencing depth across samples means that normalisation is required to make meaningful comparisons across genes and samples. The dispersion characteristics can change through normalisation.
- Longer genes are more likely to accumulate more reads purely due to their length, influencing both the mean and variance of the read counts attributed to these genes, impacting dispersion characteristics.

Visualising Gene Expression with an MA plot



$$M = \log_2(R/G) = \log_2(R) - \log_2(G)$$

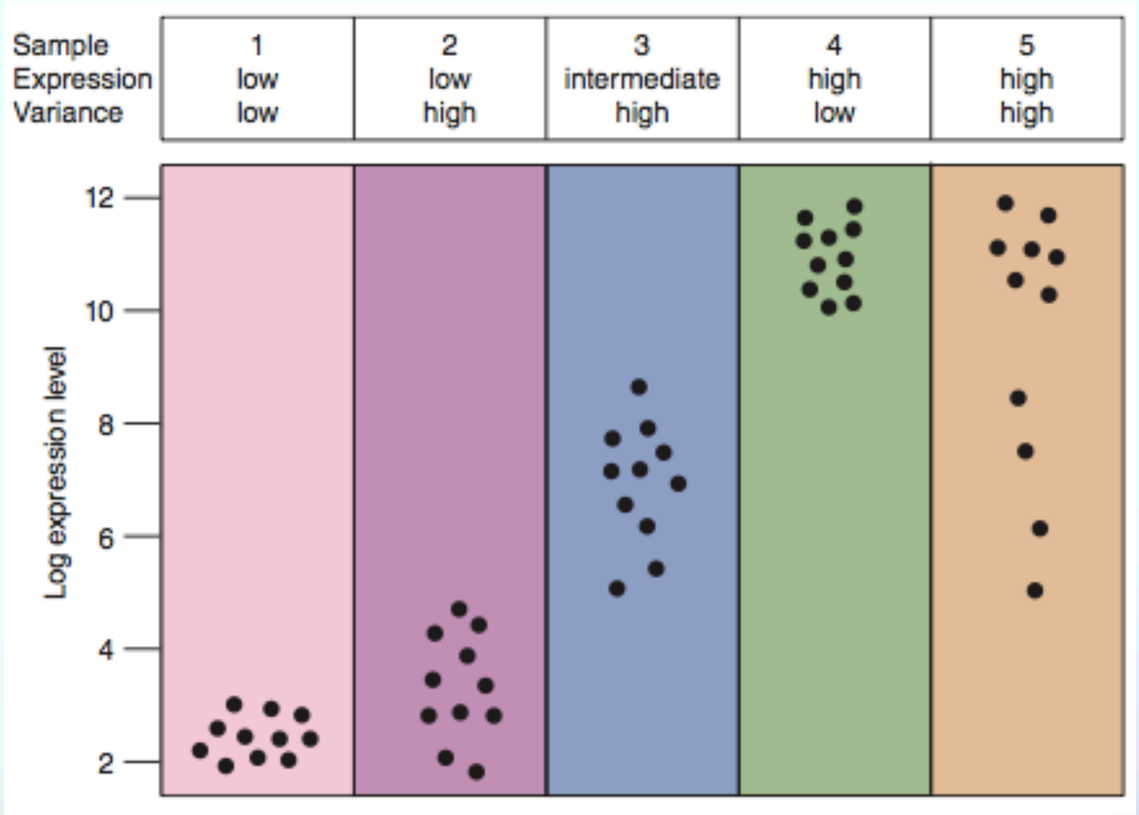
$$A = \frac{1}{2}\log_2(RG) = \frac{1}{2}(\log_2(R) + \log_2(G))$$

Scatter plot used to visualise differences between samples

The "MA" in MA plot stands for "M" (log ratio) and "A" (mean average):

- M : log ratio of the expression levels of genes between two conditions where M = log2(Expression Condition 1 / Expression Condition 2). This metric highlights the change in gene expression due to the treatment or experimental condition.
- A: the average log expression level of genes across the two conditions where A = (log2(Expression Condition 1) + log2(Expression Condition 2)) / 2. This metric provides a basis for normalisation and shows where the bulk data lies regardless of the change.
- Deviation from M=0 indicates differential expression and from linearity implies biased measurement of gene expression at different expression values - this needs to be corrected.

Gene & Transcript Specific Variance



Each dot is a replicate. Comparison of conditions 1 and 4 would produce a significant difference (we reject the null). Comparison of 3 versus 5 might not.

DESeq2 for Differential Gene Expression Analysis

DESeq2 is a popular tool implemented in R and Python for analysing count-based data like RNA-seq. It provides methods to test for differential expression by using negative binomial generalised linear models. The core of DESeq2's approach lies in its ability to accurately estimate variance-mean dependencies in count data, and to test for differential expression based on these estimates.

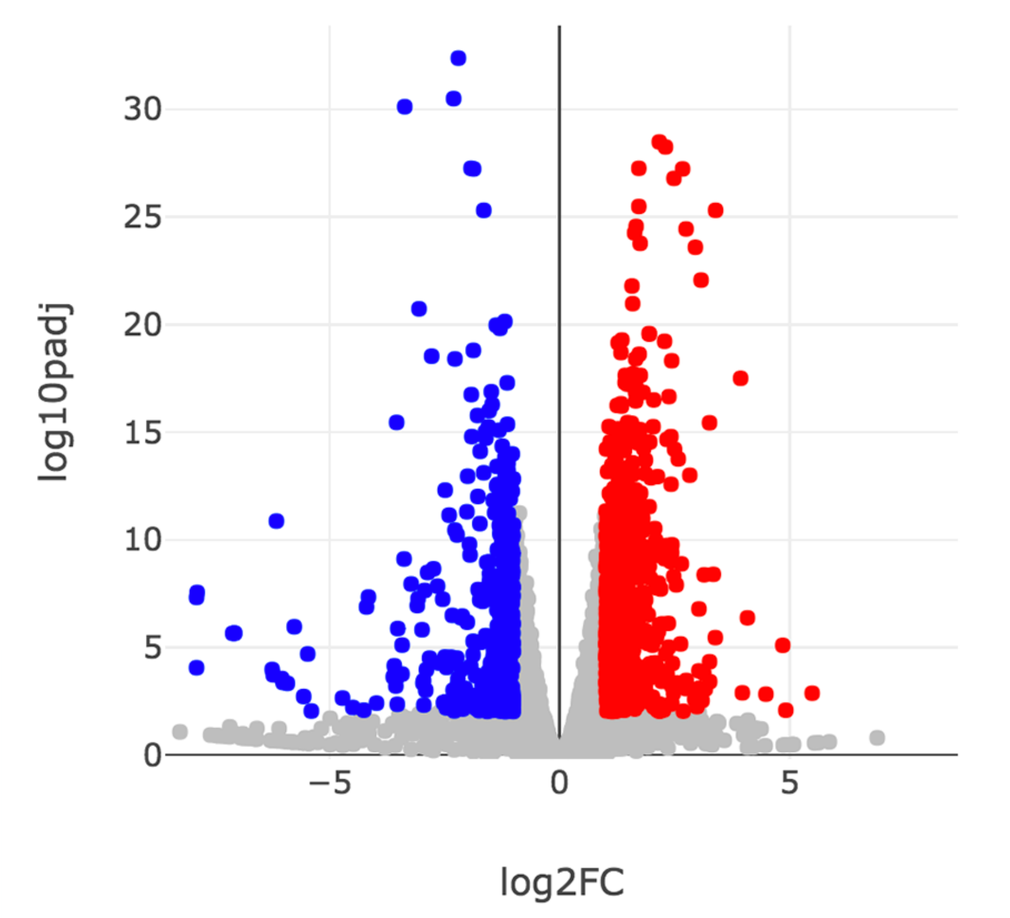
Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8

- Data is input as genes (rows) x samples (columns) matrix, along with sample information (conditions, replicates etc.).
  - **Filtering**: Low count genes are removed and gross normalisation for library size and sequencing depth are made
  - **Size Factor Calculation**: each gene's count is divided by the geometric mean of the counts for that gene across all samples
  - **Median Ratio Method**: size factors are calculated by taking the median of these ratios for each sample and these are used to normalise across samples.

- Dispersion is then estimated.
  - **Gene-wise**: each gene's dispersion is estimated from its mean and variance
  - **Shrinkage**: A Bayesian method is used to shrink the gene-wise dispersion estimates to moderate the values. This is especially important when the number of replicates is small.
  - **Fitting Dispersion-Mean Relationship**: A fit is made of dispersion estimates against mean expression to balance out gene-wise estimates and provide stable measures for all genes.

- Statistical Modelling & Hypothesis Testing
  - **Model Design**: A generalised linear model is used to incorporate information about experimental design, such as treatment conditions.
  - **Wald's Test or Likelihood Ratio Test**: DESeq2 then performs a statistical test (typically Wald's test) for each gene to determine whether it is significantly differentially expressed across conditions.
  - **Multiple Testing Correction**: Methods including Benjamini-Hochberg procedure are used to adjust p-values (often presented as an FDR - False Discovery Rate).

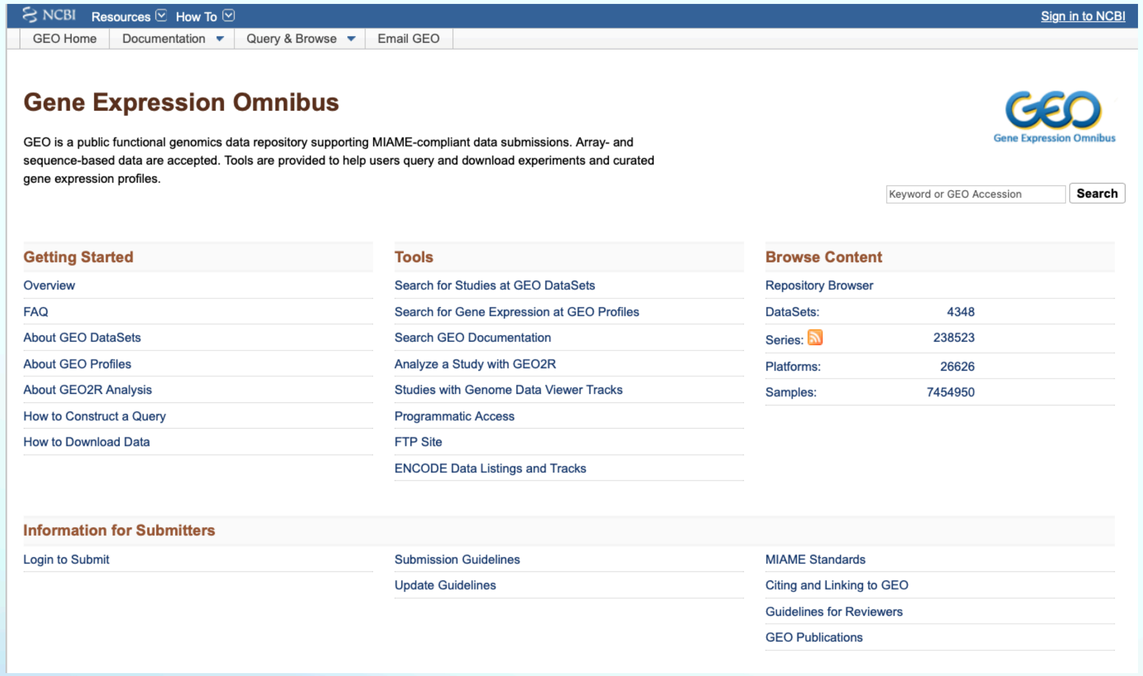
Visualising Differential Expression with a Volcano Plot

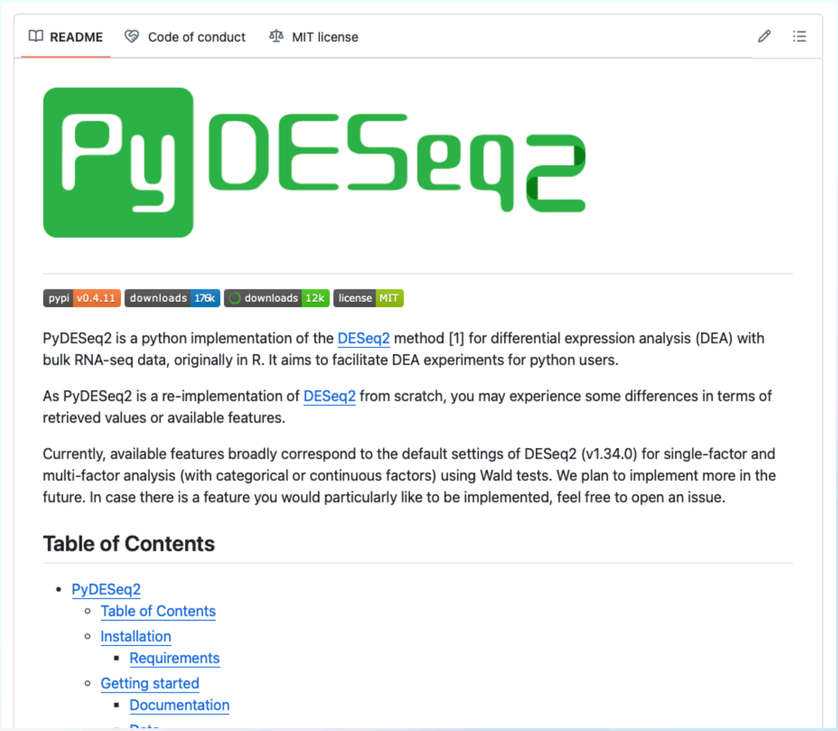


Type of scatter plot used especially for displaying results from differential gene expression. It helps to visually identify genes that are significantly differentially expressed between two biological conditions or treatments.

- X-axis: log2 scale of the fold change between two conditions. Positive values indicate up-regulation, and negative values indicate down-regulation in one condition relative to another.
- Y-axis: negative log10 of the p-value of the statistical test used to evaluate differential expression. Higher values represent lower p-values, suggesting greater statistical significance.
- Each point typically represents a gene (or other measured entity, like transcripts or proteins)
- Commonly presented with genes that are below an absolute FC value and above a certain adjusted p-value greyed out

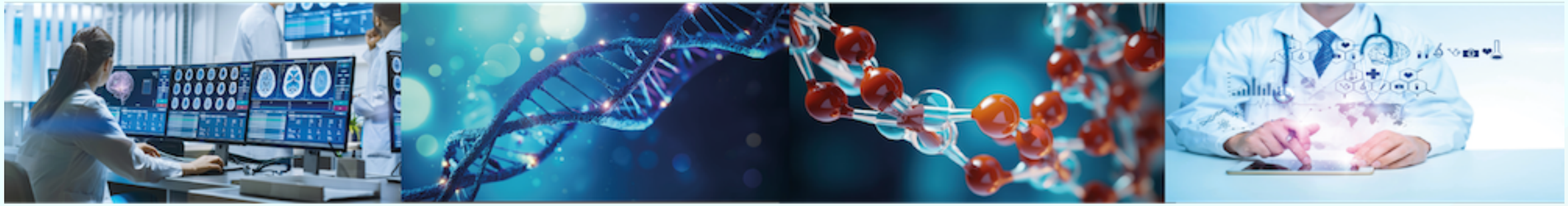
Differential Gene Expression Analysis Using NCBI-GEO & PyDESeq2





https://www.ncbi.nlm.nih.gov/geo/

https://github.com/owkin/PyDESeq2



Programming for Biomedical Informatics

Next Lecture this Thursday - “Differential Gene Expression”

Please Bring your Laptop!

Ask Questions on the EdStem Discussion Board

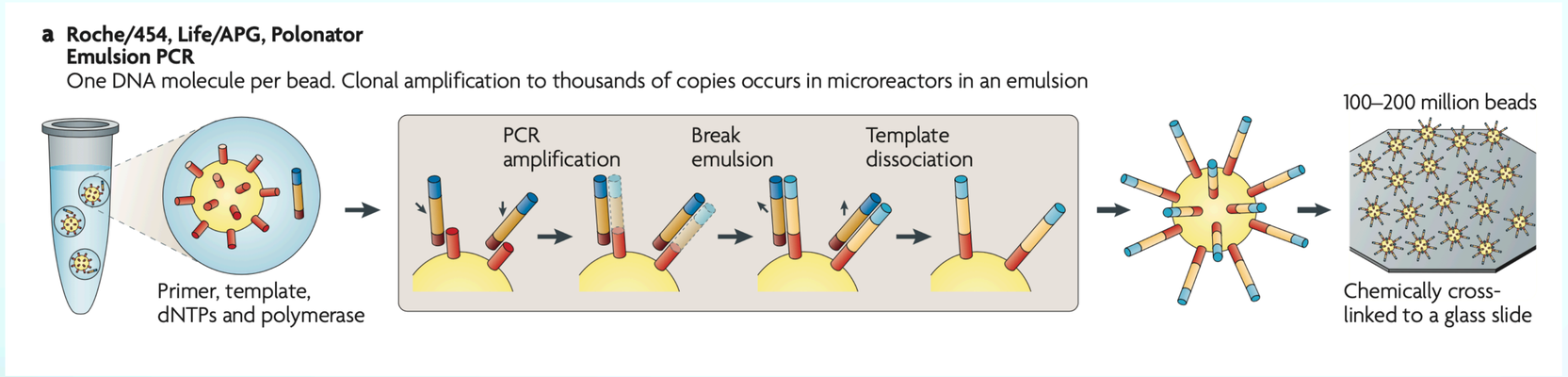
https://github.com/tisimpson/pbi

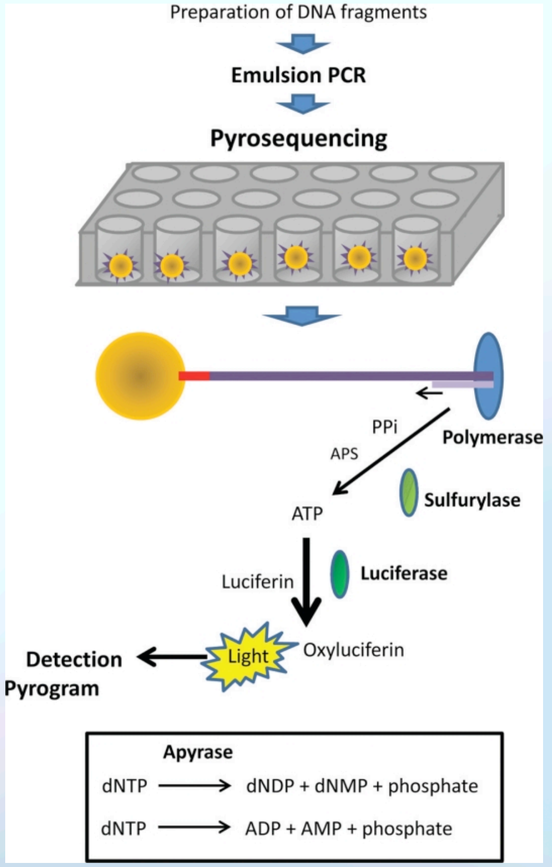
Coding

THE UNIVERSITY of EDINBURGH informatics

Extra Slides - DNA Sequencing Technologies

454 Sequencing (Emulsion PCR - Pyrosequencing)

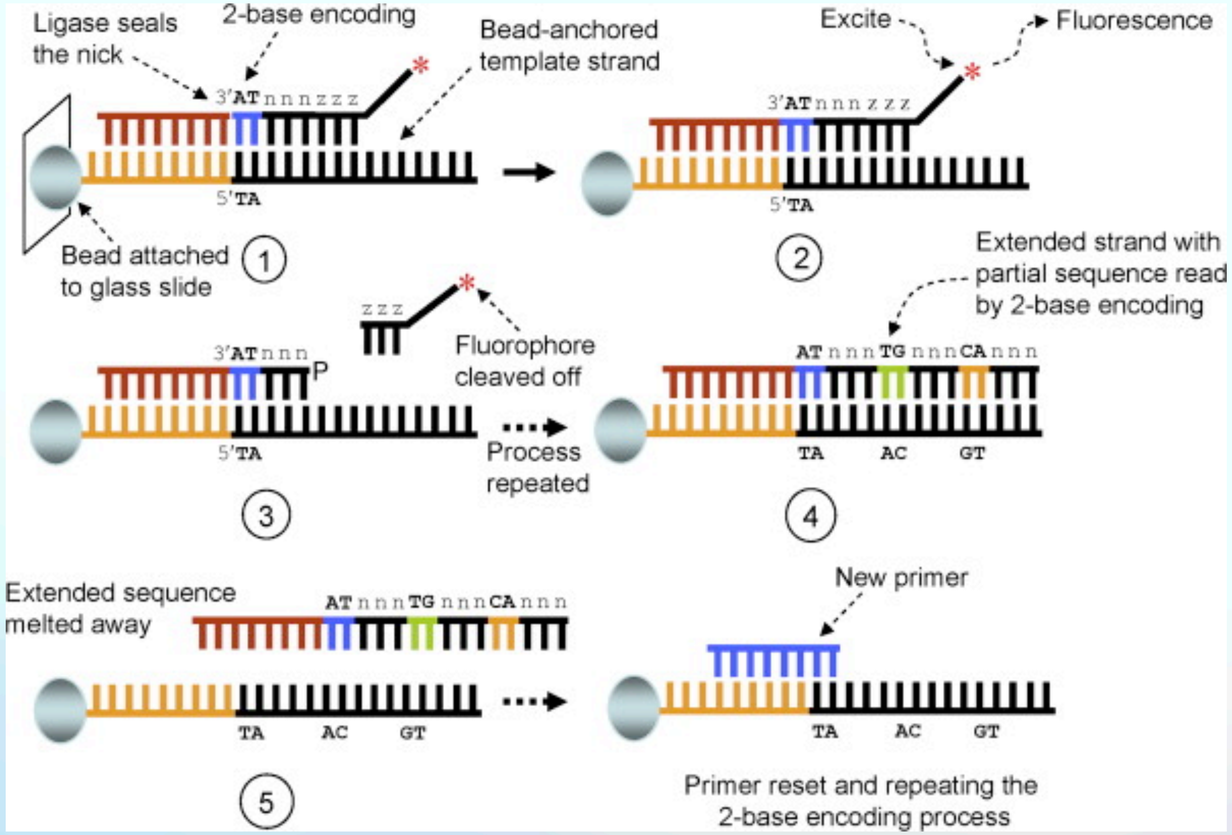




Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010)

https://www.youtube.com/watch?v=NJB-9HGSTfM

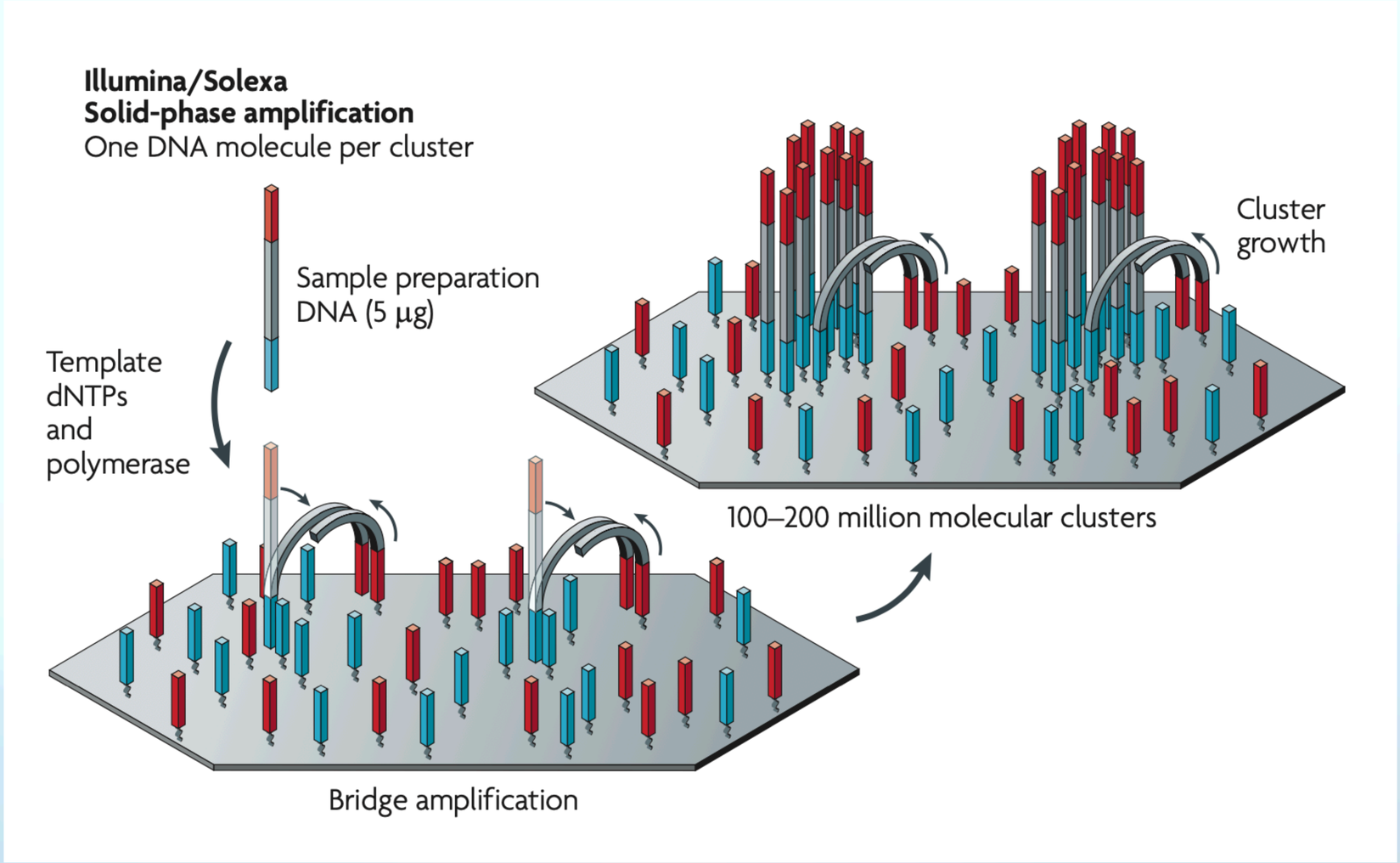
Solid Sequencing (Emulsion PCR – Sequencing by ligation)



Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

https://www.youtube.com/watch?v=nlvyF8bFDwM

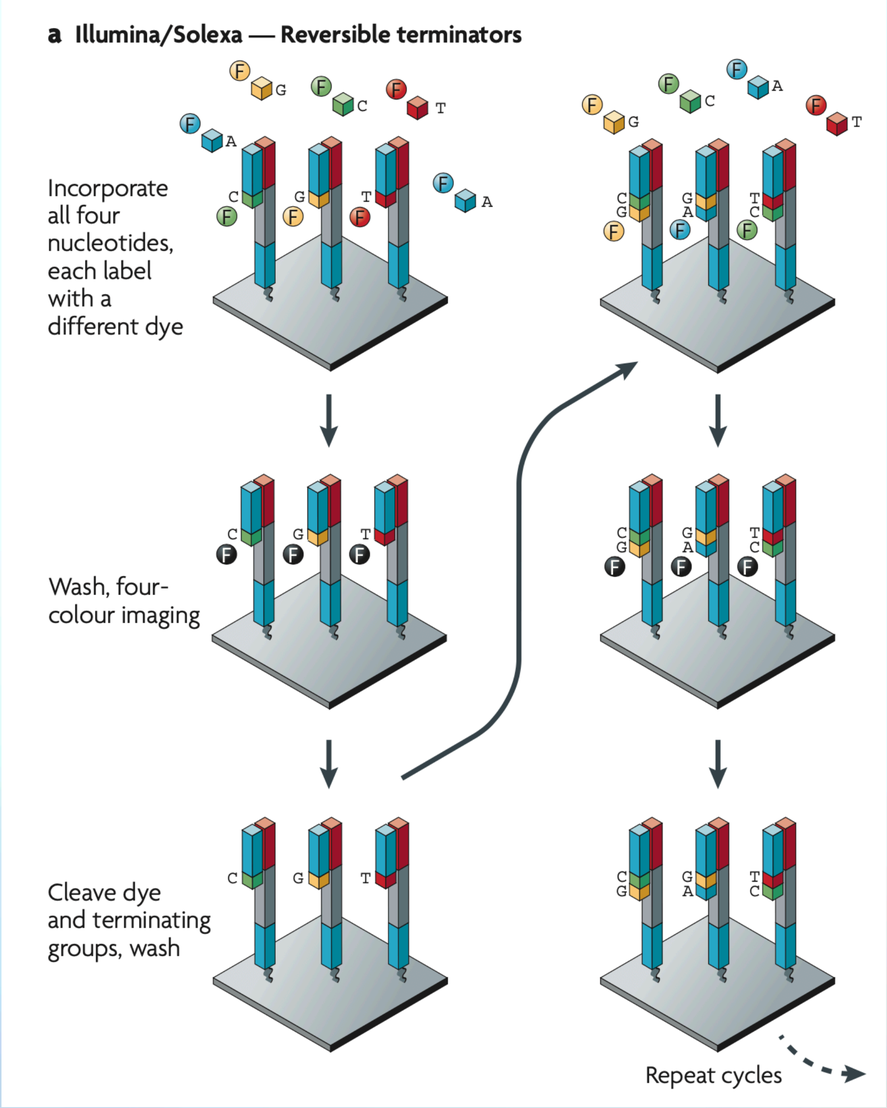
Illumina Sequencing (Bridge PCR – Sequencing by synthesis)

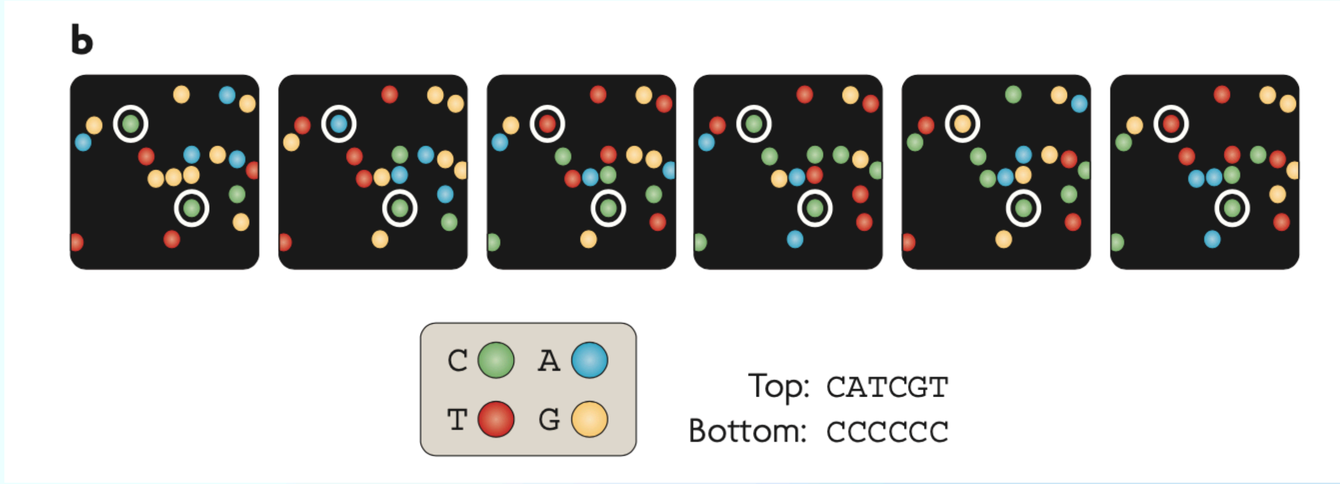


Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

https://www.youtube.com/watch?v=fCd6B5HRaZ8

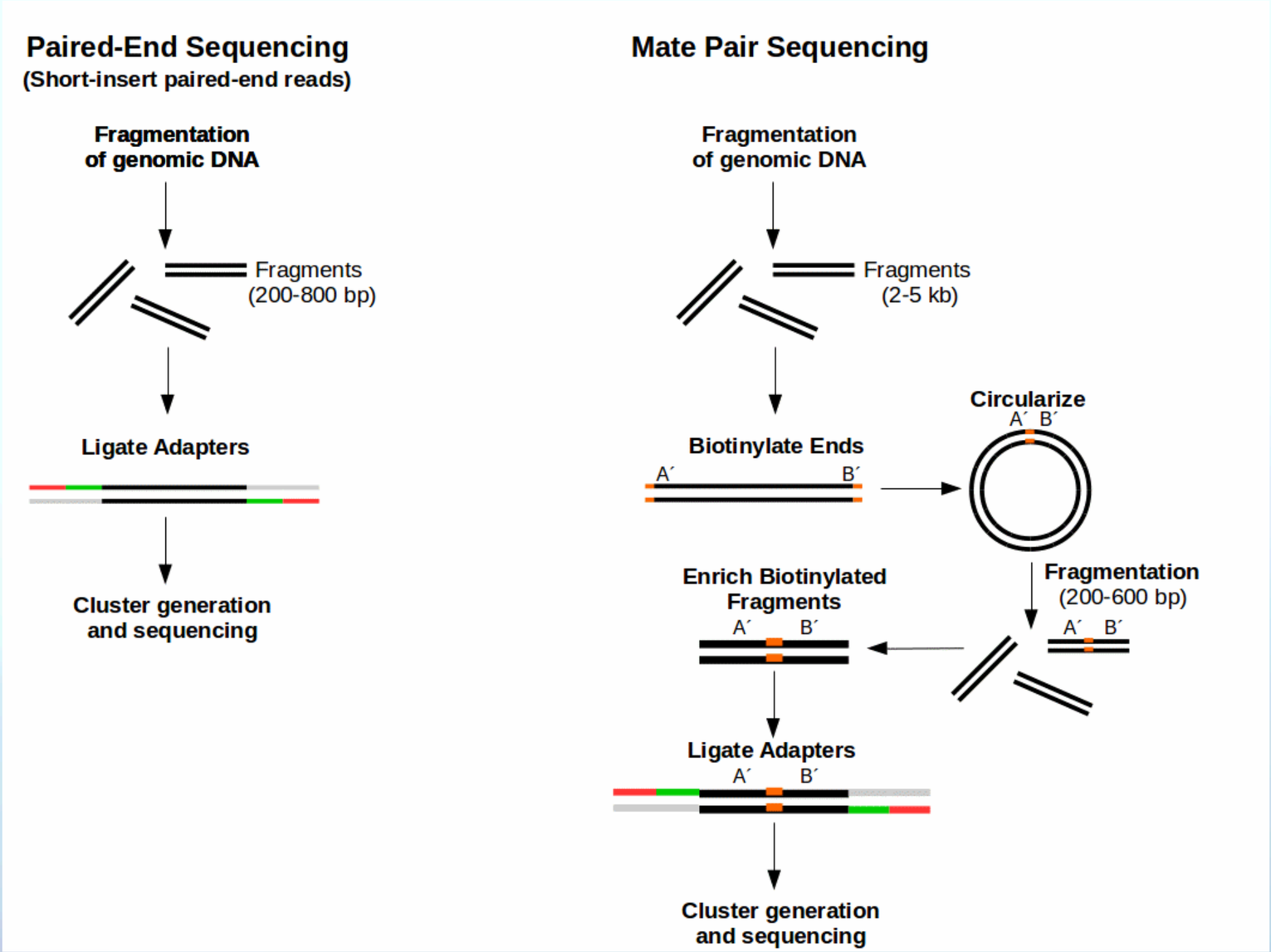
Illumina Sequencing (Bridge PCR – Sequencing by synthesis)



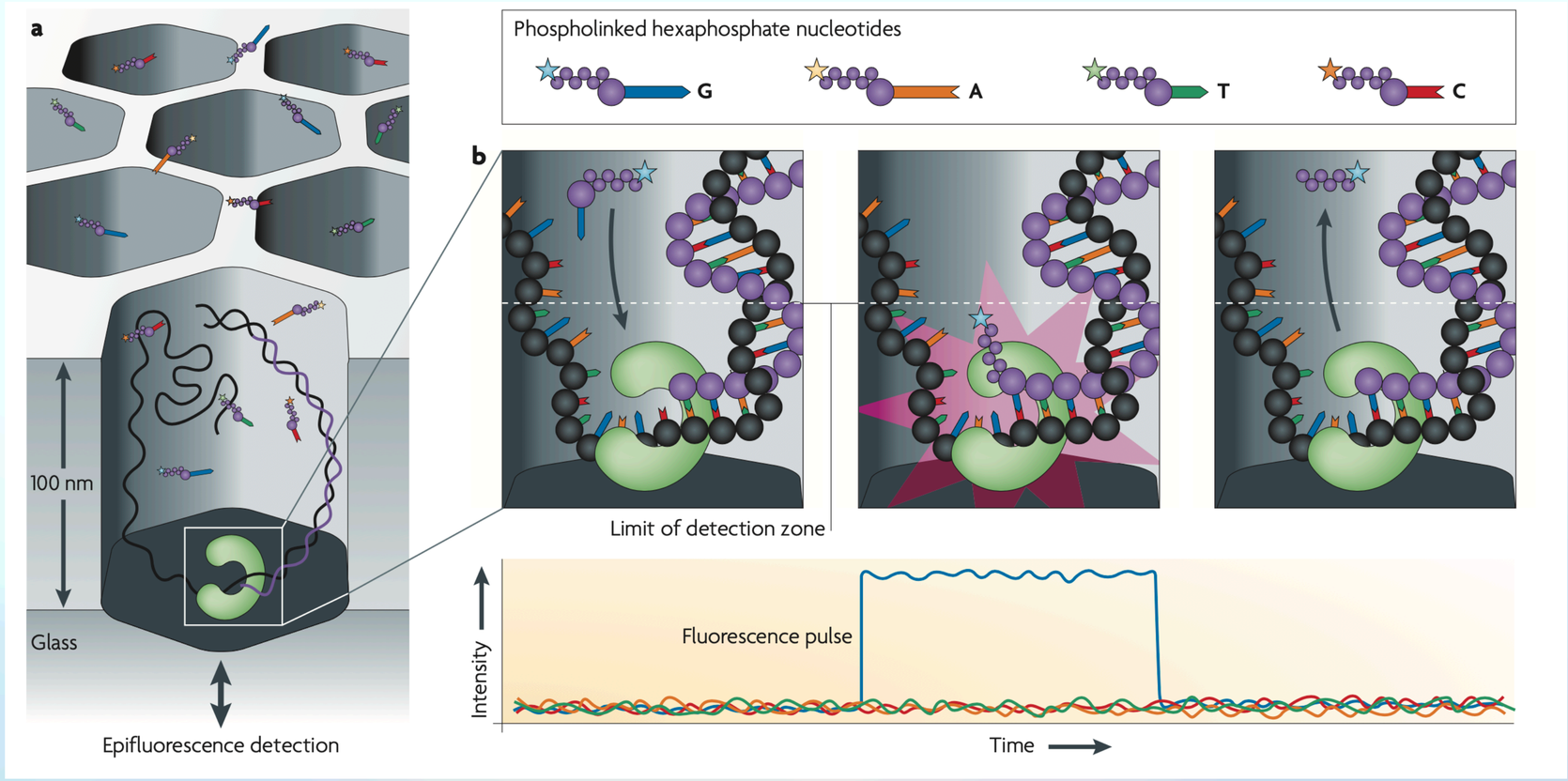


Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

Paired-End & Mate-Paired Sequencing



Pacific Biosciences Sequencing



https://www.youtube.com/watch?v=NHCJ8PtYCFc

Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

Oxford Nanopore Sequencing (ONT)

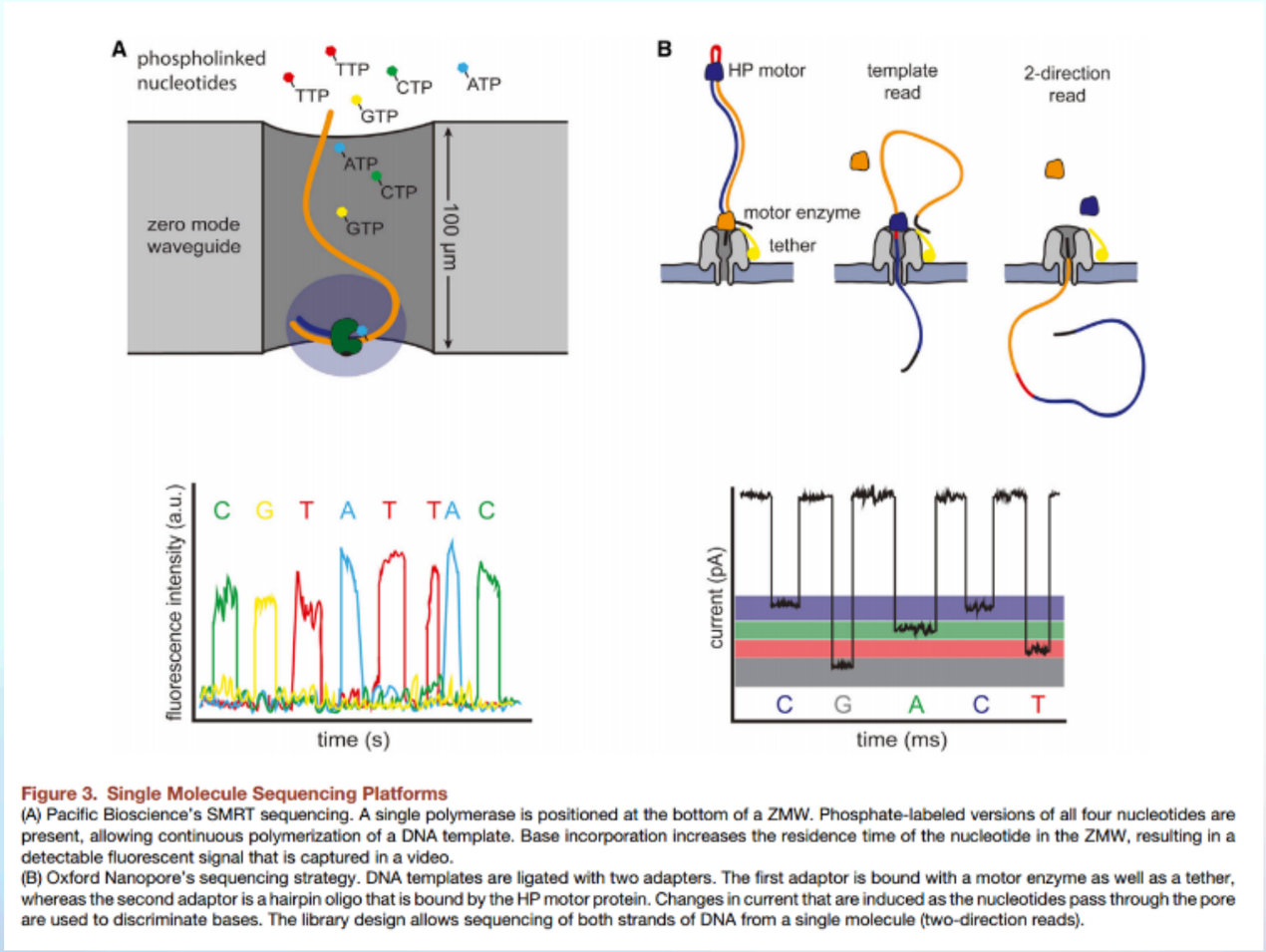


Figure 3. Single Molecule Sequencing Platforms
(A) Pacific Bioscience's SMRT sequencing. A single polymerase is positioned at the bottom of a ZMW. Phosphate-labeled versions of all four nucleotides are present, allowing continuous polymerization of a DNA template. Base incorporation increases the residence time of the nucleotide in the ZMW, resulting in a detectable fluorescent signal that is captured in a video.
(B) Oxford Nanopore's sequencing strategy. DNA templates are ligated with two adapters. The first adaptor is bound with a motor enzyme as well as a tether, whereas the second adaptor is a hairpin oligo that is bound by the HP motor protein. Changes in current that are induced as the nucleotides pass through the pore are used to discriminate bases. The library design allows sequencing of both strands of DNA from a single molecule (two-direction reads).

(MinION)

http://biochemistri.es/of-nanopores-and-isoforms

3rd Generation sequencing - Oxford Nanopore

Summary of Next Generation Sequencing Platforms

Instrument	Amplification	Run time	Millions of Reads/run	Bases / read	Reagent Cost/run	Reagent Cost/Gb	Reagent Cost/Mread	bp/run	Gbp/run	cost/Gb
Applied Biosystems 3730 (capillary)	PCR, cloning	2 hrs.	0.000096	650	$144	$2,307,692.31	$1,500,000.00	62,400	0	$2,307,692.31
454 FLX+	emPCR	20 hrs.	1	650	$6,200	$9,538.46	$6,200.00	650,000,000	0.65	$9,538.46
Illumina GA IIx - v5 PE	bridgePCR	14 days	640	288	$17,978	$97.54	$28.09	184,320,000,000	184.32	$97.54
Illumina MiSeq v3	bridgePCR	55 hrs.	22	600	$1,442	$109.24	$65.55	13,200,000,000	13.2	$109.24
Illumina NextSeq 500	BridgePCR	30 hrs.	400	300	$4,000	$33.33	$10.00	120,000,000,000	120	$33.33
Illumina HiSeq 2500 - high output v4	BridgePCR	6 days	2000	250	$14,950	$29.90	$7.48	500,000,000,000	500	$29.90
Illumina HiSeq X (2 flow cells)	BridgePCR	3 days	6000	300	$12,750	$7.08	$2.13	1,800,000,000,000	1,800.00	$7.08
Ion Torrent – PGM 318 chip	emPCR	7.3 hrs.	4.75	400	$874	$460.00	$184.00	1,900,000,000	1.9	$460.00
Ion Torrent - Proton I	emPCR	4 hrs.	70	175	$1,000	$81.63	$14.29	12,250,000,000	12.25	$81.63
Ion Torrent - Proton III (forecast)	emPCR	6 hrs.	500	175	$1,000	$11.43	$2.00	87,500,000,000	87.5	$11.43
Life Technologies SOLiD – 5500xl	emPCR	8 days	1410	110	$10,503	$67.72	$7.45	155,100,000,000	155.1	$67.72
Pacific Biosciences RS II	None - SMS	2 hrs.	0.03	3000	$100	$1,111.11	$3,333.33	90,000,000	0.09	$1,111.11
Oxford Nanopore MinION (forecast)	None - SMS	≤6 hrs.	0.1	9000	$900	$1,000.00	$9,000.00	900,000,000	0.9	$1,000.00
Oxford Nanopore GridION 2000 (forecast)	None - SMS	varies	4	10000	$1,500	$37.50	$375.00	40,000,000,000	40	$37.50
Oxford Nanopore GridION 8000 (forecast)	None - SMS	varies	10	10000	$1,000	$10.00	$100.00	100,000,000,000	100	$10.00