# Lecture 10: Power and Sample Size, Design Considerations, Emerging Issues

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2017

# Lecture Overview

1. Omnibus tests
   - 1.1 Variable Threshold Test
   - 1.2 SKAT-O
2. Weighting and Prior Knowledge
3. Design Considerations
   - 3.1 Platforms
   - 3.2 Extreme Phenotype Sampling
   - 3.3 Power and Sample Size

# Power/Sample Size calculation

- Power/Sample size calculation is essential to design future sequencing studies.
- Input information:
- Region information
  - LD structure and MAF spectrum.
  - Region size to test.

# Power/Sample Size calculation

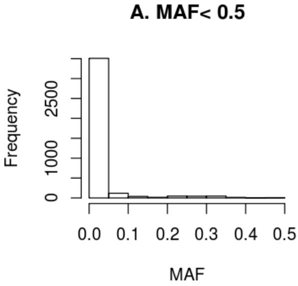
- Causal variant Information
  - Effect size (continuous traits), or Odds ratio (binary traits).
  - % of rare variants be causal.
  - % of causal variants with negative association direction.
- Binary traits
  - Case/Control Ratio.
  - Prevalence

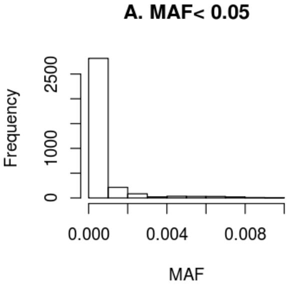
# Practical Points: SKAT Power Calculations

- Region information
  - Either simulated haplotypes or sample haplotypes from preliminary data.
  - The SKAT package provides 10,000 haplotypes over a 200 kb region generated by the coalescent simulator (COSI).

# MAF spectrum

- MAF spectrum of the simulated haplotypes
- Most of SNPs have very low MAFs.





# Practical Points: Power/Sample Size calculations

- Causal Variant Information:
  - To use $\log_{10}$ function $(-c\log_{10}(MAF))$ for the effect sizes or log odds ratio.
  - $c$ is a parameter to determine the strength of association.
    - Ex: $c = 1$
      $\beta = 2$ or $\log(OR) = 2$ for a variant with MAF=0.01
      $\beta = 4$ or $\log(OR) = 4$ for a variant with MAF=$10^{-4}$.

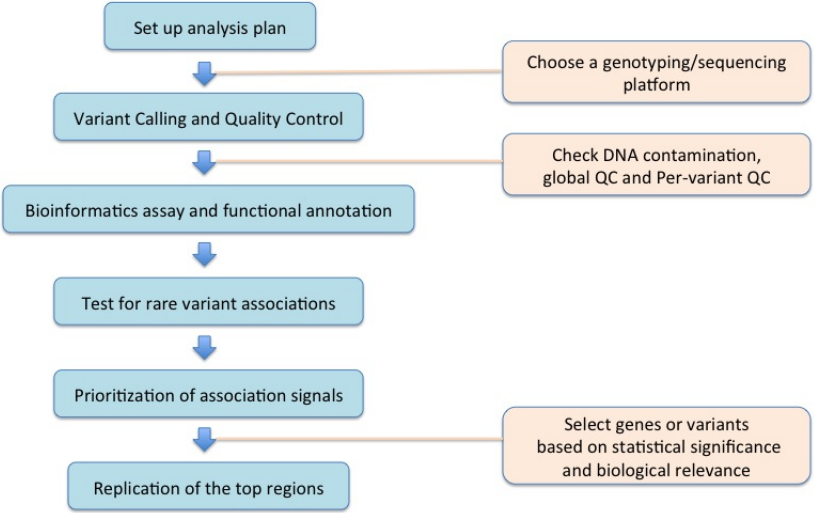
# Practical Points: Power/Sample Size calculations

- In SKAT package, you can set $c$ using the MaxOR (OR for MAF $= 10^{-4}$) or MaxBeta ($\beta$ for MAF $= 10^{-4}$).

# Practical Points: Power/Sample Size calculations

- Power depends on LD structure of the region and MAFs of the causal variants.
- We are interested in estimating power in multiple regions and multiple sets of causal variants selected from a certain disease model.
    - We estimate an average power.
    - Approximately 100 $\sim$ 500 sets of regions/causal variants are needed to estimate the average power stably.

# Data Processing and Analysis Flowchart



Set up analysis plan

Variant Calling and Quality Control

Bioinformatics assay and functional annotation

Test for rare variant associations

Prioritization of association signals

Replication of the top regions

Choose a genotyping/sequencing platform

Check DNA contamination, global QC and Per-variant QC

Select genes or variants based on statistical significance and biological relevance

# Genotyping Platforms

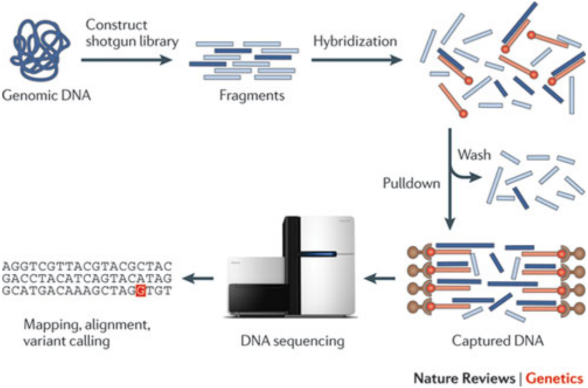
- High depth whole genome sequencing is the most informative, however it is currently expensive.
- Alternative sequencing designs and genotyping platforms
  - Low depth sequencing
  - Exome seqeuncing
  - High coverage microarrays (Exome chip)
  - Imputation

# Low depth whole genome sequencing

- Sequencing 7 $\sim$ 8 samples at low depth (4x) instead of 1 sample at high depth (30x)
- Low depth sequencing
  - Relatively affordable
  - LD based genotyping: leverage information across individuals to improve genotype accuracy.
  - 1000 Genome (4x) and UK 10K (6x) used low depth sequencing.
- Cons:
  - Subject to appreciable sequencing errors

# Exome sequencing

- Restrict to the protein coding region (1 $\sim$ 2% of genome (30 Mbps)).



Bamshad et al. (2011), NRG

# Exome sequencing

- Focus on the high value portion of the genome
- Relatively cost effective
- Cons: Only focus on the exome
  - Most of GWAS hits lie in non-exomic regions
  - Many non-coding regions have biological functions

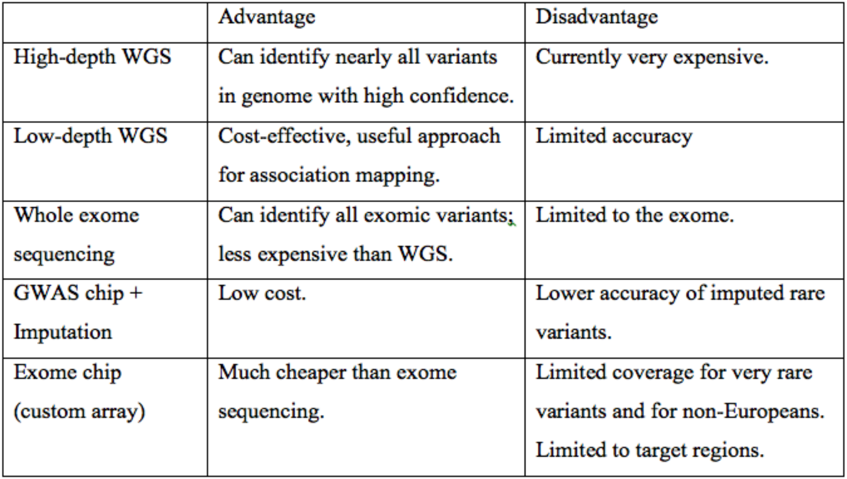
# Exome array

- Using variants discovered in 12,000 sequenced exome
- Low cost ($10 \sim 20x$ less than Exome sequencing)
  - 250K non-synonymous variants
  - 12K splicing variants
  - 7K stop altering variants
- Cons:
  - Cannot investigate very rare variants.
  - Limited coverages for non-European populations

# GWAS chip + Imputation

- Imputation: Estimate genotypes using reference samples
  - Imputation accuracy increases as the number of reference samples increases
- No additional experiment cost
- Cons:
  - Low accuracy of imputed rare variants

# Summary



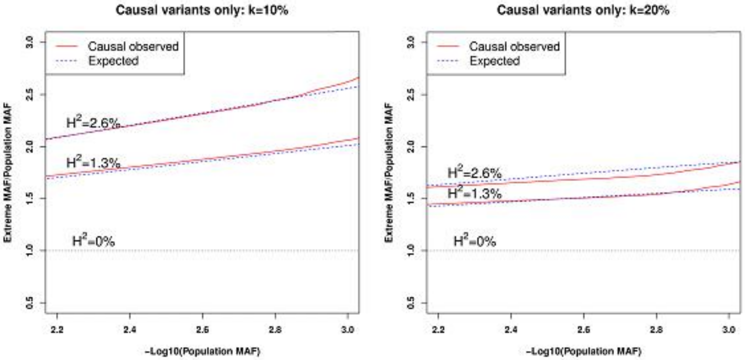
| | Advantage | Disadvantage |
|---|---|---|
| High-depth WGS | Can identify nearly all variants in genome with high confidence. | Currently very expensive. |
| Low-depth WGS | Cost-effective, useful approach for association mapping. | Limited accuracy |
| Whole exome sequencing | Can identify all exomic variants; less expensive than WGS. | Limited to the exome. |
| GWAS chip + Imputation | Low cost. | Lower accuracy of imputed rare variants. |
| Exome chip (custom array) | Much cheaper than exome sequencing. | Limited coverage for very rare variants and for non-Europeans. Limited to target regions. |

# Extreme phenotype sampling

- Rare causal variants can be enriched in extreme phenotypic samples
- Given the fixed budget, increase power by sequencing extreme phenotypic samples.

# Enrichment of causal rare variants in phenotypic extremes

- Estimated folds increase of the observed MAFs of causal variants ($k$% high/low sampling, $H^2$=Heritability).



# Extreme phenotypic sampling

- Continuous traits:
  Select individuals with extreme trait values after adjusting for covariates.
- Binary traits:
  Select individuals on the basis of known risk factors
  - Ex. T2D : family history, early onset, low BMI

# Extreme phenotypic sampling

- Extreme continuous phenotype (ECP) can be dichotomized, and then any testing methods for binary traits can be used.
- But dichotomization can cause a loss of information and can decrease the power.
- Methods modeling ECP as truncated normal distribution has been developed (Barnett, et al, 2013, Gen. Epid).

# Case Only Analysis

- Case only analysis: sequencing only cases (sporadic or familial)
- Rationale:
  - Expense
- Typical *n*:
  - 100 – 1000
  - $< 100$ or even $< 50$

## When Sample Size "Sufficient"

- Can use reference controls: 1000 Genomes, exome sequencing project, etc.
- Caution:
  - Batch effects, sequencing artifacts, processing differences
  - Relevant population: must be comparable
  - Covariate adjustment
  - Potential cases among reference

# Case Only Analysis with Modest *n*

- Small sample sizes: $n = 25$
- Potentially strong effects? High penetrance? Extremes?
- Standard case control testing may be under powered
- **Basic strategy**: Screening, filtering and bioinformatics

Reference: L. Wu, et al. (2015) *J. Med. Genet.*.

# Modest *n*: Variant Filtering

**Idea:** Prioritize variants from large scale screen

## Variant Frequency Filtering

- Use reference data, e.g. 1000 Genomes
- Remove variants with higher MAF:
  - MAF $\geq$1%
  - or Variants that appear at all in reference
- Rationale: 85% of non-synonymous and 90% of stop-gain/splice-disrupting variants are rare

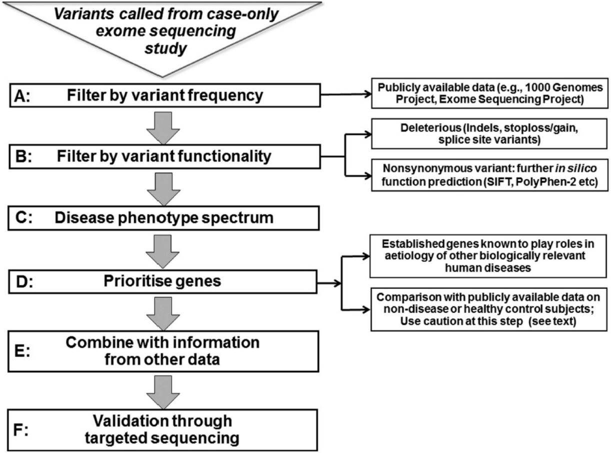
## Variant Functionality Filtering

- Functionality scores for individual variants
- SIFT PolyPhen-2, others.
- High sensitivity, but low specificity

# Modest *n*: Further Prioritization

- **Disease Phenotype Spectrum:**
- **Gene Prioritization:** knowledge on which genes play role in etiology of same or related disease
- **Publicly Available Controls:** Similar to reference data, but actual association analysis; same cautions
- **Other Genomic Data:** Integration with multiple sources of evidence
- **Validation:** Targeted sequencing of new cases and controls is only way to statistically validate findings

# Filtering Summary



*Variants called from case-only exome sequencing study*

A: Filter by variant frequency

Publicly available data (e.g., 1000 Genomes Project, Exome Sequencing Project)

B: Filter by variant functionality

Deleterious (Indels, stoploss/gain, splice site variants)

Nonsynonymous variant: further *in silico* function prediction (SIFT, PolyPhen-2 etc)

C: Disease phenotype spectrum

D: Prioritise genes

Established genes known to play roles in aetiology of other biologically relevant human diseases

Comparison with publicly available data on non-disease or healthy control subjects; Use caution at this step (see text)

E: Combine with information from other data

F: Validation through targeted sequencing

Reference: L. Wu, et al. (2015) *J. Med. Genet.*.

# Additional Concerns

- Quality control:
  - Are the observed variants really variants?
  - Batch effects
  - Some standard pipelines now in place
- Population stratification:
  - Common strategy: just use same PCs from common variant analysis to correct for PS
  - Some evidence that rare variants require special accommodation (much larger number of PCs)
- Accommodating common variants:
  - What do you do with common variants?
  - (a) Assess joint effect with rare variants
  - (b) Adjust for effect of common variants

# Additional Concerns

- Prediction
    - In a new population (sample), we're unlikely to see the same variants and we're likely to see a lot of variants not previously observed
- Prioritization of individual variants
    - How to choose individual causal variants?
    - Some work on variable selection methods, but no ability to control type I error.
    - Bioinformatics and functionality tools may be useful
- Incorporation of functional information and other genomic data

# Additional Concerns

- Design Choices
  - Want to enrich for variants (extreme phenotypes)
  - Some of these designs require specialized methods
  - Stuck with the design chosen
- Dealing with admixed populations
- Related individuals
- Tim: what is a “rare variant”?
- (Statistically) complex phenotypes