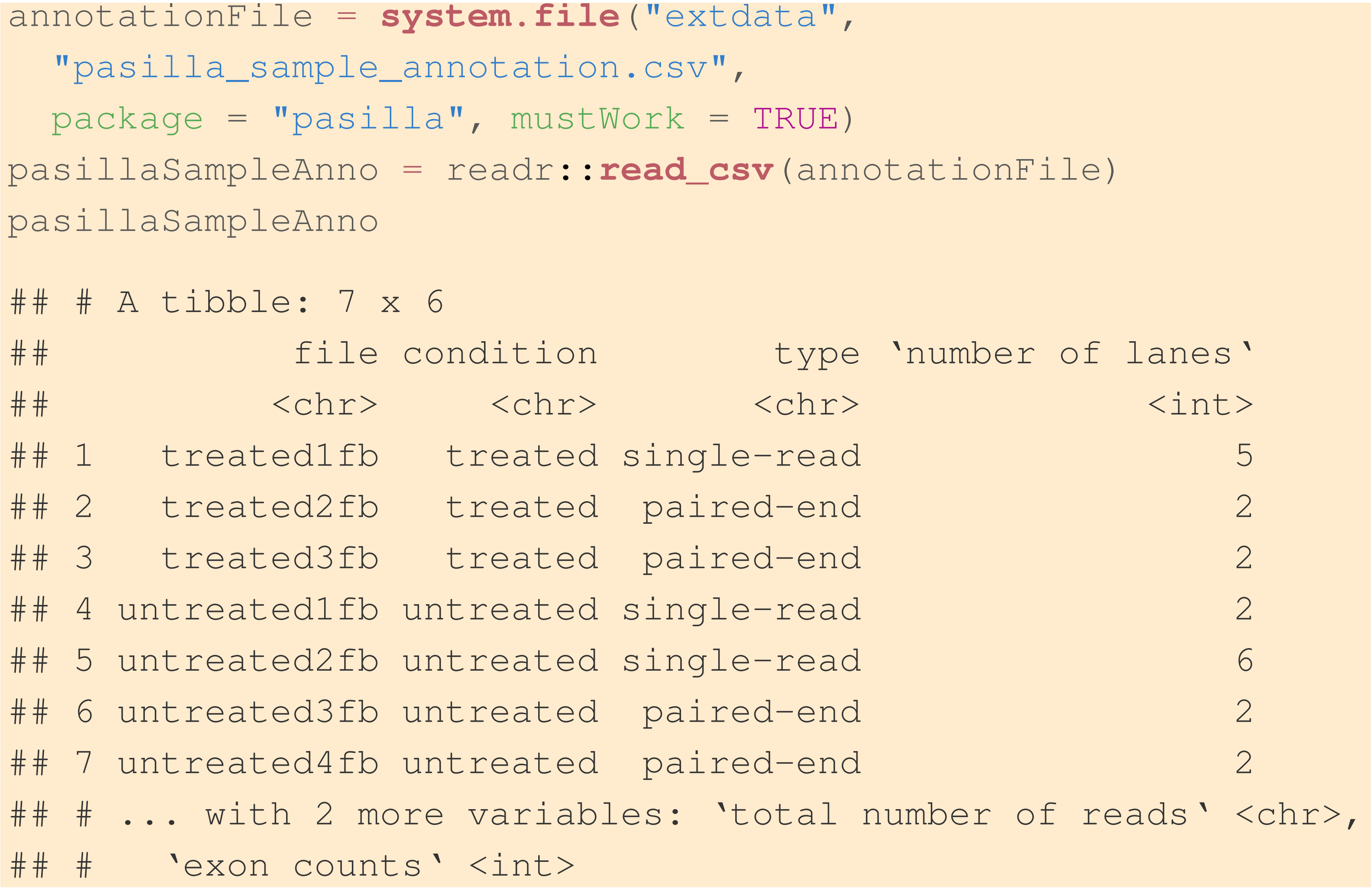
Design matrices & Co.

- We already know how to test for differential expression between two conditions and how to estimate the log-fold changes
- But reality is more complicated: factorial designs, batch effects



```
annotationFile = system.file("extdata",
  "pasilla_sample_annotation.csv",
  package = "pasilla", mustWork = TRUE)
pasillaSampleAnno = readr::read_csv(annotationFile)
pasillaSampleAnno

## # A tibble: 7 x 6
##            file condition        type `number of lanes`
##           <chr>     <chr>       <chr>             <int>
## 1   treated1fb   treated single-read                 5
## 2   treated2fb   treated  paired-end                 2
## 3   treated3fb   treated  paired-end                 2
## 4 untreated1fb untreated single-read                 2
## 5 untreated2fb untreated single-read                 6
## 6 untreated3fb untreated  paired-end                 2
## 7 untreated4fb untreated  paired-end                 2
## # ... with 2 more variables: `total number of reads` <chr>,
## #   `exon counts` <int>
```

Design: Example

Imagine we sequenced:

- 5 treated samples out of which 4 paired-end, 1 single-read
- 5 control samples out of which 1 paired-end, 4 single-read

What does it mean if a gene comes up as differentially expressed?

Imagine we have

- a cell line pair: "wild type" and BRD3-KO
- treat both with DMSO or iBET

Design

- Let us write:

$$\begin{aligned}\log_2(\mu_{\text{treat}}) &= \log_2(\mu_{\text{control}}) + \log_2(\mu_{\text{treat}}) - \log_2(\mu_{\text{control}}) \\ &= \log_2(\mu_{\text{control}}) + \log_2\left(\frac{\mu_{\text{treat}}}{\mu_{\text{control}}}\right) \\ &= \beta_0 + \beta_1\end{aligned}$$

- So we can say that for sample *i*:

$$\log_2(\mu_i) = \begin{cases}\beta_0, & \text{if control} \\ \beta_0 + \beta_1, & \text{if treated}\end{cases}$$

Design

- $$\log_2(\mu_i) = \begin{cases} \beta_0, & \text{if control} \\ \beta_0 + \beta_1, & \text{if treated} \end{cases}$$
- Now we want to include the technology (paired-end vs single-read) in the analysis as well. Let us define the log-fold change between paired-end and single-read:

$$\beta_2 = \log_2 \left(\frac{\mu_{\text{paired-end}}}{\mu_{\text{single-read}}} \right)$$

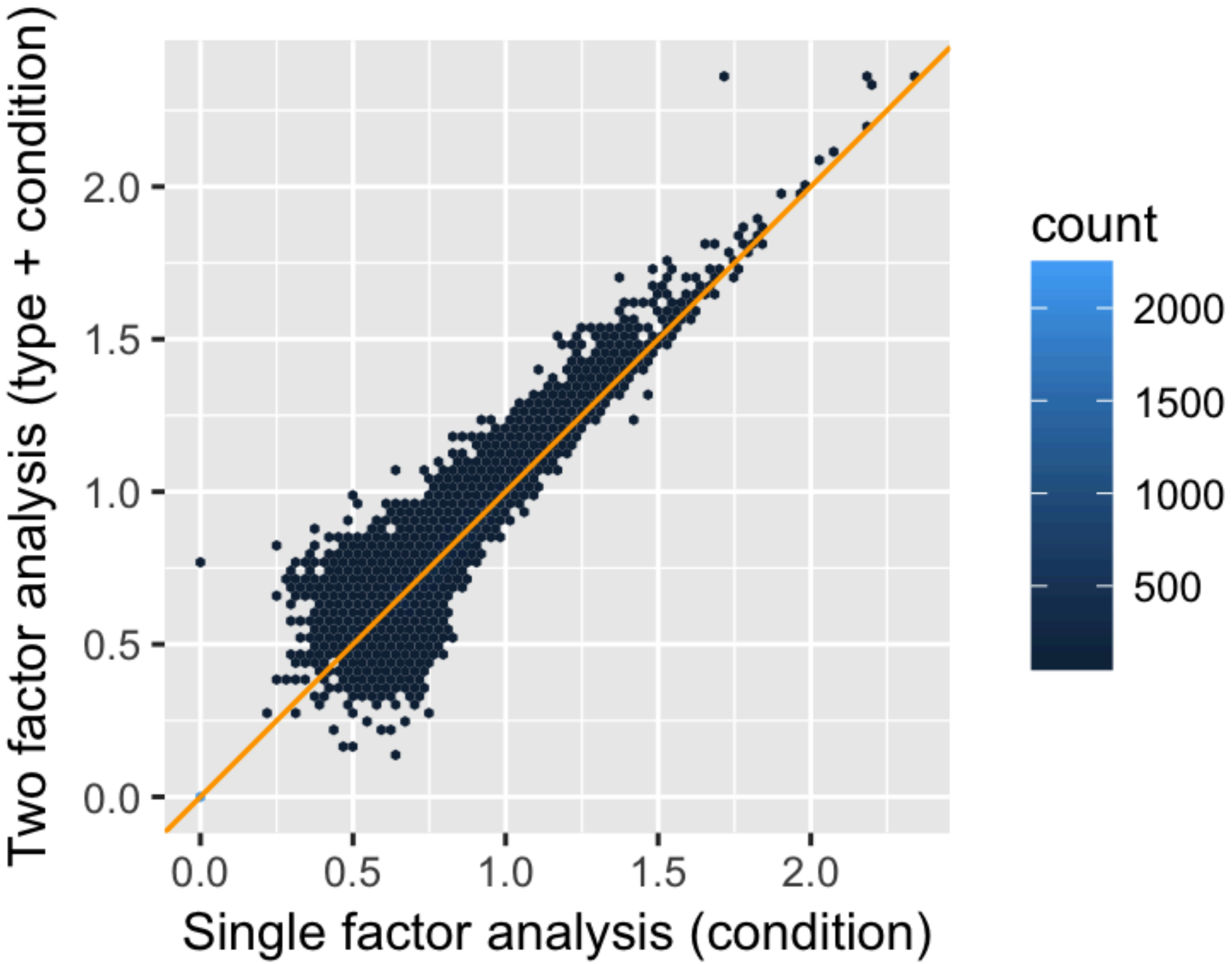
- Then

$$\log_2(\mu_i) = \begin{cases} \beta_0, & \text{if control and single-read} \\ \beta_0 + \beta_1, & \text{if treated and single-read} \\ \beta_0 + \beta_2, & \text{if control and paired-end} \\ \beta_0 + \beta_1 + \beta_2 & \text{if treated and paired-end} \end{cases}$$

Some notes on factorial designs

- We can inform DESeq2 of these designs by using the formula notation: `~ type + condition`
- If we then test for the log-fold change between treated and control, we say that we are *adjusting* or *blocking* or *controlling* for the sequencing technology.
- If every treated sample was sequenced on paired-end and every control sample was sequenced on single-read, then the model is not identifiable!

Comparison of the two analyses



On x and y-axis:
Transformation of
p-values such that
large values indicate
small p-values!

How did power increase?

Sometimes specifying the design can improve power.
Common example: Paired designs

patient	treatment
1	before
1	after
2	before
2	after
3	before
3	after
4	before
4	after

Design: Advanced

- $$\log_2(\mu_i) = \begin{cases} \beta_0, & \text{if control and single-read} \\ \beta_0 + \beta_1, & \text{if treated and single-read} \\ \beta_0 + \beta_2, & \text{if control and paired-end} \\ \beta_0 + \beta_1 + \beta_2 & \text{if treated and paired-end} \end{cases}$$

- Compact notation: Write $x_{i1} = 1$ if treated and 0 otherwise, and $x_{i2} = 1$ if paired-end and 0 otherwise, then

$$\log_2(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

Design: Generalized Linear Models

- Can generalize this even further to:

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$$

- Upshot: "Generalized Linear Models" are well studied, all methods described generalize to this setting.
- Usually expressed in terms of a design matrix

Design matrix for paired designs

patient	treatment
1	before
1	after
2	before
2	after
3	before
3	after
4	before
4	after

⟺

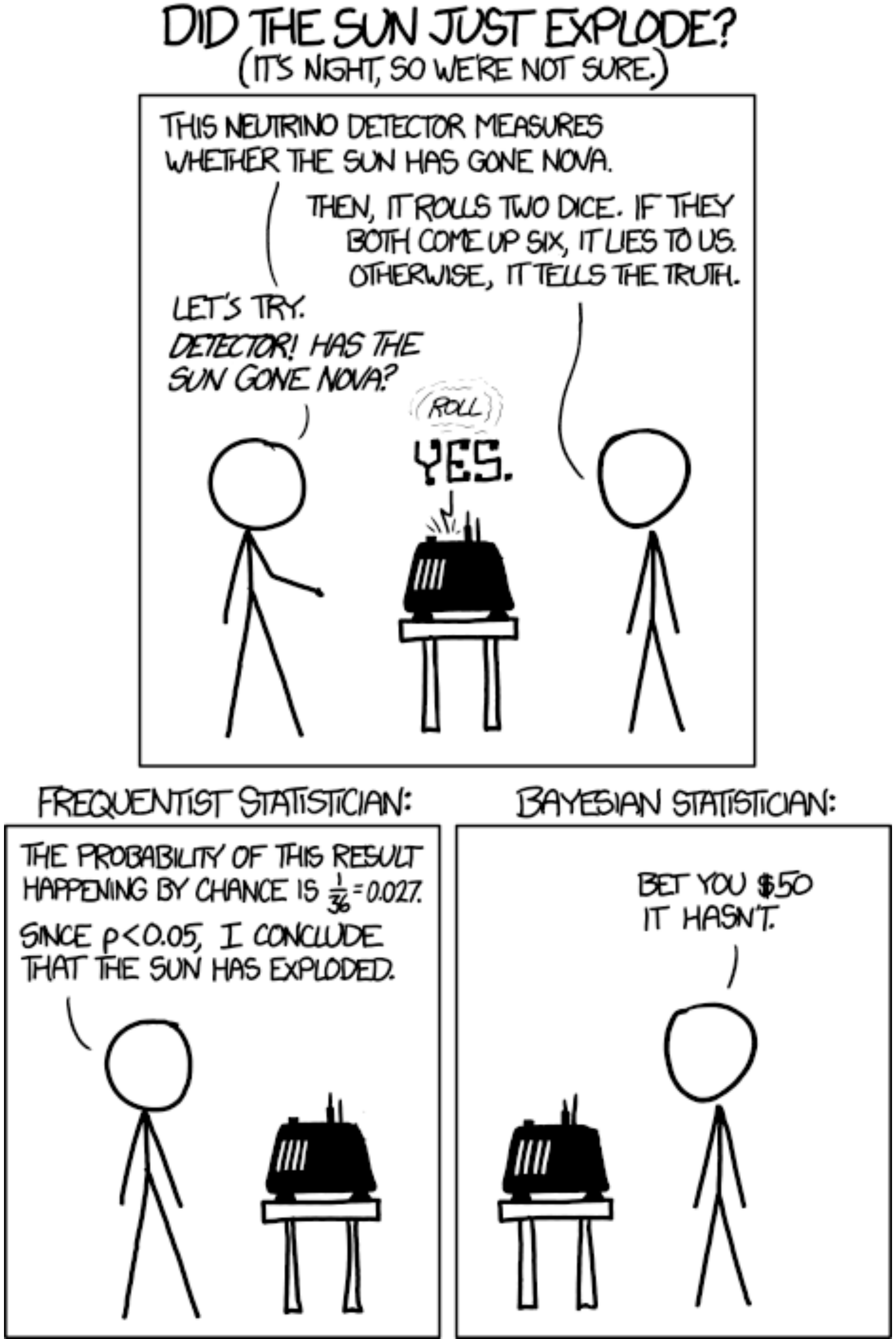
x_0	x_1	x_2	x_3	x_4
1	0	0	0	0
1	0	0	0	1
0	1	0	0	0
0	1	0	0	1
0	0	1	0	0
0	0	1	0	1
0	0	0	1	0
0	0	0	1	1

Further extensions

- Because of the flexibility of underlying GLMs, we can deal with interactions, continuous covariates, time, etc.
- `DESeq2` workflow, `limma` vignette, Bioconductor support forum
- There are also methods that try to infer batch-effects/confounders when we did not actually measure them:
  - RUV-Seq (Remove Unwanted Variation from RNA-Seq Data)
  - SVA (Surrogate Variable Analysis)

Extra: empirical Bayes - stabilising per-gene estimates in a linear model by “sharing” across genes

Bayesian statistics


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027.
SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

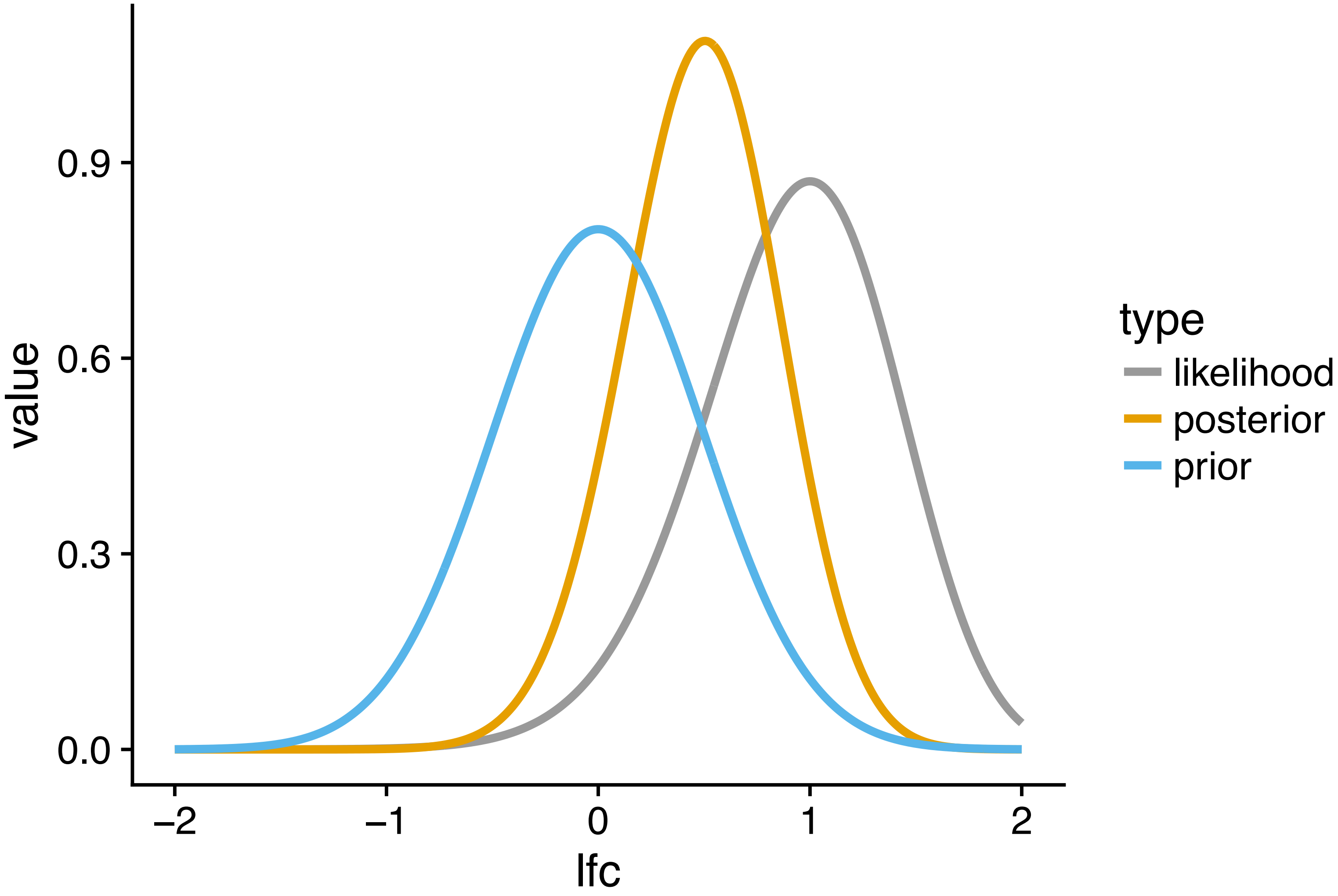
Bayesian approach

- Define prior on lfc

$$\text{lfc} \sim \mathcal{N}\left(0, \sigma^2\right)$$

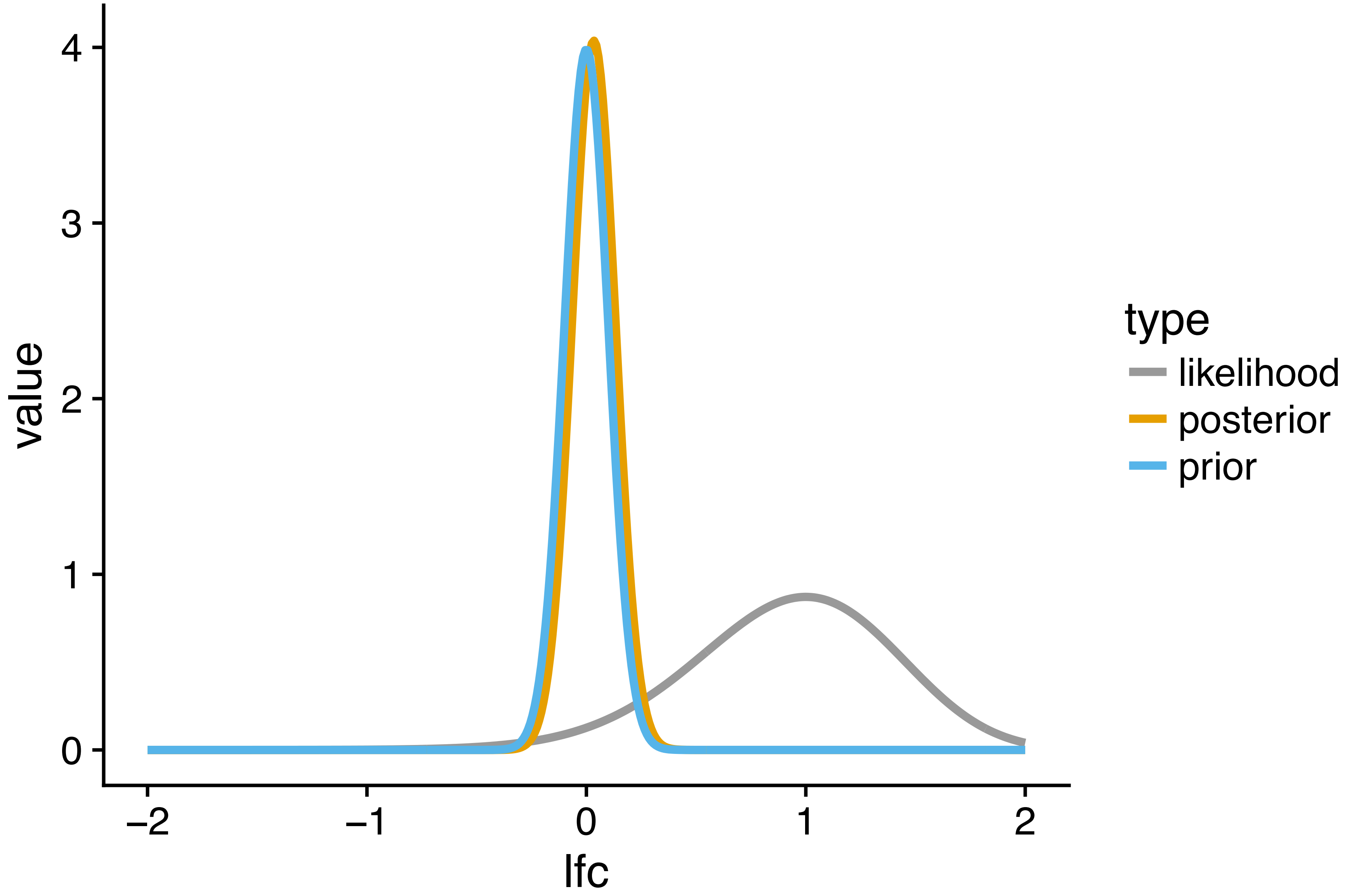
- Recall posterior = prior*likelihood
- Do not look at the maximum of the likelihood
- Look at the maximum of the posterior instead

- Example: σ=0.5,
- 5 counts for control
- 10 counts for treatment



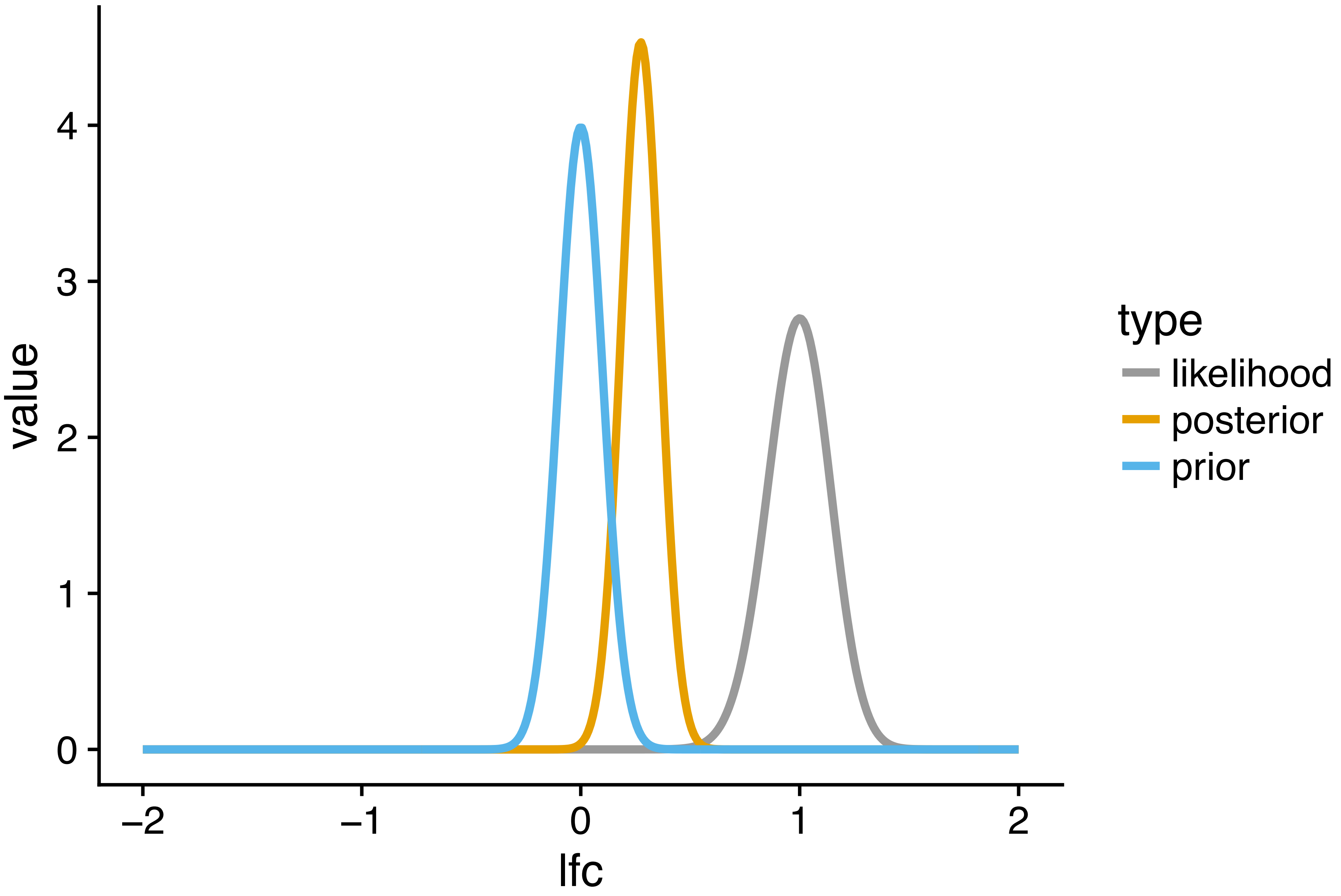
Stronger prior

- Example: σ=0.1,
- 5 counts for control
- 10 counts for treatment



More informative data

- Example: σ=0.1,
- 50 counts for control
- 100 counts for treatment

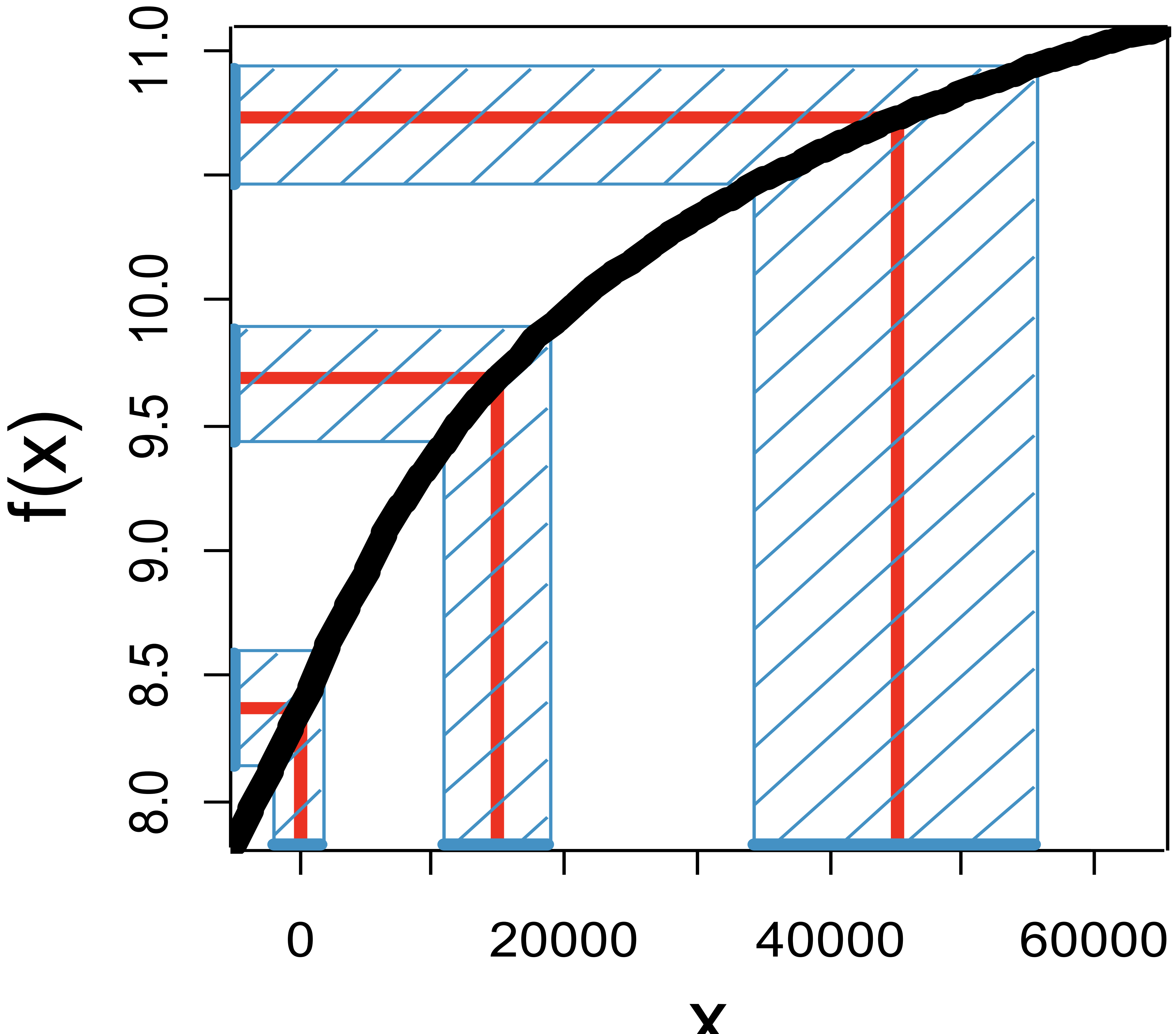


Remarks on Bayesian approach

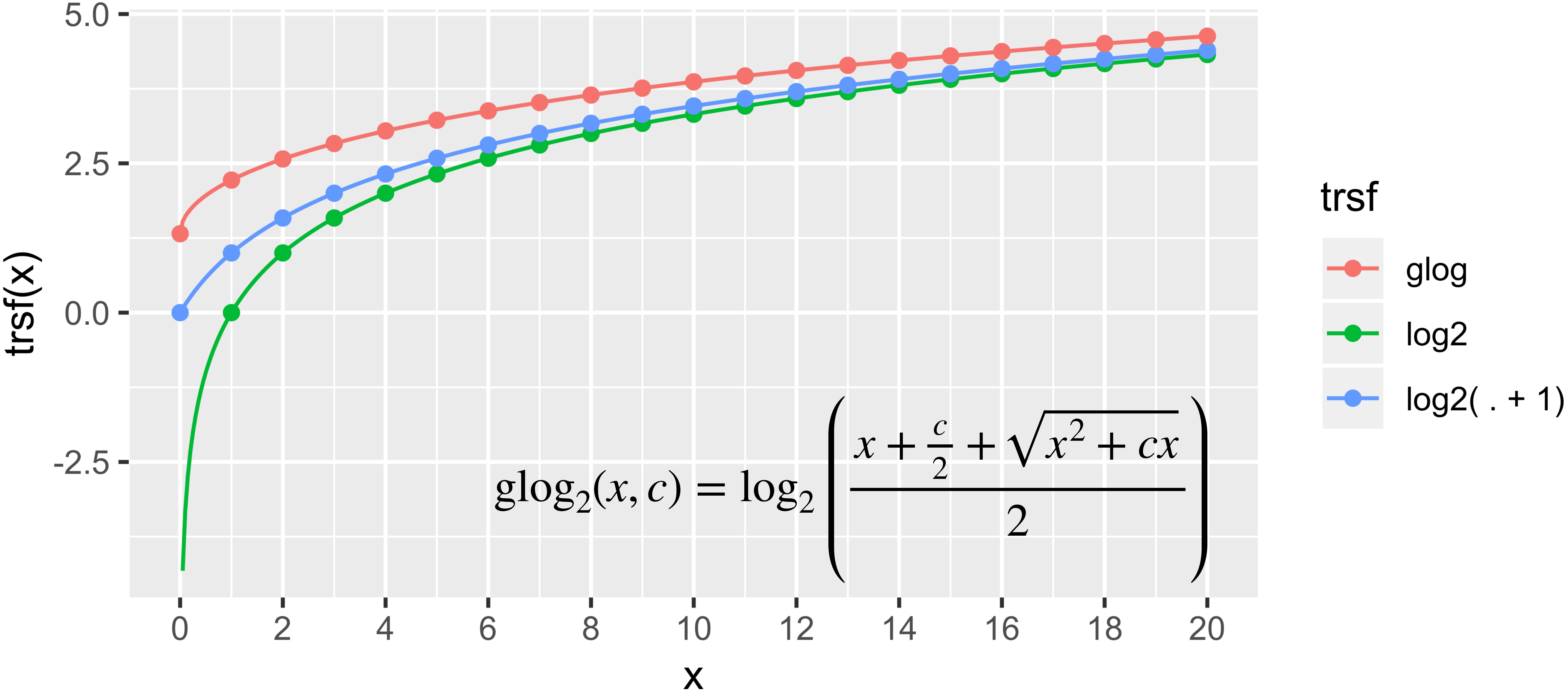
- Once we chose prior, adaptive to signal in the data.
- But how to choose the prior?
- (Or how to choose the pseudocounts?)

Extra: transformations

Variance stabilizing transformation

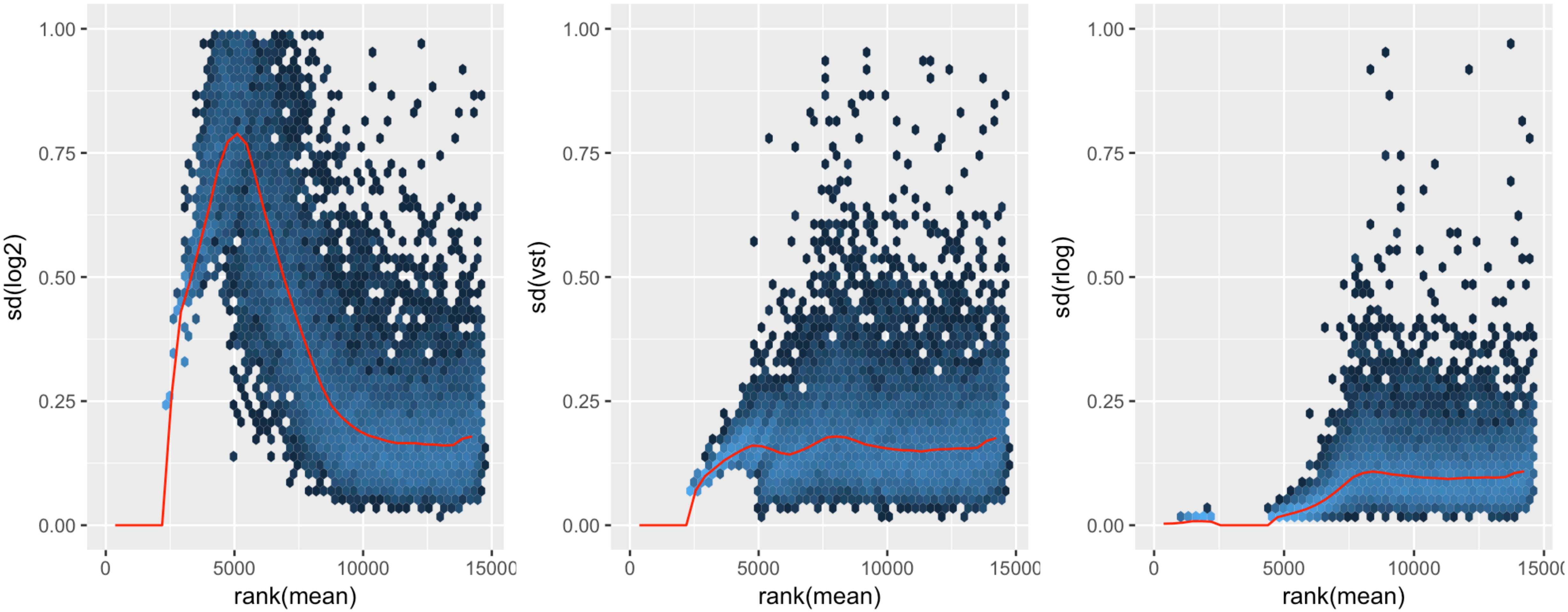


Variance-stabilizing transformation interpolates between √ and log$_2$

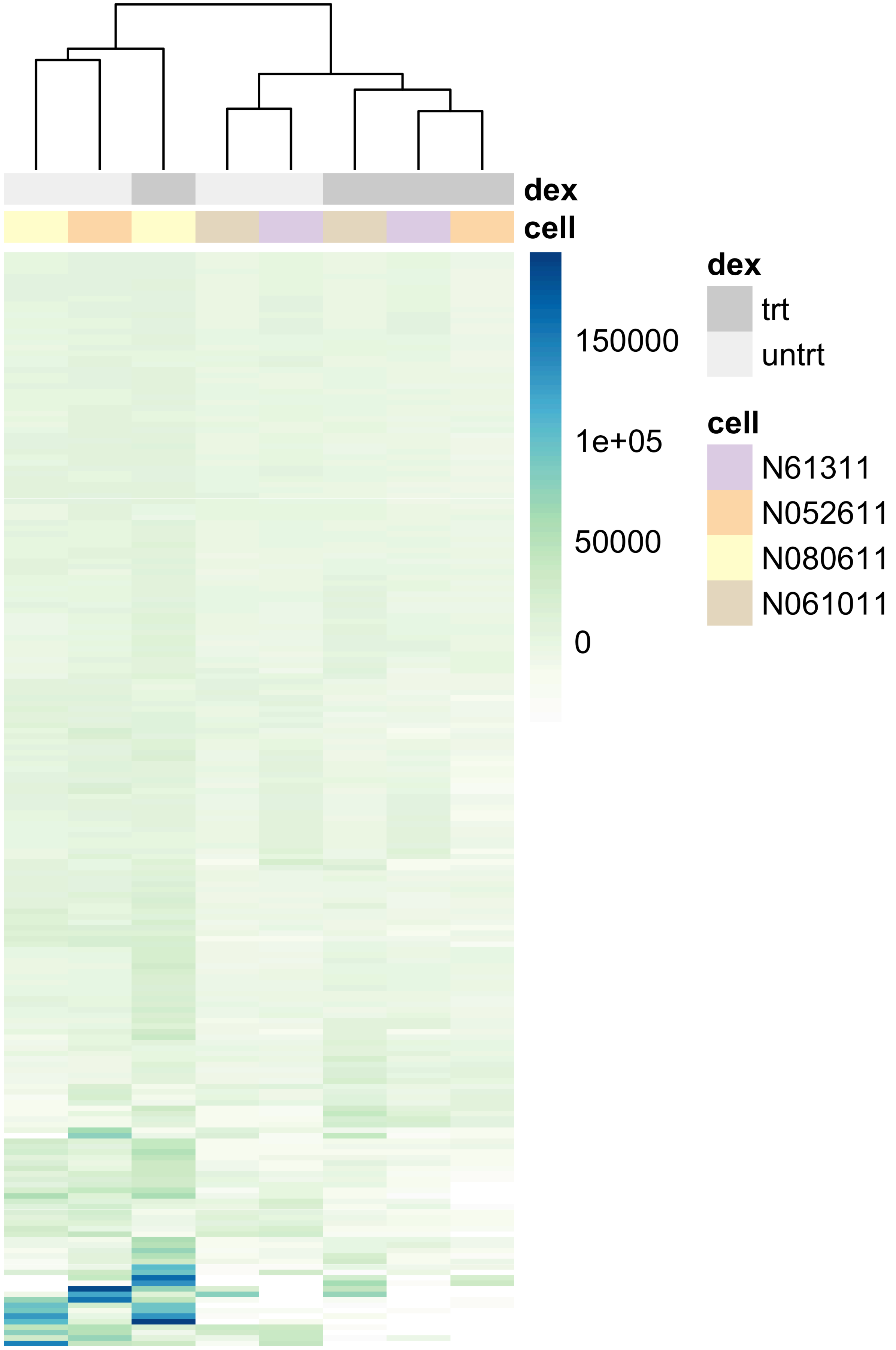


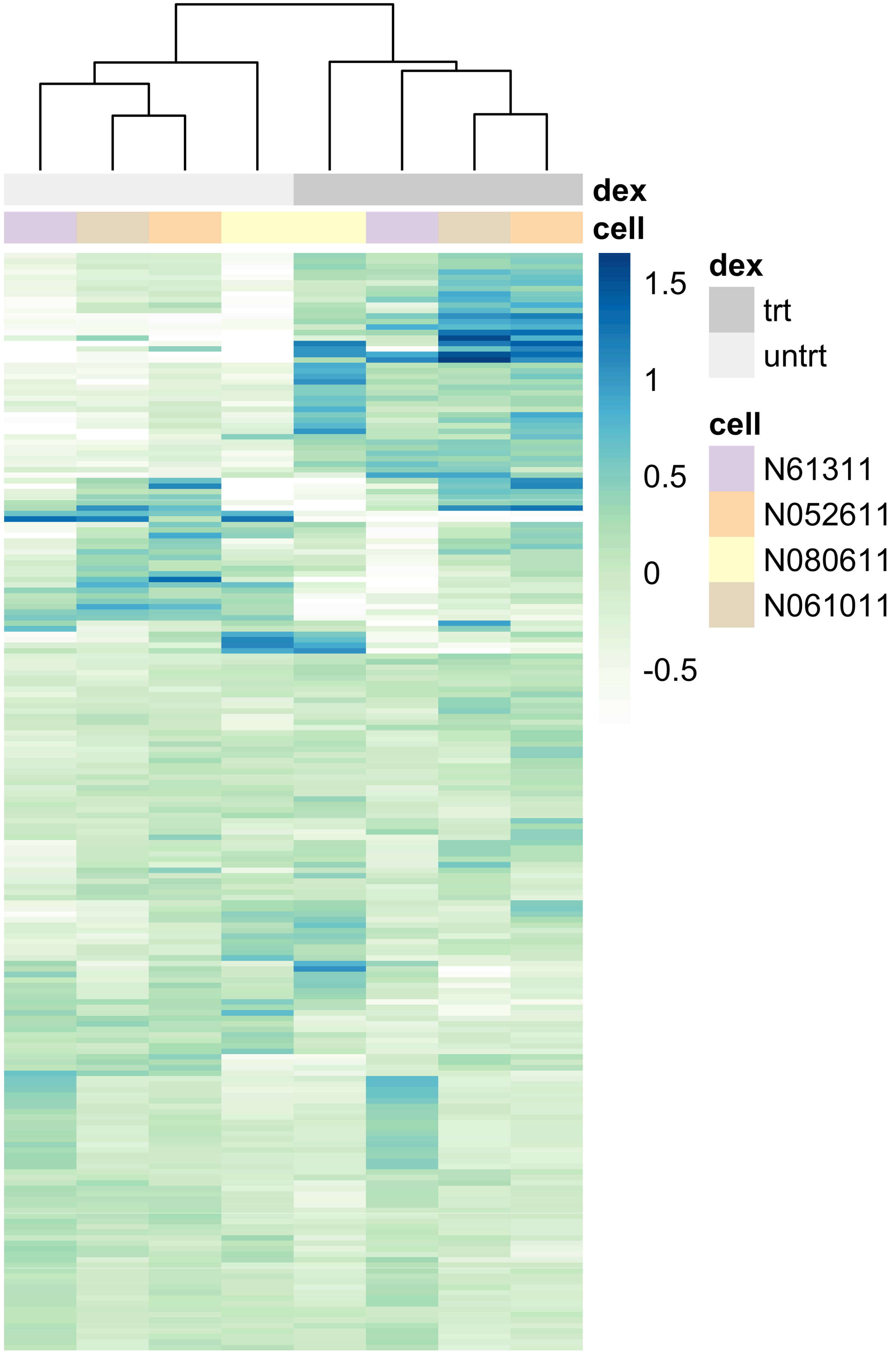
$$\mathrm{glog}_2(x, c) = \log_2\left(\frac{x + \frac{c}{2} + \sqrt{x^2 + cx}}{2}\right)$$

Variance stabilization



Variance-stabilizing transformation for color aesthetic





Recap: Transformations

Choosing the right transformation for your data is crucial.

The scale at which your data are recorded is not necessarily the one at which they should be visualised, analysed.

There is more than the logarithm.

Often, the variance-mean relationship is a good guide.

PS Such awareness exists in physics (radius vs volume, dezibels, Richter scale, critical fluctuations, Lyapunov exponents)