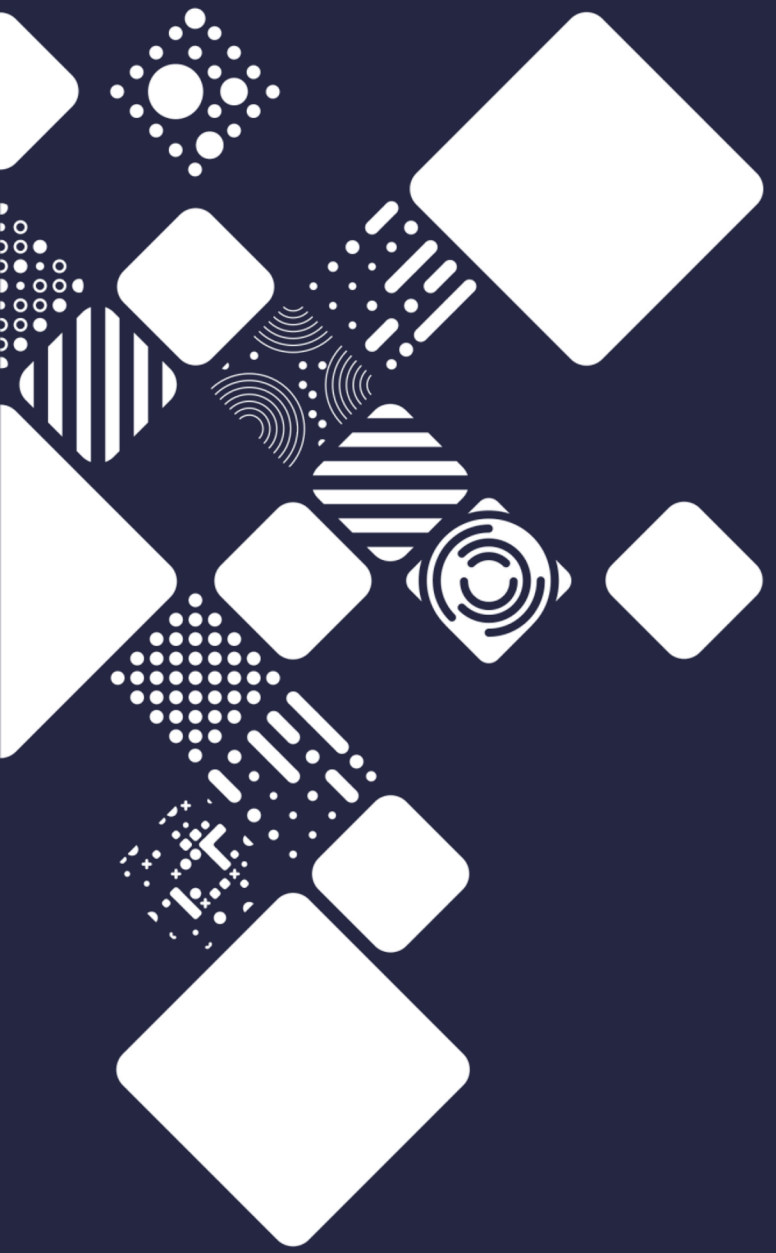


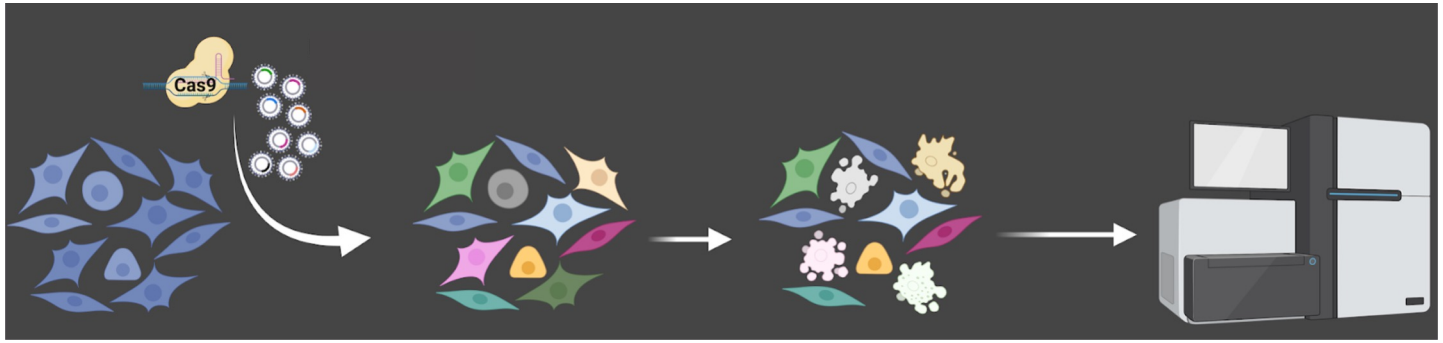
CRISPR screening analysis

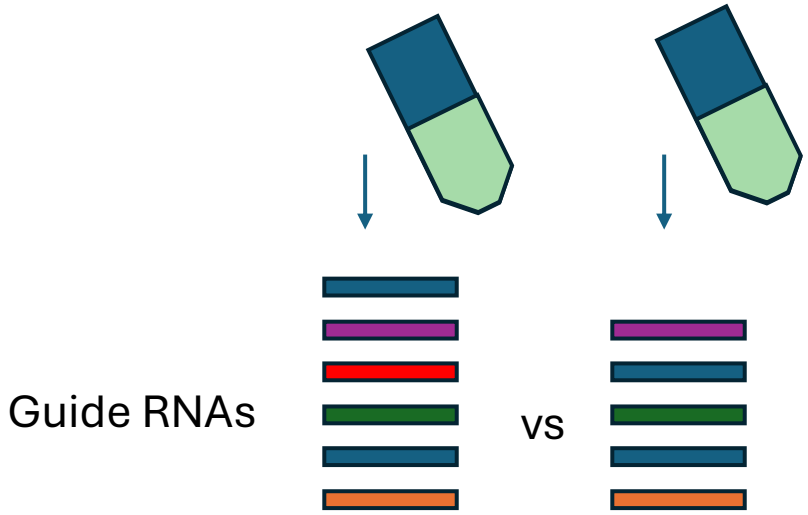
Dr Jamie Billington
Principal Bioinformatician
Adams Faculty
10/10/24

wellcome sanger institute

CRISPR screening

- Deliver guides into a population of cells expressing Cas9 or a dCas9-fusion.
- Conduct some sort of phenotypic screen.
- Sample from the population at different timepoints.
- Isolate guide RNAs and compare their abundance at different timepoints using NGS.

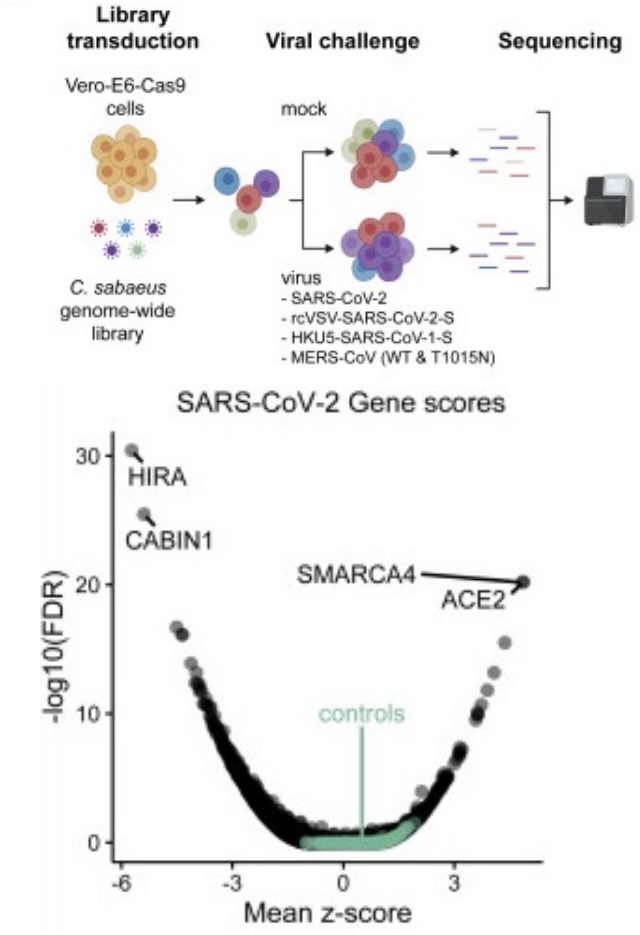




wellcome sanger institute

Application of CRISPR screens

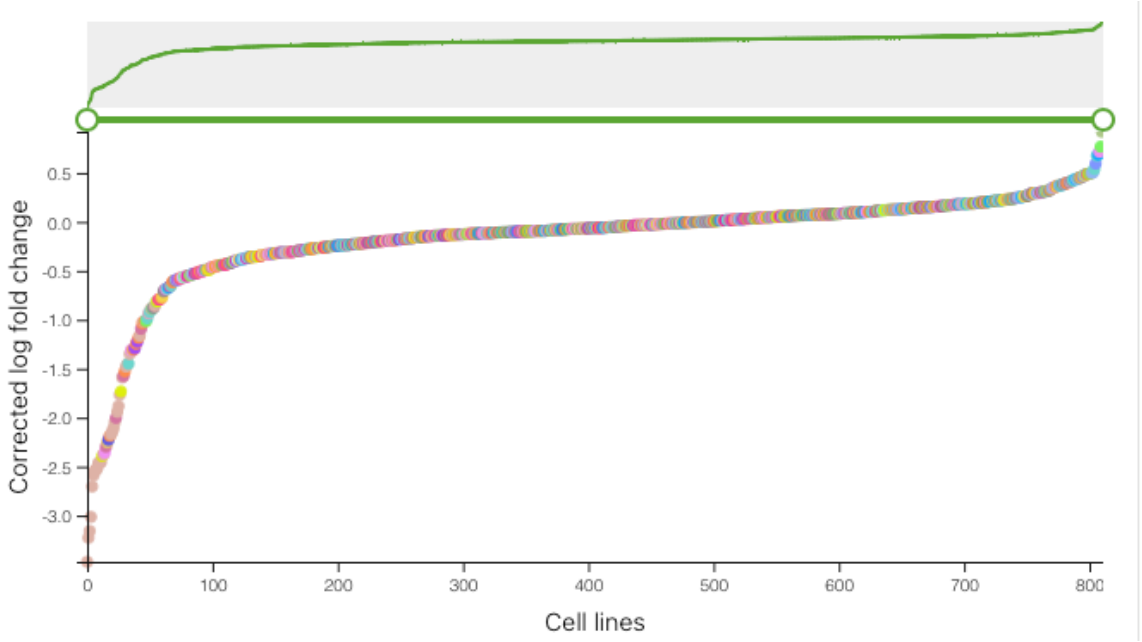
Hypothesis generation -> identify candidate genes



Adapted from Wei, Jin et al. “Genome-wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection.” *Cell* vol. 184,1 (2021): 76-91.e13.

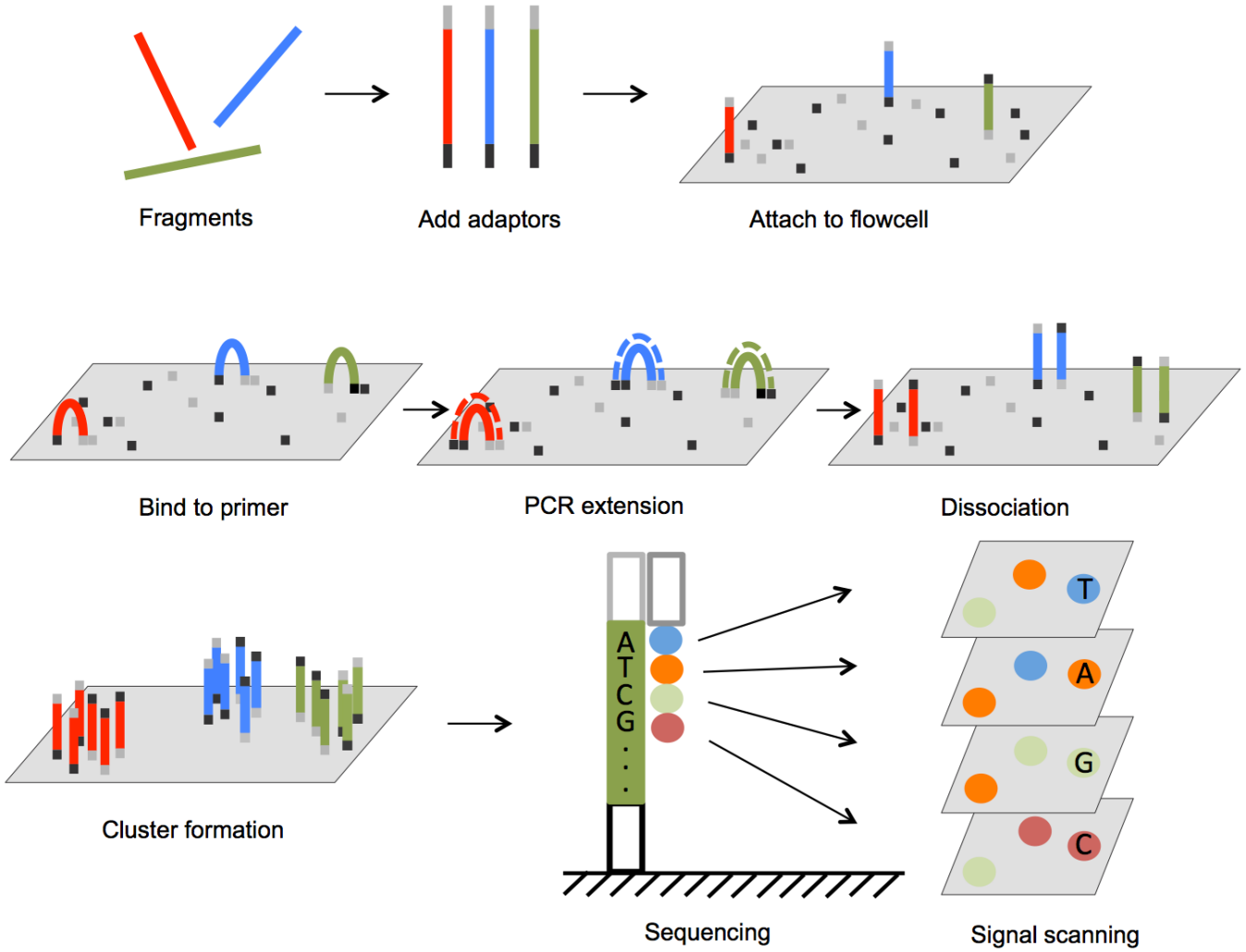
Drug development -> cancer tractability

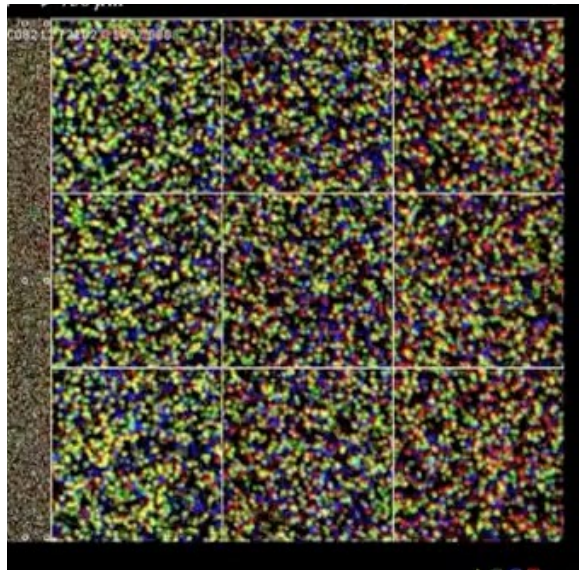
BRAF Log Fold changes in > 800 cancer cell lines - reflecting sensitivity to knockout



Adapted from the Project Score Cancer Dependency Map

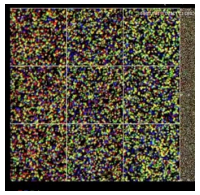
High throughput sequencing





Next Generation Sequencing-Advances, Applications and Challenges, Lu et al 2015

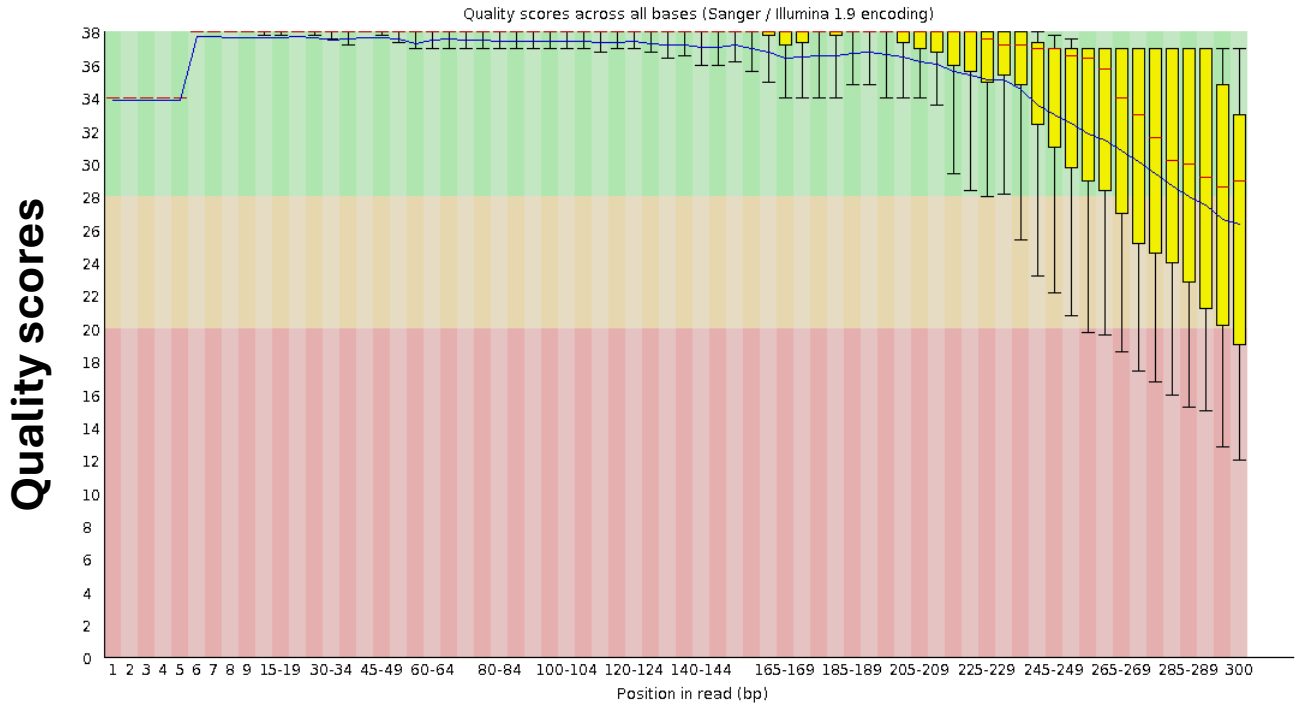
Anatomy of an NGS read



Read ID
Sequence

Quality scores

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEE
```



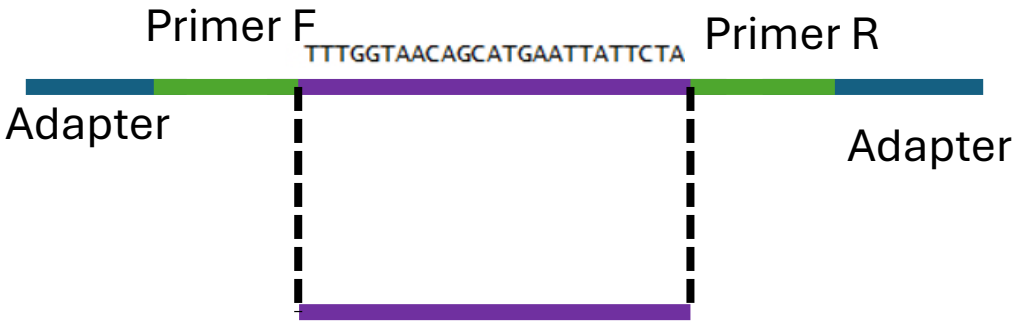
Typically, millions of reads per sample

Each base has a quality score
Typically, a decline in base quality moving along the read from 5’ -> 3’

Position along read

From reads to counts

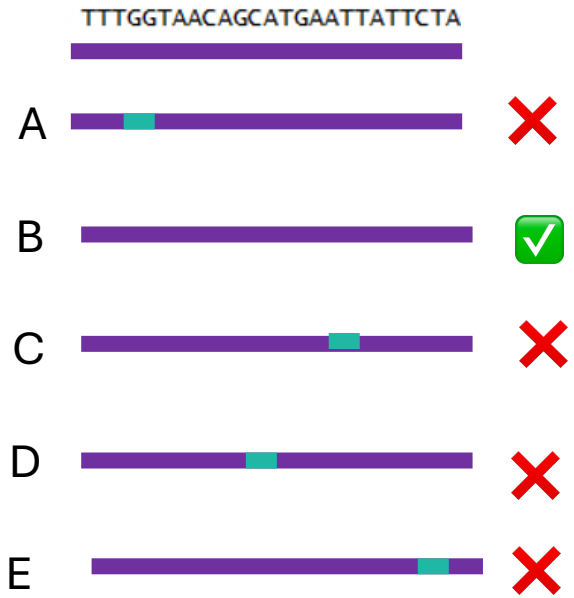
Read processing



Trim adapter sequences from reads to isolate region of interest

Mapping/alignment

Find matches between read and expected features



Tallying

Repeat for all reads to get counts per "feature"

Variant	Counts
A	100
B	120
C	145
D	30
E	150

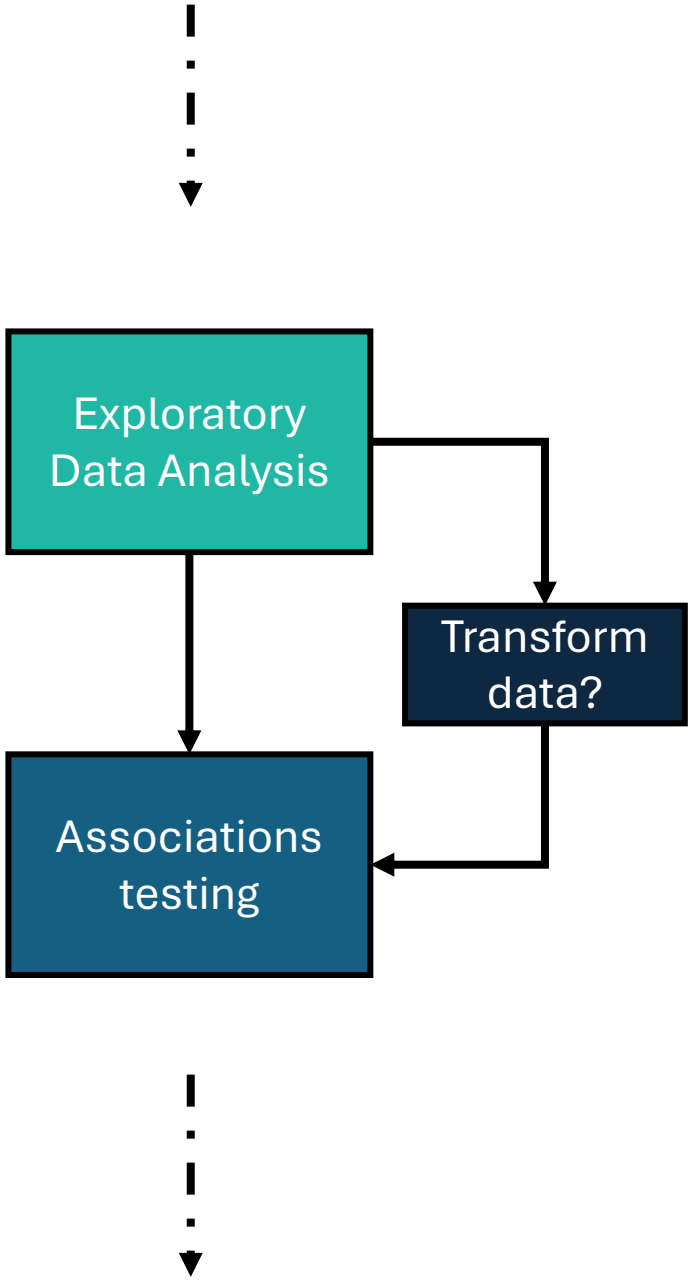
Analysing high throughput counts

	Replicate 1			Replicate 2			Replicate 3		
	Day 4	Day 7	Day 15	Day 4	Day 7	Day 15	Day 4	Day 7	Day 15
Guide A	100	90	60	102	88	29	102	88	29
Guide B	120	111	112	114	150	100	114	150	100
Guide C	145	40	100	145	42	42	145	42	42
…	…	…	…	…	…	…	…	…	…

Start from a **counts matrix:** a mathematical object with:

Rows as **features** (genes, guides, proteins, species of bacteria …)
Columns as **samples** (replicates, experimental conditions)
High dimensional (1000's -> 100,000's of features)

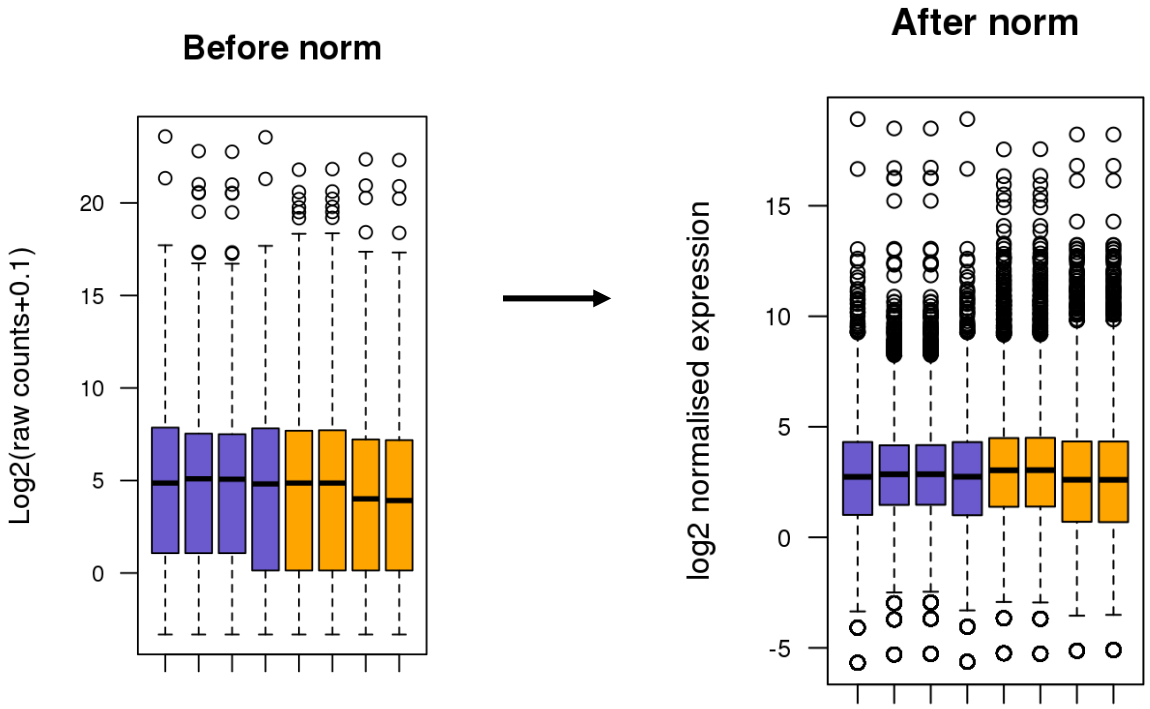
General set of numerical methods for the analysis of different kinds of datasets



Dataset transformation

Normalise to account for technical variability in:

- **Sequencing depth between samples** (1 million in Sample A vs. 2 million reads in Sample B)
- **Sample composition** (highly abundant feature can distort measurements of other features when they change)
- **Feature characteristics** (e.g., gene length in RNA seq)



Raw counts

	Replicate 1			Replicate 2			Replicate 3		
	Day 4	Day 7	Day 15	Day 4	Day 7	Day 15	Day 4	Day 7	Day 15
Variant A	100	90	60	102	88	29	102	88	29
Variant B	120	111	112	114	150	100	114	150	100
Variant C	145	40	100	145	42	42	145	42	42
...	...	...	...	...	...	...	...	...	...

Normalised counts

	Replicate 1			Replicate 2			Replicate 3		
	Day 4	Day 7	Day 15	Day 4	Day 7	Day 15	Day 4	Day 7	Day 15
Variant A	100	90	60	102	88	29	102	88	29
Variant B	120	111	112	114	150	100	114	150	100
Variant C	145	40	100	145	42	42	145	42	42
...	...	...	...	...	...	...	...	...	...

Associations testing

Features whose counts are “associated” with a condition

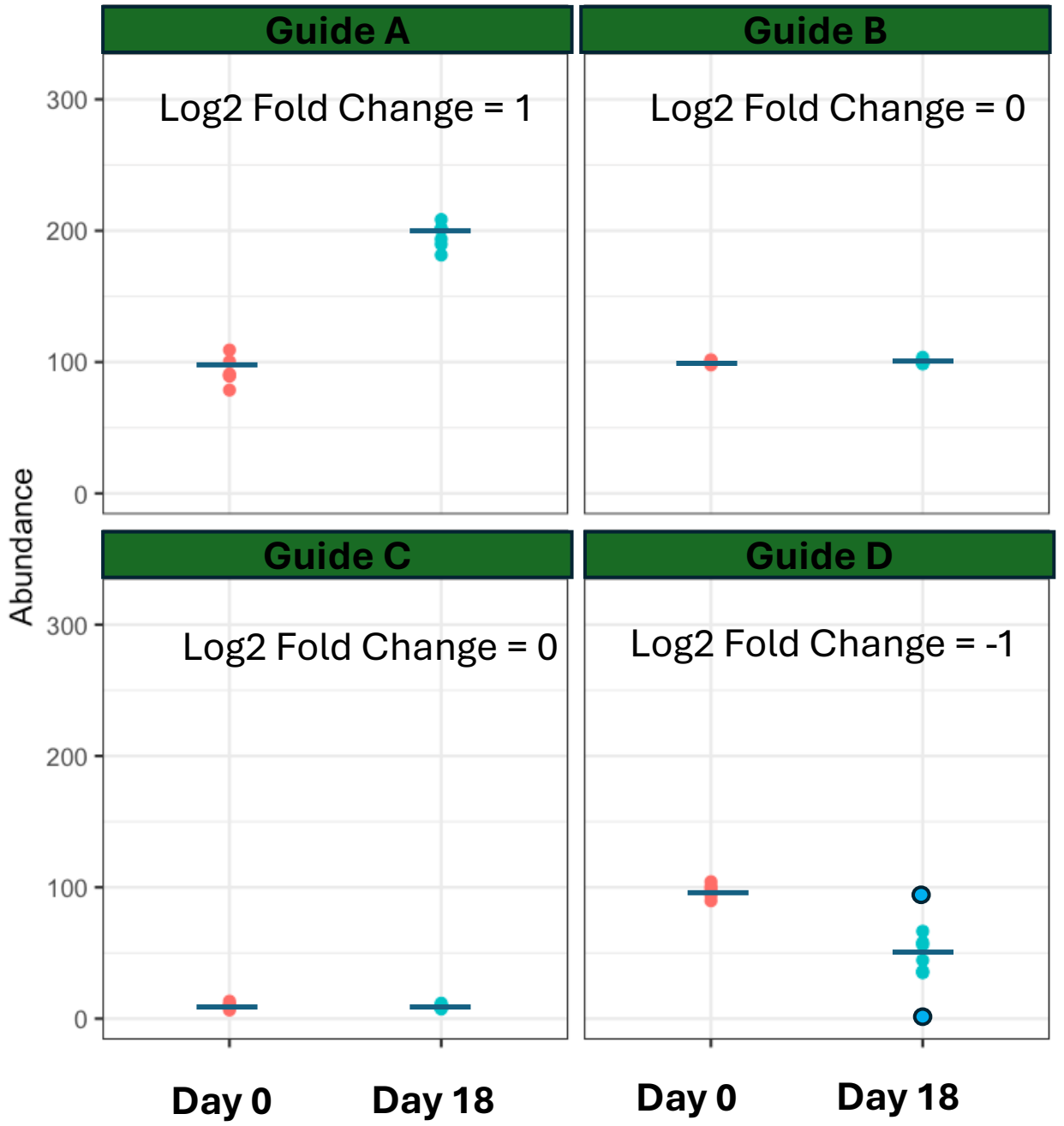
Normalised counts

	Replicate 1			Replicate 2			Replicate 3		
	Day 4	Day 7	Day 15	Day 4	Day 7	Day 15	Day 4	Day 7	Day 15
Variant A	100	90	60	102	88	29	102	88	29
Variant B	120	111	112	114	150	100	114	150	100
Variant C	145	40	100	145	42	42	145	42	42
...	...	...	...	...	...	...	...	...	...

Guides *associated* with Day 18 timepoint
Proteins *associated* with drug response

Going to perform statistical tests to compare between timepoints of Day 0 vs Day 18

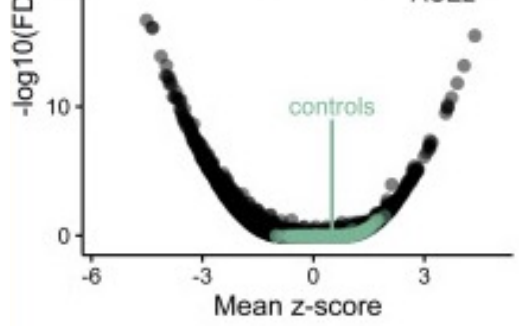
With appropriate modifications for
i) counts data ii) large numbers of comparisons

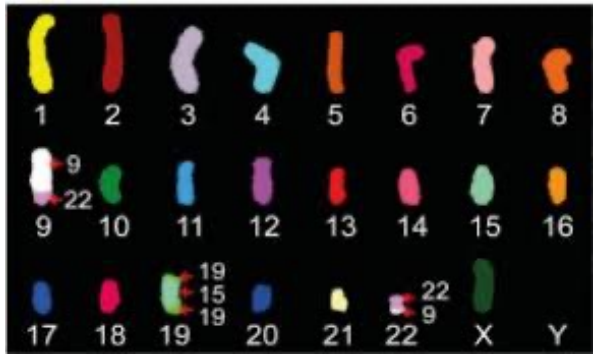


This practical

Understanding the process of analyzing CRISPR screening data from counts matrix -> gene hits

	Day 0	Day 7	Day 15	Day 0	Day 7	Day 15	Day 0	Day 7	Day 15
Guide A	100	90	60	102	88	29	102	88	29
Guide B	120	111	112	114	150	100	114	150	100
Guide C	145	40	100	145	42	42	145	42	42
…	…	…	…	…	…	…	…	…	…





Finding essential genes in the HAP1 cell line

wellcome sanger institute