UNIVERSITÄT LEIPZIG

Multiple Testing in RNA-Seq experiments

O. Muralidharan et al. 2012. Detecting mutations in mixed sample sequencing data using empirical Bayes.

Bernd Klaus
Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE),
Universität Leipzig

http://bernd.strimmerlab.org/

12 June 2012

1. Emperical Bayes Analysis of Microarray Data
2. False Discovery Rates in RNA Seq Experiments
3. Application: Modelling of Sequencing Error Rates
4. Sources of Variation for Sequencing Error Rates
5. Modelling the Variation
6. Results
   - Virus Data
   - Tumor Data
7. Conclusions

Empirical Bayes Analysis in High Dimensional Microarray Data

- Empirical Bayes analysis has a long "tradition" in microarray data
- It usually based on normal distribution assumptions
- However RNA Seq data is discrete ...
- Can methods from Microarray data be adapted?

Analysis Steps

Typical Workflow in a Nutshell

1. Normalize Data
2. Compute (**regularized**) t-scores between sample groups (e.g. healthy and tumor tissue)
3. Asses differential Expression $\Rightarrow$ **Multiple Testing**

Example: Golub-Data

REPORTS

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

Example: Golub-Data

- Data with $n = 38$ samples (Chips) and $p = 3051$ Genes
- Additional Information: 2 different Leukemia subtypes: ALL and AML
  - ALL: acute lymphoblastic leukemia ($n_0 = 27$)
  - AML: acute myeloid leukemia ($n_1 = 11$)
- Original aim of the study: Define a molecular signature of the subtypes (ALL, AML)

Differential Expression

- Given: $k \leq 2$ groups
- Which genes are differentially expressed between ALL and AML?
- How to quantify differentiall expression?
  1. Fold Change: Difference of means between two groups
  2. t-score: Standardized difference of means between two groups
  3. regularized t-score: t-Score with modified variance estimation, e.g. limma score ("Smyth - Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments 2004")
  4. Wilcoxen-score: Nonparametric version of the t-score
- Result: Ranking of genes according to differential expression

Problem: Multiple Testing

- Similar to ANOVA post-hoc tests, but usually a lot more tests are performed!
- Ex.: Two tests with Type I ("α –error") Error (reject H_0, although H_0 is true) of 0.05:

Multiple Testing, $\alpha = 0.05$

Prob of **no** false rejection in both tests:

$$0.95 * 0.95 = 0.9025 \quad < \quad 0.95$$

$\Rightarrow$ The more you test the more probable it is to falsely reject H_0!

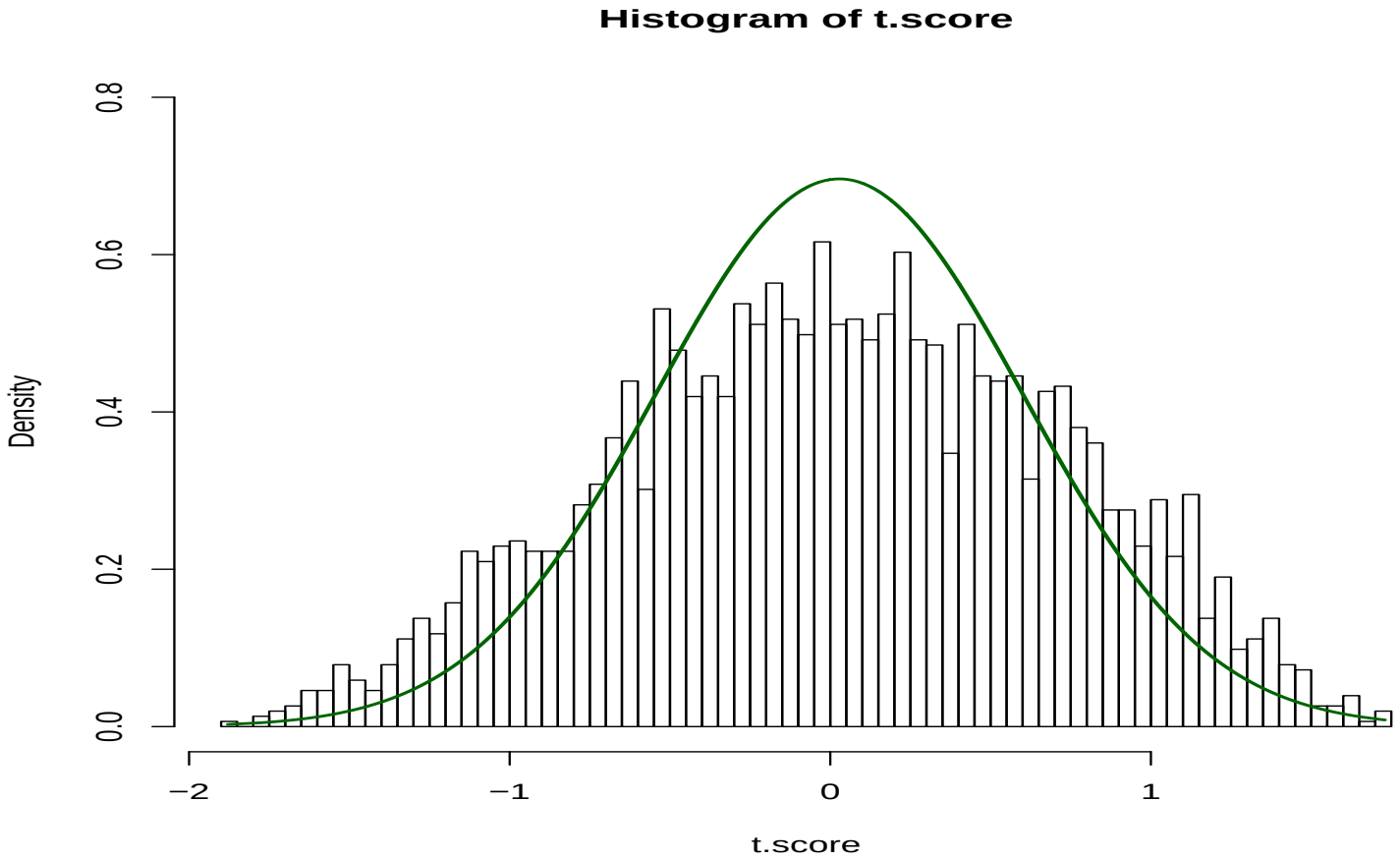
Empirical Bayes Analysis of the Golub Data I

Computing a t-score between the two Leukemia subtypes yields a ranked list of genes

- "C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds"
- "Macmarcks"
- "RETINOBLASTOMA BINDING PROTEIN" P48
- "TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) "
- "Inducible protein mRNA"
- "CCND3 Cyclin D3"
- "MYL1 Myosin light chain (alkali)"
- "SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)"
- ...

Empirical Bayes Analysis of the Golub Data II

In order to find significant genes "interesting" ones need to be seperated from "null" genes

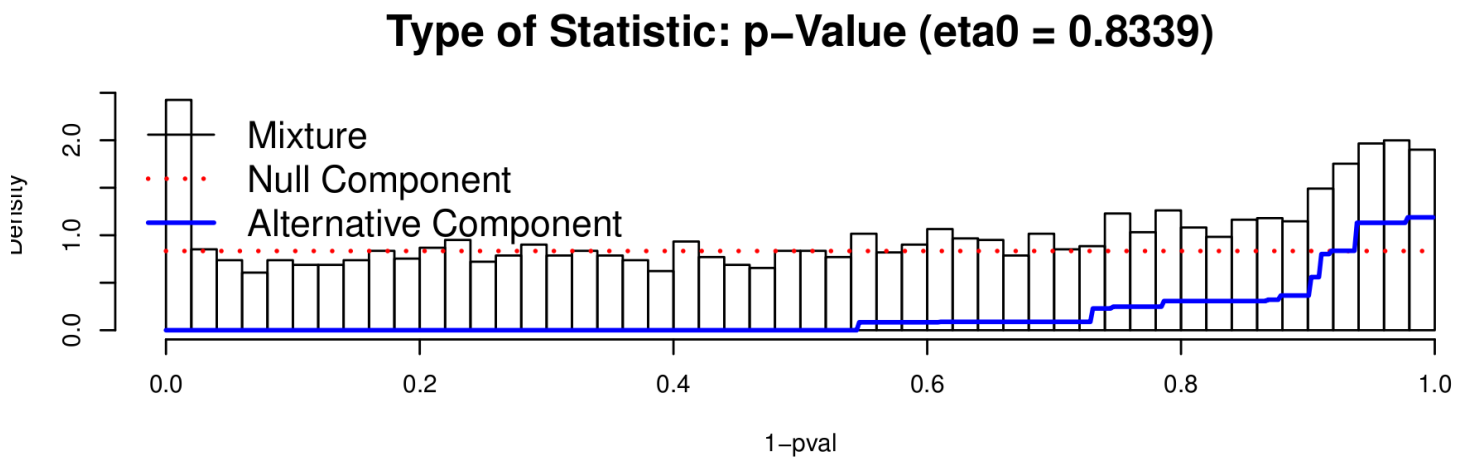


Empirical Null Modelling

- The green lines shows the density of a $N(0, 0.6^2)$ distribution which serves a model for the "null" cases – F_0
- It can be used to compute p-values: $pval = 2 - 2 * F_0(|x|)$
- They are uniformely distributed under the null!

Histogram of p-values

- There is a peak at 1, $\Rightarrow$ **a "signal" is present**
- There are differentially expressed genes!
- The peak at 1 is a technical artifact here



False Discovery Rates

Given

- the p-values
- the proportion of the null statistics η_0
- the overall marginal density $f(p)$

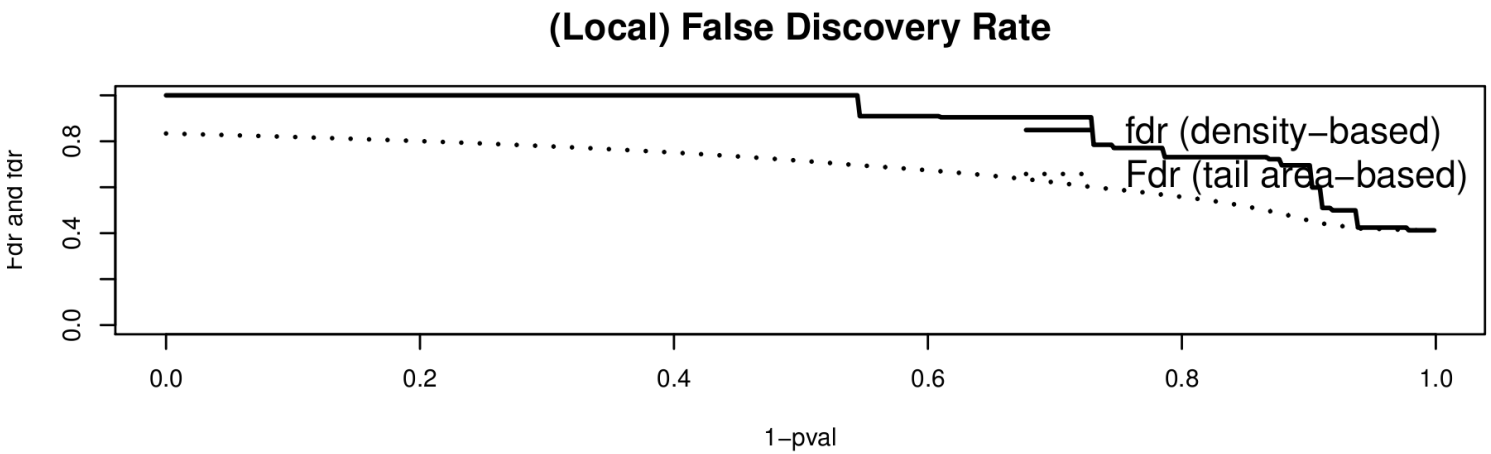
the local false discovery rate can be computed and a cutoff can be chosen:

local false discovery rate

$$\text{fdr(p)} = \text{P(“null”}|p) = \frac{\eta_0}{f(p)} = \frac{\eta_0}{f(p)}$$

Finally, include e.g. all genes with fdr < 0.2

Local False Discovery Rate



False Discovery Rates in RNA Seq Experiments

- In RNA-Seq experiments the uniformity under the null is lost!

discrete p-values correct null.pdf

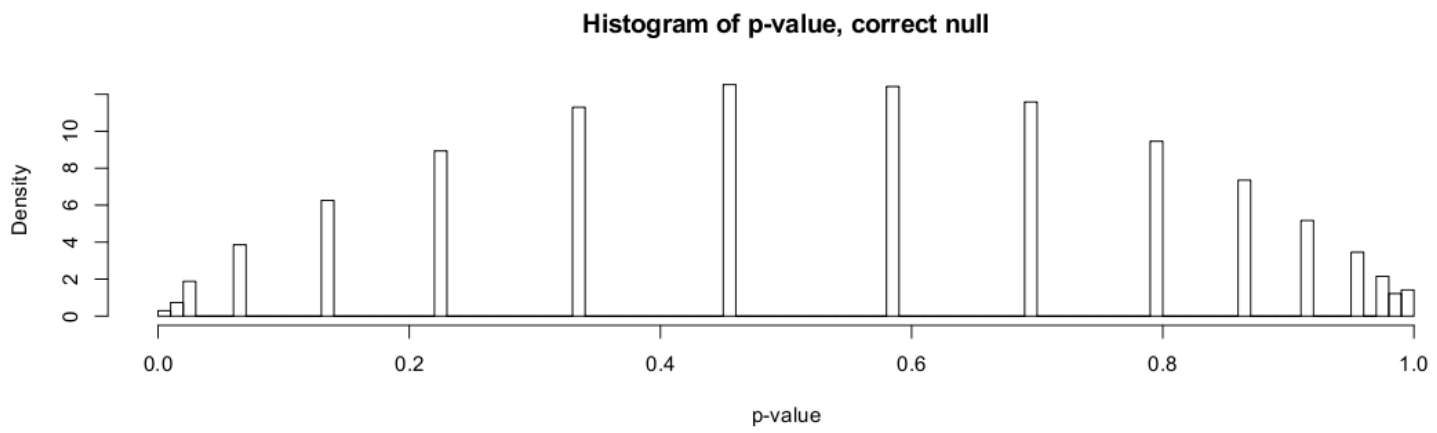


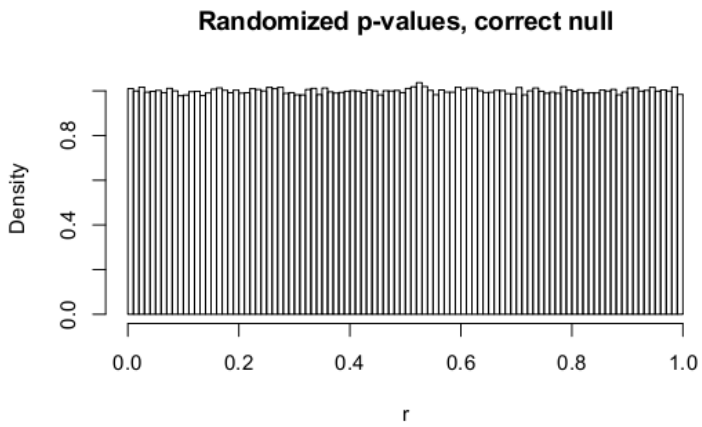
FIG 1. *p-values* $p_i = F_i(x_i)$, *where* $x_i \sim F_i = Poisson(10)$.

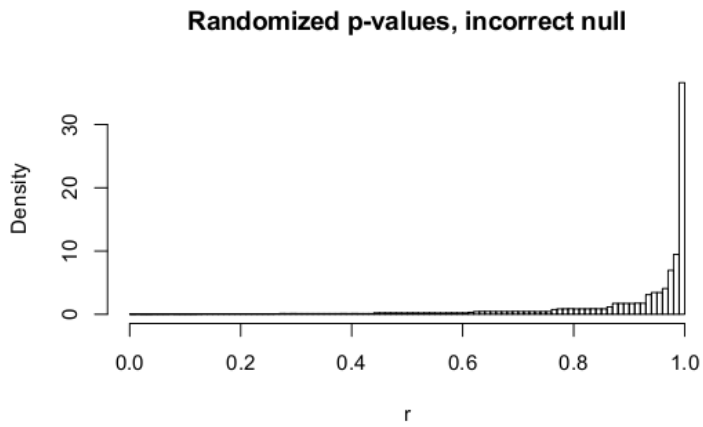
Solution: Randomized p-values

- Let $p_{(1)}, \ldots, p_{(n)}$ be the **ordered** p–values, the modified p-values $r_{(i)}$ are:

Randomized p–values

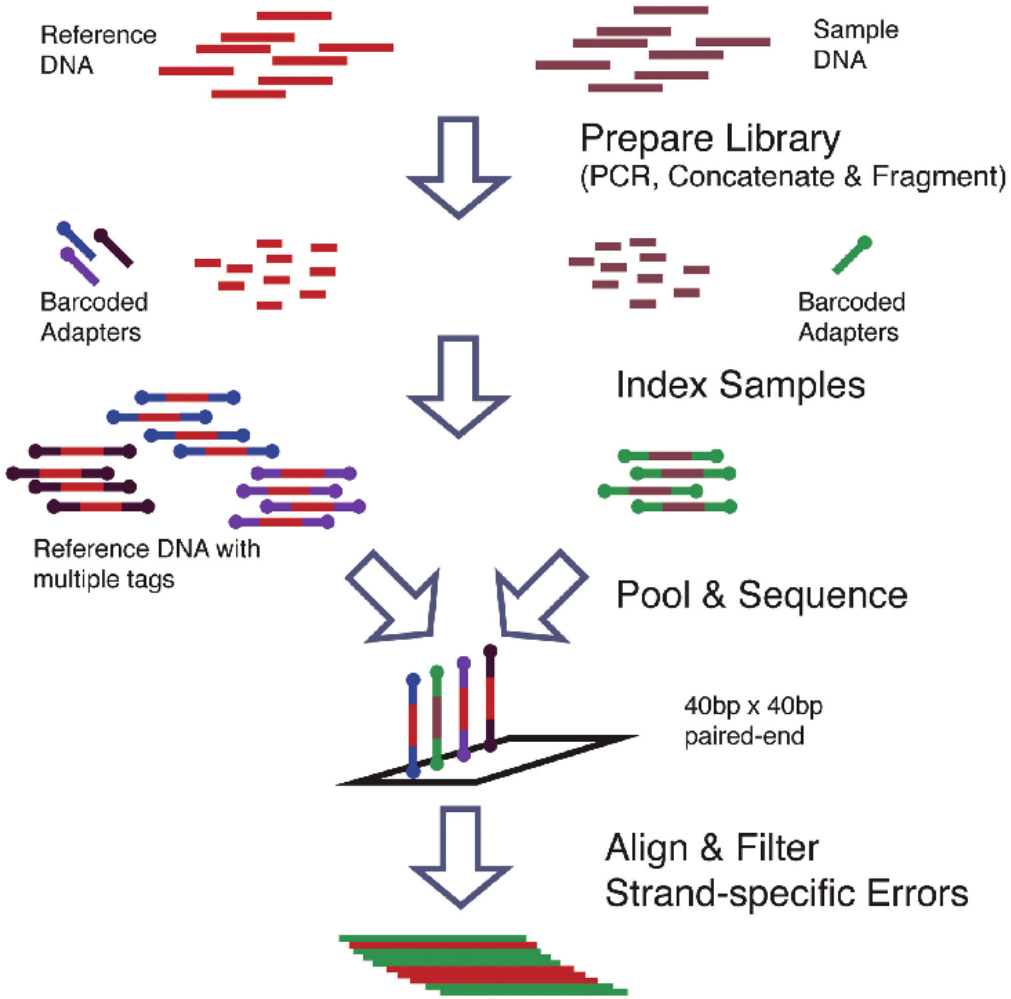
$$r_{(i)} = p_{(i-1)} + \text{Unif}(p_{(i-1)}, p_{(i)})$$


Randomized p-values, correct null
Density
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
r


Randomized p-values, incorrect null
Density
0
10
20
30
0.0
0.2
0.4
0.6
0.8
1.0
r

Experimental Design I: Finding Mutations in a Virus Sequence

- Starting point: a reference Virus sequence and a mutated sequence (synthetic data)
- Alignment yields non reference counts ("errors")
- $\Rightarrow$ There is an **error count** x_{ij} and a **sequencing depth** N_{ij} for each position i and each sample j
- Mutations in a sequence then appear as unusually large error rates x_{ij}/N_{ij}


Reference
DNA
Sample
DNA
Prepare Library
(PCR, Concatenate & Fragment)
Barcoded
Adapters
Barcoded
Adapters
Index Samples
Reference DNA with
multiple tags
Pool & Sequence
40bp x 40bp
paired-end
Align & Filter
Strand-specific Errors

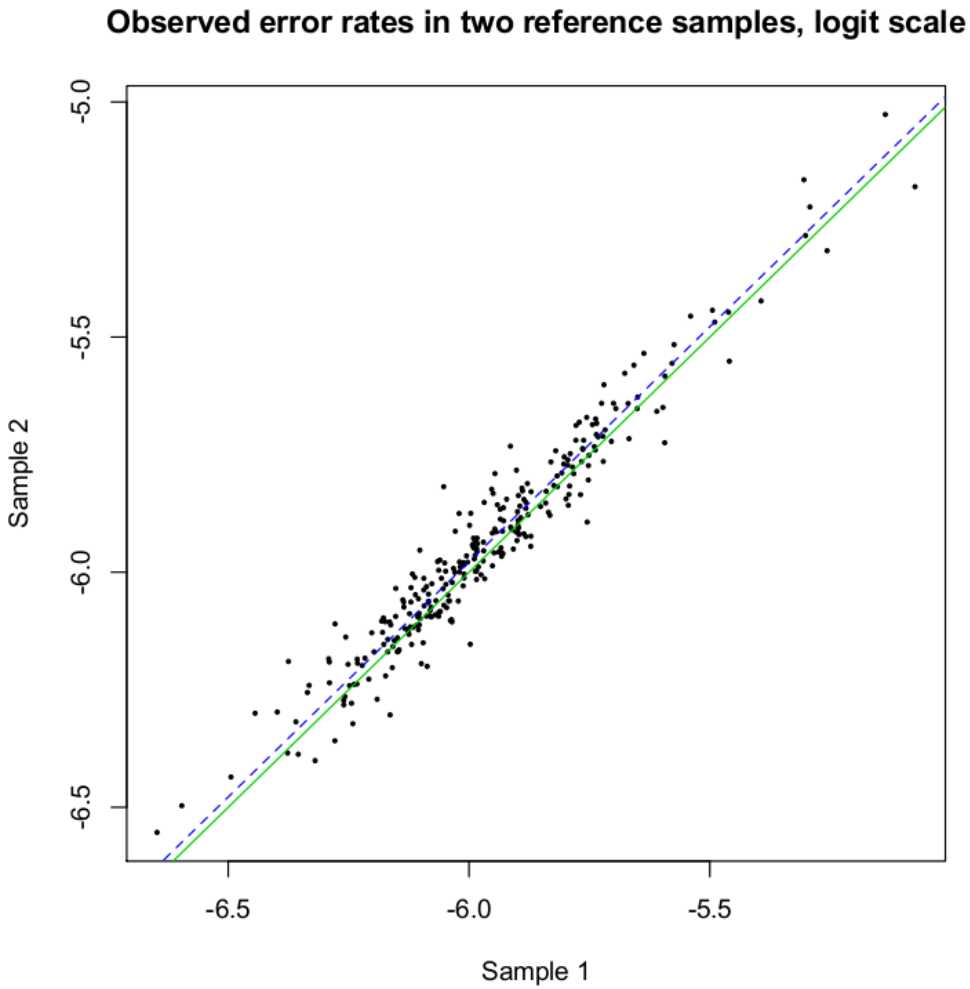
Experimental Design II: Comparison of Tumor and normal Tissue

- Cancer and healthy tissue from Lymphoma patients (paired samples!)
- Starting point: a tumor (x) and a control sequence (y) aligned to a reference
- $\Rightarrow$ As in the virus sample error rates x_{ij}/N_{ij} and y_{ij}/N_{ij}
- Biologically interesting positions appear as significant differences between x_{ij}/N_{ij} and y_{ij}/N_{ij}

Sources of Variation

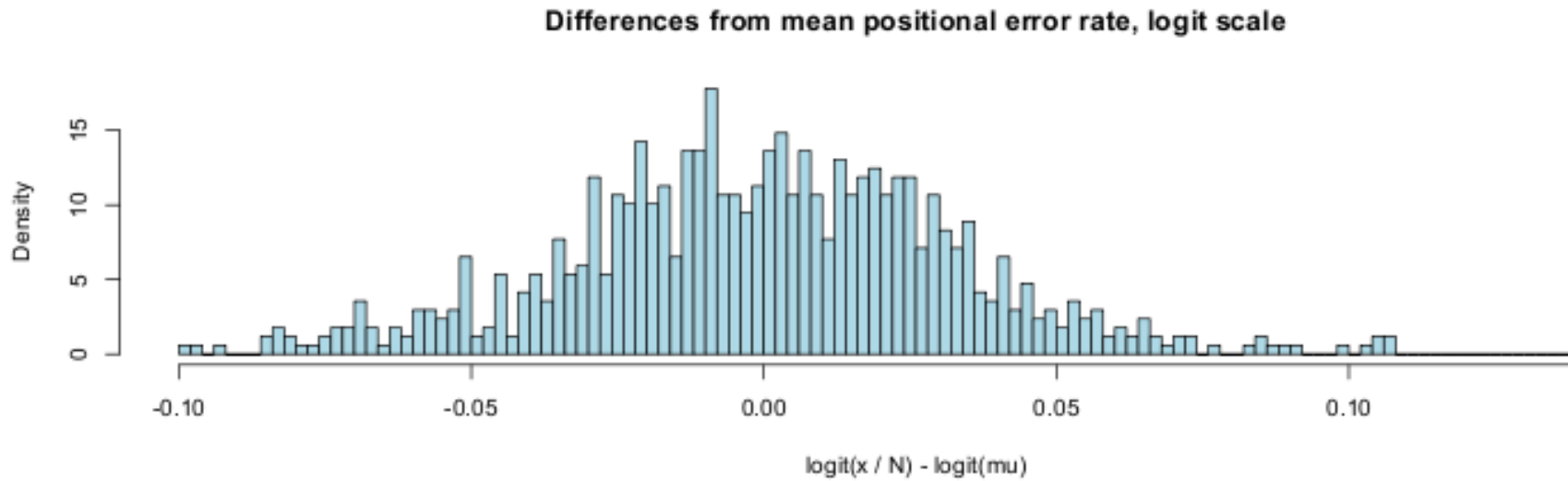
1. **Natural:** Finite sequencing depth N for each genome position
   - $x \sim \text{Binomial}(N, p)$
   - The error rate x has a "natural" variation due to finite sequencing depth!
2. **Positional:** The sequencing error rate will vary from position to position
   - Aggregate across samples to estimate baseline error
3. **Across samples:** The sequencing error rate will vary across samples
   - Aggregate across positions to estimate sample effects

Variation - tumor data - two reference samples



Distribution of error rates

- Figure shows difference of observed positional error rate $\frac{x_{ij}}{N_{ij}}$ and mean positional error rate of reference samples $\mu_i = \frac{1}{3} \sum_{j \in \text{ref-samples}} \frac{x_{ij}}{N_{ij}}$ on the logit scale:
- logit $\frac{x_{ij}}{N_{ij}}$ − logit μ_i



Statistical model for the virus data

This suggests the following model for the virus data

Model for the virus data

$$\text{logit } p_{ij} \sim N(\text{logit } \mu_i + \delta_j, \sigma_j^2)$$
$$x_{ij}|p_{ij} = \text{Binomial}(N_{ij}, p_{ij})$$

- p_{ij} - postional error rate for sample j
- μ_i = postional error rate
- δ_j = sample specific error rate bias (constant across positions)
- σ_j = sample specific noise (constant across positions)

$\Rightarrow$ **Now we can compute p-values, fit a marginal density $f(p)$ and compute false discovery rates!**

A similar model holds for the tumor data

Model for the tumor data with matched normals

$$\text{logit } p_{ij} \sim N(\text{logit } \mu_i + \delta_j, \sigma_j^2)$$
$$\text{logit } q_{ij}|p_{ij} \sim N(\text{logit } p_{ij} + \eta_j, \tau_j^2)$$
$$x_{ij}|p_{ij}, q_{ij} = \text{Binomial}(N_{ij}, p_{ij})$$
$$y_{ij}|p_{ij}, q_{ij} = \text{Binomial}(N_{ij}, q_{ij})$$

- p_{ij} / q_{ij} - postional error rate for normal / tumor sample j
- μ_i = postional error rate
- δ_j / η_j= sample specific error rate bis (constant across positions)
- σ_j / τ_j = sample specific noise (constant across positions)

$\Rightarrow$ **We can fit a null distribution for $\frac{y_{ij}}{N_{ij}} | \frac{x_{ij}}{N_{ij}}$, compute p-values, fit a marginal density $f(p)$ and compute false discovery rates!**

p values for virus data

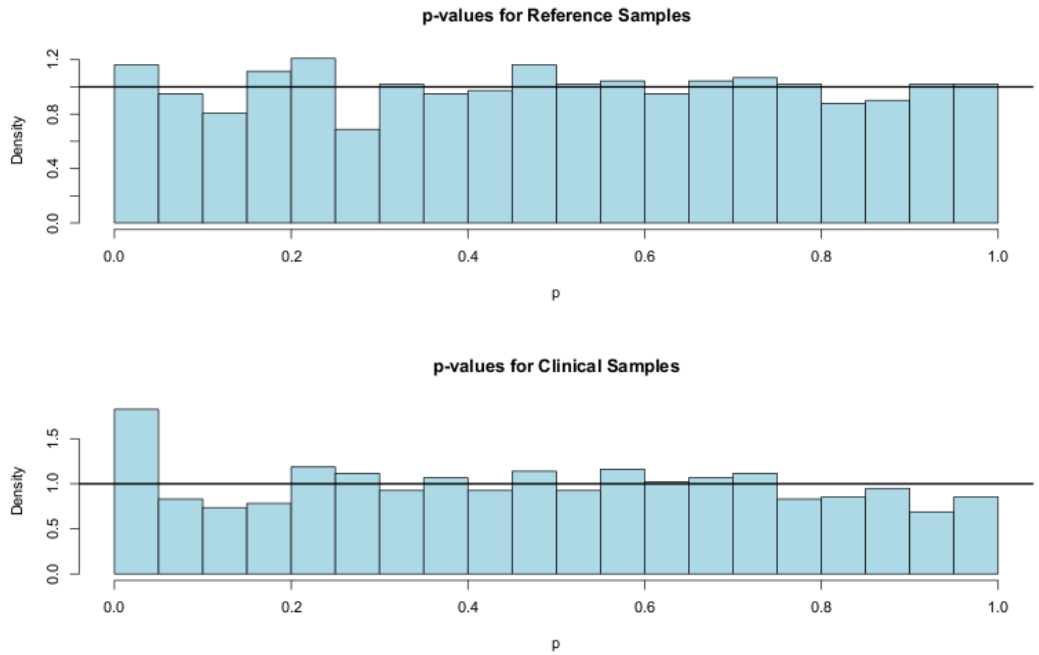


FIG 5. *Histogram of p-values for the virus data, reference samples (top plot) and clinical samples (bottom plot).*

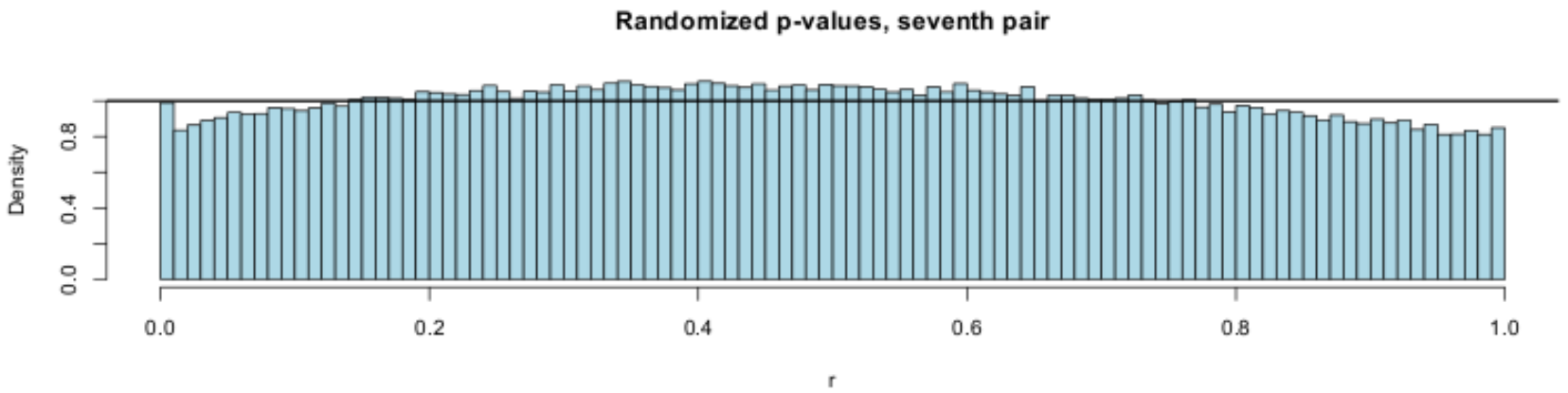
p Discoveries for the virus data

	Our method, $f\hat{d}r \leq 0.1$	Our method, $f\hat{d}r \leq 0.01$	Flaherty et al.
True Positives (of 42)	42	39	42
False Positives	1	0	10
Power	**100%**	**93%**	**100%**
False Positive Rate	**2.32%**	**0%**	**19.23%**

TABLE 1

Detection results on clinical samples of the synthetic virus data.

p values for the tumor data - sample pair 7



- The underdispersion in the figure shows that the null distributions are systimatically to wide !
- Use the normal transform $\Phi^{-1}(r_{ij})$ to obtain z-values and fit an empirical null

p values for the tumor data - sample pair 7 - empirical null

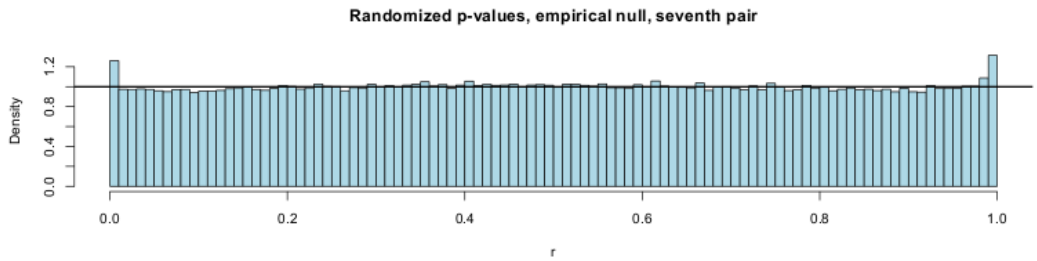


FIG 9. *Empirical null randomized p-values $\tilde{r}_i$ for the seventh normal tumor pair. The empirical null yields much more uniform p-values (compare to Figure 7).*

- much better, but "enriched" near both 0 and 1
- ... the logit scale "exaggerates" differences between p_{ij} and q_{ij} near 0 and 1
- A difference of 0.002 might correspond to 0.25 on the logit scale

Finding mutated positions

- Estimate the **change in error rate** Δ_{ij} at each position for each sample

Change in error rate given the data

$$\Delta_{ij} = P(p_{ij} \neq q_{ij}|x, y)\left(\frac{y_{ij}}{N_{ij}} - \frac{x_{ij}}{N_{ij}}\right)$$
$$= \text{fdr}_{ij}\left(\frac{y_{ij}}{N_{ij}} - \frac{x_{ij}}{N_{ij}}\right)$$

- A position with $\text{fdr}_{ij} < 0.1$ and $|\Delta_{ij}| > 0.25$ will be called interesting

Comparison to an earlier analysis of the tumor data

- 427 out of ca. 309 000 positions are called mutated
- 22% are in repetitive regions compared to 36% from an earlier analysis of Natsoulis et. al. (2011)
- **Repetitive regions provide problems e.g. for mapping step and have are large false positive rate**
- $\Rightarrow$ They are often excluded before searching for mutated regions
- Proposed method has higher number of calls in non-repetitive regions than Natsoulis et. al. (2011)
- $\Rightarrow$ may indicate a higher power of the empirical Bayes method

Conclusions

- Empirical Bayes ideas can be adapted to RNA-Seq data
- Methods for continuous data can be applied to randomized p-values
- The fitting process however is complicated