Simulating High-Dimensional Molecular Data

Axel Benner

Division of Biostatistics, German Cancer Research Center (DKFZ)
Heidelberg, Germany

September 18, 2017

STRATOS
INITIATIVE

Topic Group 9: Subtopics

(In context of large number of predictor or outcome variables[1])

1. Data pre-processing
2. Exploratory data analysis
3. Data reduction
4. Multiple testing
5. Prediction modeling/algorithms
6. Comparative effectiveness and causal inference
7. Design considerations
8. **Data simulation methods**
9. Resources for publicly available high-dimensional data sets

[1] Number of variables **p** is much larger than sample size **n**

Simulation experiments

- to study efficacy of algorithms /statistical methods over a range of differing situations
- to identify appropriate algorithms / statistical methods in specific situations
- to perform sample size / power calculation

Issues specific to high-dimensional data

- Underlying (biological) mechanism not well understood
- Difficult to simulate realistic correlation structure and suitable multivariate distributions

Subtopic 8: Data Simulation Methods

Typical Approaches

- Simulations based on assumed distributions (e.g. Poisson or negative binomial for count data)
- Simulations based on assumed distributions, using extracted parameters from pilot data
- Simulations using real data

Note:

- The way in which data are generated has a strong impact on the results of the subsequent statistical analyses
- Simulation techniques with completely synthetic data cannot capture the complex correlation structure among covariates in high-dimensional data

Methylation: Array Data

Methylation:
Infinium HumanMethylation450 BeadChip (Illumina)

The methylation status of roughly 485000 CpGs is derived by measuring the intensities of methylated (M) and unmethylated alleles (U) at each CpG site.

- Beta value: $beta_j = M_j/(M_j + U_j)\ ,\ j = 1, \ldots, p$
- Beta distribution seems "natural" since beta values represent proportions between 0 and 1.

Checking distributional assumptions

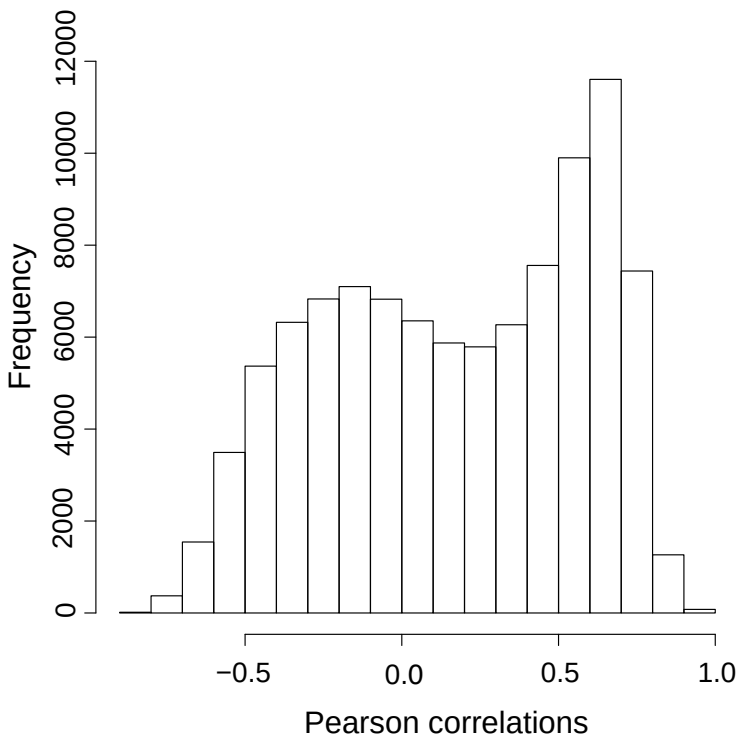
If the intensities M and U are independent, gamma distributed random variables with the same scale parameter

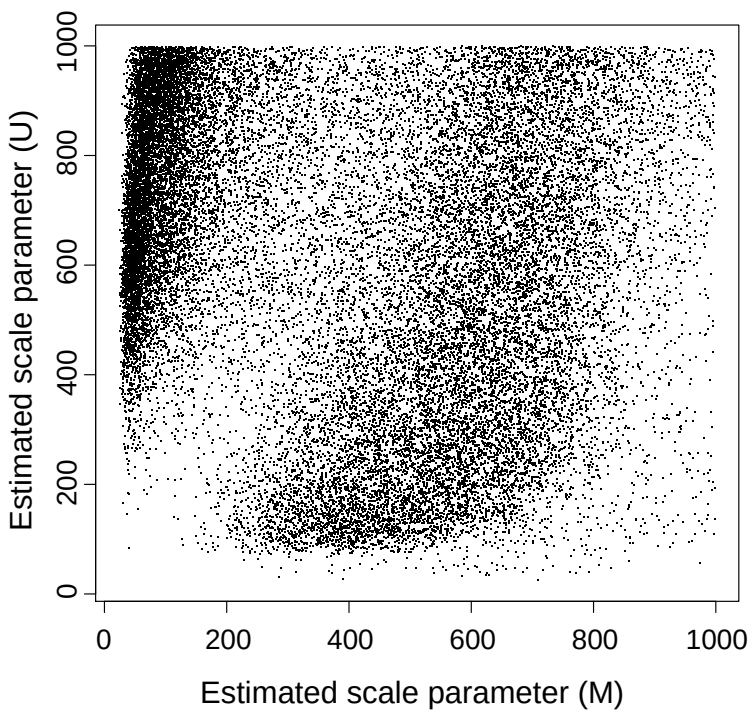
$\Longrightarrow$ beta values $(M/(M+U))$ are beta distributed.

Checking distributional assumptions

If the intensities M and U are independent, gamma distributed random variables with the same scale parameter

$\Longrightarrow$ beta values $(M/(M+U))$ are beta distributed.





Plasmode Simulation

Plasmode (from plasm=form, and mode=measure)

- is a real (i.e., from actual biological specimens) data set for which some aspect of the truth is known (Mehta et al., Physiological Genomics 2006)

Approaches

- Manipulate the biological samples (e.g. Affycomp's spike-in benchmark data (Irizarry et al., Biostatistics 2003))
- Permute samples of real datasets to generate null distribution; add 'realistic effect'

Advantage

- Distributions / correlations are taken directly from real data

Input or Output?

- Molecular data as the dependent variables.

  Univariate Screening:

$$X_j = Model(age, gender, \ldots), \; j = 1, \ldots, p$$

- Molecular data as the explanatory variables.

  Multivariable Regression Model:

$$Y = Model(X_1, \ldots, X_p, age, gender, \ldots)$$

Plasmode Simulation: Confounding Variables

Cohort Study with High-Dimensional Confounding

Since treatments are not randomized, addressing confounding is the primary methodological challenge

Objective:

- To compare the performance of
  - high-dimensional Propensity Score (hd-PS) variable selection
  - Ridge regression of the outcome on all potential confounders
  - Lasso regression of the outcome on all potential confounders
- The goal is maximum reduction in confounding bias

Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Analysis 2014; 72: 219-226.

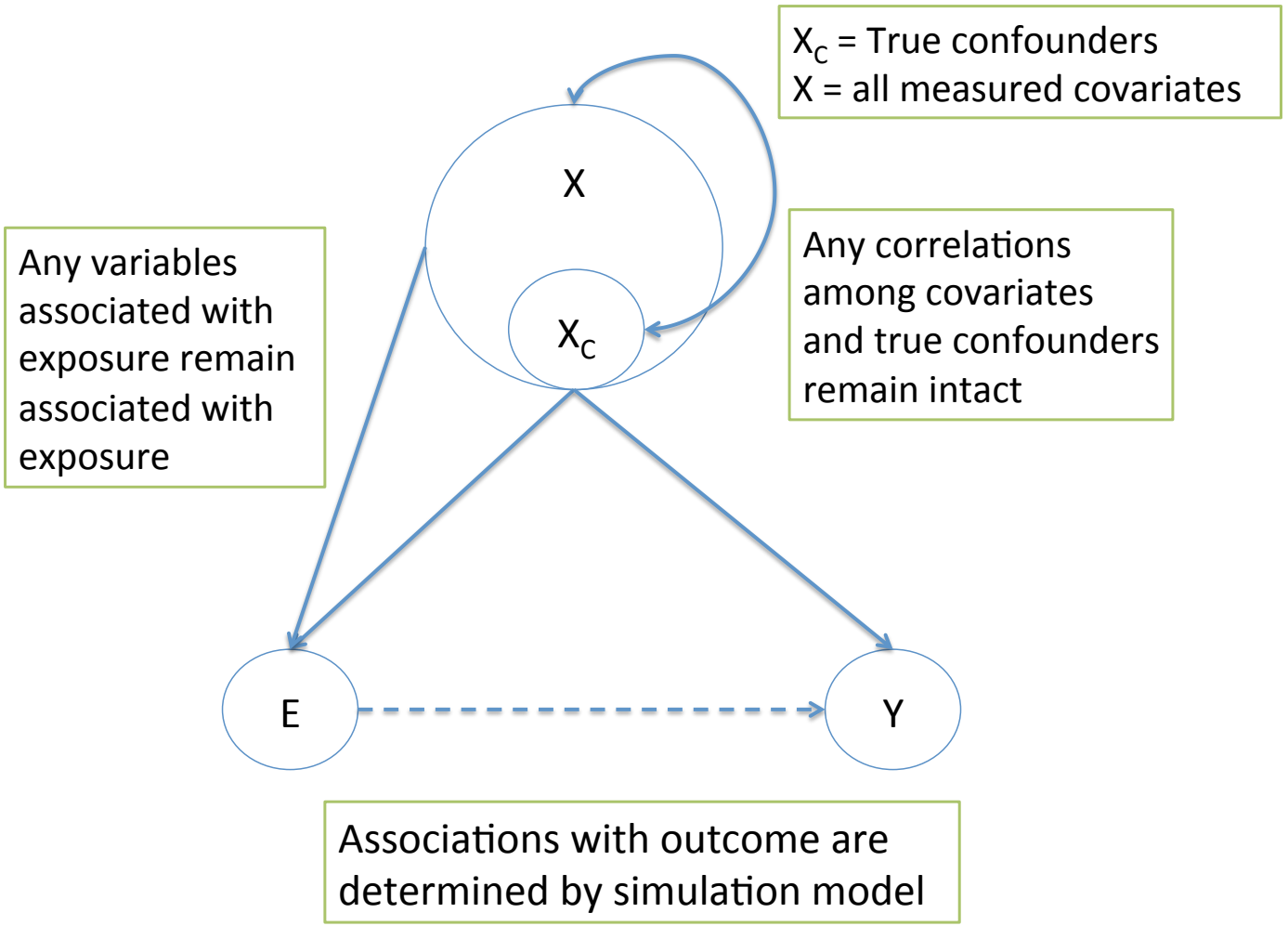
Plasmode Simulation: Confounding Variables

- Sample with replacement from cohort data to get desired sample size n and exposure prevalence
- Calculate $p_i = P(Y_i = 1|E_i, X_{ic}),\ i = 1, \ldots n$, using investigator-specified outcome-generating model
- Simulate binary outcome status according

$$Y_i^s \ \sim \ B(1, p_i),\ i = 1, \ldots n$$

$\Longrightarrow$ Correlations among exposure, true confounders, and other covariates remain unchanged.

Plasmode Simulation: Confounding Variables



Plasmode Simulation: Variable Screening

Rank Based Sampling: Sample from real data, incorporate covariable effects using ranks.

1. Draw sample of size n at a CpG site
2. Construct a linear predictor based on covariates $\mathbf{X}$ (fixed effect) and $\mathbf{Z}$ (random effect):

$$\eta_i = x_i\zeta_1 + \varepsilon(z_i\zeta_2), \qquad i = 1, ..., n$$

3. Assign methylation value to patient i using the rank of his individual η_i within the linear predictor sample η.

$\Longrightarrow$ Distribution of the methylation data is unchanged, but samples with higher values of $\mathbf{X}$ will tend to have higher methylation values at affected CpG sites.

Saadati M, Benner A. Statistical challenges of high-dimensional methylation data. Statistics in Medicine 2014; 33: 5347-5357.

Plasmode Simulation: Animal Study

Filtered RNAseq data (strain B6 vs. strain D2; Illumina)

Two factorial design (experiment, strain).

1. Analyse with edgeR (glm approach)) $\Rightarrow$ logFCs, q-values
2. Build set of effects
   - Select p_1 transcripts from total p, e.g. with $q < 0.05$
   - Set S_1: Sample w/o replacement $s = \pi p$ from p_1, $s < p_1$; π prop diff expr
3. Generate a partition of samples:
   - Select the samples from 'reference' strain B6
   - Within each of the experiments select two samples and randomly assign 'group' A or B
4. Add effects to group B:
   - Compute log-transform. of counts (c): $z = log2(c + 1)$ for samples in B
   - Add logFC of set S_1 to z of corresponding differentially expressed genes in samples labeled B
5. Back-transform values obtained in (4): $c = 2^z - 1$
6. Repeat n times step (2) through (5)

Reeb P, Steibel J. Evaluating statistical analysis models for RNA sequencing experiments. Frontiers in Genetics 2013; 4: 178.

Plasmode Simulation: Sample Size Calculation

Identification of prognostic biomarkers for time-to-event outcomes

We consider two proposals

- Tibshirani R. A simple method for assessing sample sizes in microarray experiments. BMC Bioinformatics 2006, 7: 106.
  - Uses a permutation-based algorithm using pilot data
  - Implemented in R package *samr*

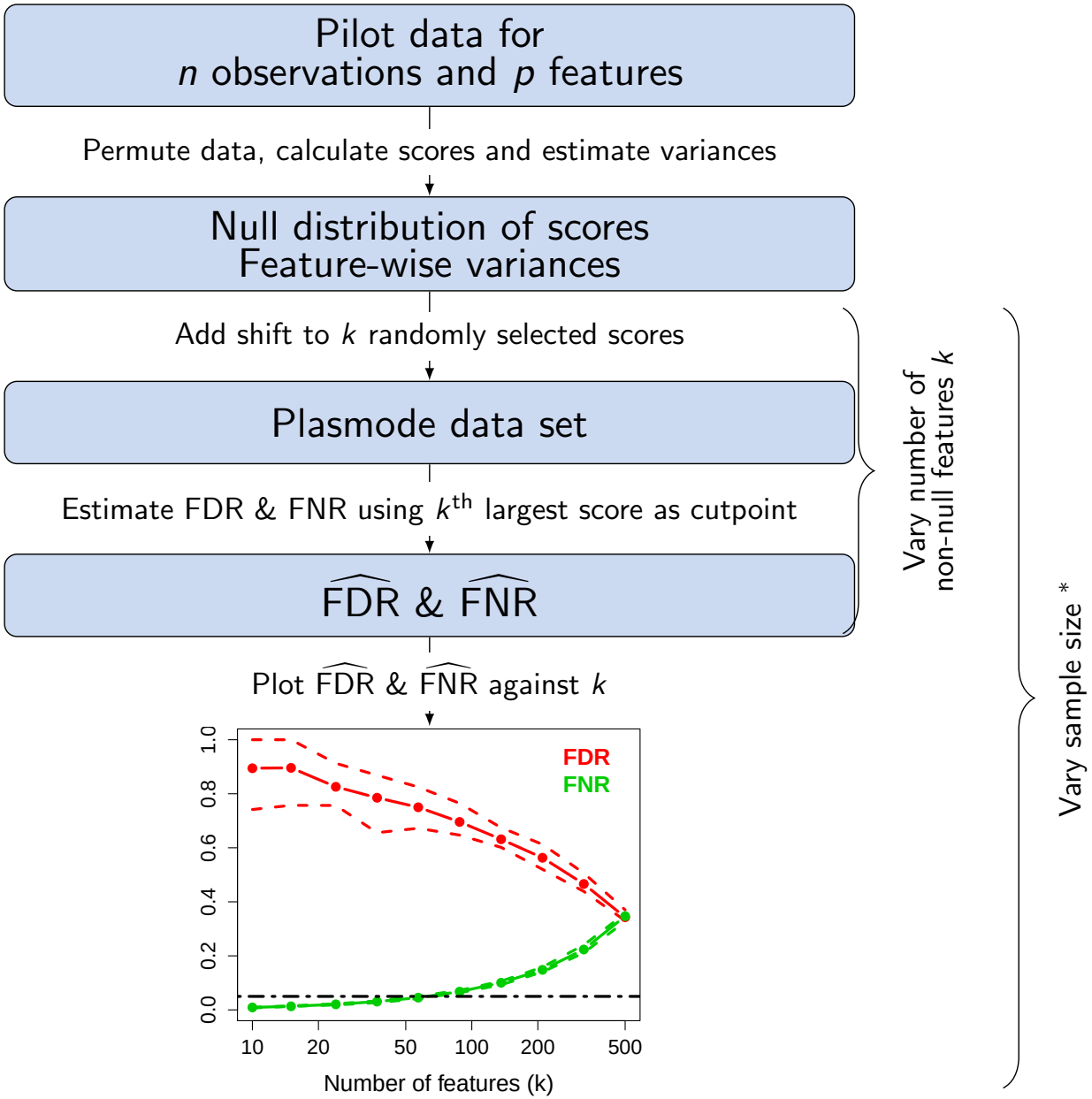
- Lin W-J, Hsueh H-M, Chen JJ. Power and sample size estimation in microarray studies. BMC Bioinformatics 2010, 11: 48.
  - Modification/extension of Tibshirani's approach

Plasmode Simulation: Sample Size Calculation

Tibshirani (2006)

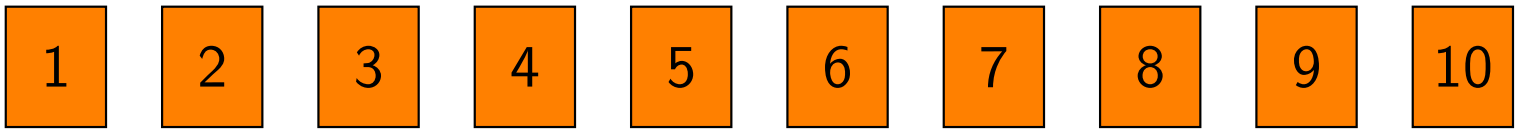
Hypotheses	Not rejected	Rejected	Total
True	U	V	m_0
False	T	S	m_1
Total	$m - R$	R	m

- Estimates false discovery rate FDR $= \frac{V}{R}$ and false nondiscovery rate FNR $= \frac{T}{m-R}$
- For simplicity, choose rule so that $R = m_1$
- Now 1 − power = FDR and type 1 error = FNR

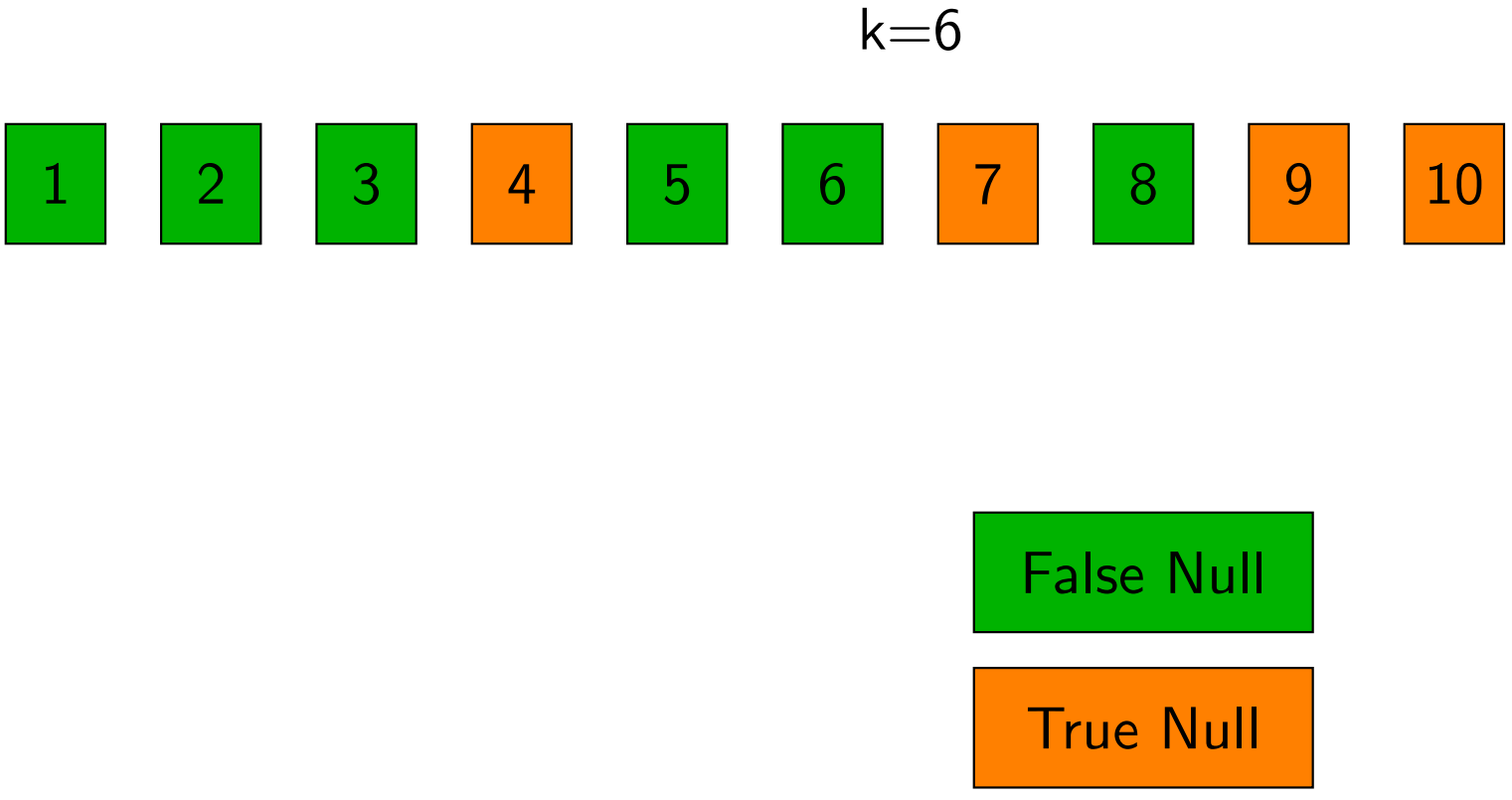


* sample size affects shifts

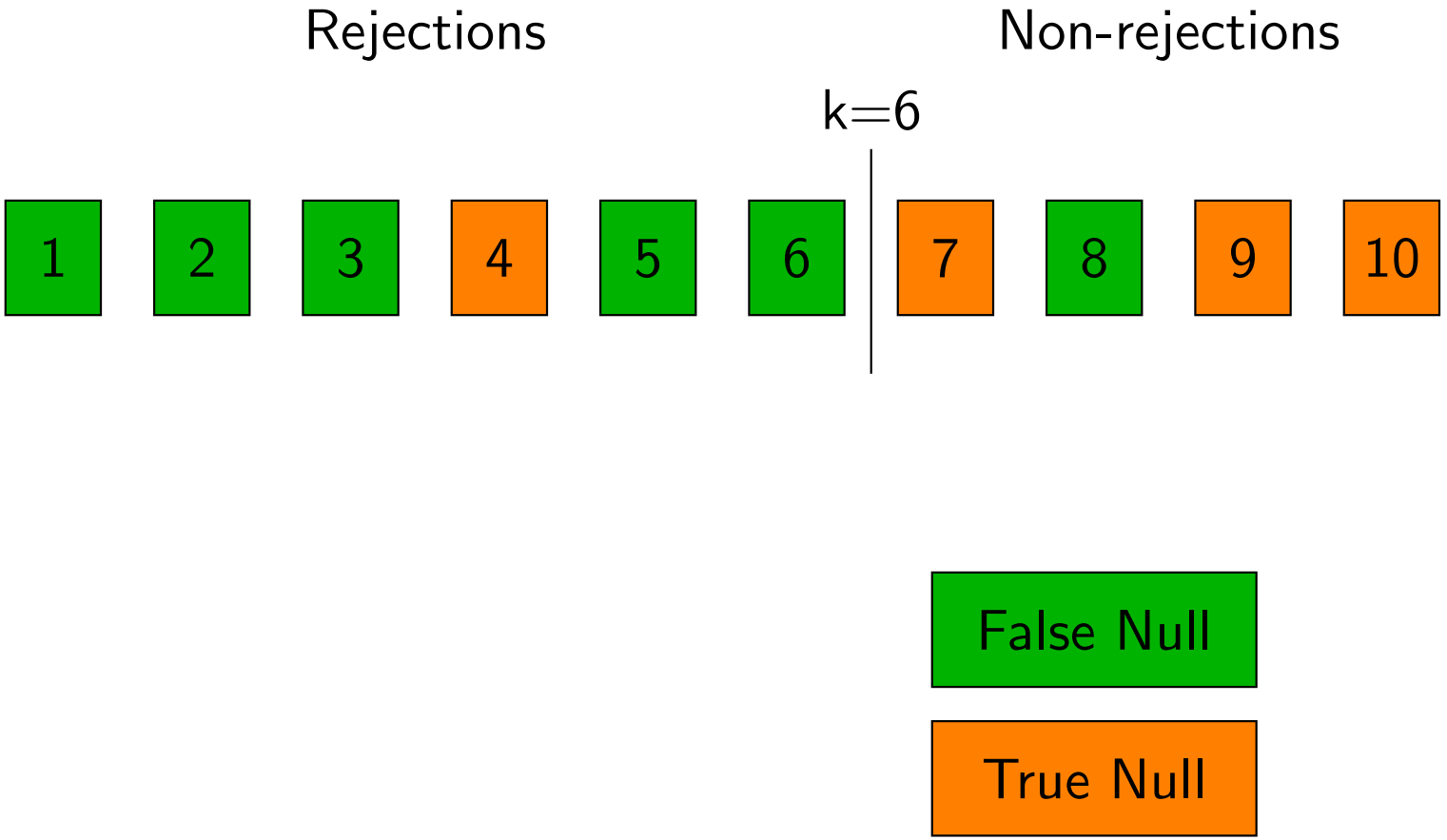
Plasmode Simulation: Sample Size Calculation



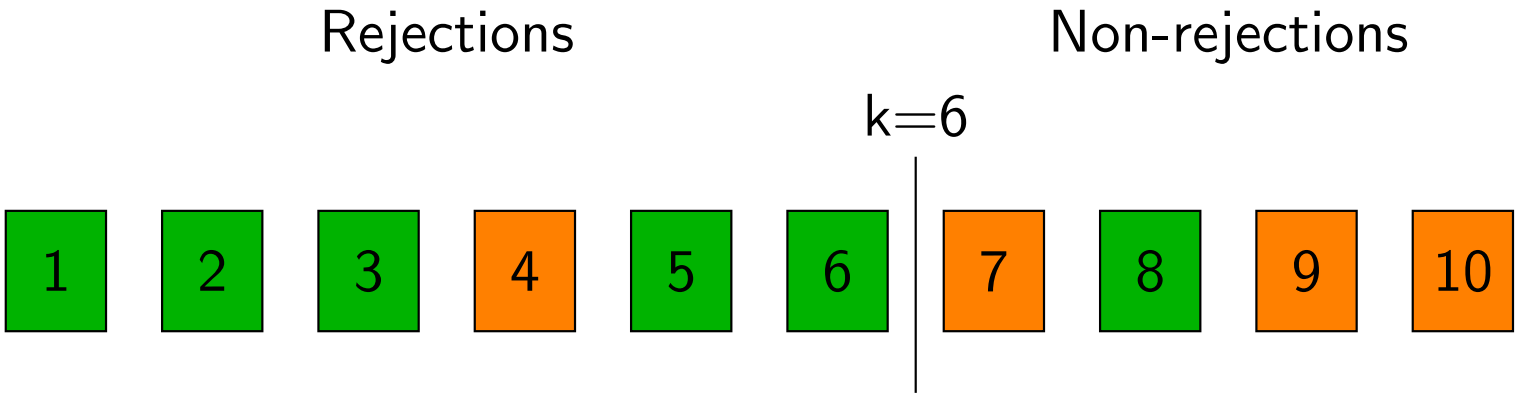
Plasmode Simulation: Sample Size Calculation



Plasmode Simulation: Sample Size Calculation



Plasmode Simulation: Sample Size Calculation



$\rightarrow$ FDR $= \frac{1}{6}$

$\rightarrow$ FNR $= \frac{1}{4}$

False Null

True Null

Plasmode Simulation: Sample Size Calculation

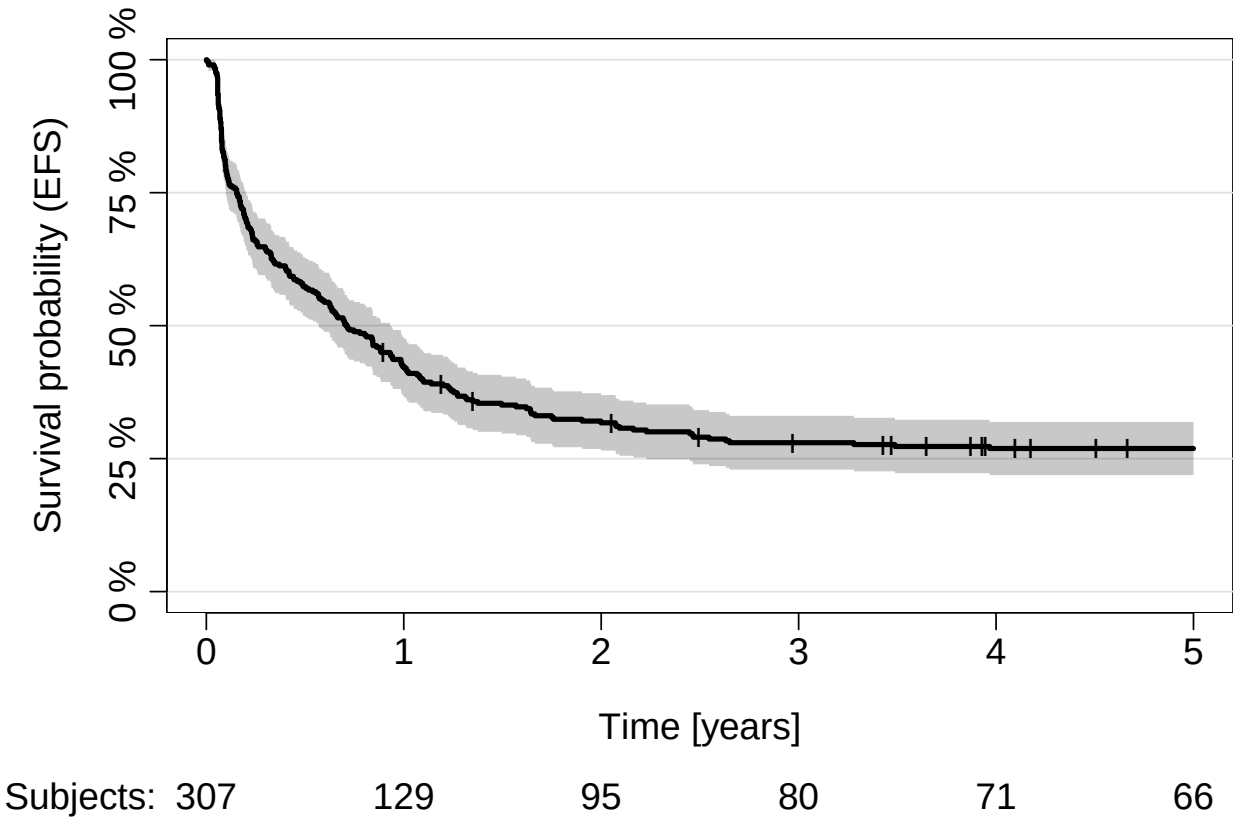
Lin et al. (2010)

Hypotheses	Not rejected	Rejected	Total
True	U	V	m_0
False	T	S	m_1
Total	$m - R$	R	m

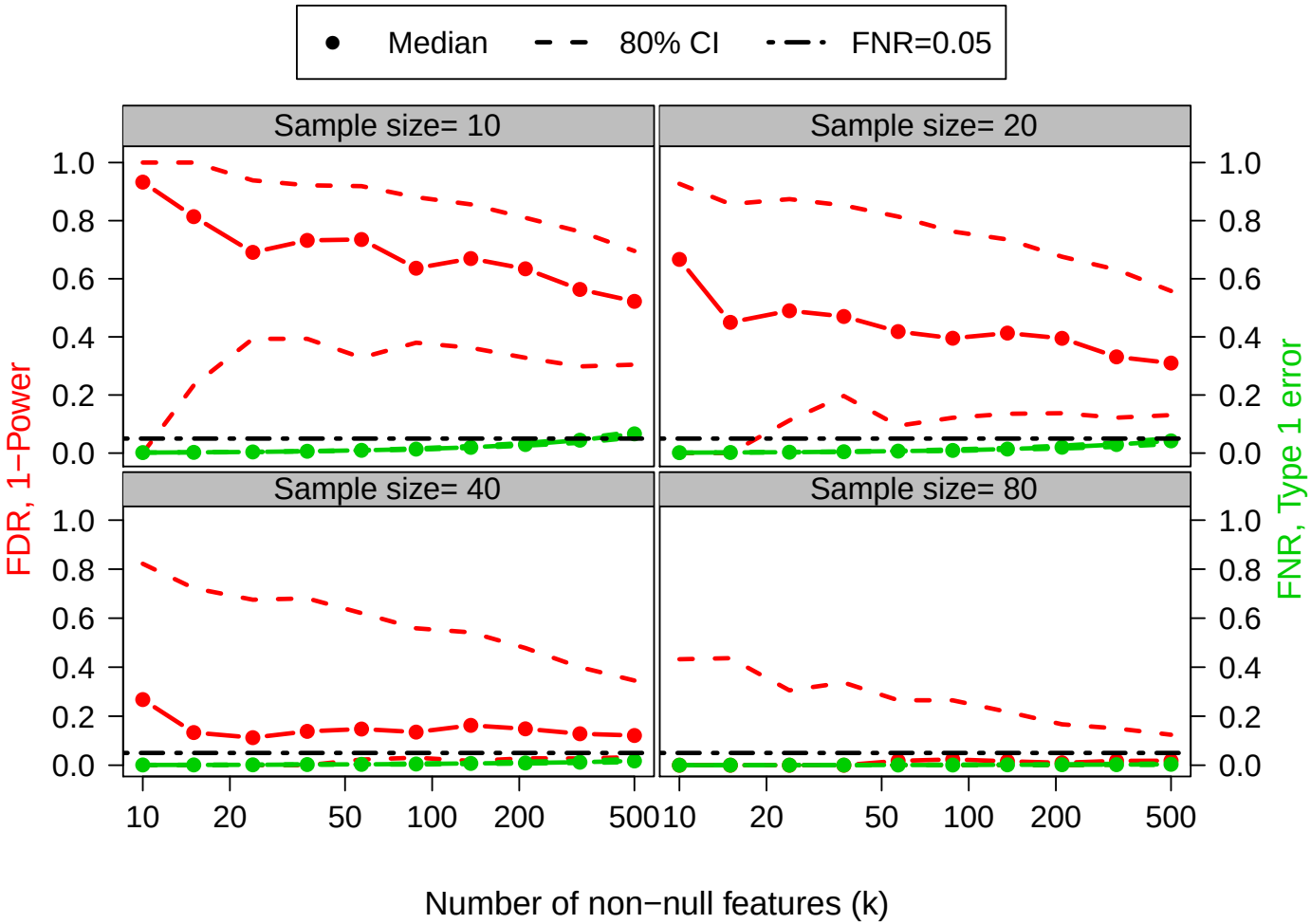
- Modification of the approach of Tibshirani (2006)
- Add adjustment factor to avoid bias due to small pilot data sets
- Revise definition of the cut-off
- Calculates sample size for specified TPR $= \frac{S}{m_1}$ (power)
- FDR $= \frac{V}{R}$ is controlled

Application

- 307 newly diagnosed acute myeloid leukemia (AML) patients
- Clinical data + gene expression data for $\approx$ 5000 genes
- Endpoint: Event-free survival (EFS)

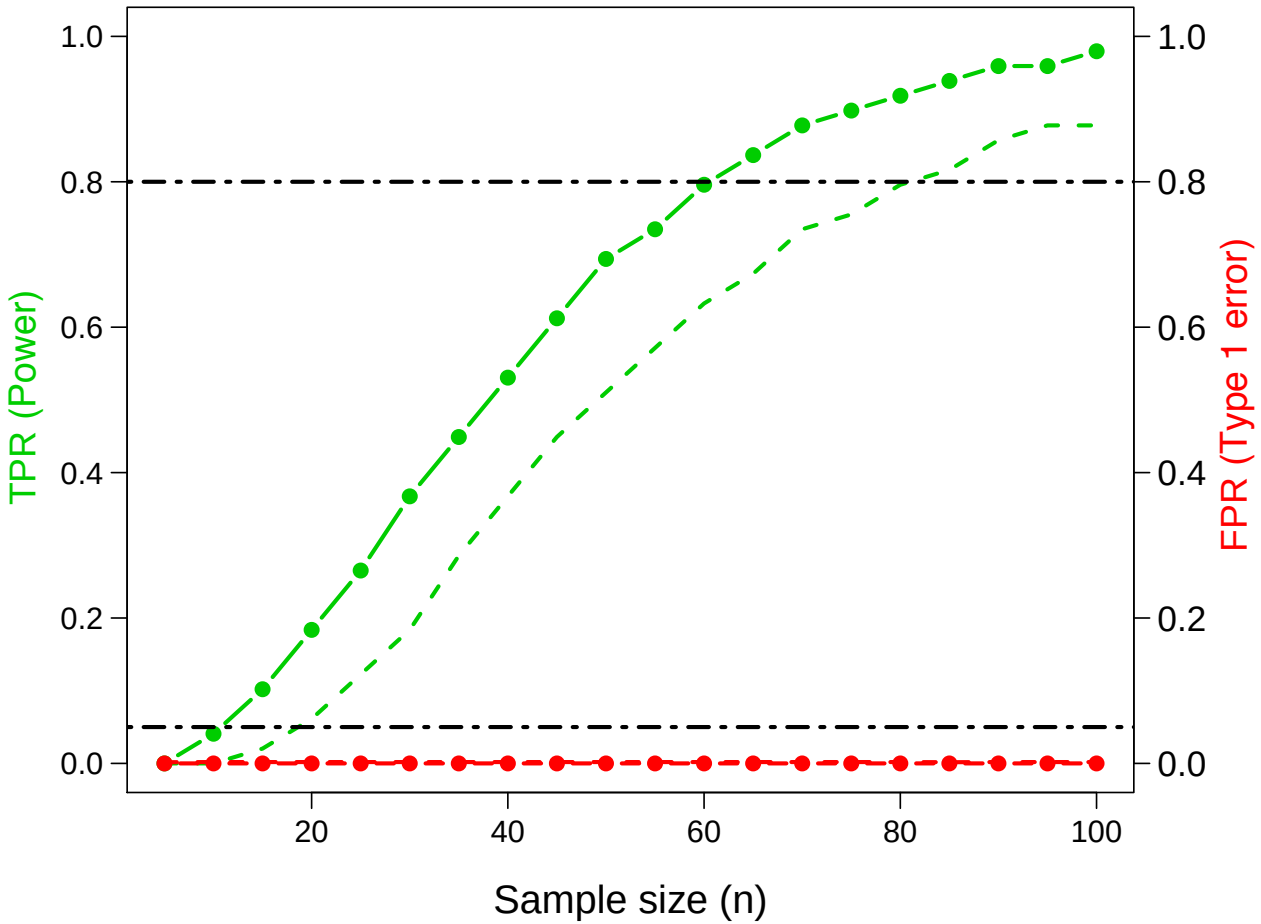


Application - Tibshirani (2006)



Application - Lin et al. (2010)

$k = 50$



Summary

- Plasmodes can be an alternative to synthetic data

but

- No one-fits-all solution
- Depend on availability of appropriate real data sets

- Of course: Work in progress

Acknowledgements

- Maral Saadati
- Julia Krzykalla

- Alicia Poplawski
- Harald Binder

Contact

- benner@dkfz.de