# Assessing gene essentiality using pooled CRISPR screens

Katharina Imkeller, DKFZ and EMBL, Heidelberg, Germany

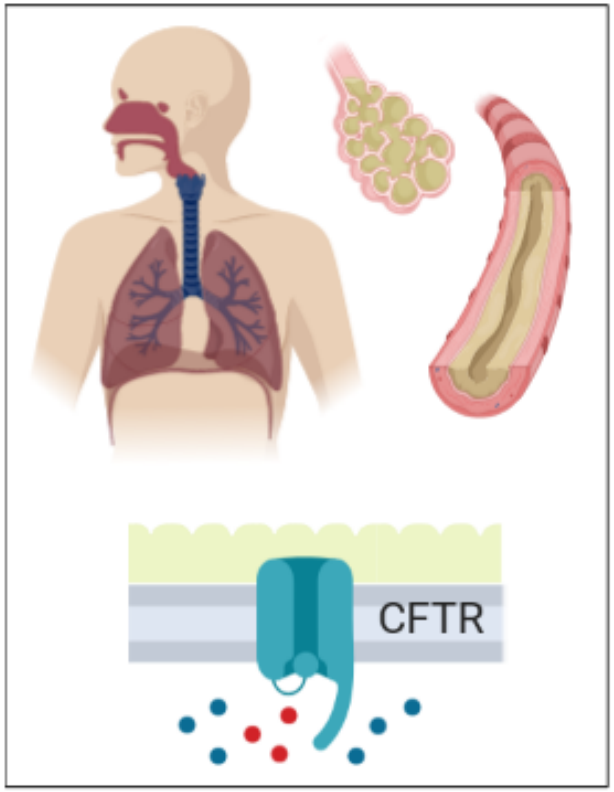
June 25th 2019

CSAMA 2019, 17th edition


@K_Imkeller @Boutroslab

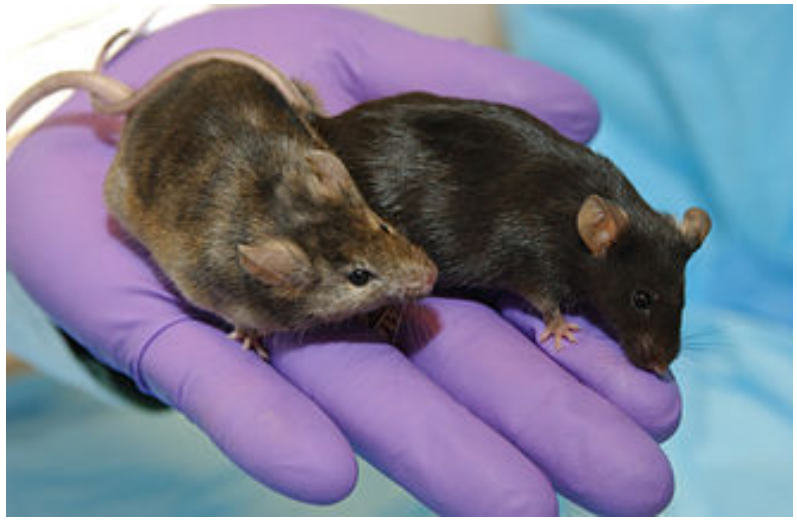
# Reverse vs. forward genetics

**Forward genetics:**
Find the genetic basis for a specific observed phenotype.




*Discovery of CFTR gene mutation causing Cystic fibrosis.*

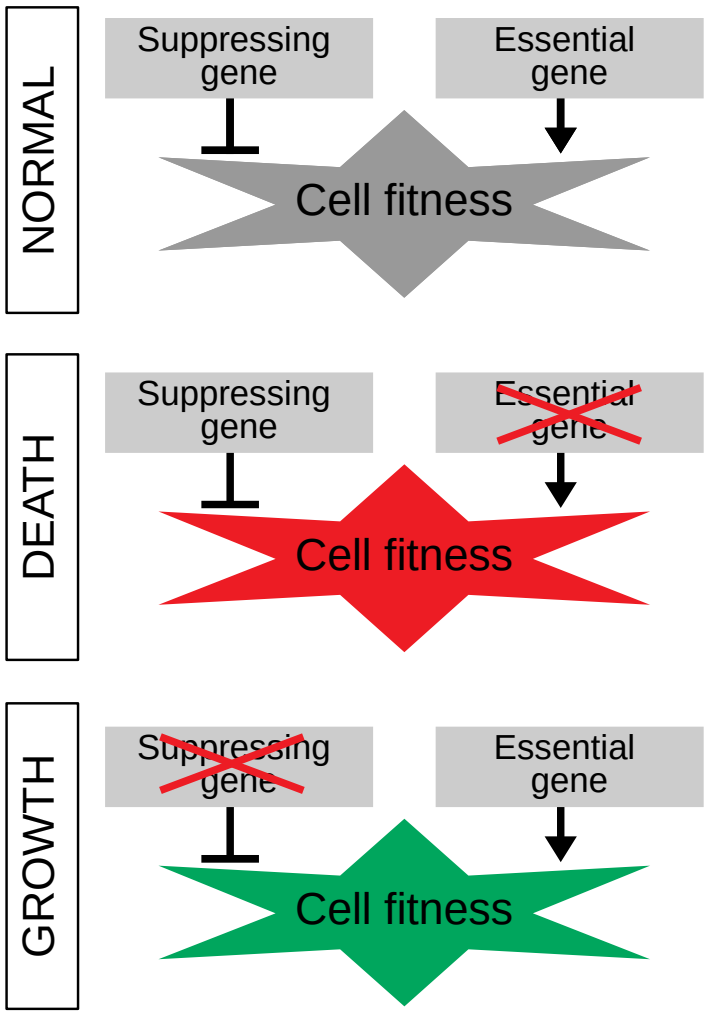
**Reverse genetics:**
Modify gene sequence and analyzes the resulting phenotype.



Wikipedia

*Knockout of gene affecting hair growth.*

# Biological motivation for reverse genetics screens




**Core essential genes:**

- *RPL13* - ribosomal component
- *POLR1B* - RNA polymerase

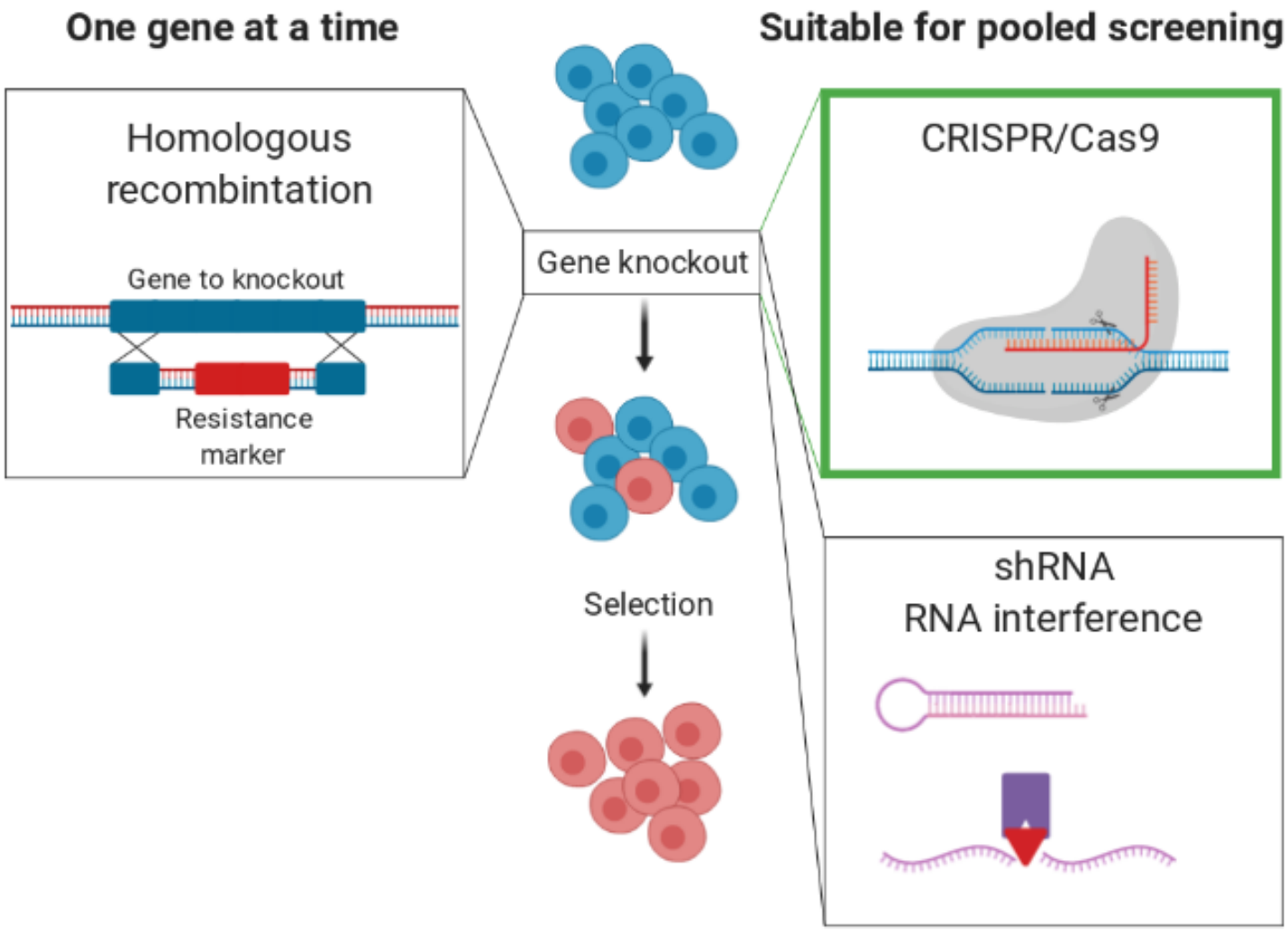
**Growthsupressing genes:**

- "tumor supressor"
- *TP53*
- *BRCA1*

**Synthetic lethality to target tumor cells:**

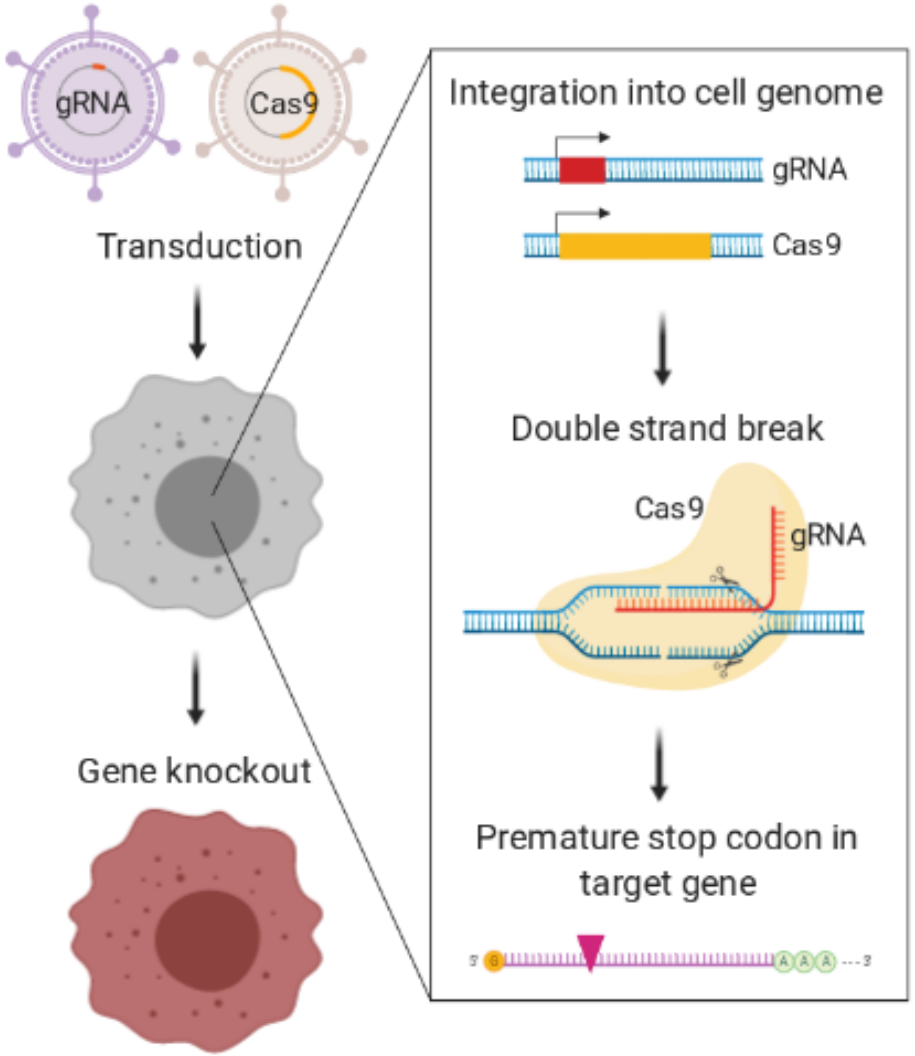
- *PARP* in *BRCA1* mutated tumors
- *BRAF* in *KRAS* mutated tumors

# Advantages of using CRISPR-Cas9 for gene knockout




* shRNA based screens have problems with off-target effects and weak phenotypes.

# Guide RNA (gRNA) simultaneously serves as perturbagen and barcode

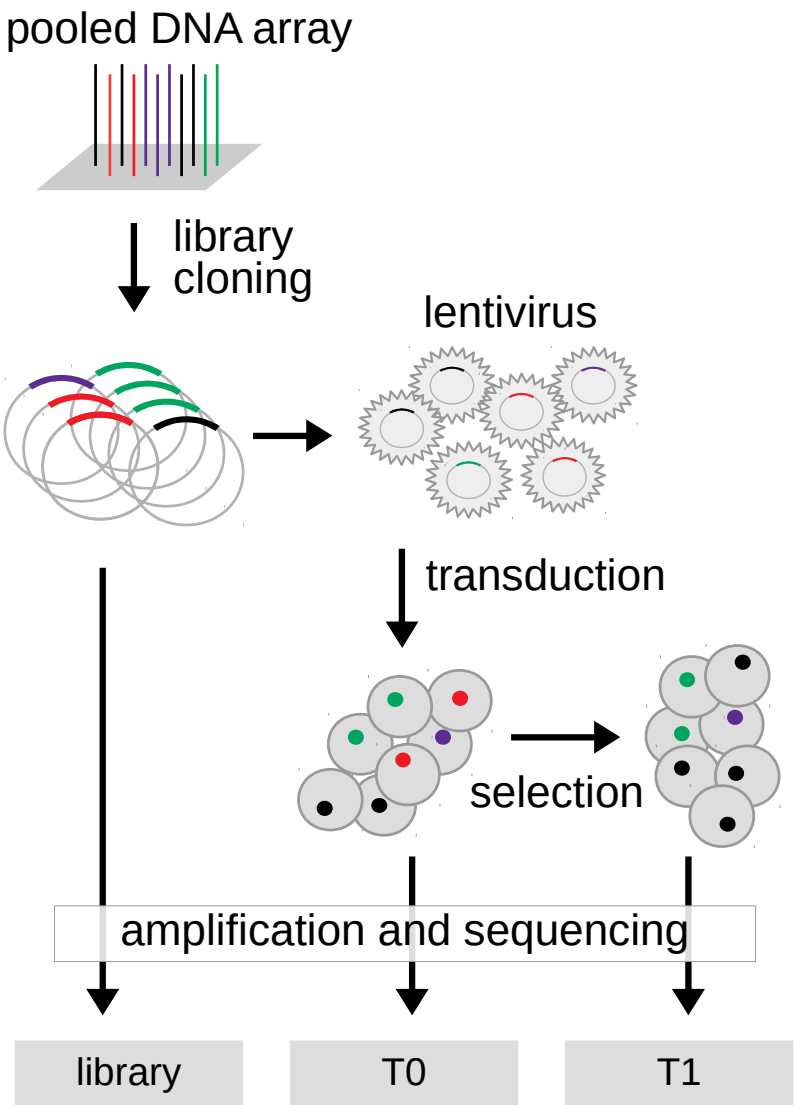



- gRNA can be PCR amplified from genome
- serves as a proxy for gene knockout

# Different types of CRISPR mediated genetic perturbations

| **Name** | **CRISPR associated enzyme** | **perturbation** |
|---|---|---|
| CRISPR-KO | Cas9 | gene knockout |
| CRISPRi | dCas9 + transcription inactivator | expression inhibition |
| CRISPRa | dCas9 + transcription activator | expression activation |
| CRISPR-BE | dCas9 + base editor | base editing (C-G, A-T) |

# Experimental procedure of pooled CRISPR screens

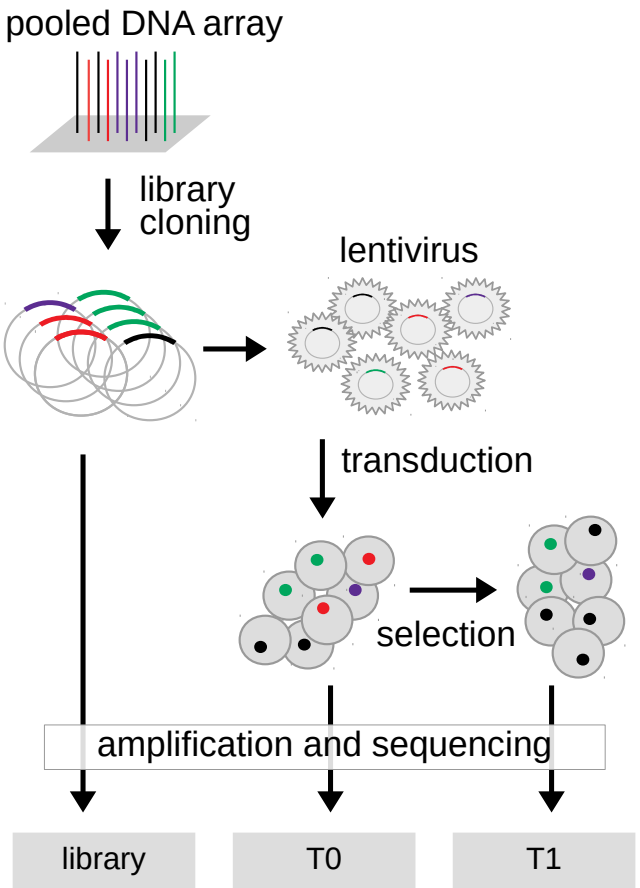


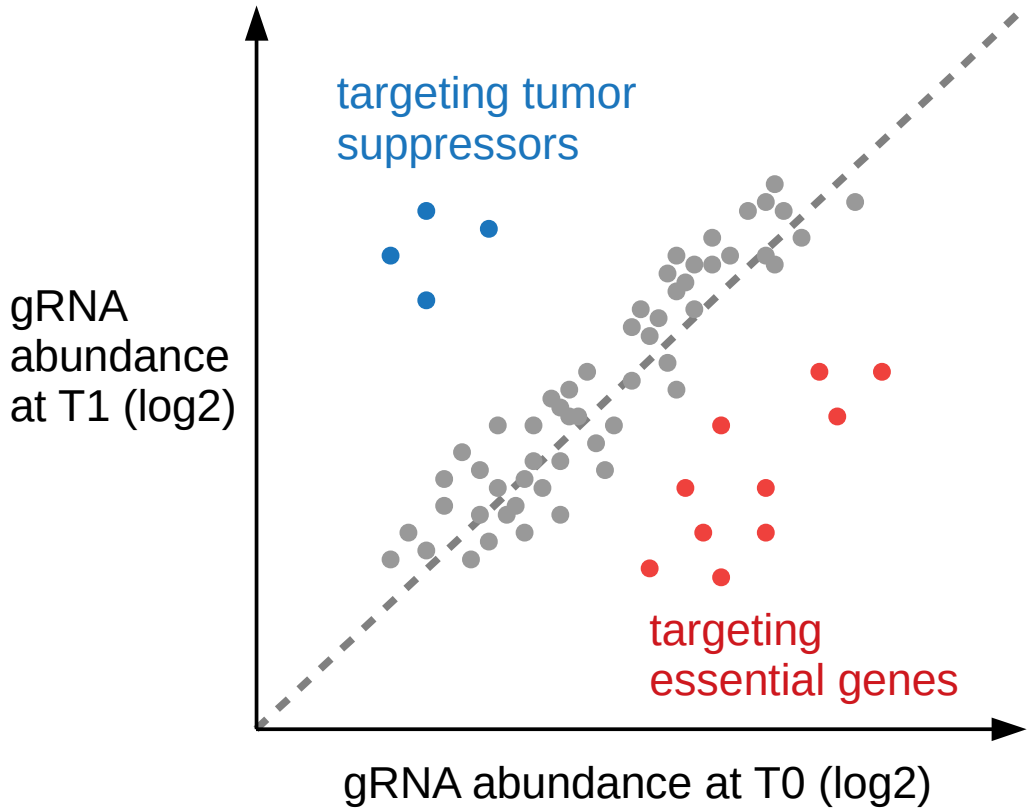

**Experimental design principle**

- **guide RNA/ gRNA**: perturbagen and barcode
- library size around 100K gRNAs
- perturbed cells growing in a pool
- individual growth rate depends on gene knockout
- compare abundance T0 vs. T1

*For protocol see e.g. Joung et al. 2017*

# Phenotype detection in pooled CRISPR screens



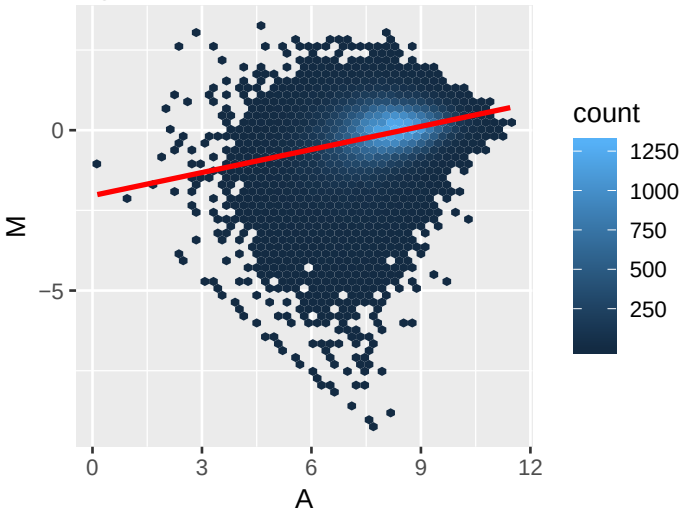
# Differences between RNA-seq and CRISPR screening data

**M-A plot**

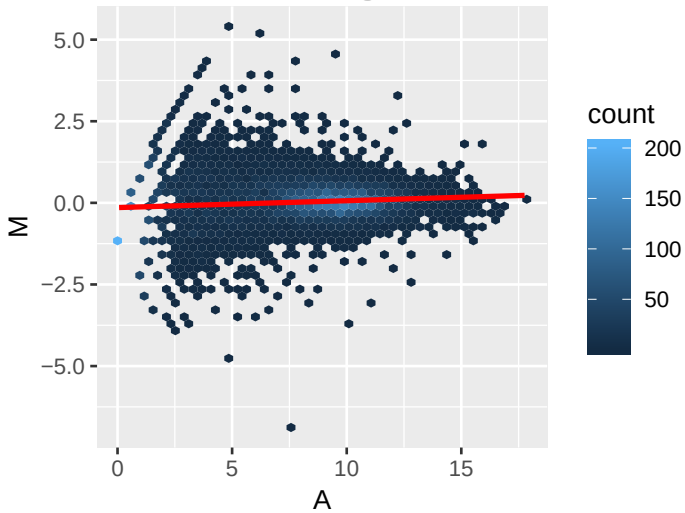
Logarithmic fold change: $M = log_2(\frac{S_1}{S_2})$

Mean abundance: $A = \frac{1}{2} log_2(S_1 S_2)$

**CRISPR screen**



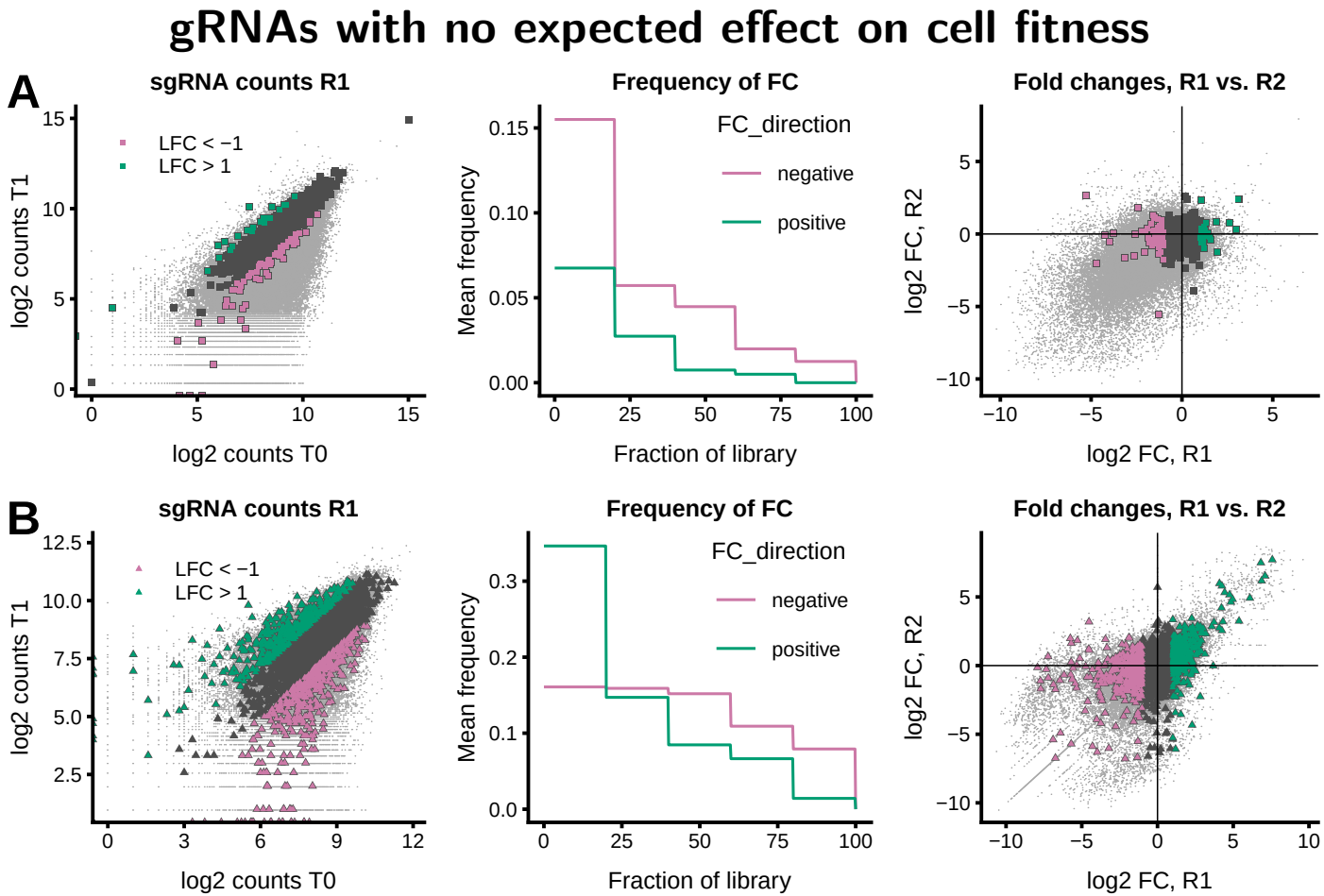

**RNA sequencing**



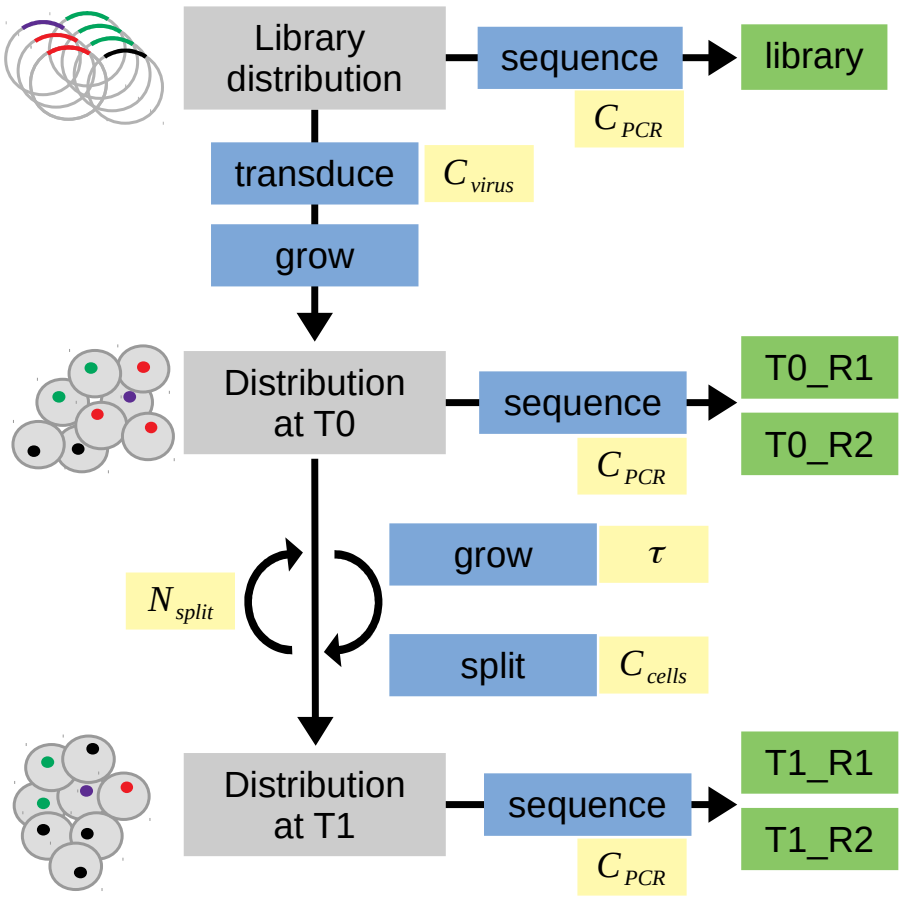
# Screening data is skewed towards negative fold changes

## ASYMMETRY: T0 vs. T1 gRNA abundance

- negative logFC are more frequent
- especially for gRNAs that have low initial frequency

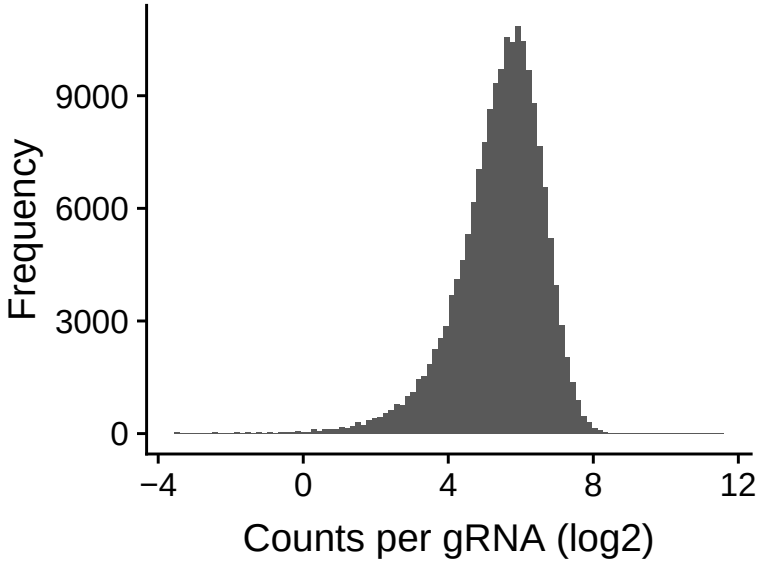


# Computational simulation of screen to test influence of experiment design

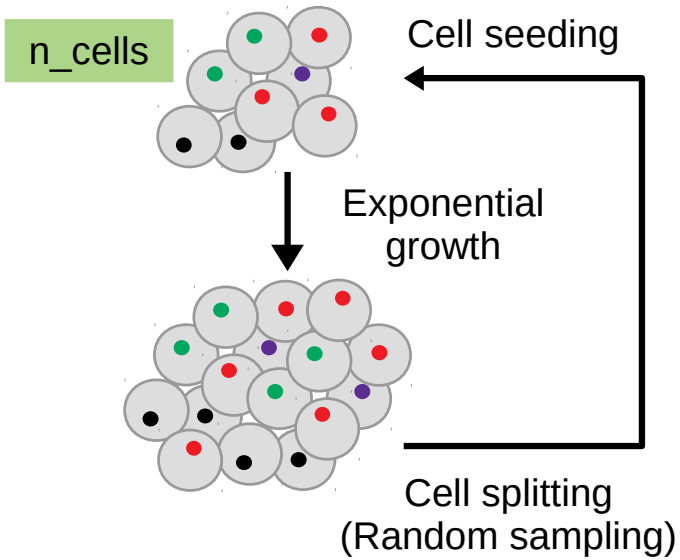



- gRNA counts modelled as a tuple of integer numbers
- result: counts after sequencing
- functions modify the counts (multiplication, random sampling)
- number of cell splittings $N_{split}$, cell duplication time $\tau$
- "coverages" $C_{PCR}$, $C_{virus}$, $C_{cells}$

# Mean gRNA coverage in pooled CRISPR screens determines cell number




Experiments assume a narrow distribution and are designed based on mean gRNA coverage.



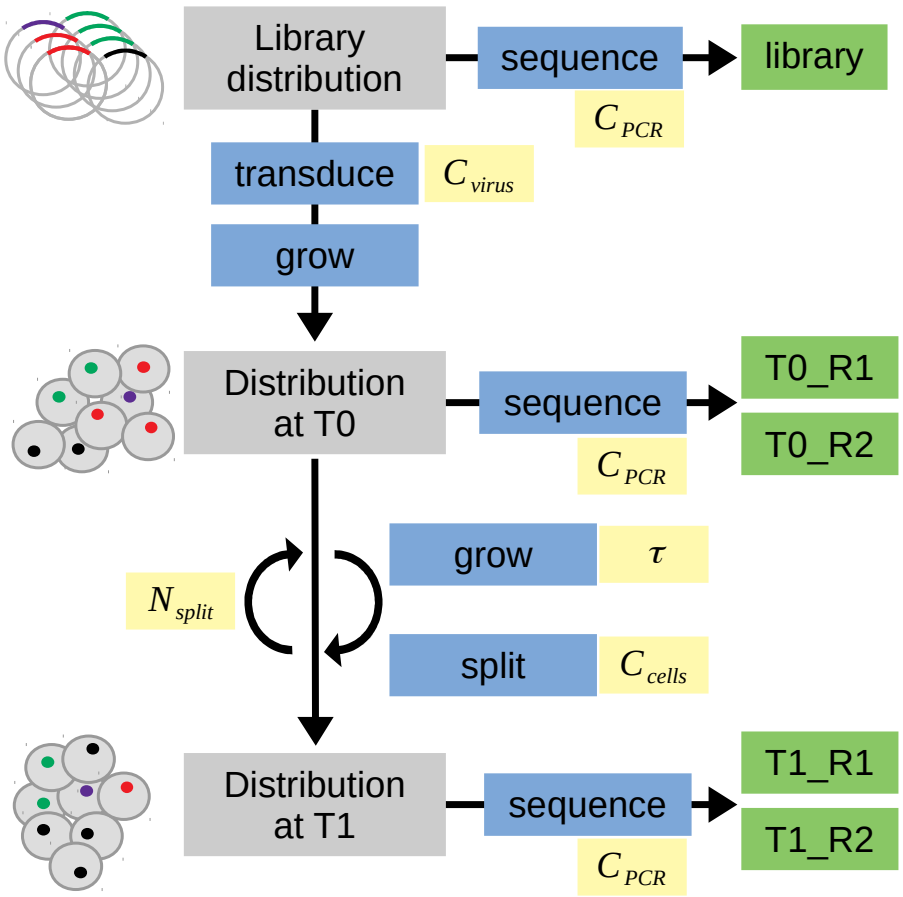

## Mean coverage of gRNAs:

how many times is one gRNA on average represented in a pooled experiment?

$$\text{coverage} = \frac{n_{cells}}{n_{gRNAs}}$$

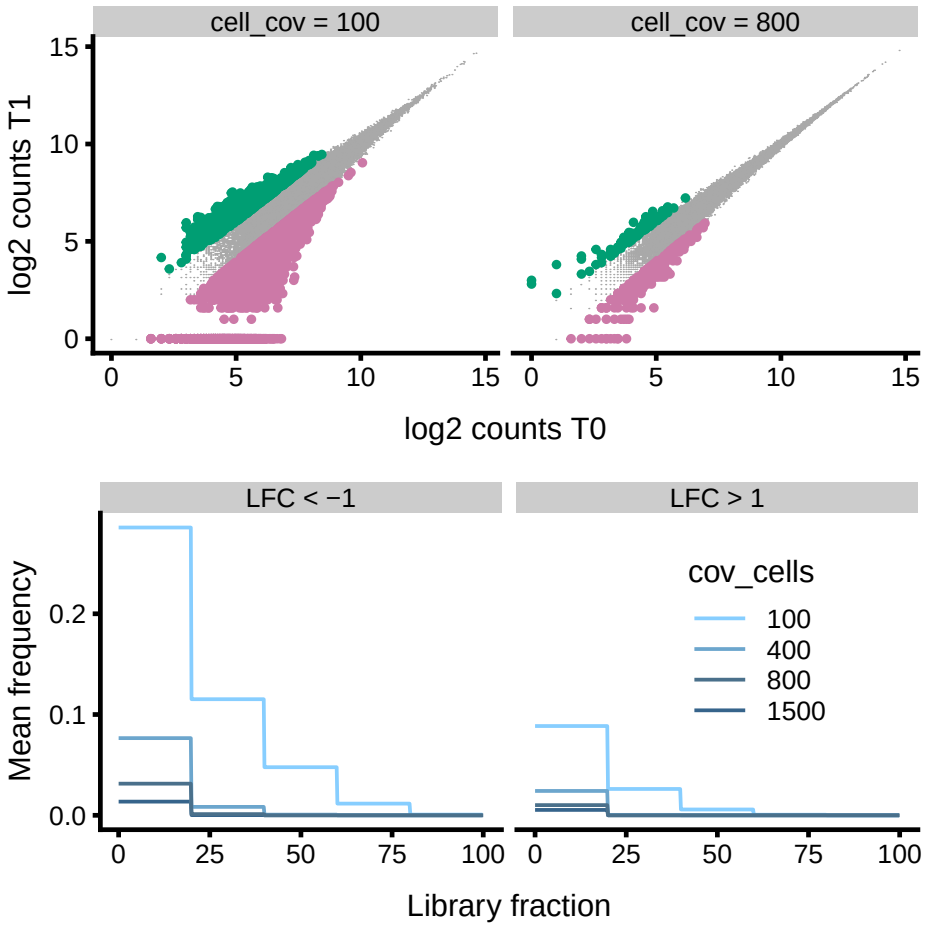
For example (coverage 500):
$10^7 \text{gRNAs} \times 500 = 5 \times 10^9 \text{cells}$

# Computational simulation of screen to test influence of experiment design



# Cell splitting causes asymmetry in gRNA abundance changes

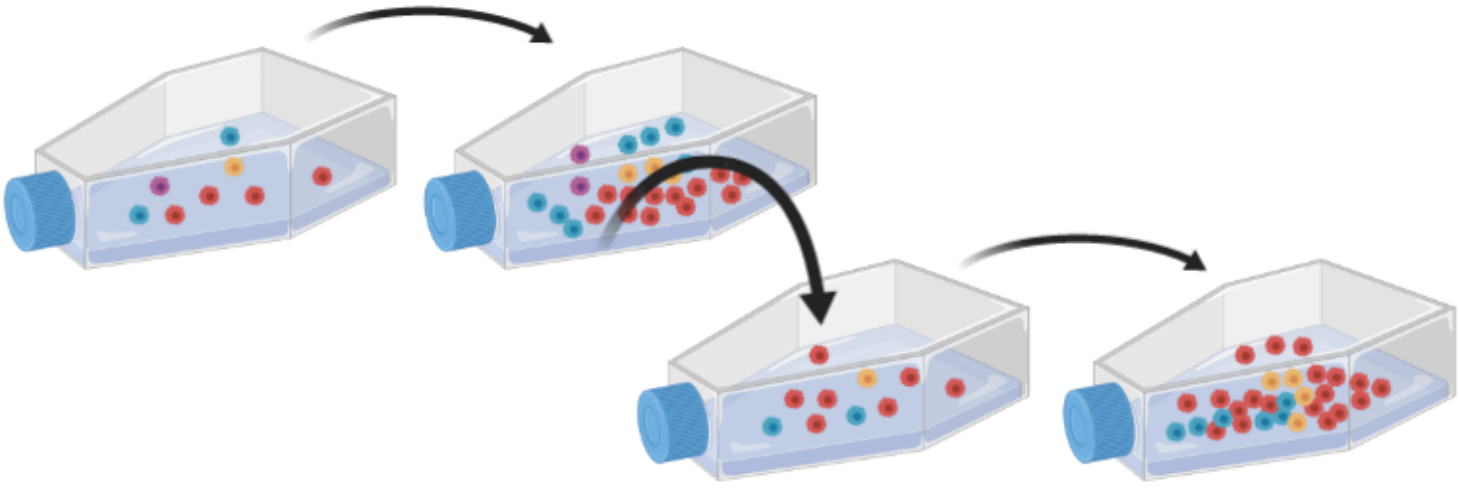
## Simulation with different coverage




**Lower gRNA coverage increases asymmetry of gRNA abundance changes.**

# Asymmetry is caused by repetitive cell splitting

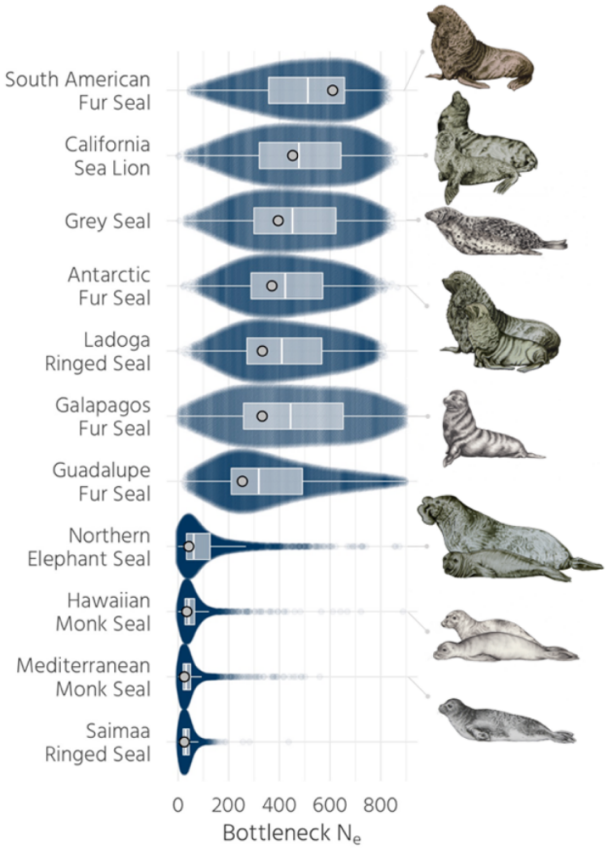
## Bottle neck effect



| | | | |
|---|---|---|---|
| purple: 1 | purple: 2 | purple: 0 | purple: 0 |
| red: 4 | red: 12 | red: 7 | red: 16 |
| blue: 2 | blue: 7 | blue: 2 | blue: 6 |
| yellow: 1 | yellow: 3 | yellow: 1 | yellow: 4 |

# Population bottlenecks in the Northern Elephant Seal

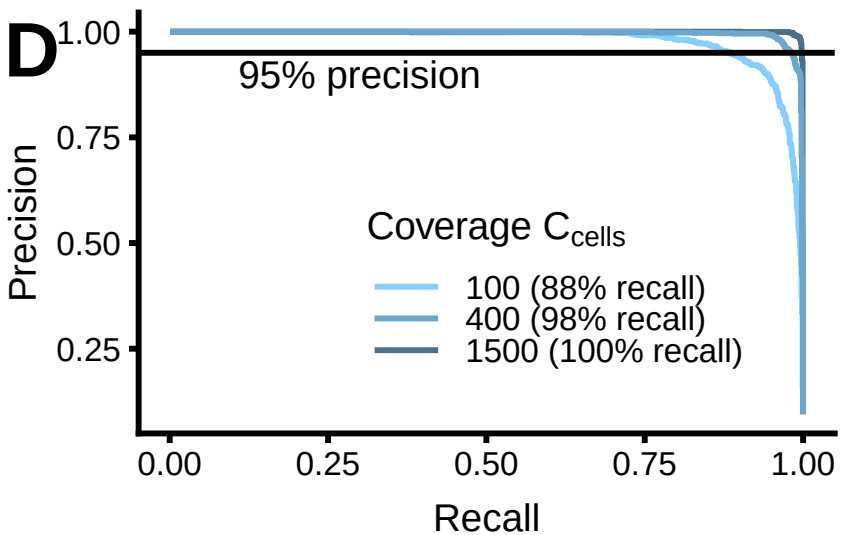
**Bottle neck event**: Hunting in 19th century, reduction of population size to 20 individuals. Today's 30,000 seals have a strongly reduced genetic diversity.

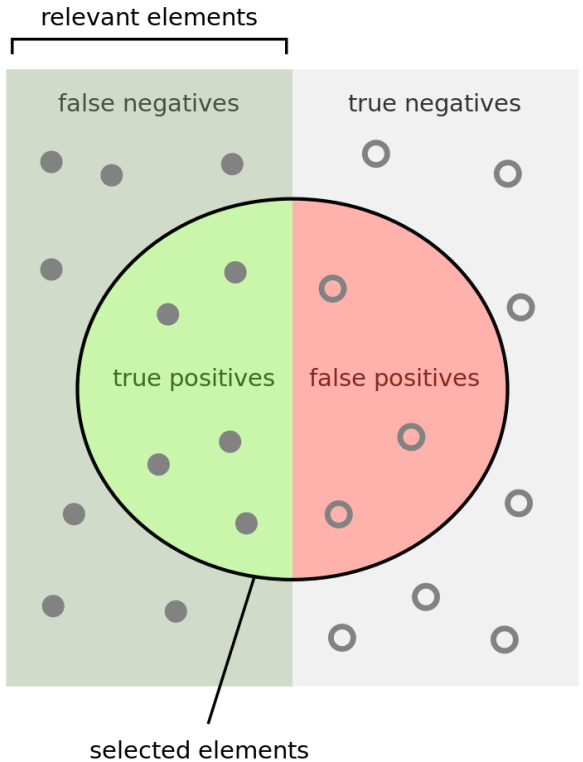



# It is not OK to assume symmetry of null-distribution!

**Current analysis tools loose detection power when asymmetry increases.**

**Detection of essential genes by MAGeCK-RRA (Li et al. 2014)**



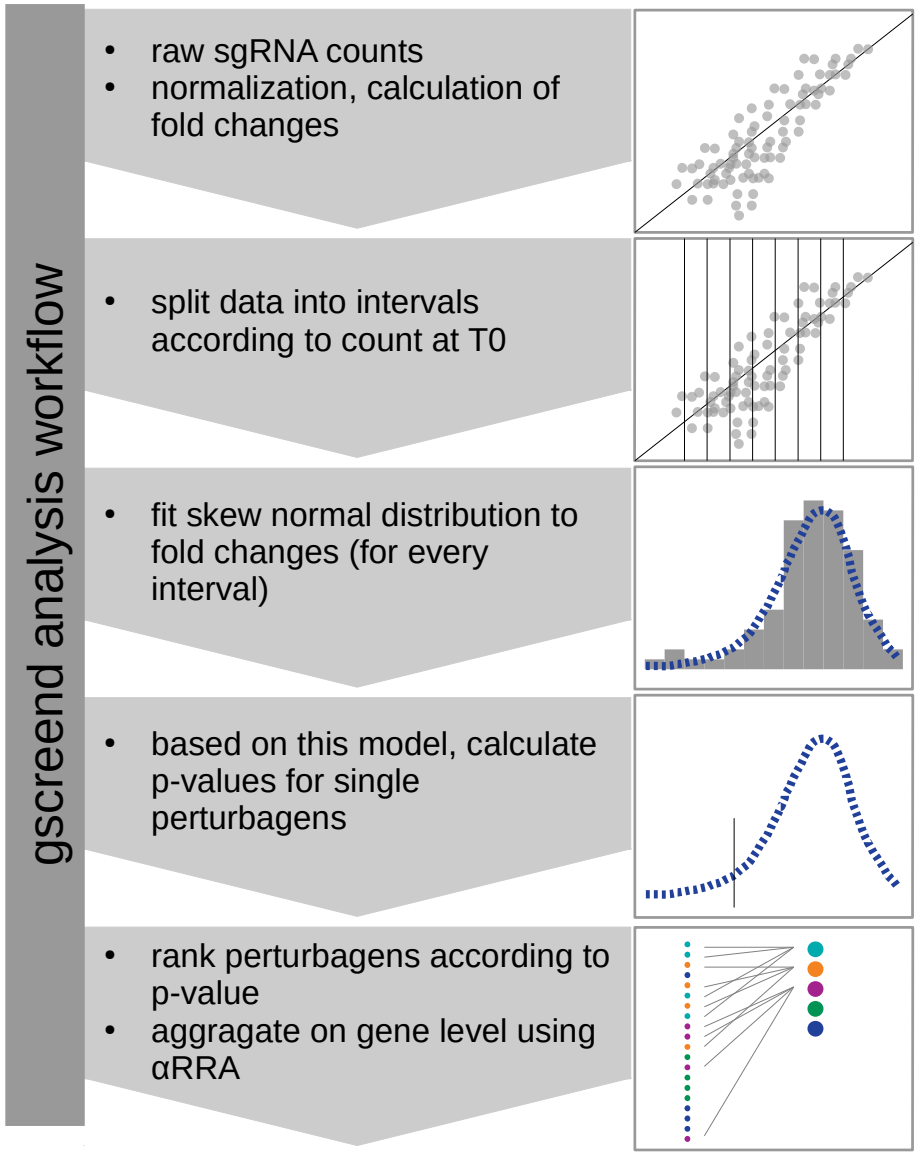

How many selected items are relevant?

Precision =

How many relevant items are selected?

Recall =

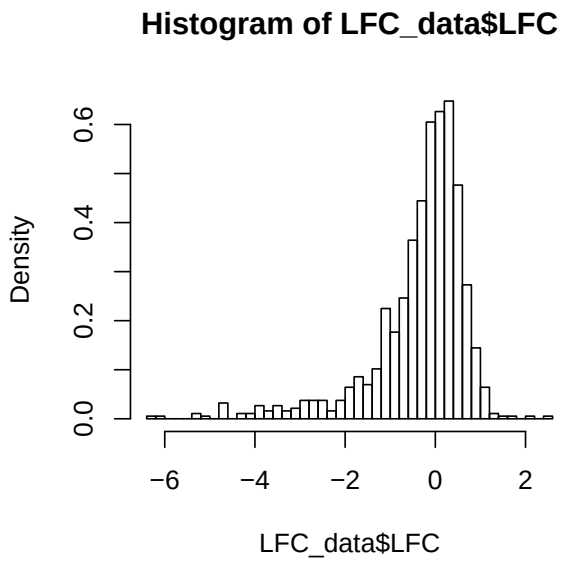
# Software package gscreend with improved statistical test



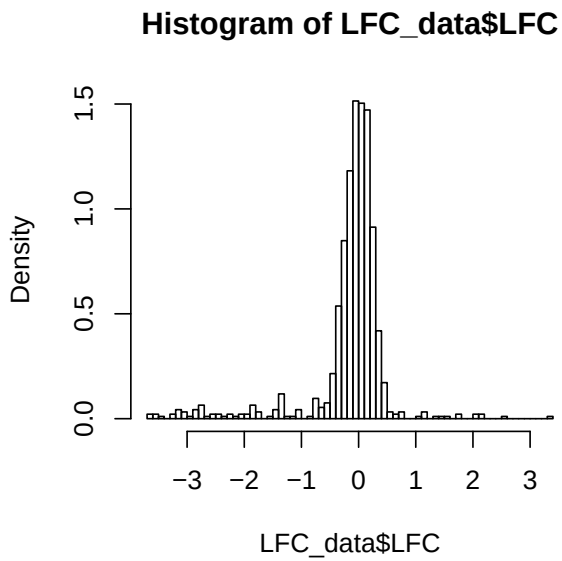
# Step 1: Data preparation

- Normalization or scaling to the total counts in the reference sample.
- Calculation of logarithmic fold changes, addition of pseudo-counts:
  $LFC = log_2(\frac{n_{T_1} + 1}{n_{T_0} + 1})$
- Partitioning into groups according to abundance in reference sample.

Low abundant gRNAs
(20-30% percentile)

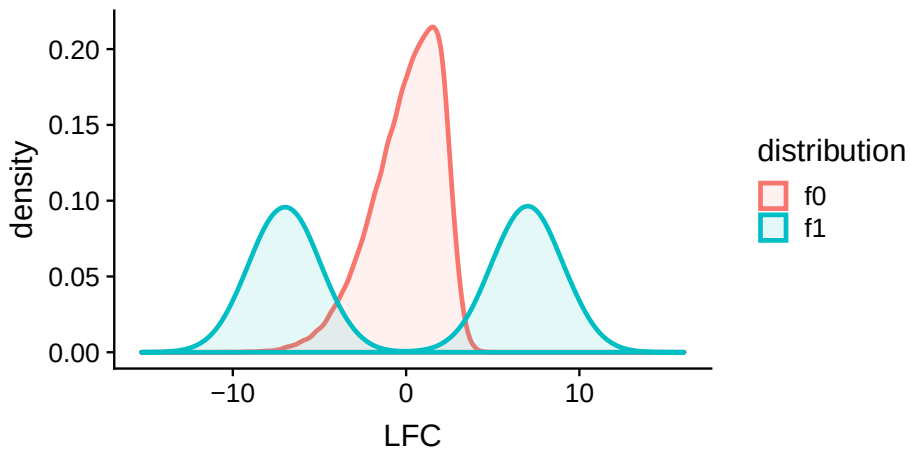



High abundant gRNAs
(80-90% percentile)



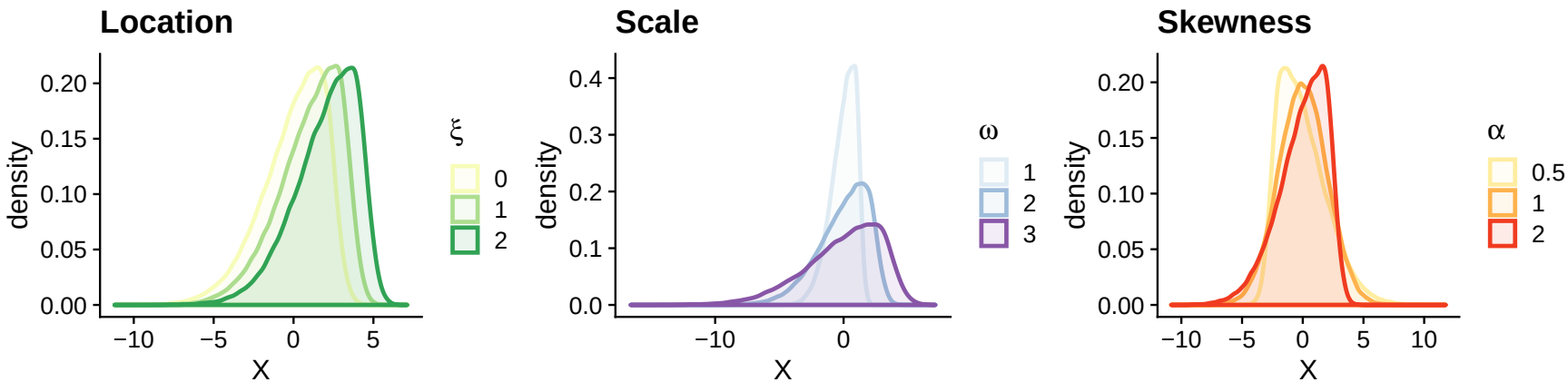
# Step 2: Statistical modelling of gRNA level data

- Modeling of the data as a mixture of null-distribution $f_0$ and unknown distribution $f_1$ of the gRNAs with fitness effect:
  $f(x) = (1 - \lambda) f_0(x) + \lambda f_1(x)$



# Step 2: Statistical modelling of gRNA level data

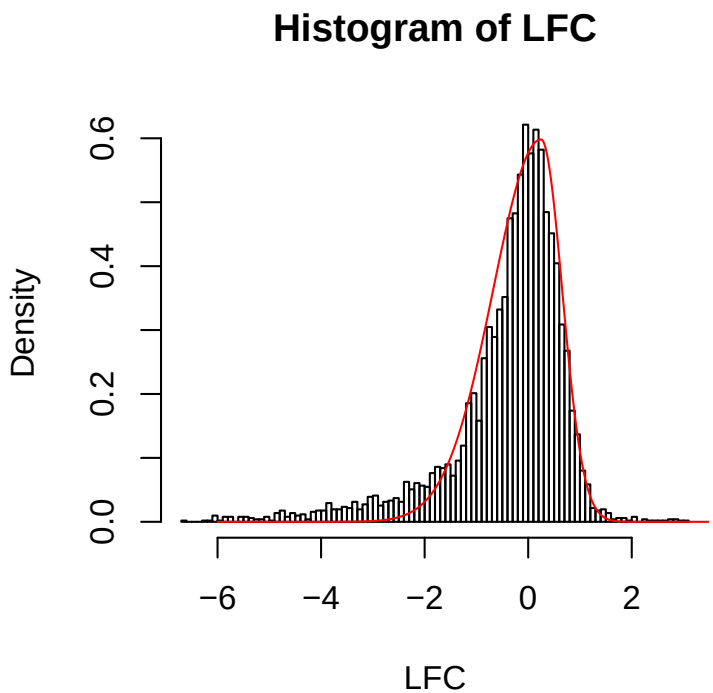
- Modeling of the data as a mixture of null-distribution $f_0$ and unknown distribution $f_1$ of the gRNAs with fitness effect:
  $f(x) = (1 - \lambda) f_0(x) + \lambda f_1(x)$
- $f_0$ is a skew normal distribution with 3 parameters:
  location $\xi$, scale parameter $\omega$, skewness parameter $\alpha$



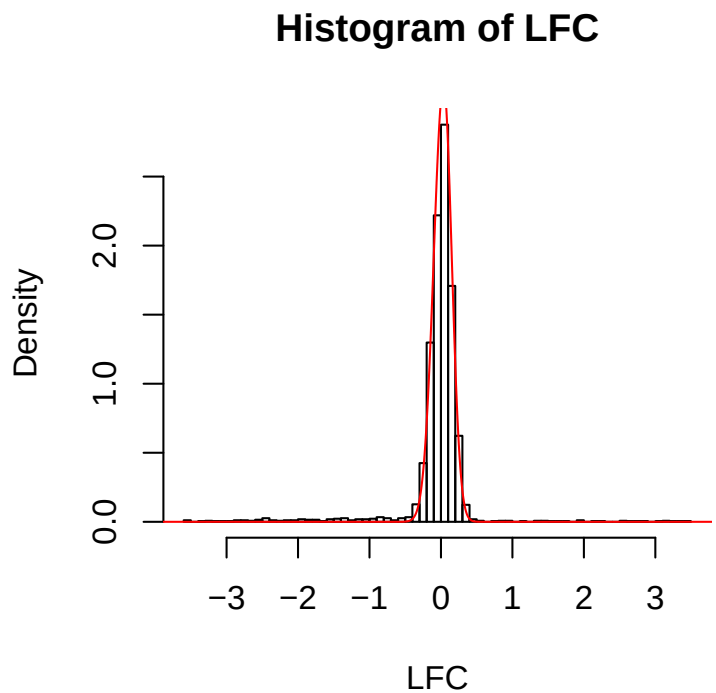
# Step 3: Fitting the null-distribution

- Fit $\xi$, $\omega$, and $\alpha$ from the actual LFC data.
- Ignore strong positive or negative LFCs, only consider the central 90% data point (using approach derived from least quantile of squares regression (*Rousseeuw et al. 1987*).

Low abundant gRNAs
(20-30% percentile)
$\xi = 0.16$, $\omega = 0.69$, $\alpha = 1.57$

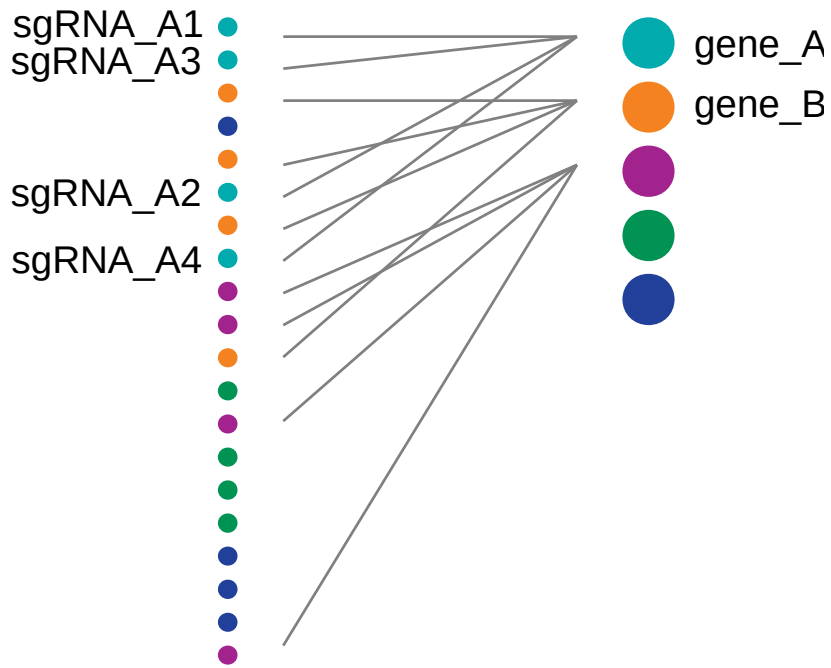



High abundant gRNAs
(80-90% percentile)
$\xi = -0.02$, $\omega = 0.13$, $\alpha = 1.09$



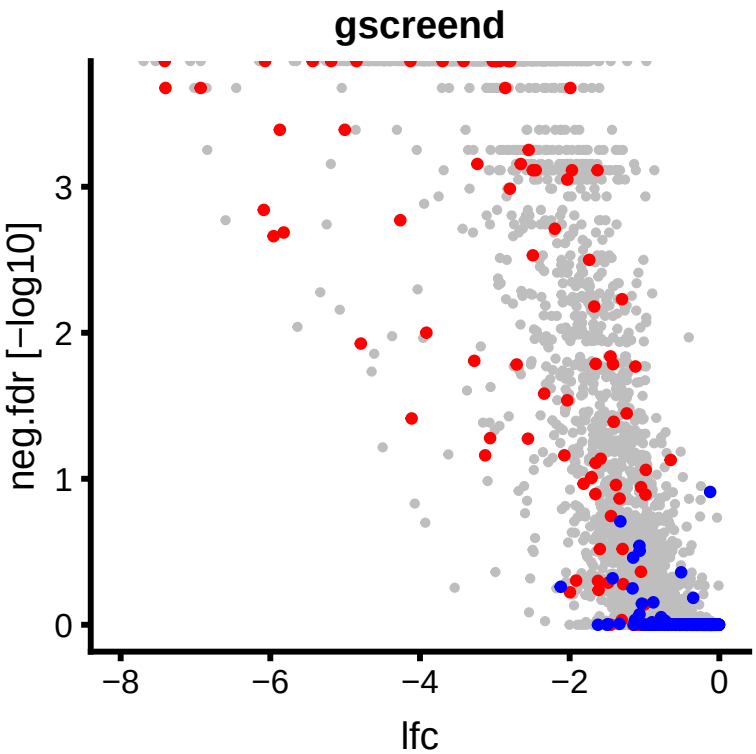
# Step 4: Aggregation of gRNA level data to gene level

- Calculation of p-value for every gRNA.
- Ranking of gRNAs according to p-values.
- Robust rank aggregation (*Kolde et al. 2012*) to aggregate on gene level (typically 3-10 gRNAs per gene).
- Do the observed gRNA ranks for a given gene lie significantly outside of what you would expect by random sampling? (Permutation test)



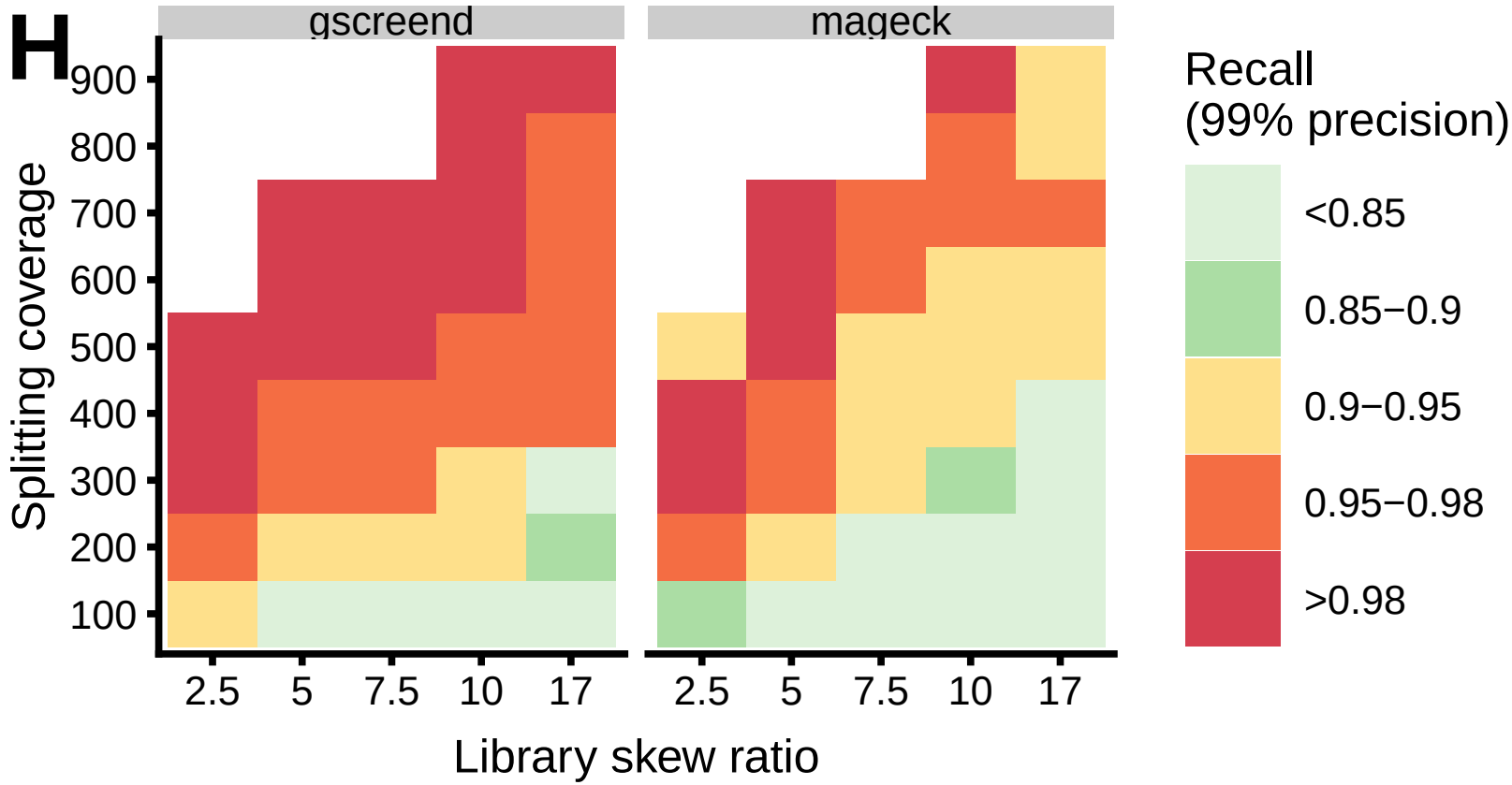
# Results from a screen performed in HCT116 cells

components of the ribosome
non-essential genes



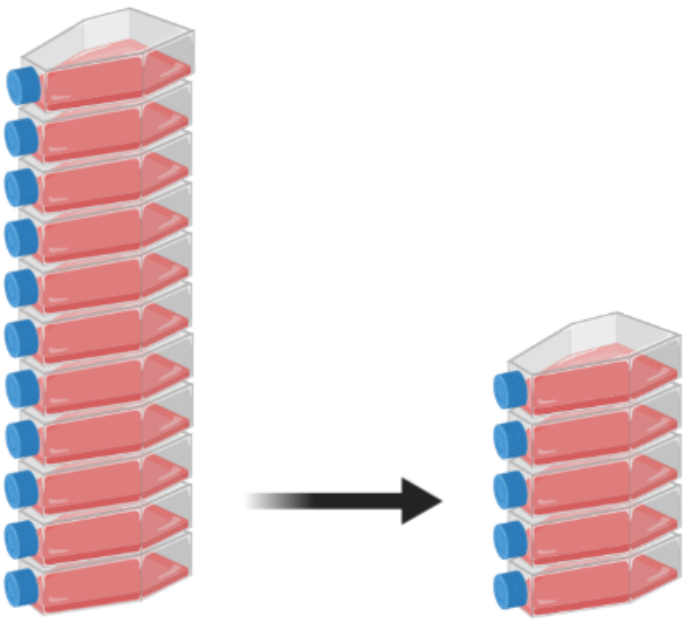
# gscreend performance on simulated data

Ranking accuracy is improved using **gscreend** compared to other method.



# This has major implications for experiment design

- We can predict the minimal necessary experiment size.
- gscreend allows reduction of experiment size by up to 50%.



## Conclusions

- Understand the data from an experimental point of view!
- Changes in gRNA abundance are asymmetric in pooled CRISPR screens (unlike RNA-seq data).
- We provide recommendation for optimal experimental design.
- **gscreend**: more accurate phenotype detection at smaller experiment size.

**gscreend** (in preparation for Bioconductor submission):

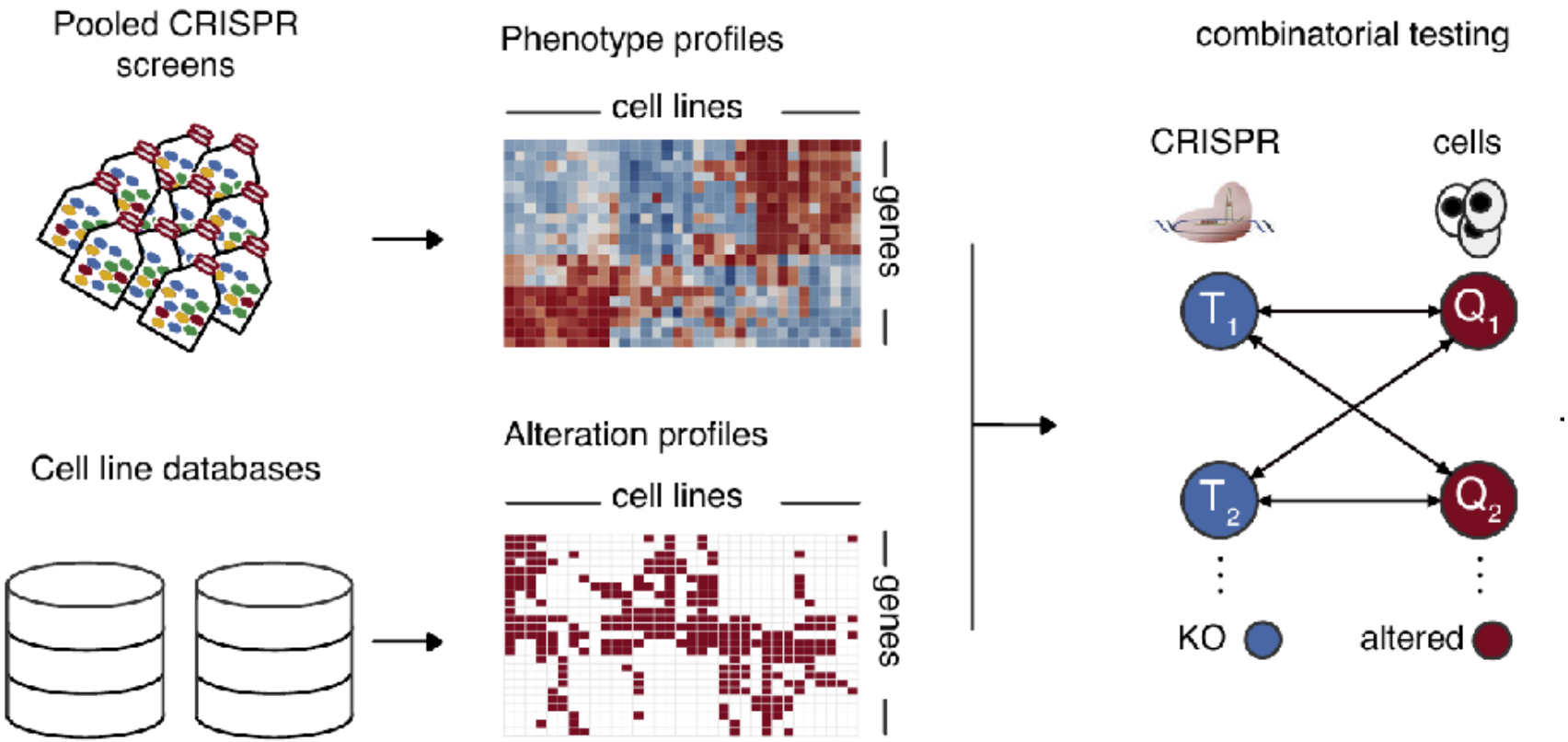
https://github.com/imkeller/gscreend

**bioRxiv**
Modelling asymmetric count ratios in CRISPR screens to decrease experiment size and improve phenotype detection
*Katharina Imkeller, Giulia Ambrosi, View ORCID ProfileMichael Boutros, Wolfgang Huber*
doi: https://doi.org/10.1101/699348


**Example for applications of CRISPR screens...**

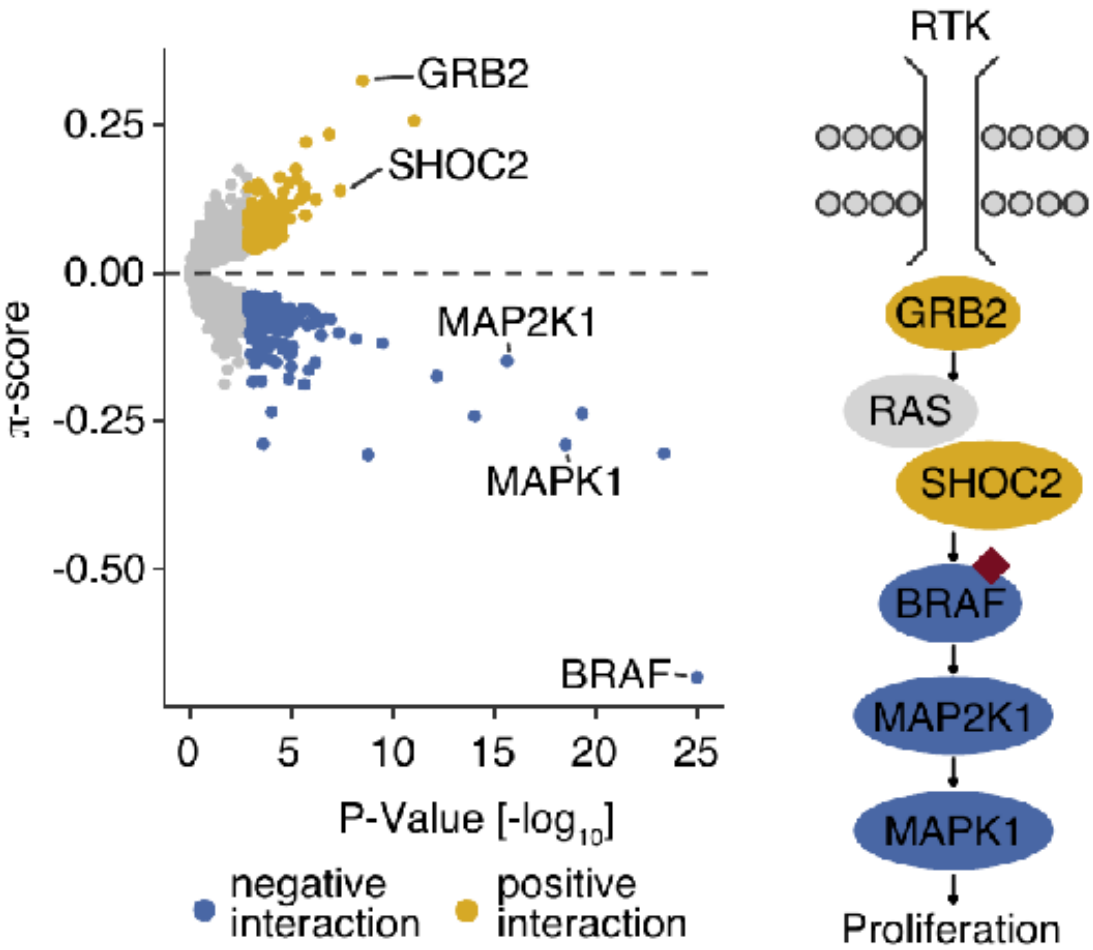
# Context dependent lethality

Cancer dependency map: https://depmap.org/portal/




*Rauscher et al. 2018, Henkel et al. 2019*

# BRAF mutation context dependency




*Rauscher et al. 2018, Henkel et al. 2019*

# Resources

- Some of the graphics in this presentation were generated with Biorender (www.biorender.com)
- Rousseeuw, PJ, Leroy, AM. Robust regression and outlier detection. Wiley Series in Probability and Statistics 329 (1987).
- Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012. 10.1093/bioinformatics/btr709
- Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, Irizarry RA, Liu JS, Brown M, Liu XS. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014. 10.1186/s13059-014-0554-4
- Joung J, Konermann S, Gootenberg JS, Abudayyeh OO, Platt RJ, Brigham MD, Sanjana NE, Zhang F. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat Protoc. 2017. 10.1038/nprot.2017.016
- Rauscher B, Heigwer F, Henkel L, Hielscher T, Voloshanenko O, Boutros M. Toward an integrated map of genetic interactions in cancer cells. Mol Syst Biol. 2018. 10.15252/msb.20177656
- Henkel L, Rauscher B, Boutros M. Context-dependent genetic interactions in cancer. Curr Opin Genet Dev. 2019. 10.1016/j.gde.2019.03.004