DEPARTMENT of Biostatistics

University of Pittsburgh | Graduate School of Public Health

SeqDesign: A Framework for RNA-Seq Genome-wide Power Calculation and Experimental Design Issues

Masaki Lin, Serena G. Liao, Yongseok Park, George C. Tseng

Department of Biostatistics

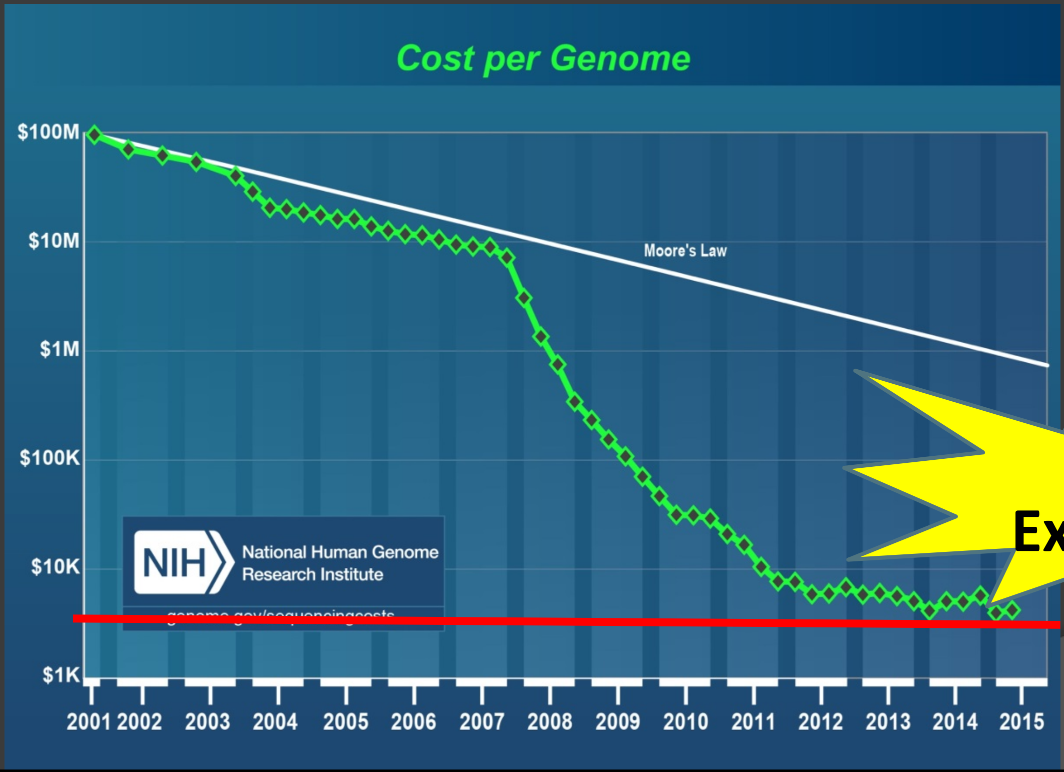
University of Pittsburgh

Outline

- Introduction
- Proposed method
- Simulation study
- Conclusion

Introduction

- Wide application of NGS technology
  - DNA, RNA, Methylation, …etc
- The cost of sequencing

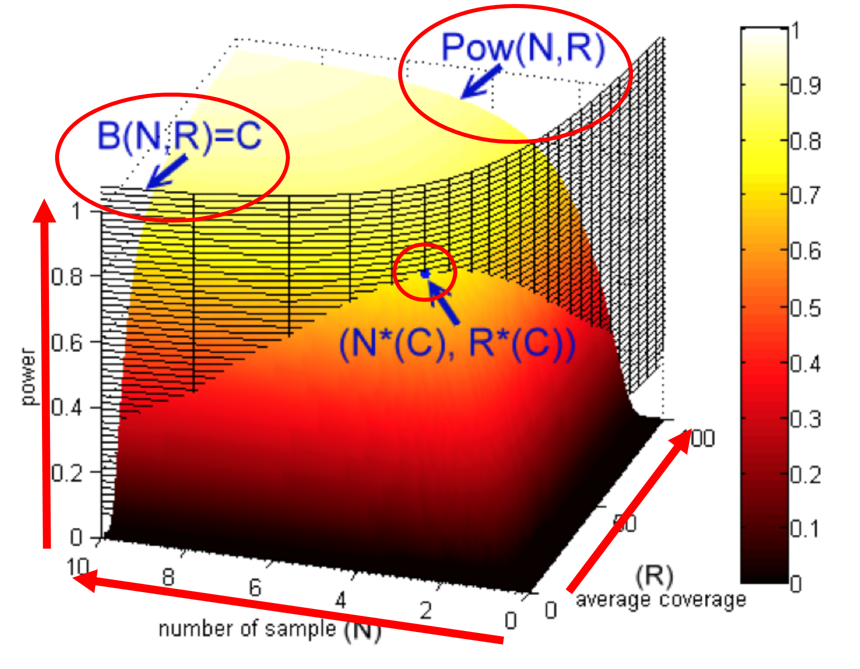


Still Expensive!!

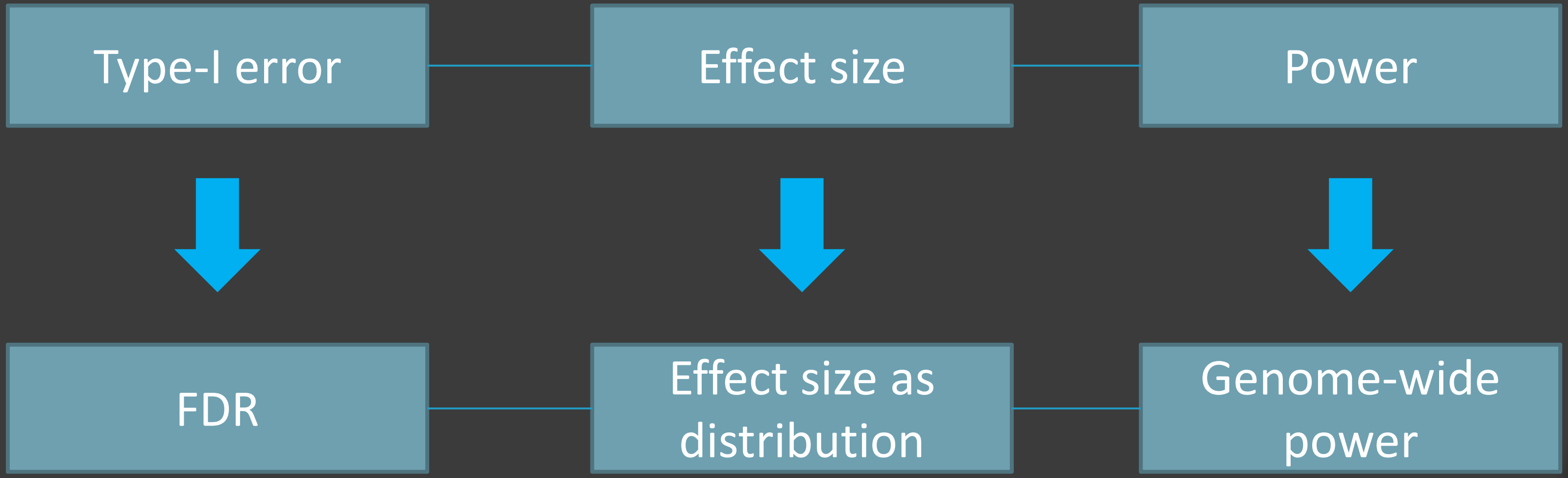
We need power calculation tool!!

Introduction

- ***RNA-seq***
  - compare to microarray platform
    - not only sample size (N)
    - but also read depth (R)
  - Two-dimensional optimal design



Introduction



Existing power calculation approach in RNA-seq

$$n = 2(z_{1-\frac{\alpha}{2}} + z_{\beta})\frac{1/\mu + \sigma^2}{ln(\Delta^2)}$$

Features	Poisson model*	RNASeqPower**	Scotty***
Pilot data			Partial
Model count data adequately		√	
Sequencing depth			√
Multiple comparison (FDR)	√		
Genome wide power calculation			√
Cost function by N and R			√

None of them satisfies practical settings!

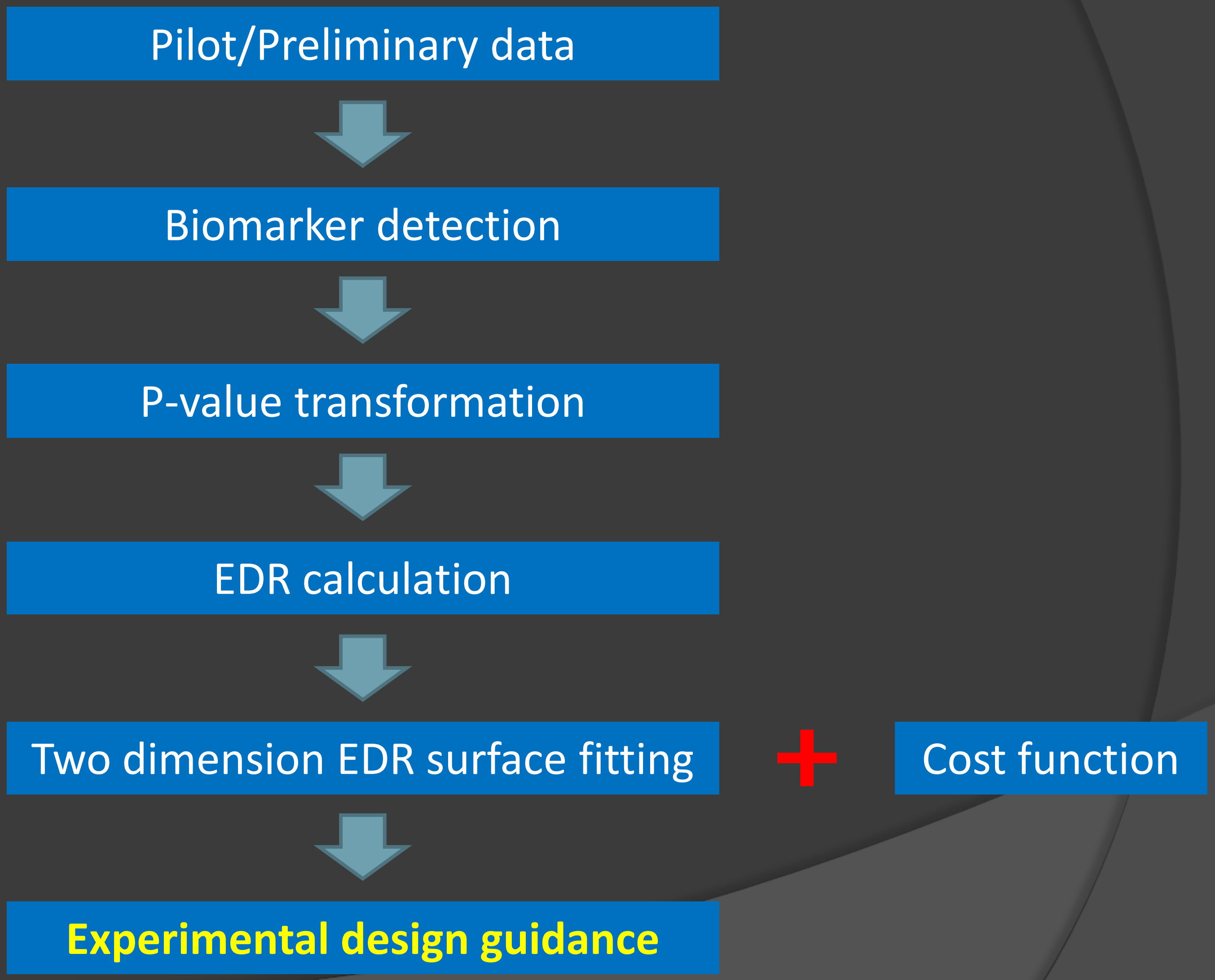
* Lee et al, 2013 ** Hart et al, 2013 *** Busby et al, 2013

Method comparisons

Features	Poisson model*	RNASeqPower**	Scotty***	SeqDesign
Pilot data			Partial	√
Model count data adequately		√		√
Sequencing depth			√	√
Multiple comparison (FDR)	√			√
Genome wide power calculation			√	√
Cost function by N and R			√	√

* Lee et al, 2013 ** Hart et al, 2013 *** Busby et al, 2013

Flow chart of SeqDesign



Pilot/Preliminary data

Biomarker detection

P-value transformation

EDR calculation

Two dimension EDR surface fitting + Cost function

Experimental design guidance

Model and test statistics

- Negative binomial regression for count data (GLM)
  - $Y_{gij} \sim NB(\mu_{gij}, k),\ \mu_{gij} = R_{ij} p_{gj},$
  - $\log(\mu_{gij}) = \log(R_{ij}) + \beta_{g0} + \boxed{\beta_{g1}} x_{ij}$

H_0: $\beta_{g1} = 0$ vs. H_1: $\beta_{g1} \neq 0$

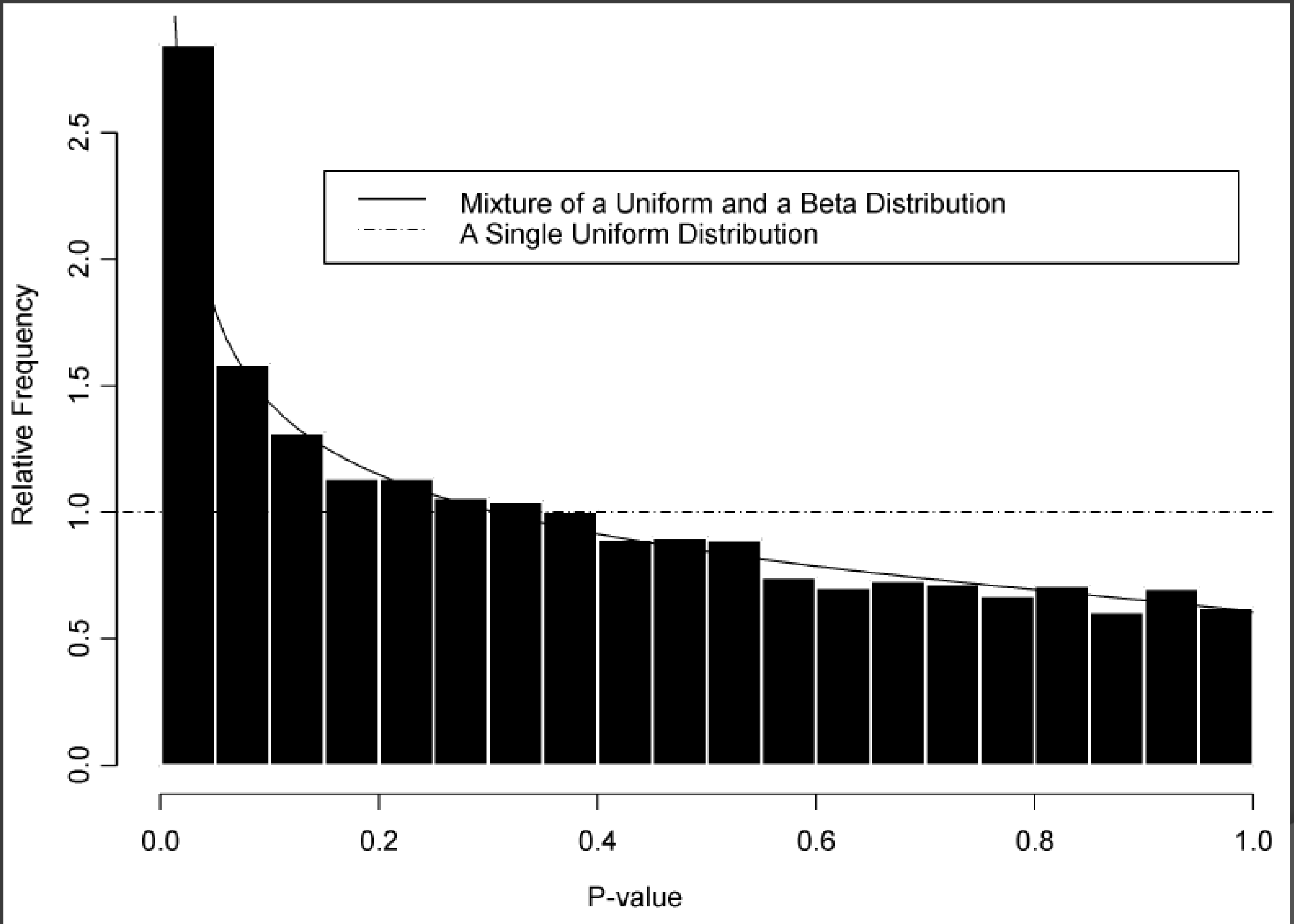
- Assume total reads are the same

$$\widehat{Var}(\hat{\beta}_{g1}) = \frac{1}{n} \cdot \left(\frac{1+\theta \cdot e^{\hat{\beta}_{g1}}}{\theta \cdot R \cdot e^{\beta_{g0}+\beta_{g1}}} + \frac{(1+\theta)\delta}{\theta} \right)$$

only n changes, the rest are fixed

$$Z = \frac{\hat{\beta}_{g1}}{\sqrt{\widehat{Var}(\hat{\beta}_{g1})}} \sim N(0,1)$$

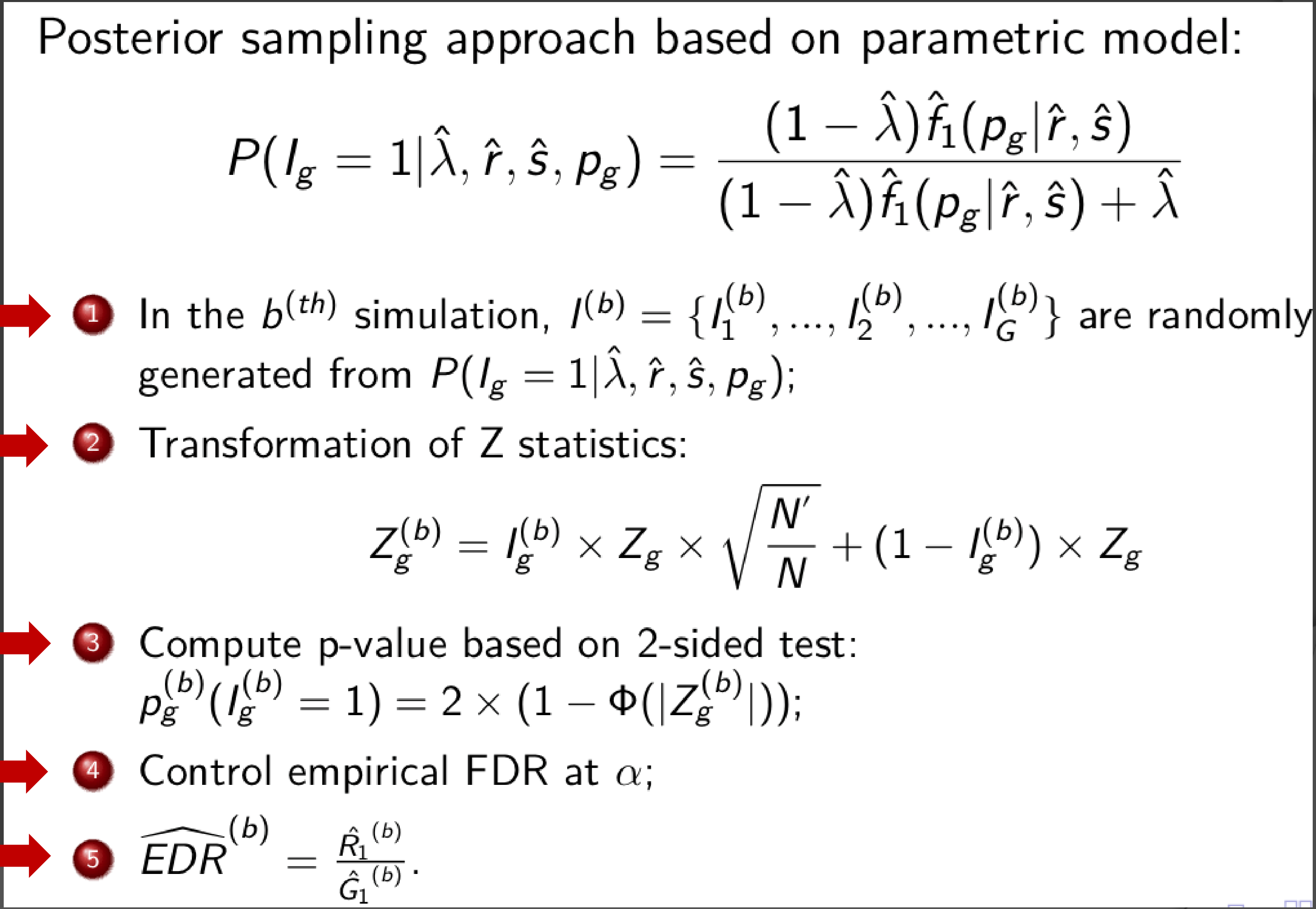
P-value distribution



Mixture model

$$f(p_g|r, s, \lambda) = \lambda + (1 - \lambda)\beta(p_g; r, s)$$

Genome-wide power prediction (changing from N to N')



Posterior sampling approach based on parametric model:

$$P(I_g = 1|\hat{\lambda}, \hat{r}, \hat{s}, p_g) = \frac{(1-\hat{\lambda})\hat{f}_1(p_g|\hat{r}, \hat{s})}{(1-\hat{\lambda})\hat{f}_1(p_g|\hat{r}, \hat{s}) + \hat{\lambda}}$$

1. In the $b^{(th)}$ simulation, $I^{(b)} = \{I_1^{(b)}, \ldots, I_2^{(b)}, \ldots, I_G^{(b)}\}$ are randomly generated from $P(I_g = 1|\hat{\lambda}, \hat{r}, \hat{s}, p_g)$;
2. Transformation of Z statistics:
$$Z_g^{(b)} = I_g^{(b)} \times Z_g \times \sqrt{\frac{N'}{N}} + (1 - I_g^{(b)}) \times Z_g$$
3. Compute p-value based on 2-sided test: $p_g^{(b)}(I_g^{(b)} = 1) = 2 \times (1 - \Phi(|Z_g^{(b)}|))$;
4. Control empirical FDR at α;
5. $\widehat{EDR}^{(b)} = \frac{\hat{R}_1^{(b)}}{\hat{G}_1^{(b)}}$.

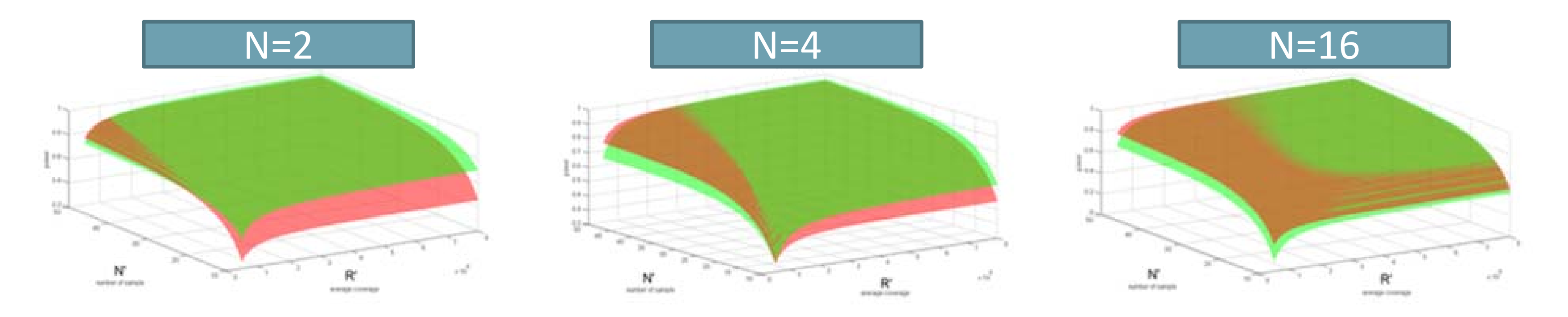
Genome-wide power prediction (changing from N to N’ and R to R’)

- The transformation step is achieved by

$$Z_g^{(b)} = I_g^{(b)} \times Z_g \times \frac{\sqrt{N'} \times \left(\frac{1+\theta e^{\hat{\beta}g1}}{\theta R e^{\hat{\beta}g0+\hat{\beta}g1}} + \frac{(1+\theta)}{\theta\hat{\delta}}\right)}{\sqrt{N} \times \left(\frac{1+\theta e^{\hat{\beta}g1}}{\theta R' e^{\hat{\beta}g0+\hat{\beta}g1}} + \frac{(1+\theta)}{\theta\hat{\delta}}\right)} + (1 - I_g^{(b)}) \times Z_g$$

Two-dimension EDR surface fitting

$$EDR = Pow(N', R') = 1 - \mathrm{a} \times N'^{-\mathrm{b}} - \mathrm{c} \times R'^{-\mathrm{d}}$$

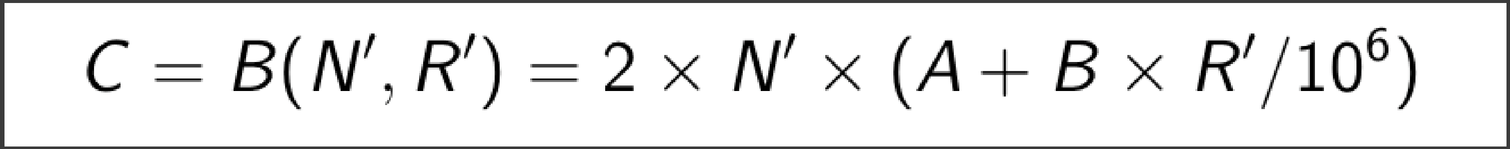


N=2

N=4

N=16

Cost function



$$C = B(N', R') = 2 \times N' \times (A + B \times R'/10^6)$$

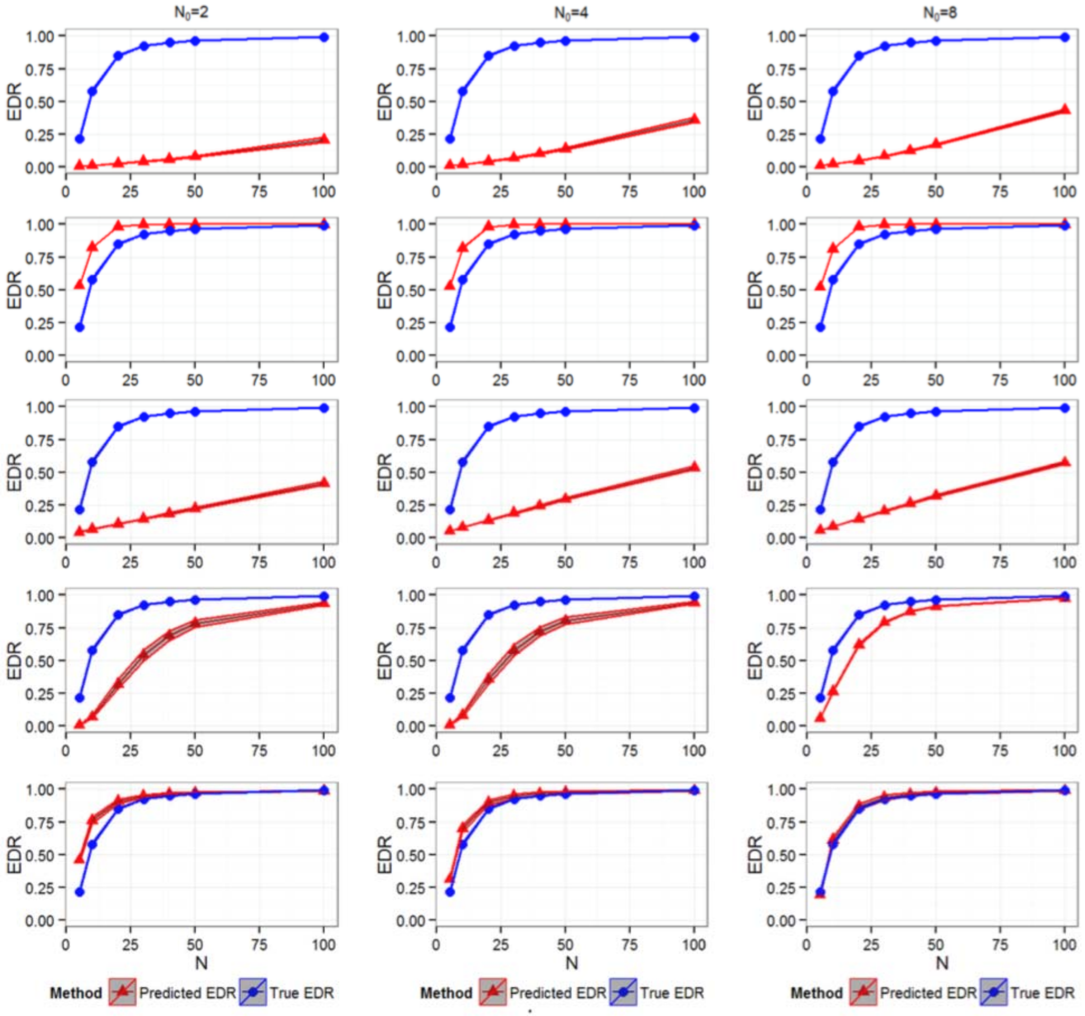
- A = 500, sample collection cost per sample
- B = 25, sequencing cost per sample per million reads

Simulation setting

- Parameters estimated from real dataset
- Effect size distribution (in log2 scale):
  - N(0, 0.04) and truncate at 0.15
  - $2^{0.15} = 1.1$
- Dispersion parameter = 50
- Number of gene = 25,000
- Proportion of DE gene = 10%
- 1 lane 60M reads

Methods comparison

FIG 3. *Method Comparison in Simulation* $I(\delta = 50$ *and* $fc \geq 1.20)$.*(A) Poisson model; (B) RNASeqPower; (C) NB exact test; (D) Scotty; (E) SeqDEsign.*



Poisson model

RNASeqPower

NB exact test

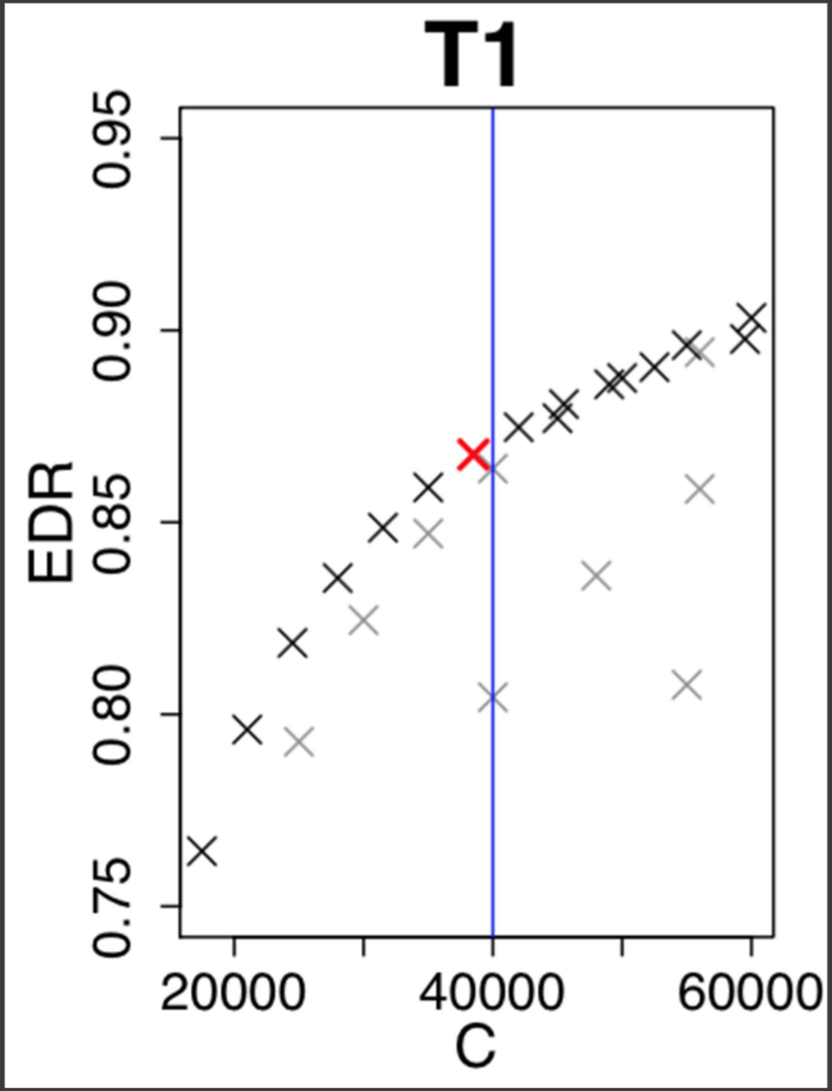
Scotty

SeqDesign

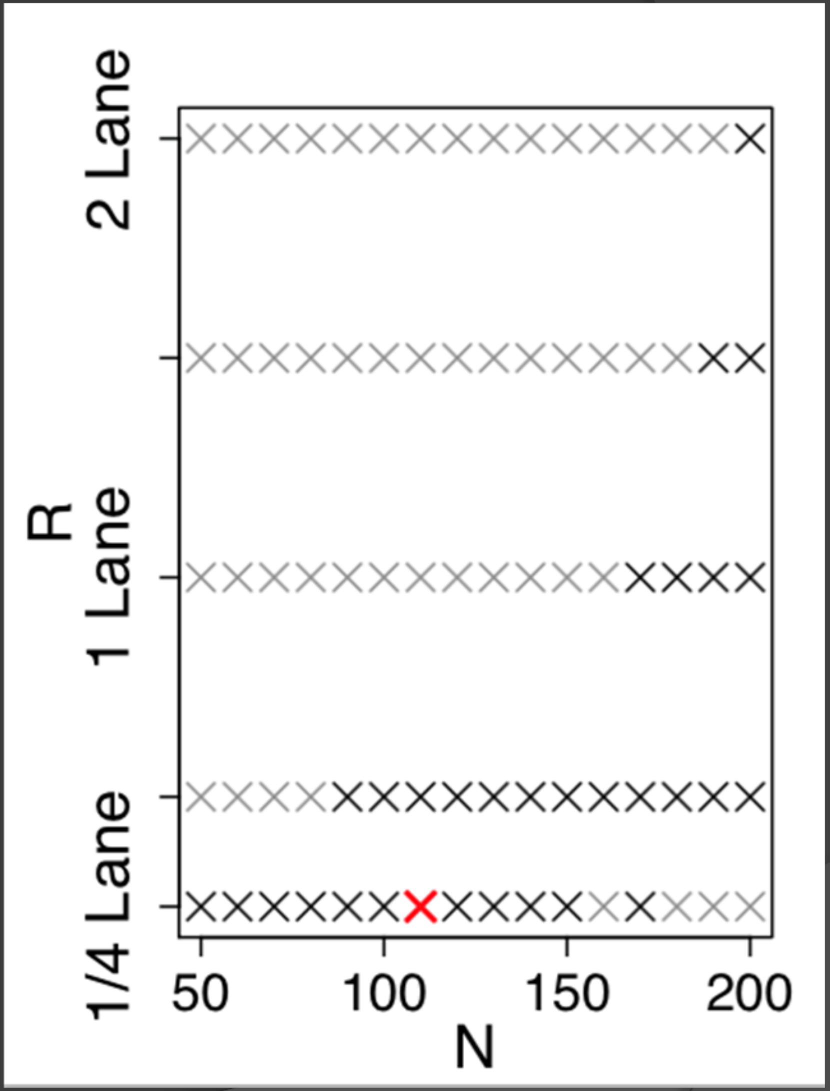
Five tasks in NGS design

	Have	Want
T1.	Budget	Optimal design
T2.	Desired power	Money
T3.	Maximum sample size	Best design
T4.	Maximum sample size	Recruit more samples?
T5.	Maximum sample size	Sequence deeper?

T1. Budget → Optimal design

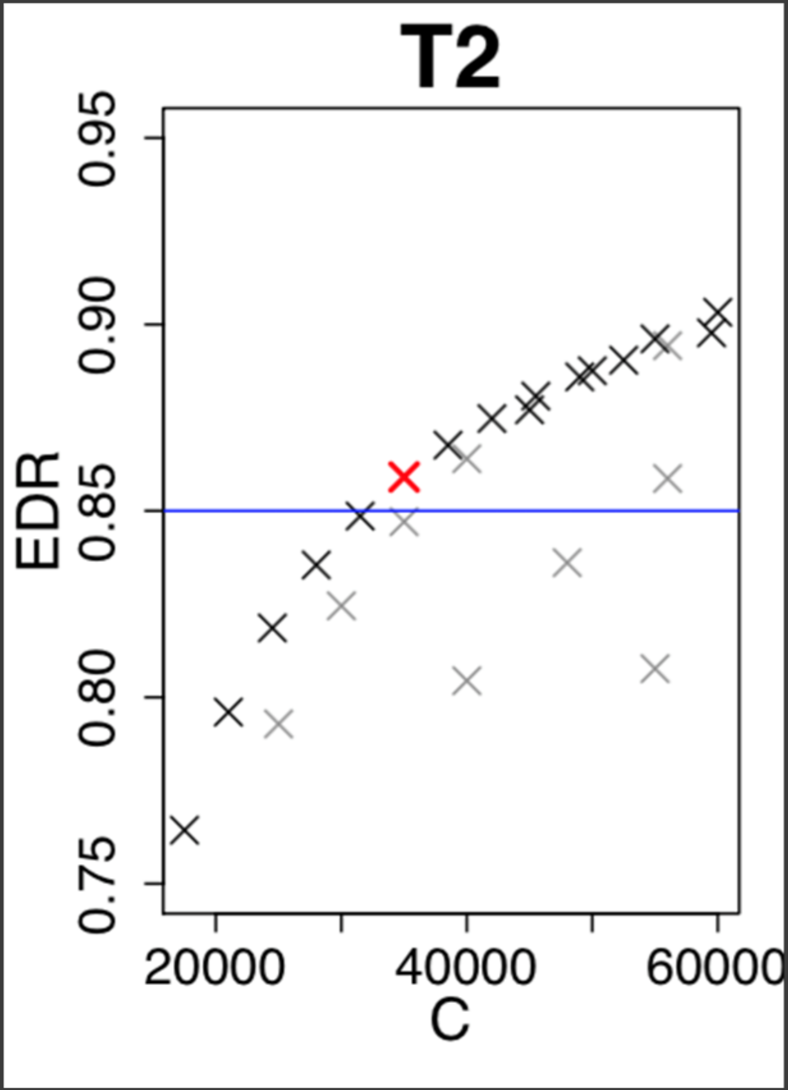


Cost benefit plot

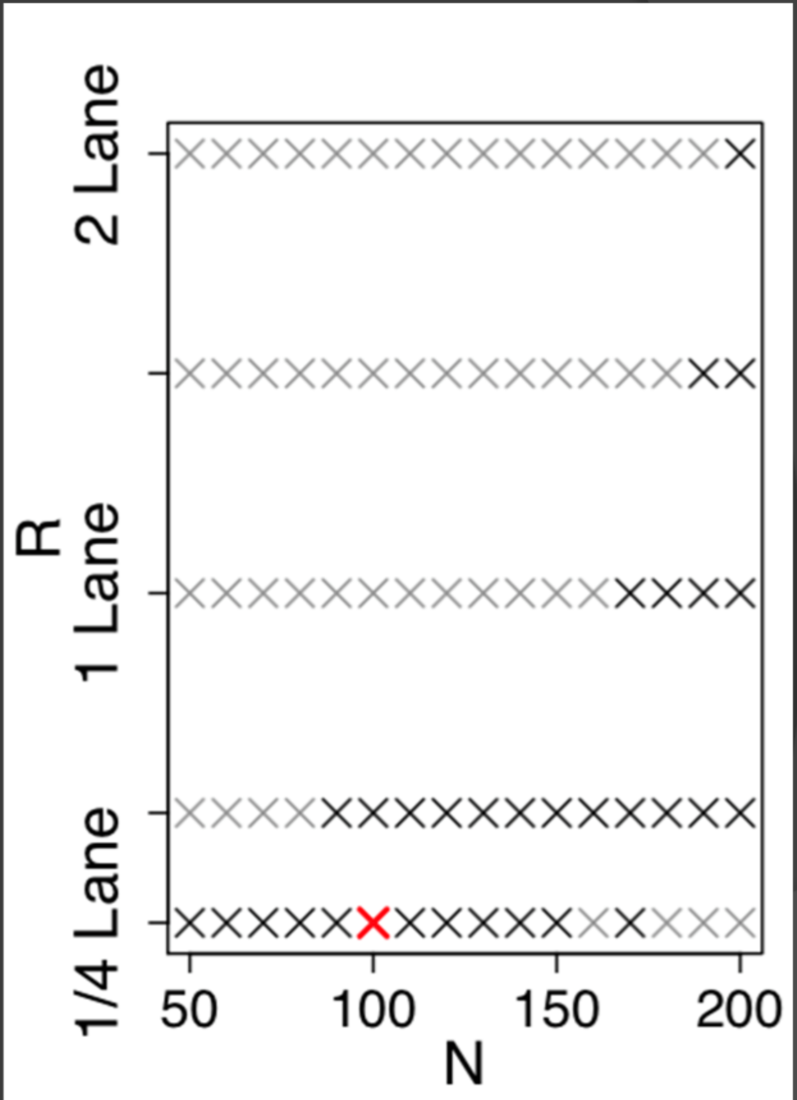


Design plot

T2. Desired power → Money

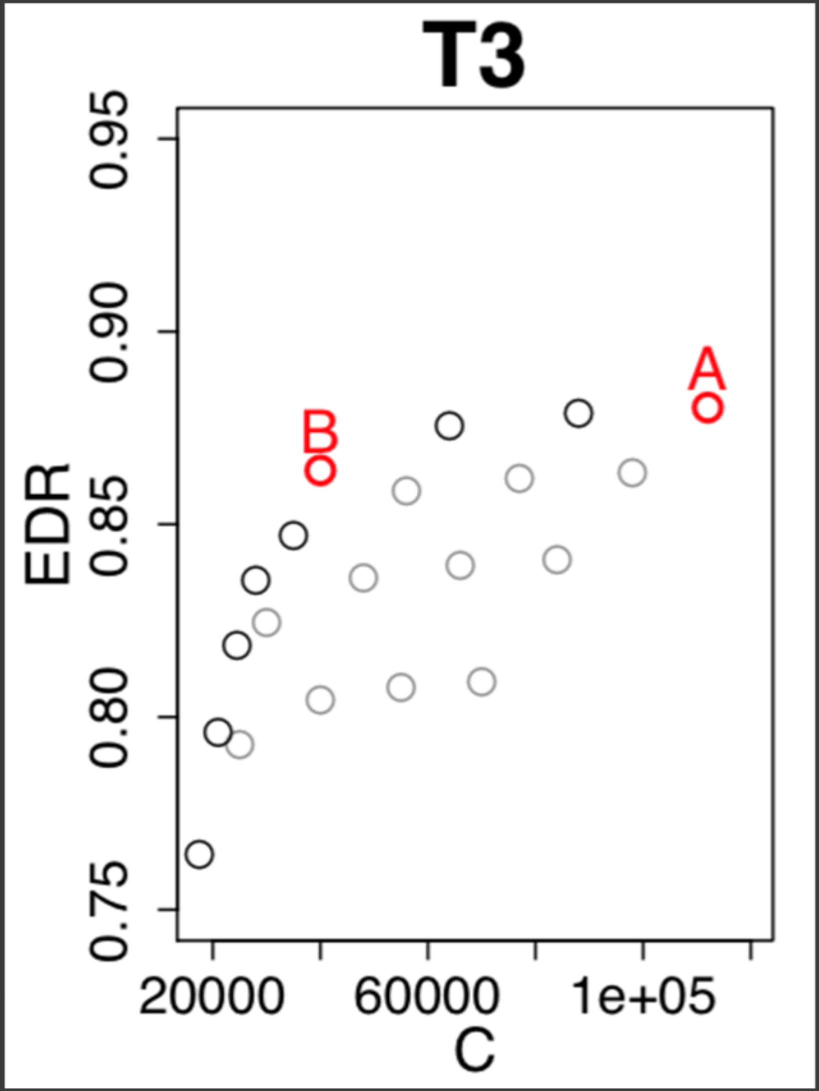


Cost benefit plot

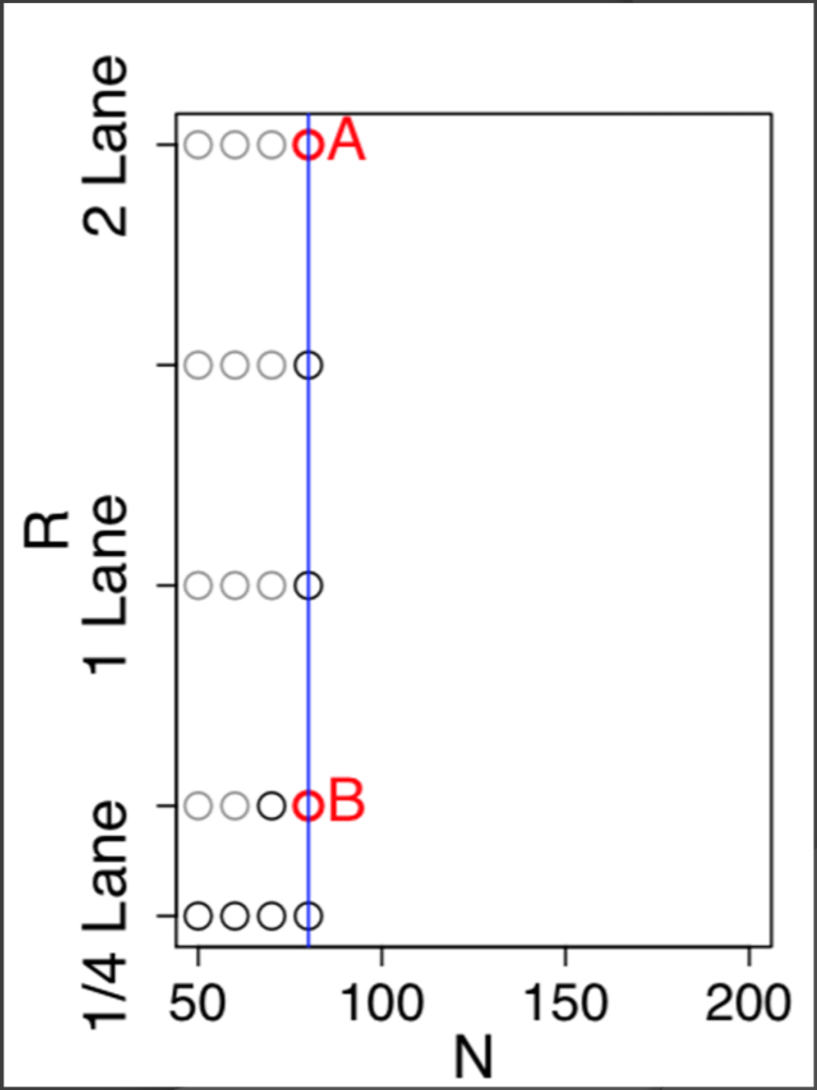


Design plot

T3. Maximum sample size → Best design



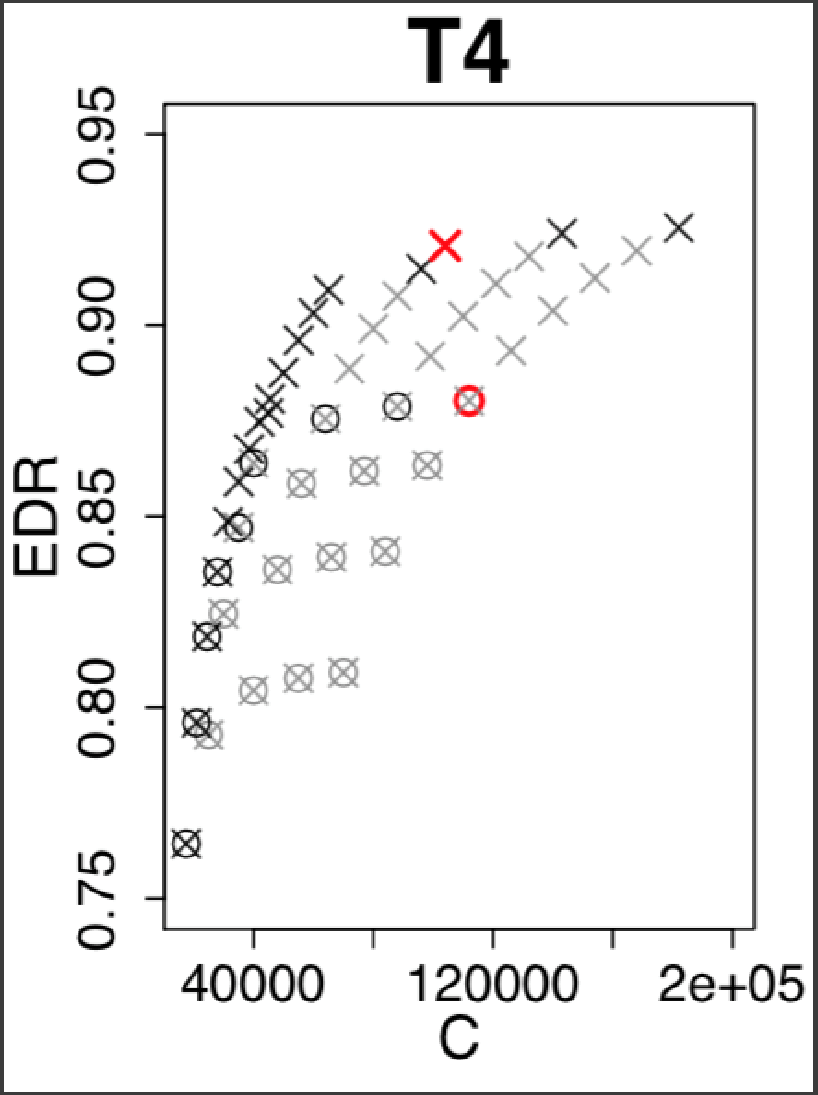
Cost benefit plot



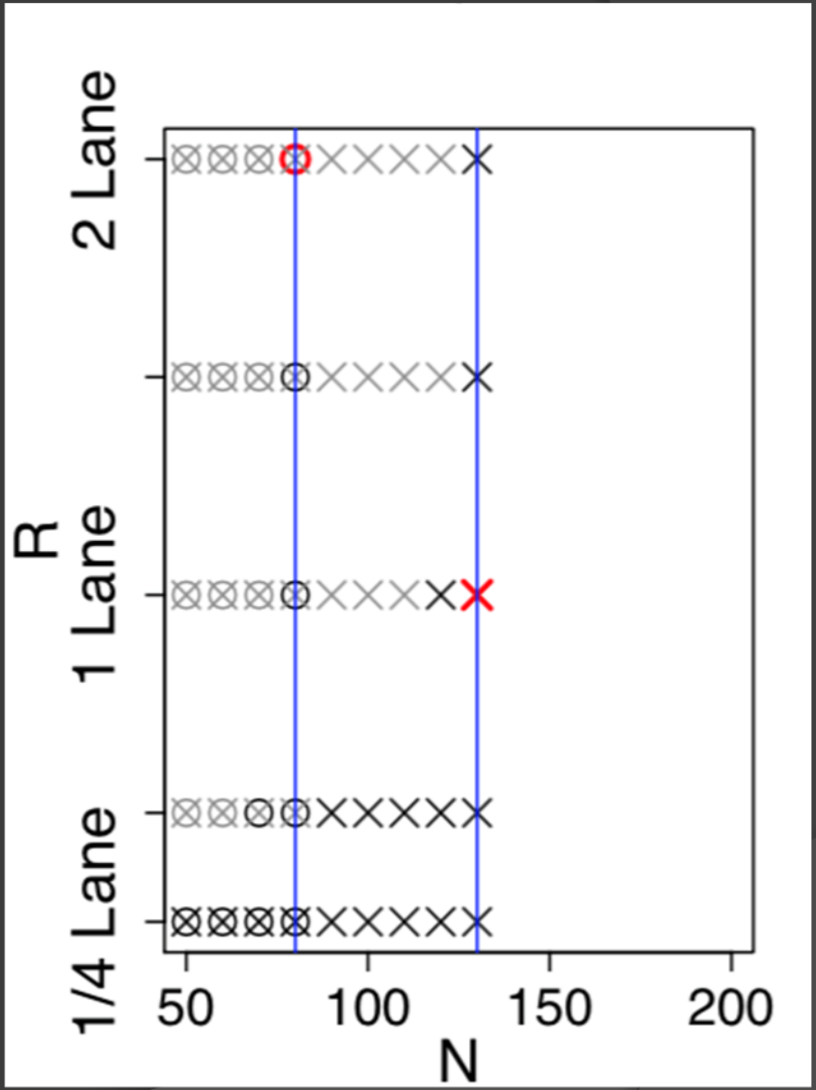
Design plot

T4. Maximum sample size

Recruit more samples?



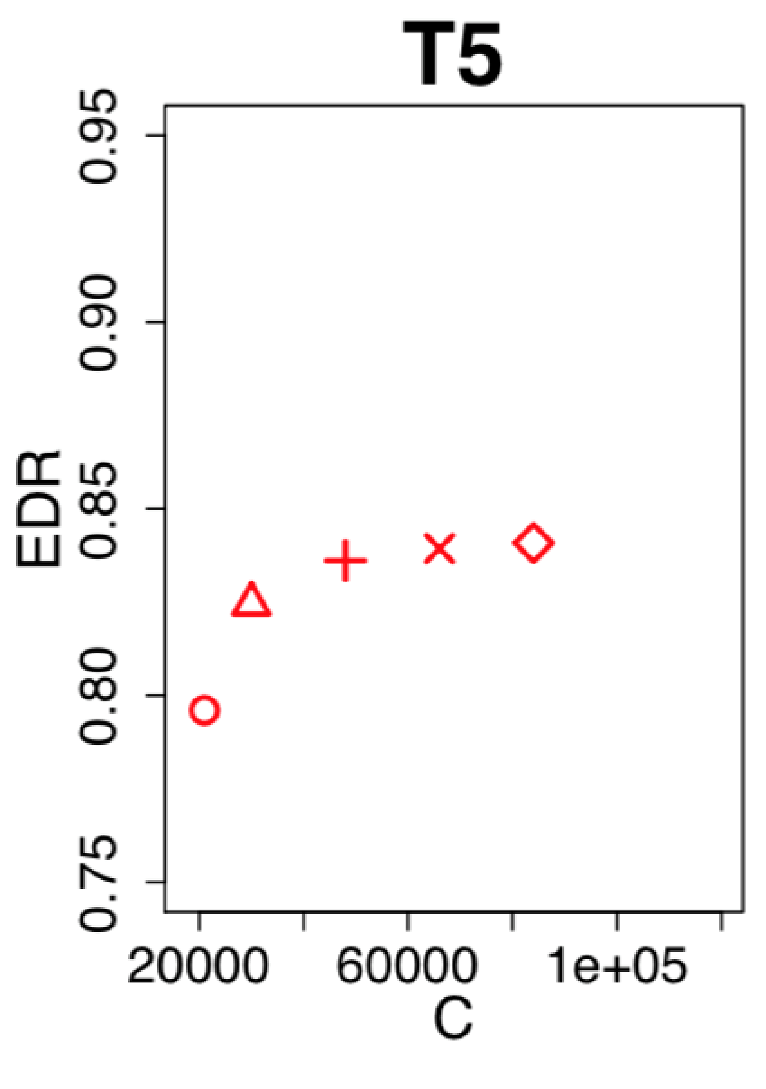
Cost benefit plot



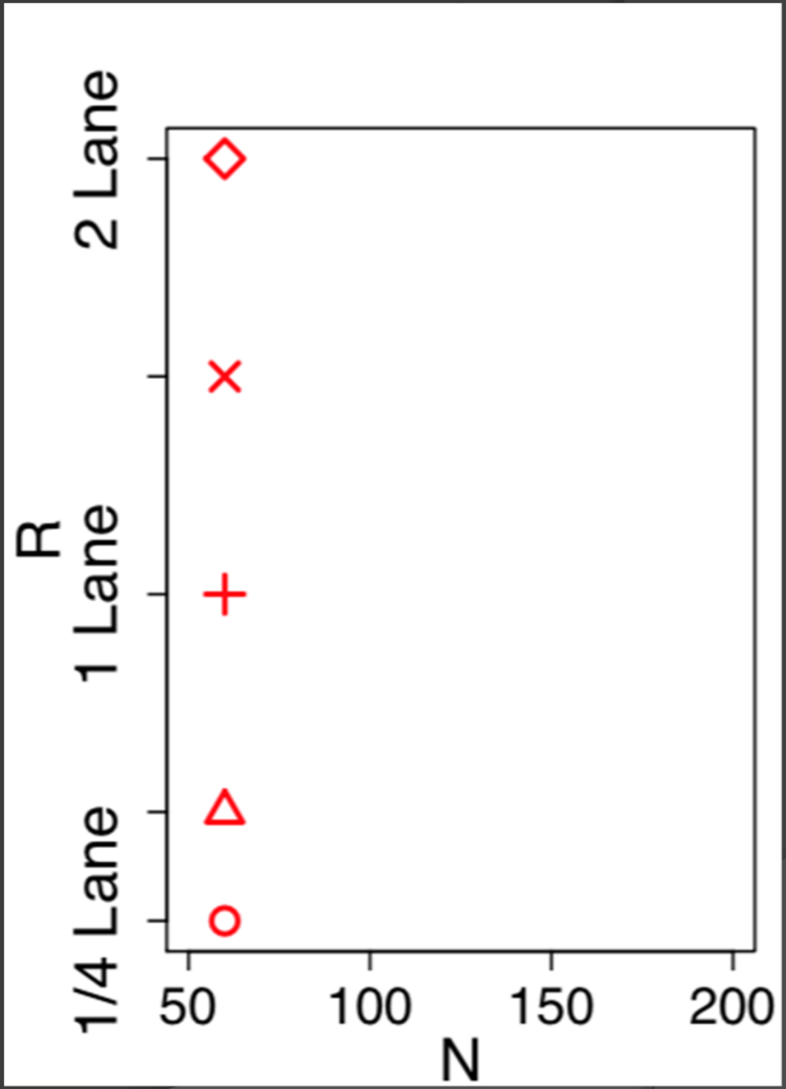
Design plot

T5. Maximum sample size

Sequence deeper?



Cost benefit plot



Design plot

Conclusion

- Better modeling
  - Count data
  - FDR
  - EDR
- Reflecting real situation
  - Pilot data
- Experimental design
  - Cost function
  - Consider both R and N