RNAseq Theory

Teo Sakel

Kallisto

De Bruijn Graph

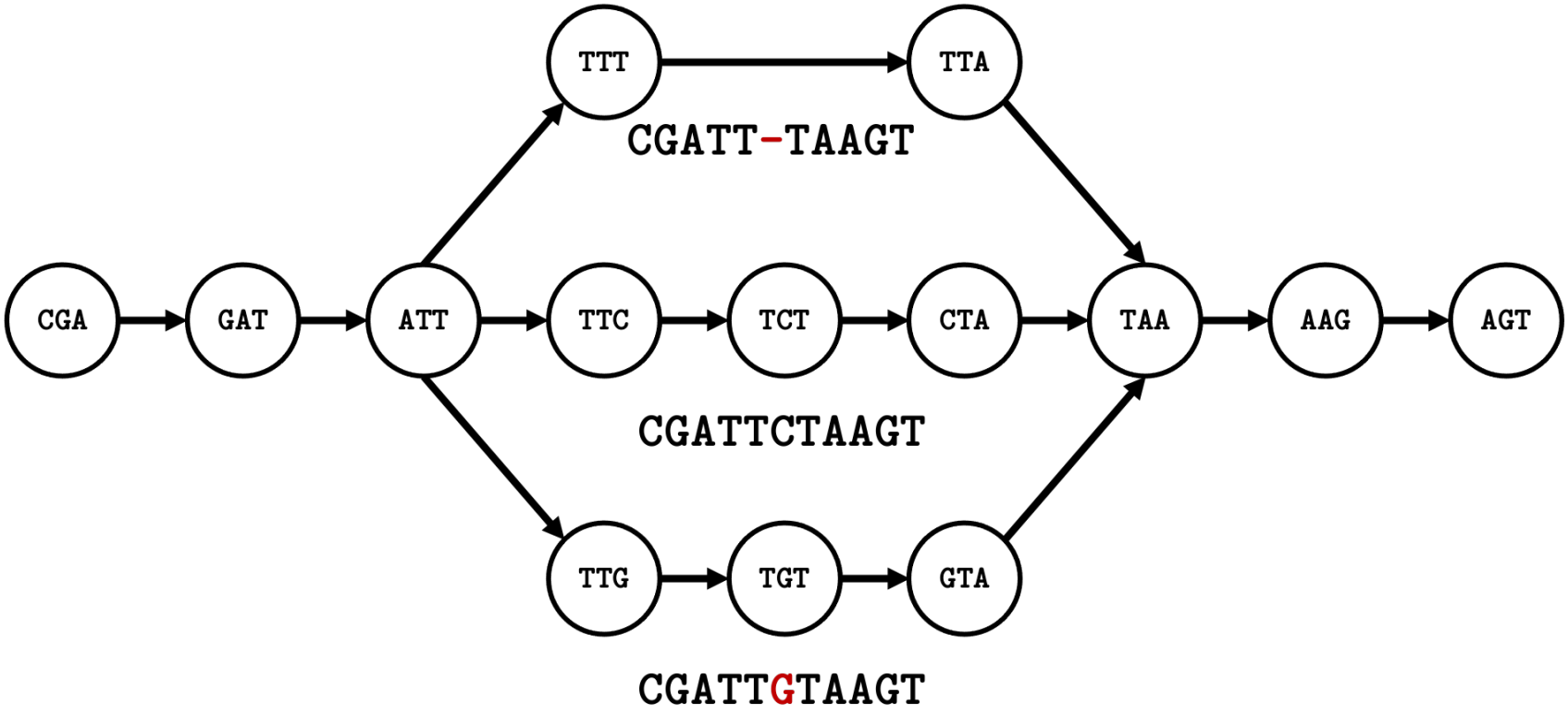
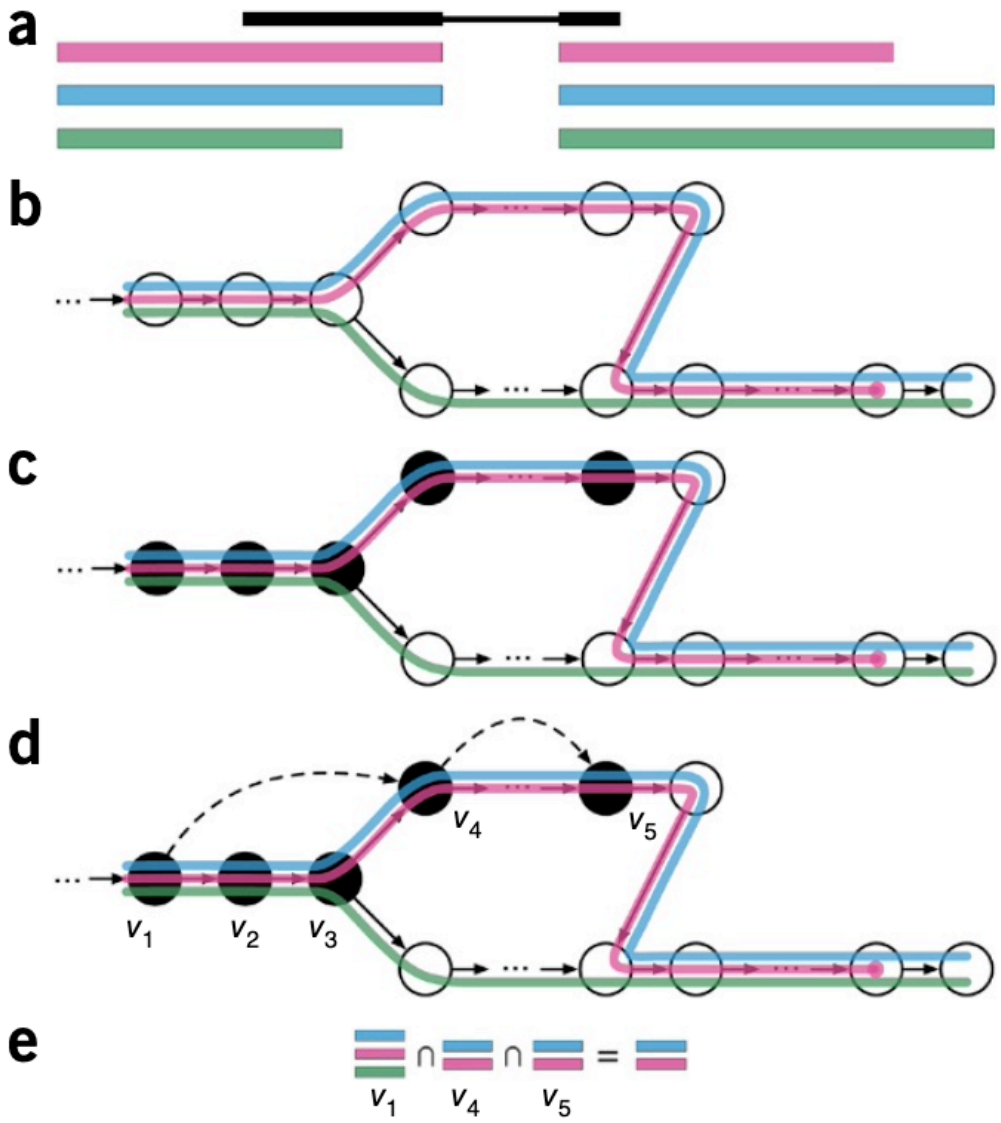


Figure 1: Leggett et al. PloS one 8.3 (2013): e60058

Kallisto Index



Transcript Abundances

Trascripts vs Fragments vs Reads

- pools of transcripts
- transcript abundances (simplex): $\sum_t \rho_t = 1$
- reads per X_t

Abundances Estimation

- effective length: $\tilde{\ell}_t = \ell_t - \bar{m} + 1$
- probability of sequencing: $a_t \propto \rho_t \tilde{\ell}_t$
- likelihood of observing X_t reads from a set of T transcripts:

$$\mathcal{L}(\rho \mid X) = \prod_{t \in T} \prod_{r \in R} P(r|t) P(f \in t) = \prod_{t \in T} \left(\frac{a_t}{\tilde{\ell}_t} \right)^{X_t}$$

RPKM

- treat $\tilde{\ell}_t$ as constant

$$\hat{a}_t = \frac{X_t}{\sum_t X_t}$$

$$\hat{\rho}_t = \frac{\hat{a}_t}{\tilde{\ell}_t} = \frac{X_t}{N\tilde{\ell}_t}$$

TPM

- Problem with RPKM:
  - $\sum_t X_t$ is not an estimate of total number of transcripts
  - $\tilde{\ell}_t$ differs from experiment to experiment.
- Estimate of transcript count: $Y_t = \frac{X_t \bar{m}}{\tilde{\ell}_t}$

$$\hat{\rho}_t = \frac{Y_t}{\sum_t Y_t}$$

Estimated Counts

- We do not know the origin of reads: $P(r \mid t)$

$$\mathcal{L}(\rho) = \prod_{r \in R} \sum_{t \in T} a_t P(r \mid t)$$

- E-step: estimate X_t
- M-step: estimate $\hat{a}_t$

Paired-end Reads

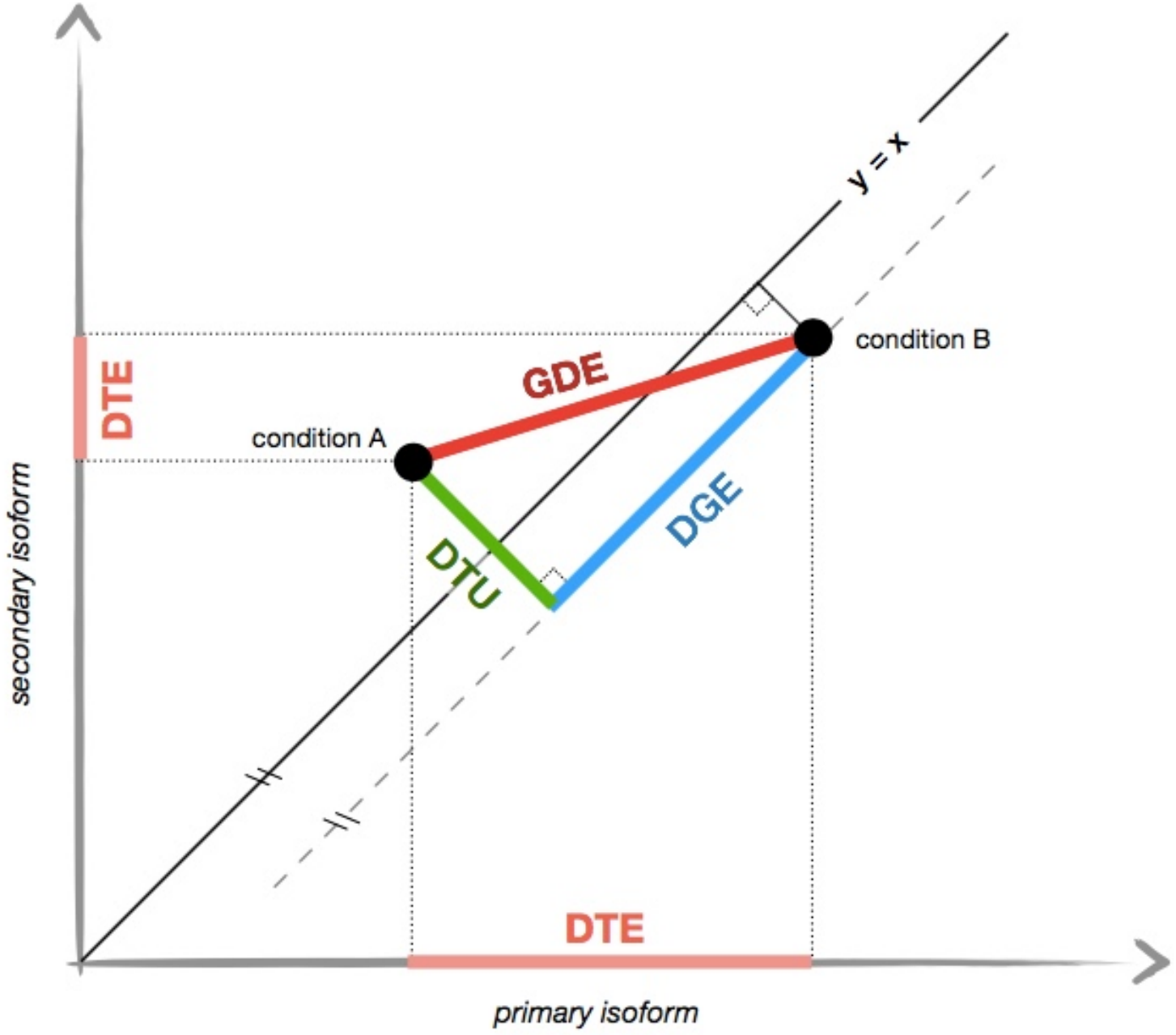
2 ends define the length of fragment:

- Fragment length distribution: $F(m)$ usually modeled as normal
- effective length: $\tilde{\ell}_t = \sum_m F(m)(\ell_t - m + 1)$
- compatibility matrix: y_{rt} 1 if r is compatible with t 0 otherwise
- probability of r mapping to t:

$$P(r \mid t) = y_{rt} \frac{F(m_r)}{\ell_t - m_r + 1}$$

Differential Expression - Part 1

$$GDE^2 = DGE^2 + DTU^2$$



Abundance Fold Change

Assuming the simplest model $\hat{\rho}_t = \frac{X_t}{N\tilde{\ell}_t}$

- Transcript:

$$\Delta\rho_t = \frac{\hat{\rho}_t^b}{\hat{\rho}_t^a} = \frac{X_t^b}{X_t^a}$$

- Gene:

$$\Delta\rho_G = \frac{\sum_{t\in G}\hat{\rho}_t^b}{\sum_{t\in G}\hat{\rho}_t^a} = \frac{N_a}{N_b}\frac{\sum_{t\in G}X_t^b/\tilde{\ell}_t^b}{\sum_{t\in G}X_t^a/\tilde{\ell}_t^a}$$

Problematic Cases

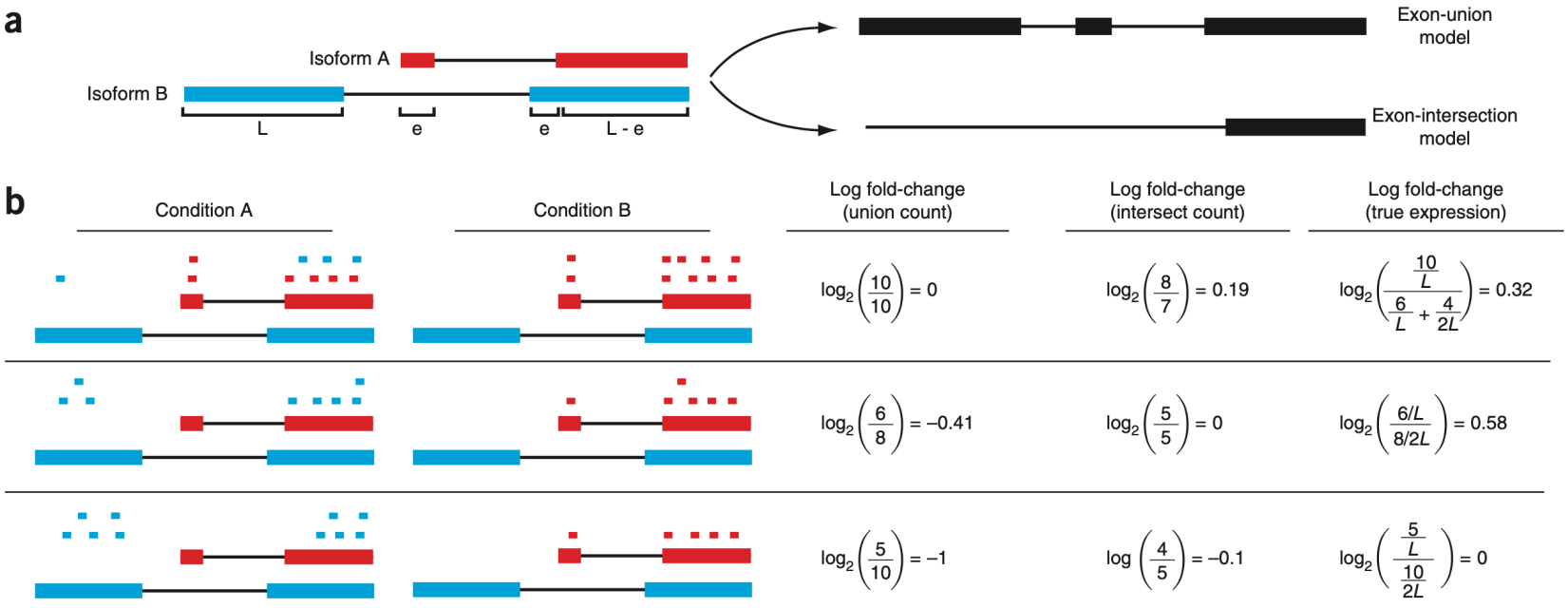


Figure 2: Trapnell et al. Nature biotechnology 31.1 (2013): 46-53.

Gene Counts

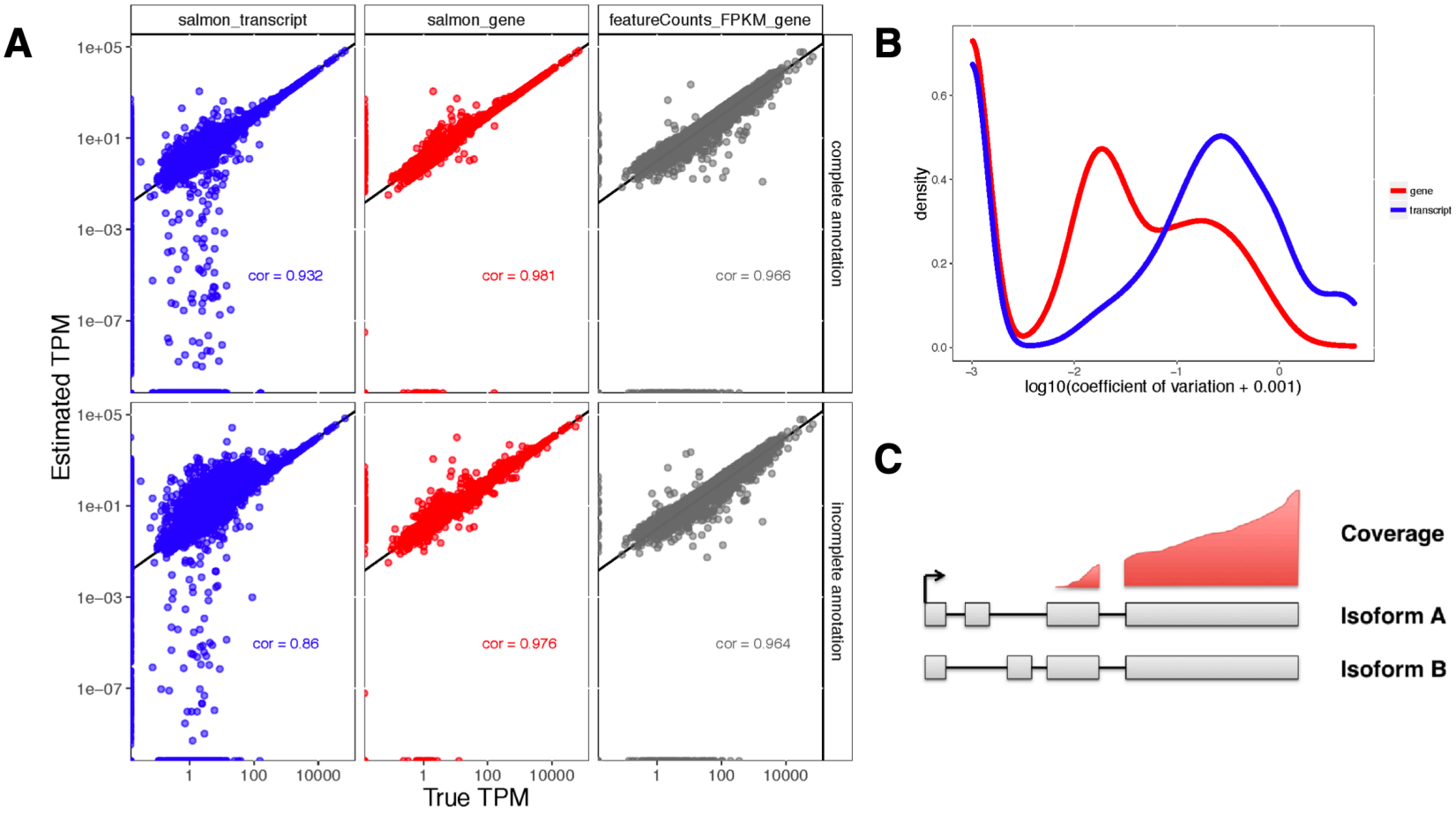


Figure 3: F1000Research 2016, 4:1521 Last updated: 18 JUL 2022

DESeq Analysis

```
dds <- DESeqDataSet(airway, ~ cell + dex)
keep <- rowSums(counts(dds) >= 10) >= 3
dds <- dds[keep,]
dds <- DESeq(dds)
```

```
estimating size factors

estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

final dispersion estimates

fitting model and testing
```

Scaling Factors

Are the differences biological or technical?

```
estimateSF <- function(K) {
    # K: count matrix genes x samples
    K[K == 0] <- NA
    k <- log(K)
    k <- k - rowMeans(k, na.rm = TRUE)
    sf <- colMedians(k, na.rm = TRUE)
    exp(sf - mean(sf))
}
```

Scaling Factor - TMM

edgeR uses a different approach

```
trim <- function(X, p) {
    p <- c(p, 1 - p)
    q <- rowQuantiles(X, probs = p, na.rm = TRUE)
    X < q[, 1] | X > q[, 2]
}

estimate_TMM <- function(K) {
    ref <- which.min(colSums2(K == 0))
    k <- log(K %*% diag(x = 1/colSums(K)))
    M <- k - k[, ref]
    A <- (k + k[, ref]) / 2
    qmask <- trim(M, 0.3) | trim(A, 0.05)
    M[qmask] <- NA_real_
    sf <- rowMeans(M, na.rm = TRUE)
    exp(sf - mean(sf))
}
```

Results

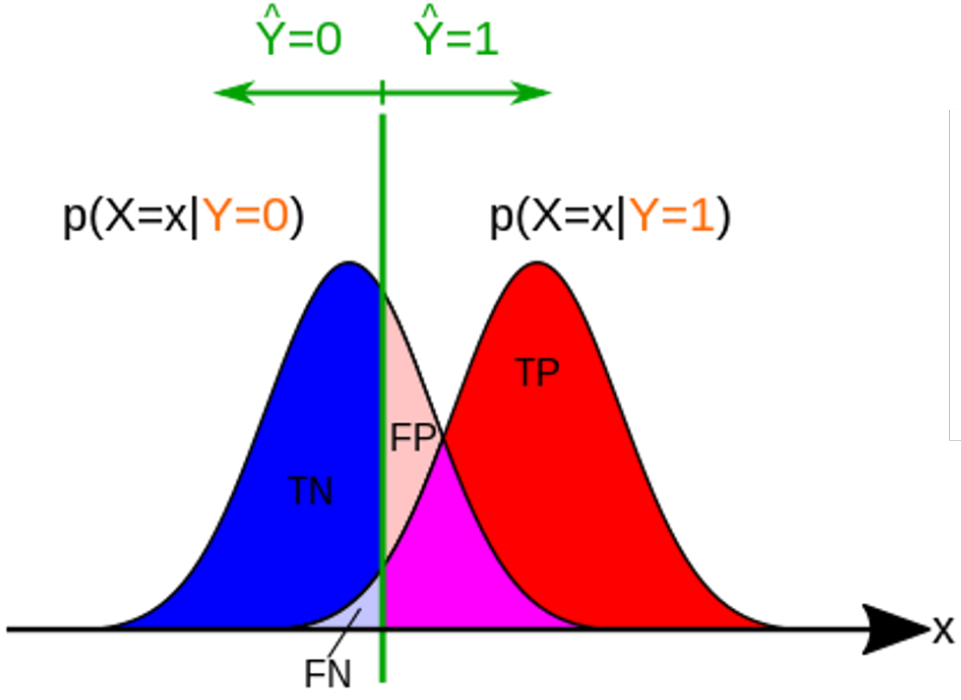
```
res <- results(dds,
               contrast = c("dex", "untrt", "trt"),
               lfcThreshold = 0,
               altHypothesis = "greaterAbs",
               alpha = 0.1)
```

Results DataFrame

	log2FC	lfcSE	pval	padj
ENSG00000000003	0.38	0.10	0.0001	0.0011
ENSG00000000419	-0.20	0.11	0.0675	0.1859
ENSG00000000457	-0.04	0.14	0.8027	0.9040
ENSG00000000460	0.09	0.28	0.7410	0.8717
ENSG00000000971	-0.42	0.09	0.0000	0.0000
ENSG00000001036	0.24	0.09	0.0066	0.0302
ENSG00000001084	0.05	0.17	0.7619	0.8827
ENSG00000001167	0.50	0.12	0.0000	0.0002
ENSG00000001460	0.13	0.18	0.4729	0.6829
ENSG00000001461	0.04	0.10	0.6725	0.8289

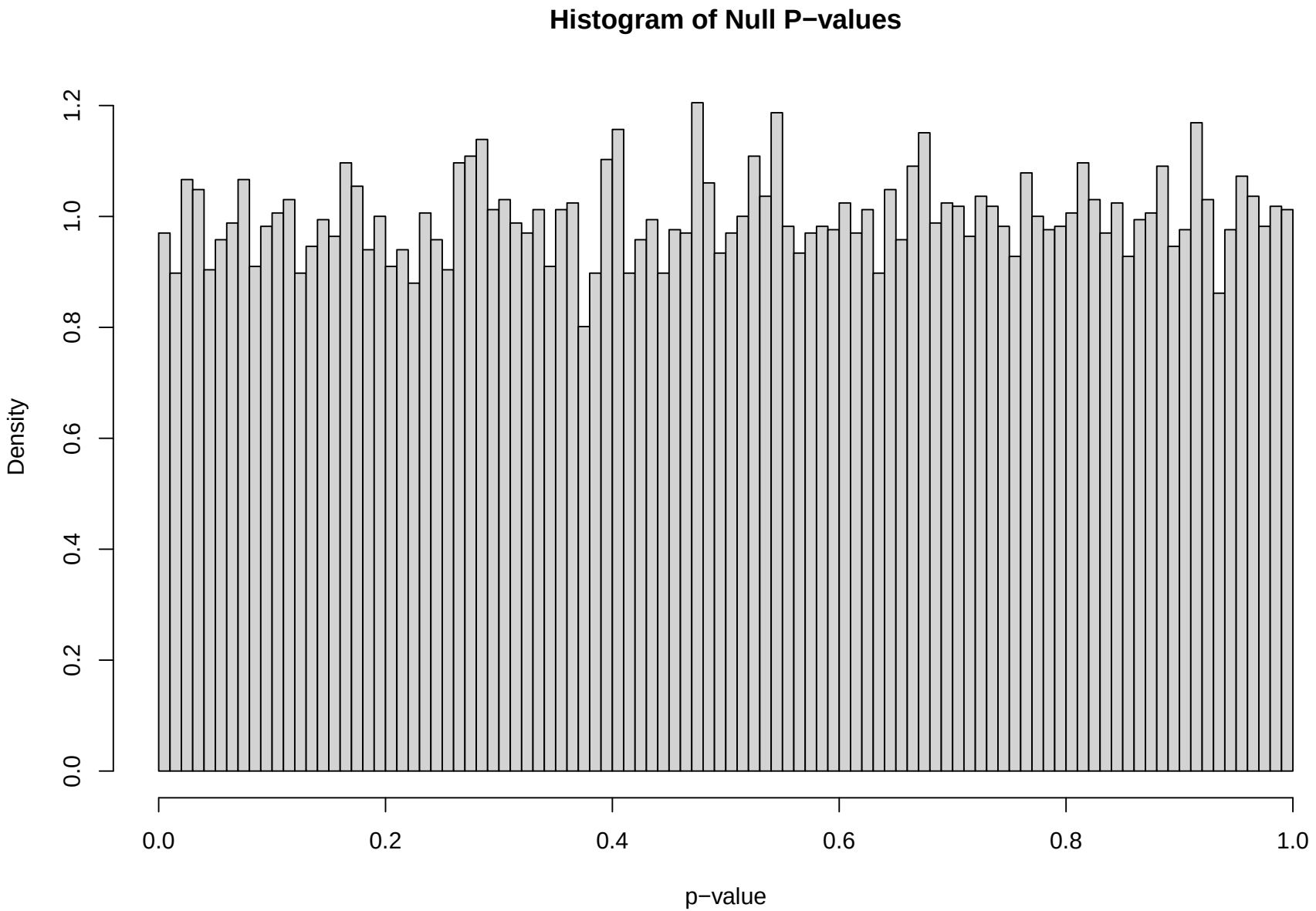
Calling Differentially Expressed Genes

$$H_0 : |\beta| \leq \beta_0$$
$$P(H_0) \leq a$$

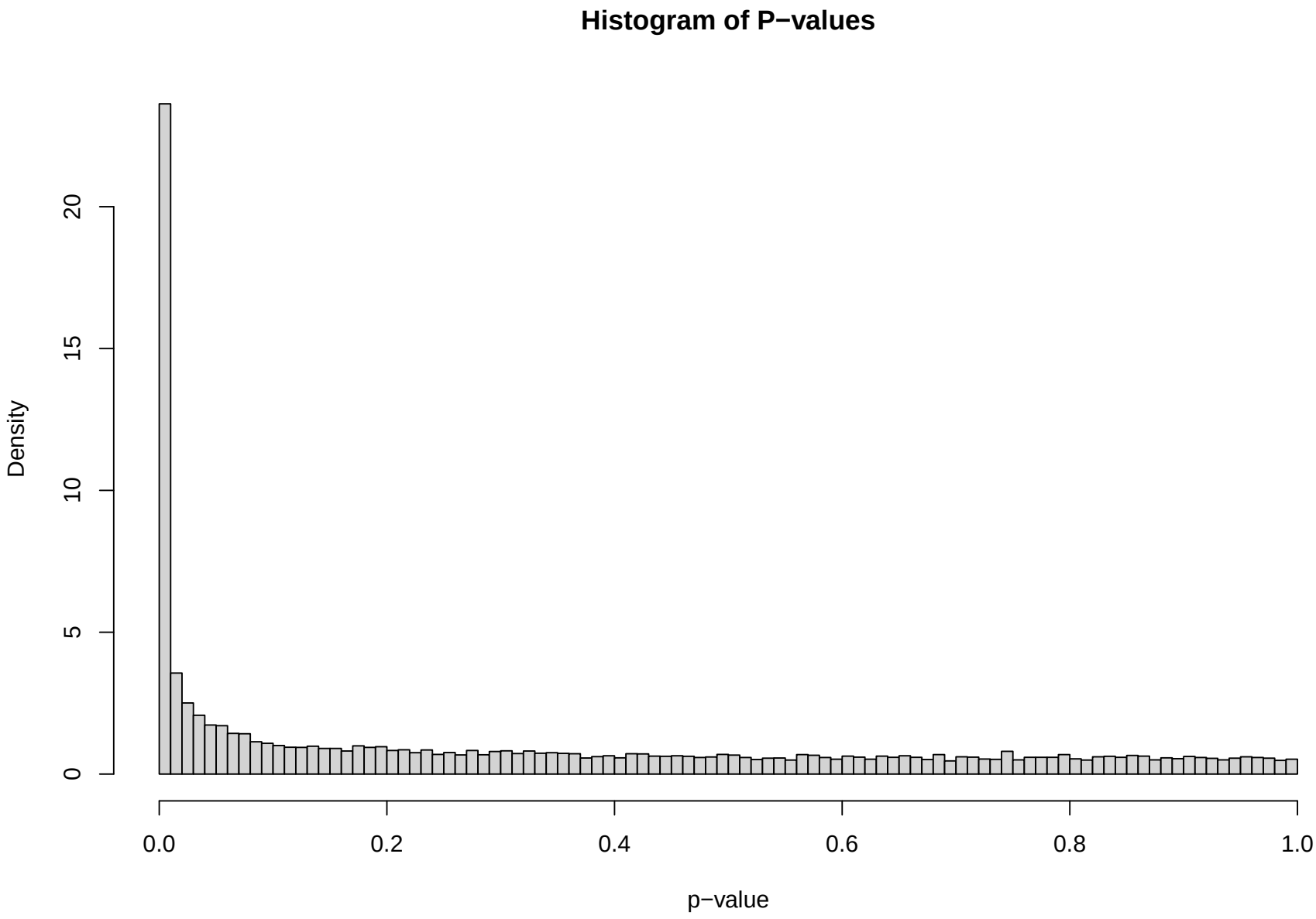


P-values under H_0

```
pval <- rnorm(nrow(res)) |> pnorm()
```



P-value Distribution of dds



Multiple Comparisons

- Probability refers to a single event
- More stringent threshold for *all* comparison
- Q-value: p-value adjusted to the new threshold

Confusion Matrix

Predict	H_0	H_1	Total
0	TN	FN	$n - R$
1	FP	TP	R
Total	n_0	n_1	n

Family-wise error rate (FWER)

$$\text{FWER} = P(\text{FP} \geq 1)$$

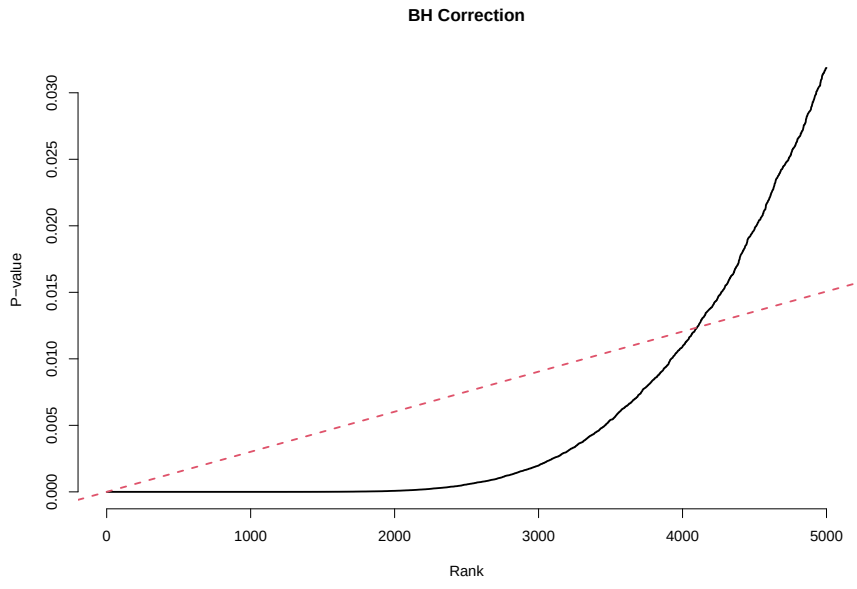
- *Bonferonni Correction*: $\text{FWER} \leq a \Rightarrow p_i \leq \frac{a}{n}$
- Q-value: $q_i = np_i$

False Discovery Rate

$$\mathrm{FDR} = E\left[\frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}\right]$$

- *Benjamini–Hochberg*: $\mathrm{FDR} \leq \frac{n_0}{n} a \leq a$
- Q-value $q_i = \min\{\mathrm{FDR}(a) \mid a \geq p_i\}$

Benjamini–Hochberg procedure




```
BH <- function(pval) {
    n <- length(pval)
    k <- order(pval)
    q <- pval[k] * n/(1:n)
    q <- rev(q) |>
        cummin() |>
        rev()
    q[order(k)]
}
```

Filtering Steps

- Marginal Independence: should not affect the null distribution
- Enrichment: we want u and t to be correlated in H_1 but not in H_0
- Correlation structure: multiple-comparison correction take advantage of the p-value correlation the filtering should not alter that much
- Threshold on variance/mean imply thresholds on logFC
- Better to filter on variance than mean (base rate can vary by a lot)
- Filtering on mean combats discreteness (low counts)

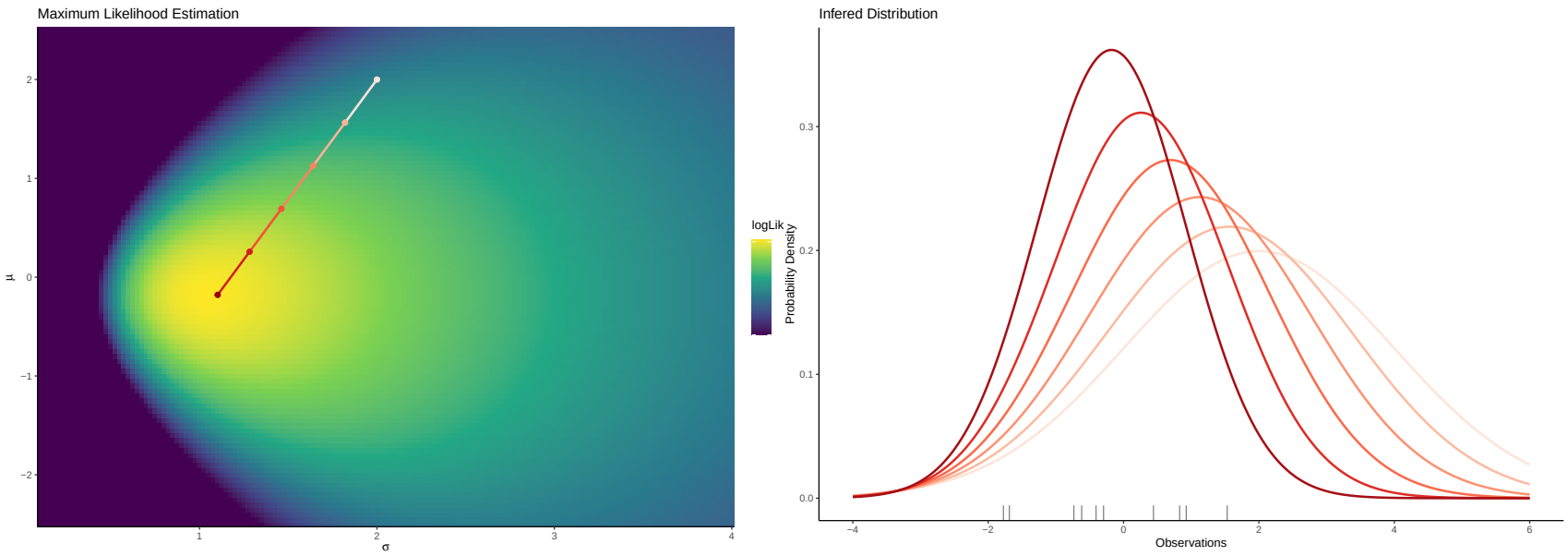
How we calculate p-values?

	log2FC	lfcSE	pval	padj
ENSG00000000003	0.38	0.10	0.0001	0.0011
ENSG00000000419	-0.20	0.11	0.0675	0.1859
ENSG00000000457	-0.04	0.14	0.8027	0.9040
ENSG00000000460	0.09	0.28	0.7410	0.8717
ENSG00000000971	-0.42	0.09	0.0000	0.0000
ENSG00000001036	0.24	0.09	0.0066	0.0302
ENSG00000001084	0.05	0.17	0.7619	0.8827
ENSG00000001167	0.50	0.12	0.0000	0.0002
ENSG00000001460	0.13	0.18	0.4729	0.6829
ENSG00000001461	0.04	0.10	0.6725	0.8289

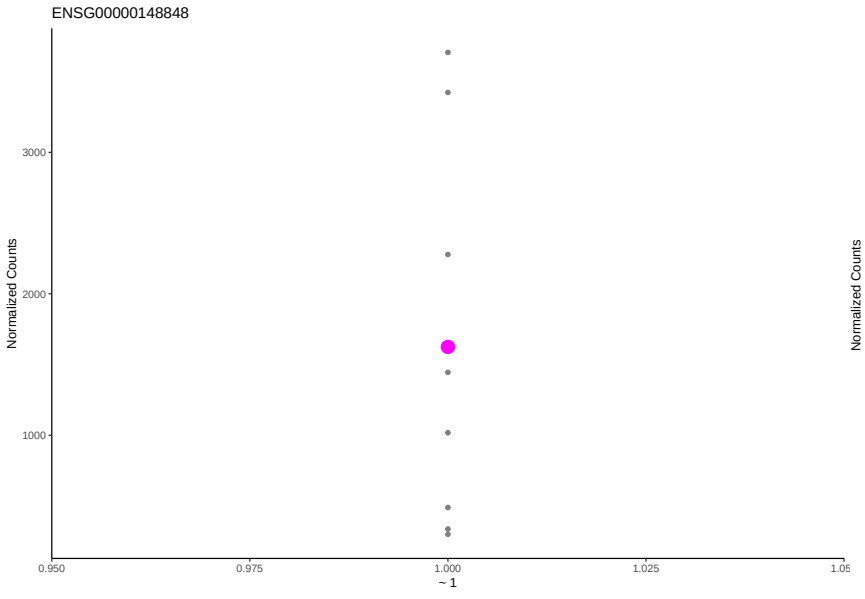
Generative Models

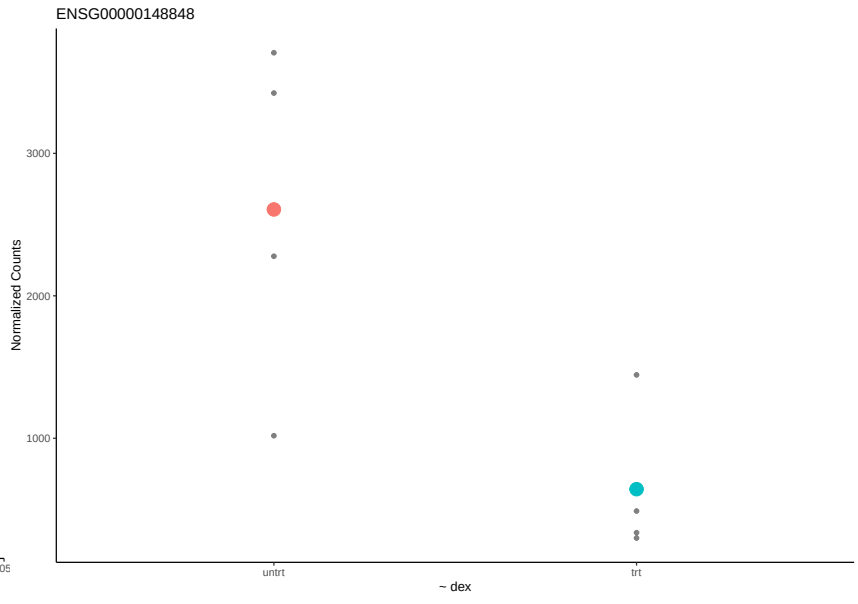
$$
\begin{aligned}
y &\sim f\left(\mu, \sigma^2\right) \\
\mu &= \eta(X\beta) \\
\beta &\sim \mathcal{N}(\hat{\beta}, \sigma^2_\beta)
\end{aligned}
$$

Maximum Likelihood Estimation

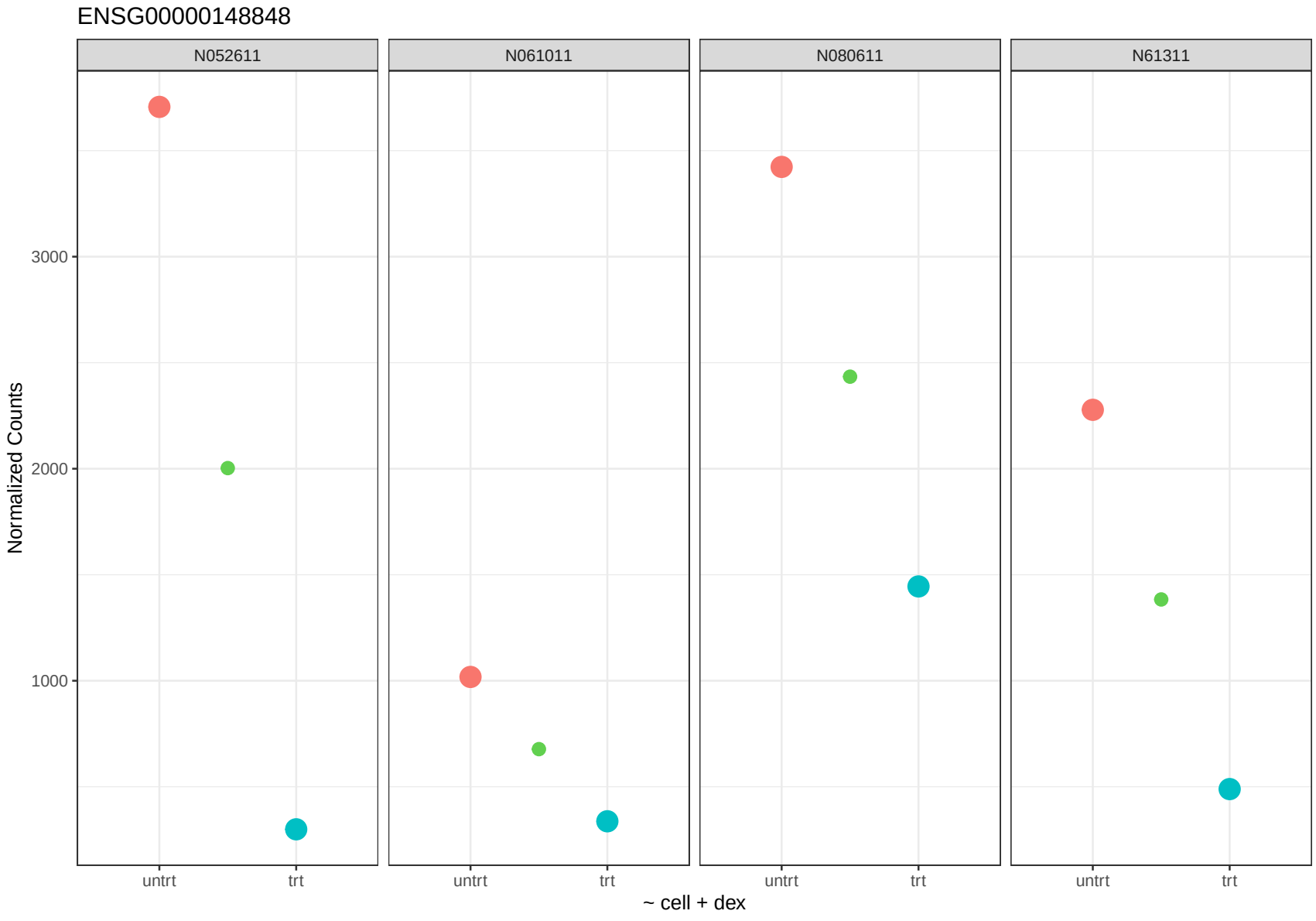


Design (X) - 1


ENSG00000148848
Normalized Counts
3000
2000
1000
0.950
0.975
1.000
1.025
~ 1


ENSG00000148848
Normalized Counts
3000
2000
1000
untrt
trt
~ dex

Design (X) - 2


ENSG00000148848
N052611
N061011
N080611
N61311
Normalized Counts
3000
2000
1000
untrt
trt
~ cell + dex

Predicting Mean vs Variance

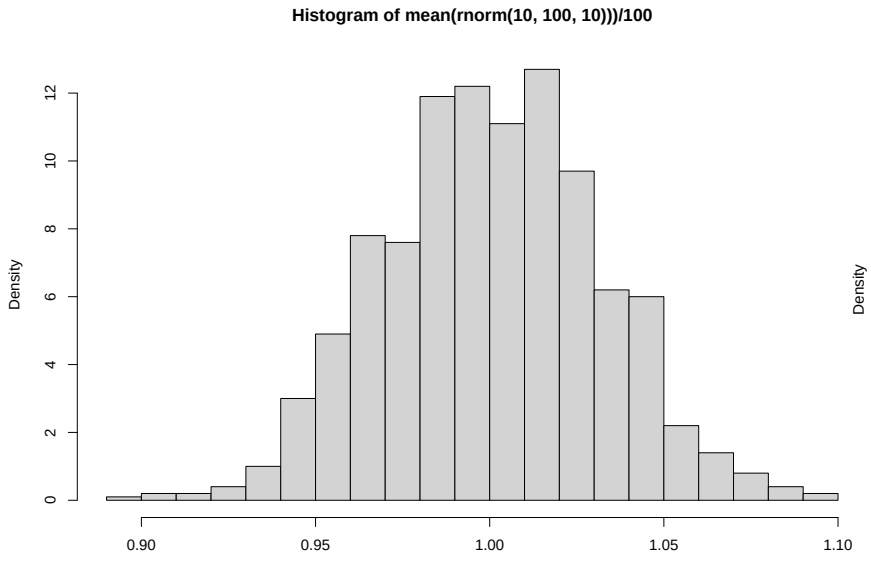


Figure 4: Mean

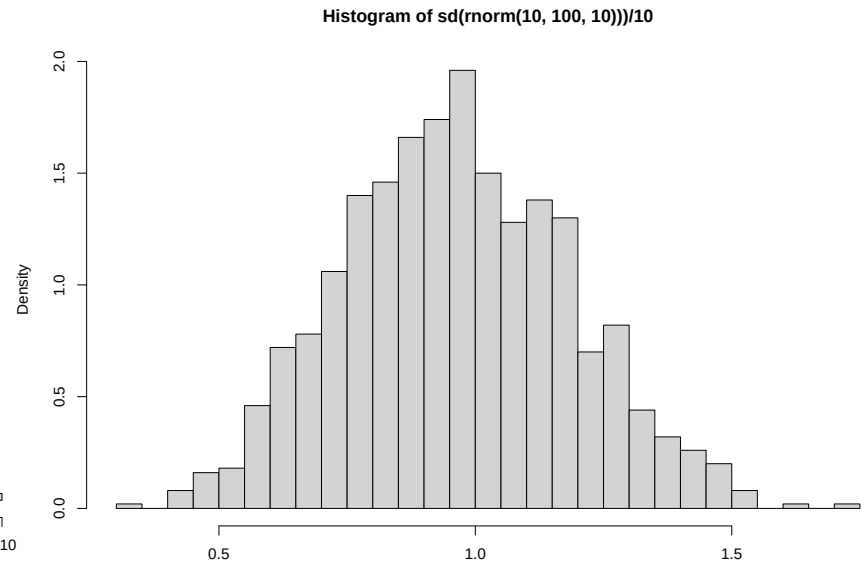
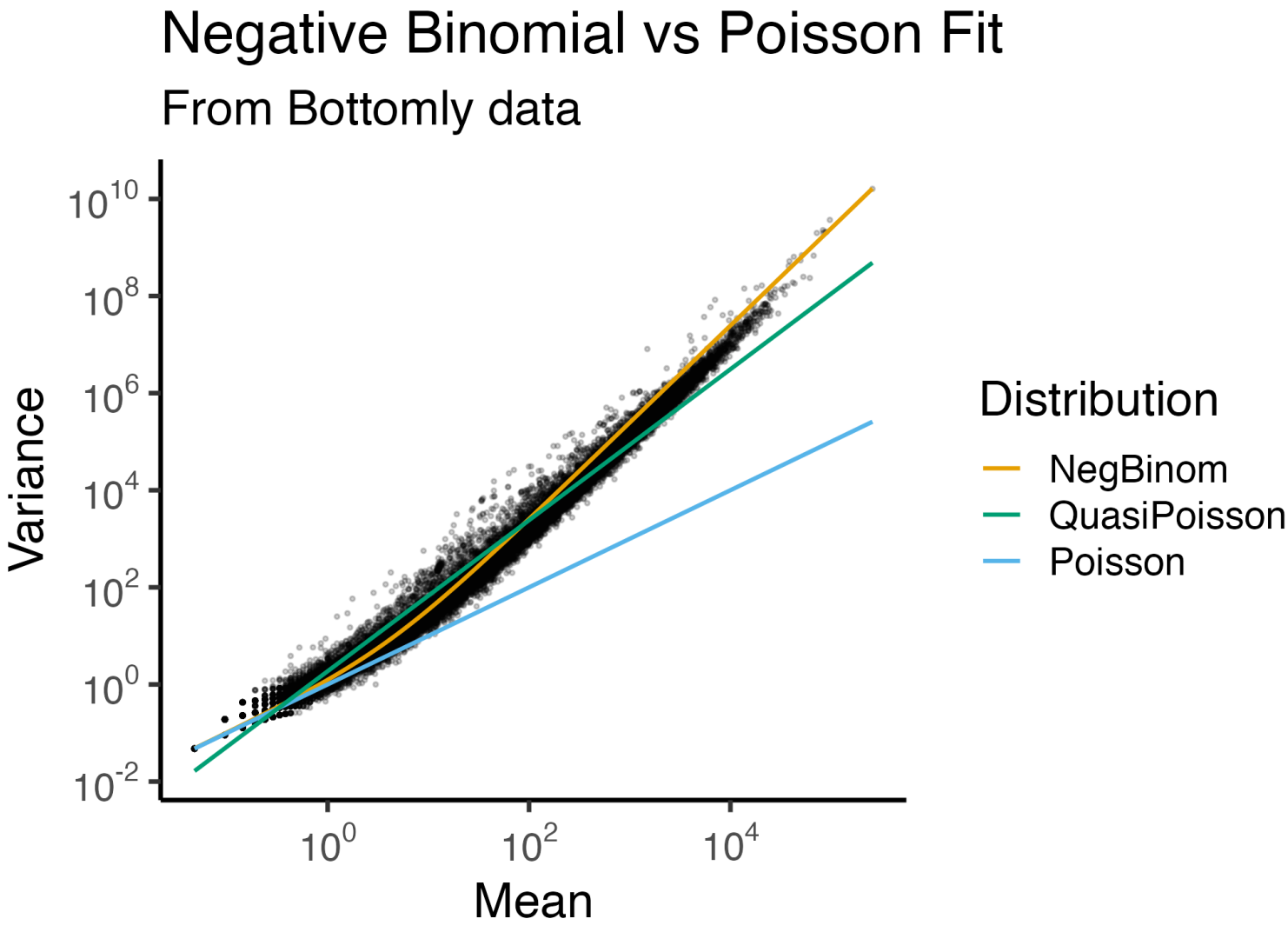


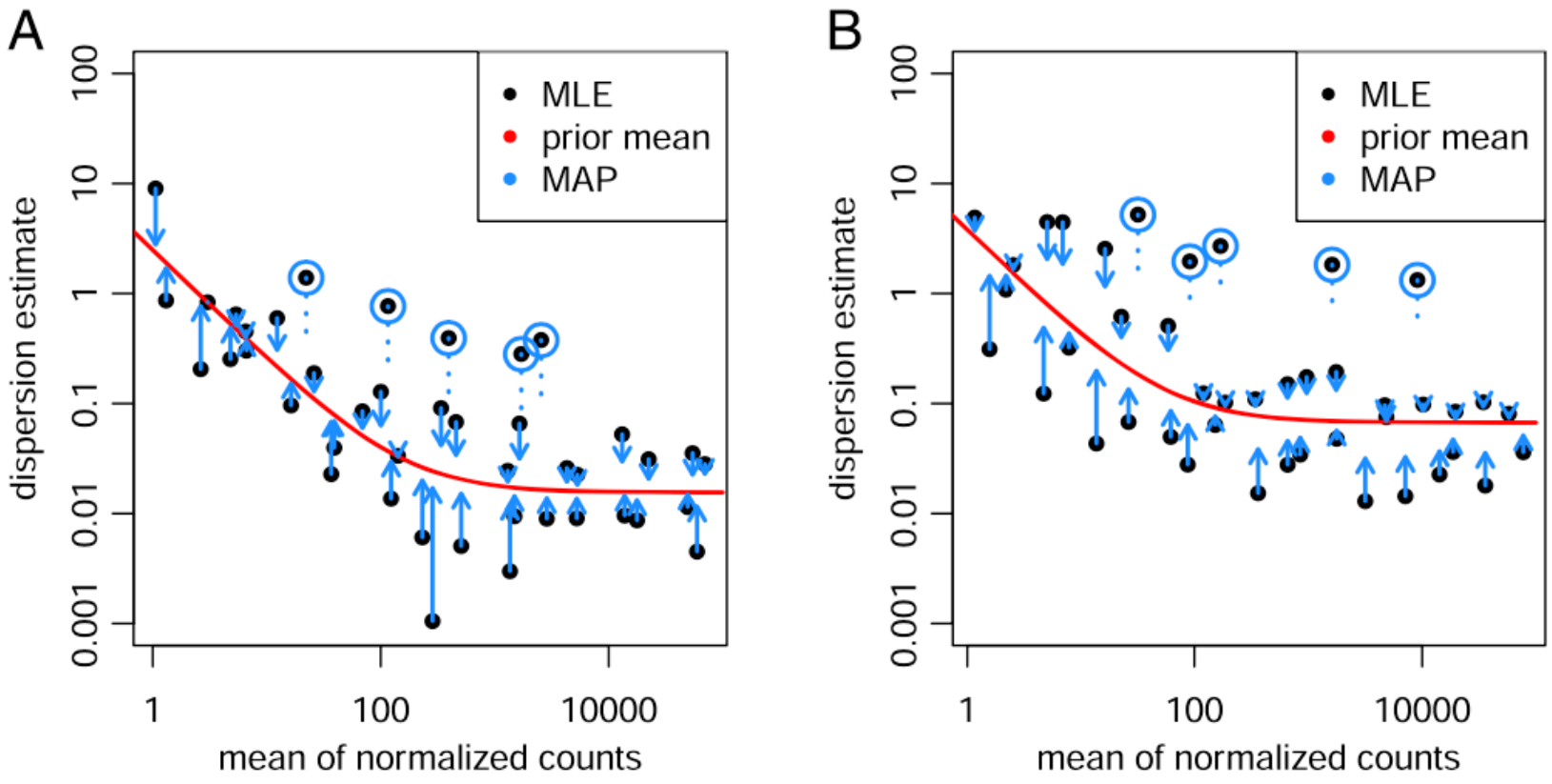
Figure 5: Variance

Overdispersion

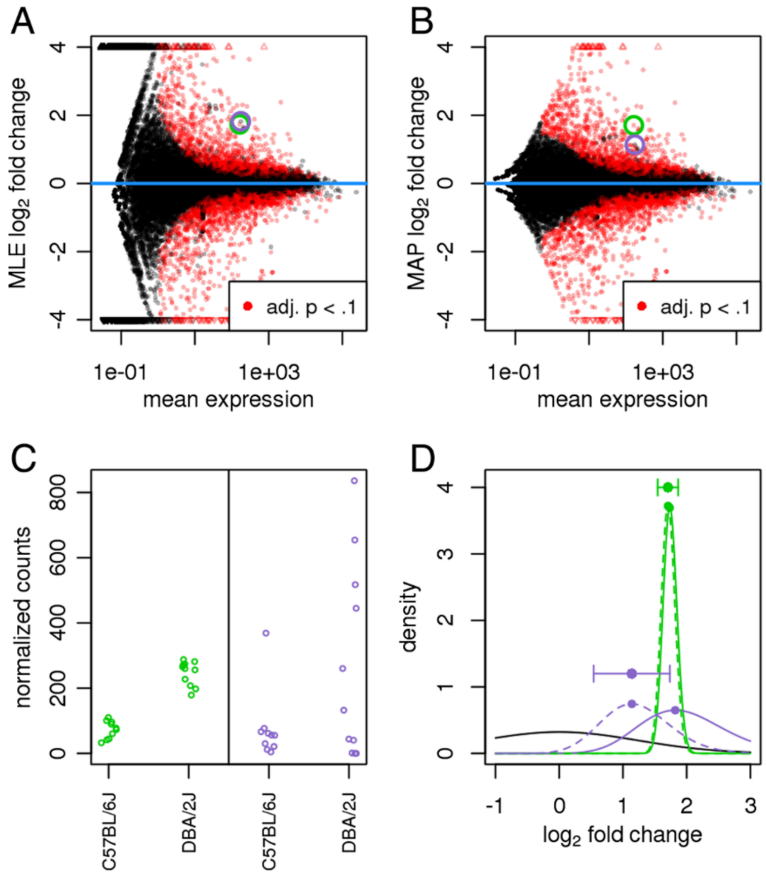


Partial Pooling

Partial Pooling



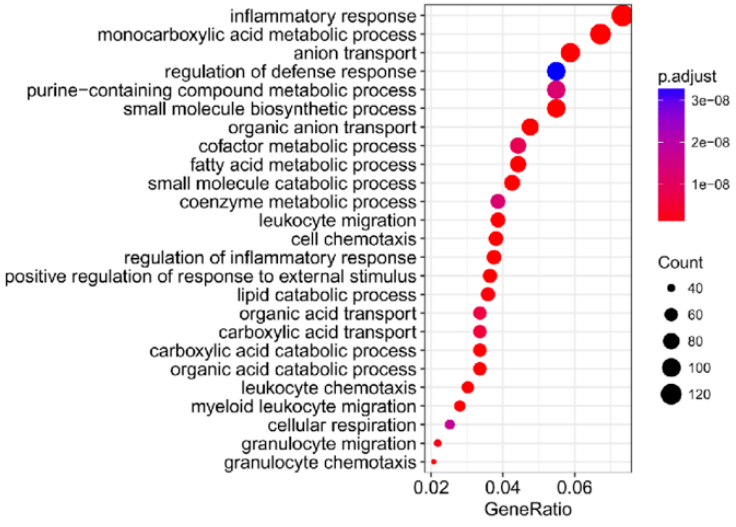
LFC Shrinkage



Pathways Enrichment

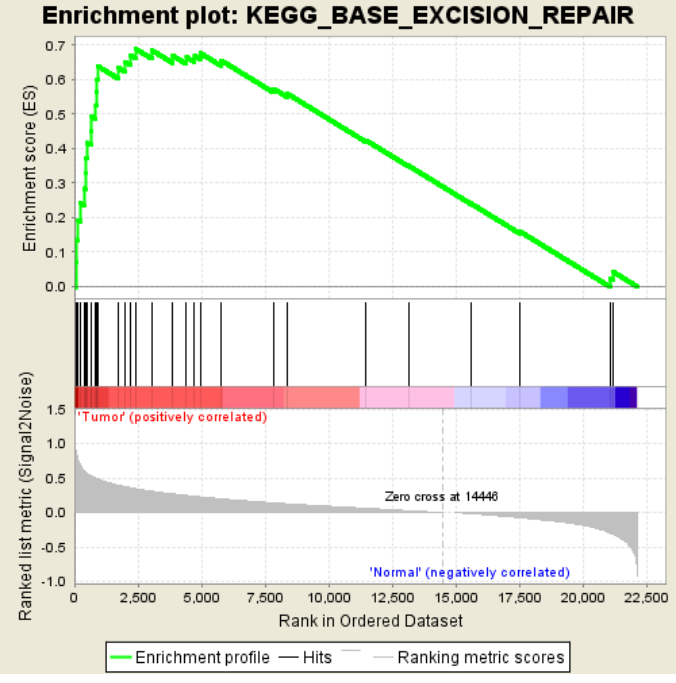
- Main Categories
  - *ORA*: over-representation analysis
  - *GSEA*: gene set enrichment analysis
  - *Network/Topology-based*
- Sources of Sets:
  - Gene Ontology
  - Kyoto Encyclopedia of Genes and Genomes
  - Molecular Signature DB
  - Reactome
  - many more...

Over-Representation Analysis




```
phyper(q - 1, m, n, k,
       lower.tail = FALSE)
```

GSEA




```
ks.test(x, x[pathway], ...)
```

Network/Topology Based Methods

Table 1 Overview of tested pathway enrichment methods

From: A comparative study of topology-based pathway enrichment analysis methods

Method	Null hypothesis	Gene *p*-value thresholding	Expression data	Pathway	R/Bioconductor
Pathway-Express	Competitive	Optional	No	Topology	`ROntoTools` 2.10.0
SPIA	Competitive	Yes	No	Topology	`graphite` 1.28.2
NetGSA	Self-contained	No	Yes	Topology	`netgsa` 3.1.0
topologyGSA	Self-contained	No	Yes	Topology	`topologyGSA` 1.4.6
DEGraph	Self-contained	No	Yes	Topology	`DEGraph` 1.34.0
CAMERA	Competitive	No	Yes	Membership	`limma` 3.38.3
CePa	Competitive	Yes	No	Topology	`CePa` 0.6
PRS	Competitive	Yes	No	Topology	`ToPASeq` 1.16.1
PathNet	Competitive	Yes	No	Topology	`PathNet` 1.22.0

All in vain?

From: Exaggerated false positives by popular differential expression methods when analyzing human population samples