Power Analysis for Single and Rare Variant Aggregate Association Analyses

Why Estimate Sample Sizes and/or Power?

- Not wasting your time and money
  - Carrying out a study for which you will never find a true association due to inadequate sample sizes
- Almost always necessary for grant proposals
  - Usually will be denied funding if cannot demonstrate planned study has adequate power

Power and Sample Size Estimation for Case-Control Data

- The correct α must be use for sample size estimation/power analysis
- Type I (α) the probability of rejecting the null hypothesis of no association when it is true
- Due to multiple testing a more stringent value than α=0.05 is used in order to control the Family Wise Error Rate

Power and Sample Size Estimation for Case-Control Data

- GWAS of common variants where each variant is test separately
  - $\alpha = 5 \times 10^{-8}$ (Bonferroni Correction for testing 1,000,000 variant sites)
  - Shown to be a good approximation for the effective number of tests
    - Valid even when more than 1,000,000 variant sites tested
  - Effective number of tests is dependent of the LD structure
- Analysis of individual variants for whole genome sequence data
  - More rare variants than common variants
    - Also have lower levels of LD than between common variants
  - The number of effective tests is higher than for analysis limited to common variants
  - α yet to be determined

Determining Genome-wide Significance Levels

- Using genotypes from the Wellcome Trust Case-Control Consortium
- Dudbridge and Gusnato, Genet Epidemiol 2008
- Estimated a genome wide significance threshold for the UK European population
- By sub-sampling the genotypes at increasing densities and using permutation to estimate the nominal p-value for 5% family-wise error
- Then extrapolating to infinite density
- The genome wide significance threshold was estimated to be ~$7.2X10^{-8}$
- Estimate is based on LD structure for Europeans
  - Not sufficiently stringent for populations of African Ancestry

Power and Sample Size for Aggregate Rare Variant Tests

- For gene based methods a Bonferroni correction for the number of genes/regions tested is used
  - e.g. 20,000 genes significance level $\alpha = 2.5 \times 10^{-6}$
    - Can use a less stringent criteria
      - Not all genes have two or more variants
        - » Divide 0.05 by number of genes tested
    - If units other than genes used may have to use a more stringent
- Little LD between variants in separate genes
  - Little to no correlation between tests
    - Bonferroni correction is not overly stringent

Power and Sample Size for Replication Studies

- For replication studies can base the significance level (α)
- On the number of genes/variants being brought from the discovery (stage I) study
- To replication (stage II)
- For example is hypothesized that 20 genes and 80 independent variants will be brought to stage II
  - A Bonferroni correct can be made for performing 100 tests
    - An $\alpha = 5.0 \times 10^{-3}$ cab be used for a family wise error rate of 0.05

Estimating Power/Sample Sizes For Individual Variants

- Can be obtained analytically
- Information necessary
  - Prevalence
  - Risk allele frequency
  - Effect size (odds ratio-for case control data)
  - Genetic model for the susceptibility variant
    - Recessive ($\gamma_1=1$)
    - Dominant ($\gamma_2=\gamma_1$)
    - Additive ($\gamma_2=2\gamma_1-1$)
    - Multiplicative ($\gamma_2=\gamma_1^2$)

Estimating Power/Sample Sizes For Individual Variants

- Usually information on disease prevalence is known from epidemiological data
- A range of risk allele allele frequencies and effect sizes are used
- A variety of genetic models are also used
  - Dominant
  - Additive
  - Multiplicative

Armitage Trend Test

- Power and Sample size
  - Calculated under different models
    - Where γ is the relative risk
      - Multiplicative
        - » $\gamma_2=\gamma_1^2$
      - Additive
        - » $\gamma_2=2\gamma_1-1$
      - Dominant
        - » $\gamma_2=\gamma_1$
      - Recessive
        - » $\gamma_1=1$

Gamma is the Relative Risk

- Many programs work with the relative risk (ɣ)
- Relative risk only approximates odds ratio when disease is rare
  - Not appropriate for common trait
- Example risk variant and marker allele frequency 0.01
  - D' and $r^2=1$

Disease Prevalence	1/2 RR=1.5	2/2 RR=1.5
0.01	1.51	1.51
0.10	1.59	1.59
0.20	1.71	1.71

Disease Prevalence	1/2 RR=1.5	2/2 RR=2.25
0.01	1.51	2.28
0.10	1.59	2.61
0.20	1.71	3.25

Armitage Trend Test - Power Calculations

- Information need
  - Population prevalence
  - Genetic Model
  - Risk allele frequency
- Tools
  - http://ihg.gsf.de/cgi-bin/hw/power2.pl
  - Reference Slager and Schaid 2001

Armitage Test for Trend

Sample size approximations for Armitage's test for trend:

Disease prevalence	0.01
High risk allele frequency	0.05
Type 1 error (alpha)	0.00000005
Power (1- beta)	0.8
Gamma 1	2
Gamma 2	2
Cases / (cases + controls)	0.5

Cases necessary = 1502

Controls necessary = 1502

Cases and controls necessary = 3004

submit Reset

Gamma (genotypic relative risk):
Under a multiplicative model, gamma2 = gamma1^2; under a additive model, gamma2 = 2 * gamma1 - 1; under a dominant model, gamma2 = gamma1; under a recessive model, gamma1 = 1.

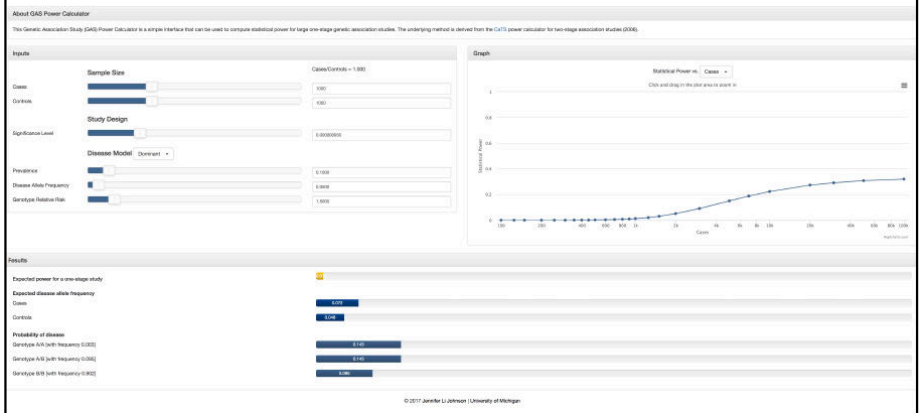
Adapted from:
Slager SL, Schaid DJ: Case-control studies of genetic markers:
Power and sample size approximations for Armitage's test for trend.
Hum Hered 52, 149-153 (2001).
and
Freidlin B, Zheng G, Li Z, Gastwirth JL:
Trend tests for case-control studies of genetic markers:
Power, sample size and robustness.
Hum Hered 53, 146-152 (2002).

Tim M. Strom

Genetic Association Study (GAS) Power Calculator

- http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/index.html
- A one-stage study power calculator
  - Which was derived from CaTs
    - Which is to perform two-stage genome wide association studies
      - Skol et al. 2006
- Cochran Armitage Trend Test
- Displays Graphs of results

GAS Power Calculator



Genetic Power Calculator

- http://zzz.bwh.harvard.edu/gpc/
- S Purcell & P Sham
- Uses the methods described in Sham PC et al. (2000) Am J Hum Genet 66:1616-1630
  - VC QTL linkage for sibships
  - VC QTL association for sibships
  - VC QTL linkage for sibships conditional on the trait
  - TDT for discrete traits
  - Case-Control for discrete traits
  - TDT for quantitative traits
  - Case-Control quantitative traits
- Although input is relative risk
  - Displays odds ratios

Genetic Power Calculator

Case - control for discrete traits

```
High risk allele frequency (A)   : 0.01   (0 - 1)
Prevalence                       : 0.2    (0.0001 - 0.9999)
Genotype relative risk Aa        : 1.5    ( >1 )
Genotype relative risk AA        : 1.5    ( >1 )

D-prime                          : 1      (0 - 1)
Marker allele frequency (B)      : 0.01   (0 - 1)

Number of cases                  : 10000  (0 - 10000000)
Control : case ratio             : 1      ( >0 )
                                          ( 1 = equal number of cases and controls)
                                  [x] Unselected controls? (* see below)

User-defined type I error rate   : 0.00000005 (0.00000001 - 0.5)
User-defined power: determine N  : 0.80   (0 - 1)
(1 - type II error rate)
```

Process Reset

Created by Shaun Purcell 24.Oct.2008

PAWE

- Power Association With Errors
  - Will give same results for case-control studies of discrete traits as Genetic Power Calculator when calculations are done without errors
- Four different error models can be used
  - See online documentation for complete explanation
- Can either perform:
  - Power calculations for a fixed sample size
  - Sample size calculations for a fixed power
- The genotype frequencies can be generated either using a:
  - Genetic model free method or
  - Genetic model based method

Quanto

- Provides sample size and power calculations for
- Genetic and environmental main effects
- Interactions
  - Gene x gene
  - Gene x environment
- Sample & power calculations can be carried for:
  - Case-control
    - Unmatched
    - Matched
  - Case-sibling
  - Case-parent (trios)
    - Quantitative
    - Qualitative
  - Independent sample of individuals
    - Quantitative traits
      - Assumption sampled from a random population

Linkage Disequilibrium (LD)

- Power will be reduced if causal variant is not in perfect LD (r^2=1) with the tag SNP
- Can adjust sample size when r^2 <1 to increase power to the same level as when r^2=1
- Can estimate sample size when $r^2 \neq 1$
  - $N/r^2=N'$
  - Valid only for multiplicative model
  - (Pritchard and Przeworski, 2001)
- Power calculation almost always assume that r^2=1

Power Analysis for Rare Variant Aggregate Association Tests

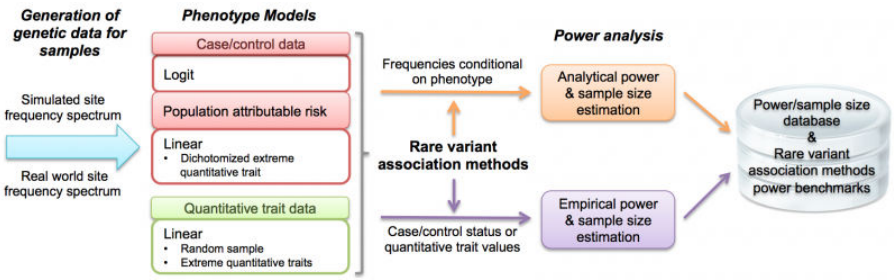
- Many unknown parameters must be modeled
  - Allelic architecture within a genetic region
    - Varied across genes and populations
  - Effects of variants within a region
    - Fixed or varied effect sizes of causal variants
    - Bidirectional effect of variants
    - Proportion of non-causal variants
- Power usually must be estimated empirically
- Simplified assumptions can be made to obtain analytical estimates
  - All variants have the same effect size
  - No non-causal variants

SKAT Power Calculator

- R Library
- Provides a haplotype matrix
  - 10,000 haplotypes over 200kb region
  - Simulated using a calibrated coalescent model (cosi)
  - Mimicking linkage disequilibrium structure of European ancestry
  - User can also provide haplotype data
- Power and sample size calculations for binary and quantitative traits
- User specify proportion of variants that increase or lower risk

SEQPower

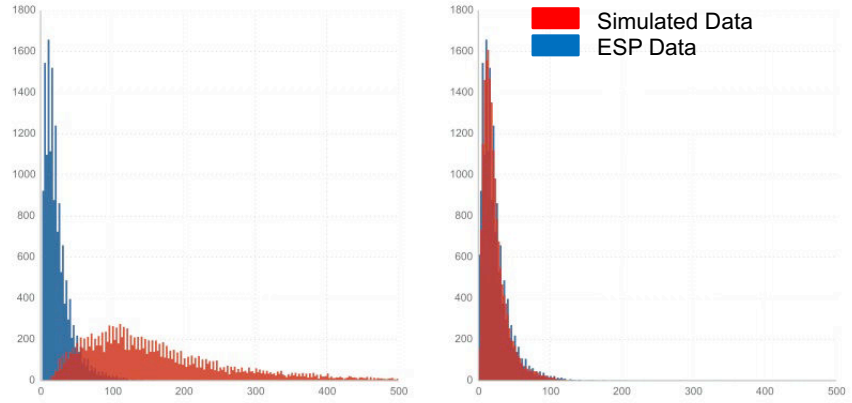
http://www.bioinformatics.org/spower/



Wang et al. 2014 Bioinformatics

Does Generating Variant Data Using the European Population Demographic Model Perform Well?

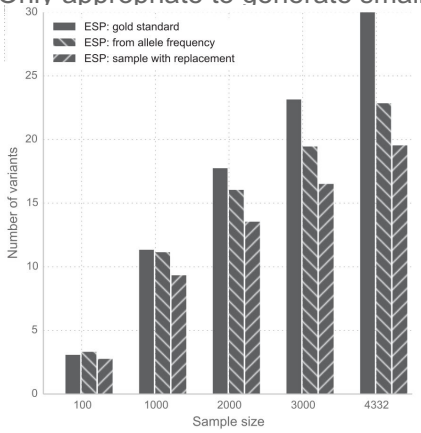
Distribution of number of variants per gene



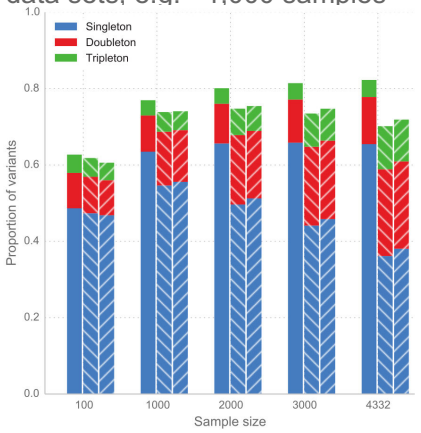
- Simulated variant counts based on the entire simulated population
- Simulated variant counts based on haplotype pool down-sampled to ESP size

Simulating Data Using Sequence Data (ESP)*

*Only appropriate to generate small data sets, e.g. <1,000 samples



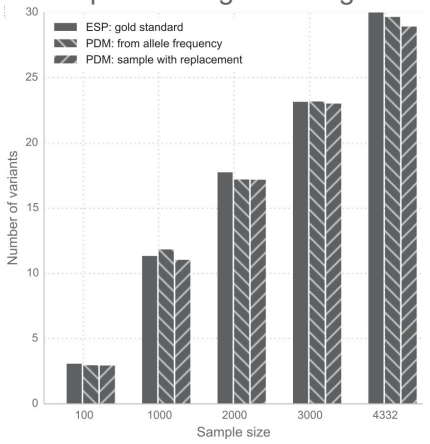
Number of Variant Sites



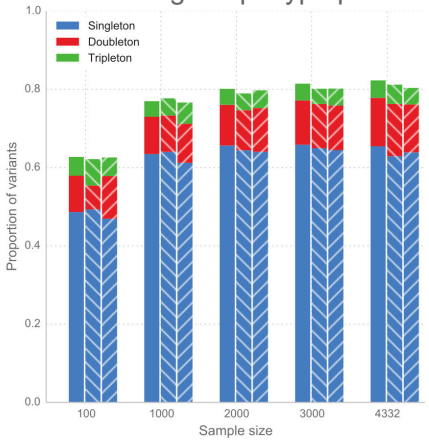
Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

Simulating Data: Using Population Demographic Models (PDM)*

*Resample or using MAF to generate data from large haplotype pools



Number of Variant Sites



Proportion of Variant Sites that are Singletons, Doubletons and Tripletons

Simulation Studies to Evaluate Power for Rare Variant Association Studies

- It is unknown which genes are important in disease etiology
  - Correct allelic architecture is unknown
- Can get a better understanding of power to detect associations by generating variants for the entire exome
- Use a variety of disease models
  - Odds ratios
  - Proportion of pathogenic variants
- Analyze of all genes
  - e.g. those with 3 or more variant sites
- Determine power as the proportion of genes that meet exome-wide significance (alpha=2.5×10^{-6})

Power Analysis

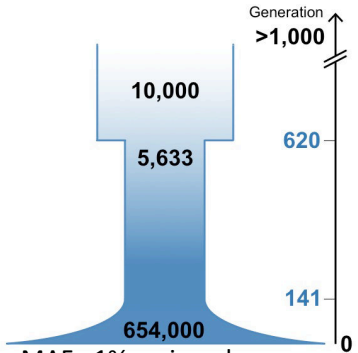
- For tests of individual variants
  - Power depended on sample size, disease prevalence, minor allele frequency, genetic model and variant effect size
- For rare variants (aggregate association tests)
  - Also dependent on the allelic architecture
    - Cumulative variant frequency within analyzed region
    - Proportion of causal variants
      - How much contamination by non-causal variants
    - Effect sizes the same the same or different across gene regions
      - Effects of variants in the same or different directions
        - » Protective and detrimental
        - » Increase and decrease quantitative trait values

Power Analysis Rare Variants (Aggregate Association Tests)

- Power will not only vary between traits greatly
- The power to detect an association will also vary drastically between genes
- For some genes even with hundreds of thousands of samples power will still be low, while for others a few thousand samples may be sufficient

How Large of a Sample Size is Necessary to Detect Rare Variant Associations?

- Data generated on 18,397 genes
- Variant data simulated using a European population demographic model
  - Gazave et al. 2013



- Every missense, nonsense and splice with a MAF≤ 1% assigned an odds ratio of 1.5
- Sample sizes to detect X number of genes determined for
  - $\alpha = 2.5 \times 10^{-6}$
  - power=0.8

Sample Sizes Necessary to Detect an Association (Case-Control Data)