Bayesian nonparametric strategies for power maximization in rare variants association studies

Lorenzo Masoero
MIT, EECS [lom@mit.edu]

Joshua Schraiber
Genome Interpretation Group
Illumina, Inc.

Tamara Broderick
MIT, EECS

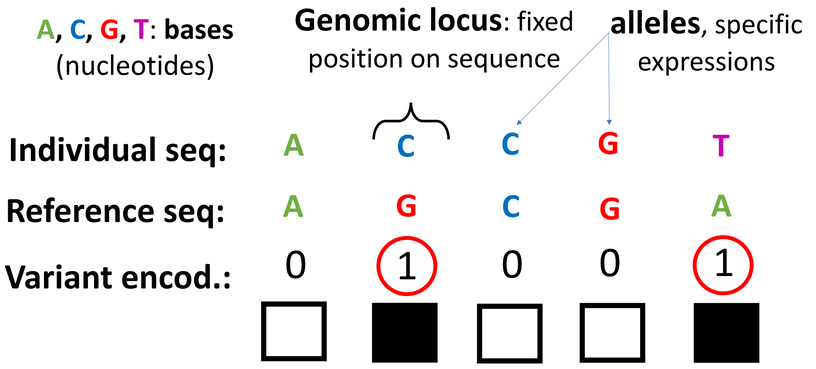
- Next generation sequencing: uncover genetic basis of disease via ***common variants*** association studies (e.g., GWAS)
- To unlock full potential of genomic-based approach, need effective ***rare variants*** association studies [RVAS]**: hard and costly to design**
  - Under fixed budget constraint need to optimally design study, trading off sequencing depth and # samples
  - We provide quantitative framework for optimal budget allocation to maximize power of statistical tests in RVAS

Data sketch & problem description

A, C, G, T: **bases** (nucleotides)

Genomic locus: fixed position on sequence

alleles, specific expressions



Variant ``called'' whenever **allele differs from reference genome**

- **Goal**: Test if rare variants are associated with disease
- **Problem**: Rare variants present in few individuals: can't analyze them individually, need many. To get more:
  - Sequence more individuals
  - Increase sequencing depth

  → Both options: greater cost of experiment
- **Our contribution**: Provide a statistical framework for:
  - Prediction ("fixed design"): # new samples needed to achieve target power π in statistical test under fixed study design
  - Experimental design ("fixed budget"): maximize power of association test, choosing sample size and sequencing depth for fixed budget ***B***

Hypothesis tests in RVAS

RVAS: are rare alleles associated with disease?

1. Collect data from ***unaffected*** & ***affected*** populations
2. Compare abundance of rare variants

Focus on **singletons burden tests**:

- Singleton: variant appearing in exactly one sample
- $\mu_j \coloneqq$ E[# singletons/patient in pop. j]
- Null hypothesis: singletons abundance same in ***unaffected*** & ***affected*** populations.

$$H_0 : \{\mu_U = \mu_A\}$$

- Alternative hypothesis: more singletons in ***affected*** than ***unaffected***

$$H_1 : \{\mu_U < \mu_A\}$$

Test H_0 with T:

- N_j datapoints
- Average # singletons $\overline{\mu_j}$
- Standard deviation s_j

$$T = \frac{(\overline{\mu_A} - \overline{\mu_U})}{\left\{\frac{s_U^2}{N_U} + \frac{s_A^2}{N_A}\right\}^{1/2}}$$

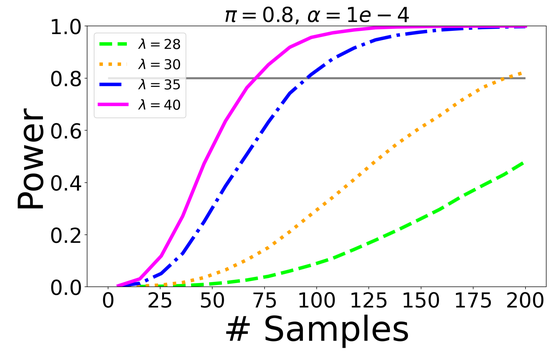
$\overline{\mu_j}$ and s_j depend on

- sample size N_A, N_U
- sequencing depth λ

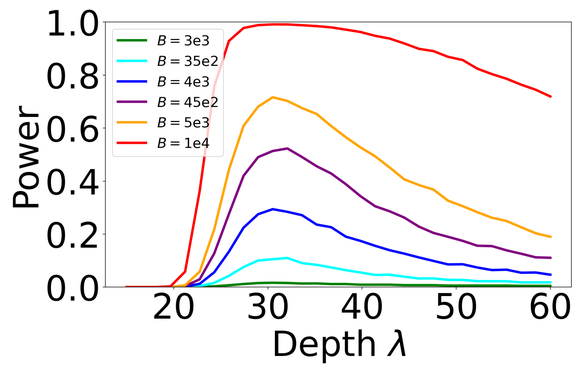
Optimal budget allocation: maximize power of T at target confidence level $1 - \alpha$

Experimental results

Fixed design: How many samples needed to achieve target power π at fixed sequencing depth λ, confidence $1 - \alpha$?



Fixed budget: What's the *optimal* sequencing depth λ, maximizing the power π, under a fixed budget ***B***?



"Optimal sequencing strategies in rare variants association studies: a hierarchical Bayesian nonparametric approach" *L.M, J. Schraiber and T. Broderick*
"More for less: predicting and maximizing genomic variant discovery via Bayesian nonparametrics" *[Biometrika, to appear; arXiv: 1912:05516]; L.M., F. Camerlenghi, S. Favaro and T. Broderick*