Rare Variant Association

Robert Porsch

March 8, 2017

Table of Contents

Why rare variants?

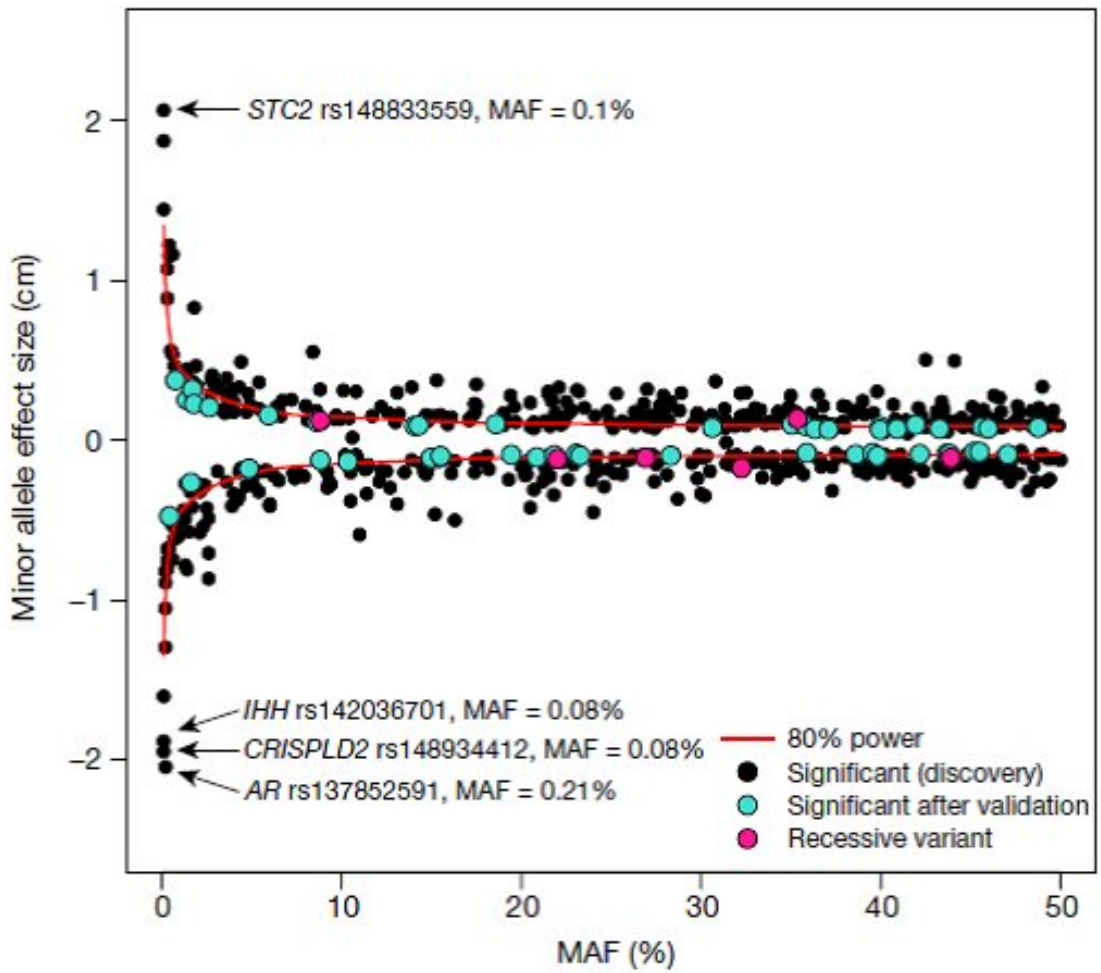
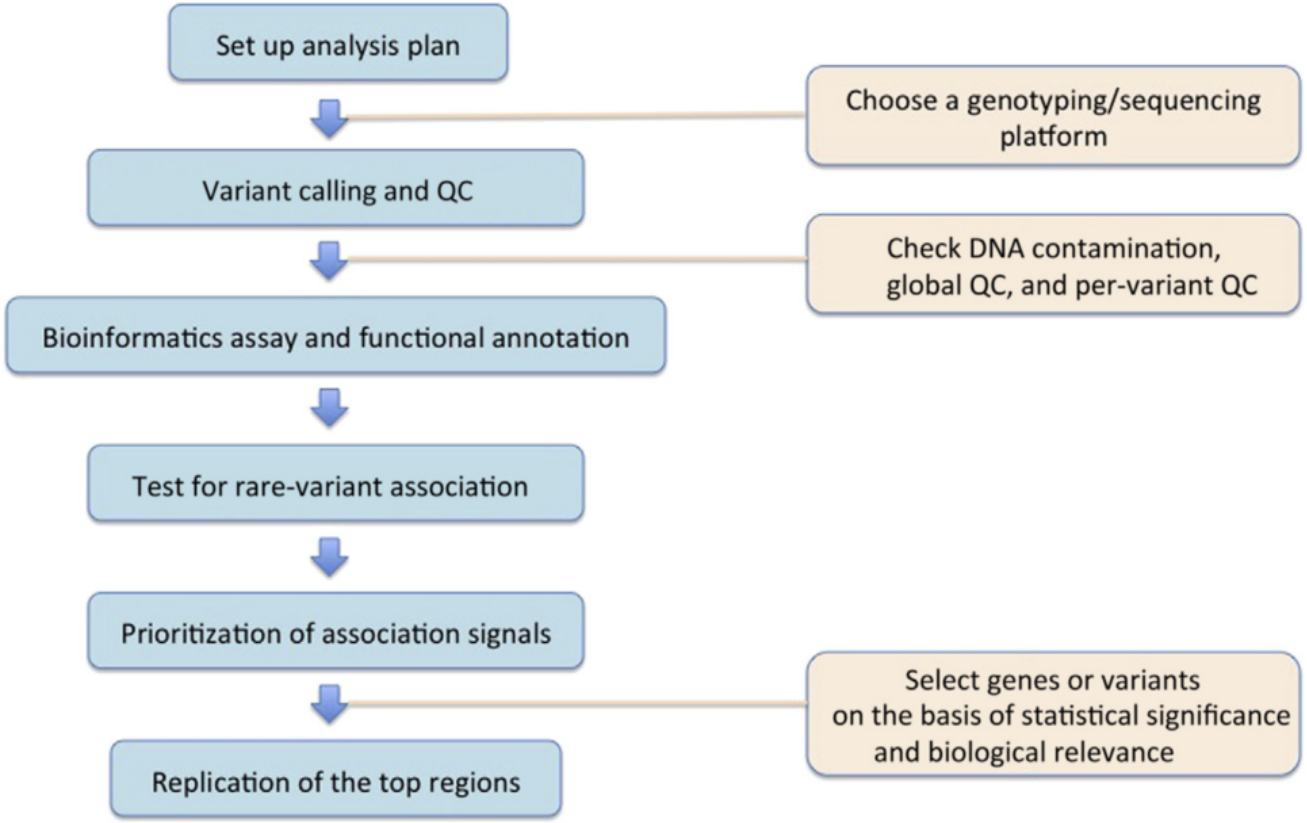


Figure: Effect size of rare and common variants in height [1]

Table of Contents

General Data-Processing and Analytics [2]



Design Strategies

Custom Genotyping Array

- cost-effective alternative to sequencing
- allows to genotype more subjects
- limited range of variants which can be targeted
- mixed experience

Exome Sequencing

- cost-effective
- successful identified a number of causal variants
- higher average depth
- focus on high-value proportions of the genome
- only coding regions

Targeted-Region Sequencing

- cost-effective approach for higher-priority regions
- prior information required
- nearly as expensive as exome sequencing

Low-Depth WGS

- covers the whole genome
- cost-effective alternative ($MAF > 0.002$ with $n = 3,000$ at 4x is more powerful as $n = 2,000$ with 30x)
- higher genotype error rate

Extreme Sampling

- improves association power
- allows for a smaller sample size
- reduces generalizability to the population
- sensitive to outliers
- sampling bias, assumption of normality
- reduction in power if variants in the two extremes have different directions

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

Ronald Fisher

Platform, Power, and others

- Rare variant associations require careful planing
- Minimize technical noise or confounders
  - Do not use multiple different platforms
- Pre-define analysis plan
- Estimate needed sample size!!!
  - SKAT R package
  - write your own simulation
  - 'What effect size you wouldn't mind missing?'

Table of Contents

Variant Calling and QC

- Multi step and error prone process
- Quality Control! Quality Control! Quality Control!
  - Contamination, quality of reads, alignment, and calling quality etc.
- quality scores can be later used to prioritize variants
- GATK and others
- A talk by itself

Outline

Functional Annotations

Goals

- predict the impact of variants
  synonymous, missense, nonsense, splicing side
- deleterious score
- MAF in reference dataset
- Define genomic areas of interest (Genes, Pathways)

Which variants should I include?

- Genes with ≥ 3 variants
- non-synonymous variants
- $MAF \leq 1\%$
- These are all suggestions!

Software

- KGGseq (advertisement)
  - provides a comprehensive framework
  - annotation, association testing, etc.
- PlinkSeq
- ANNOVAR
- and others (SNPeff, AnnTools, VARIANT)

P-value fishing

- Have a clear analysis plan
  - Pathway based analysis?
    Which pathway set?
  - Gene based analysis
- Be open about what you have done!

Table of Contents

Outline

Single Variant Test

- very large sample size needed
- sampling alleles with MAF 0.5% or 0.05% with 99% requires 460 or 4,600 subjects
- OR $= 1.4$, MAF $= \{0.1, 0.01, 0.001\}$, prevalence $= 5\%$ requires n $= \{4,600, 54,000, 540,000\}$ with $\alpha = 5 \times 10^{-8}$
- p-values might not be accurate if number of subjects ith variant is small
- might still be successful if effect is very large and sample size is large
- QQ and Manhattan plot can be used (evaluating quality and stratification)

Outline

Why region based?

- sample sizes is usually too small for single variant associations
- increase in statistical power
- α level is more achievable
- What about non-coding regions???

Pathway or Gene based testing?

- It depends!
- Prior Hypothesis?
- Sample size?

Be aware of pathways!

- overlapping pathways
- multiple different ways to define pathways
- Do not end up with more pathways than you have genes

Different classes of tests

- There are 3 major classes of tests
  - aggregating
  - variance component
  - omnibus tests
- each has important advantages and disadvantages
- it is important to understand the differences
- performance of different tests can give valuable information about the disease architecture

General Model

Lets let G be the $n \times p$ genotype matrix. For example:

$$G = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad (1)$$

Then lets also say that we have a dichotomous phenotype y and that y_i (the i^{th} subject) follows some distribution

$$h(\mu_i) = \alpha_0 + \boldsymbol{\alpha}' X_i + \boldsymbol{\beta}' G_i$$

in which $h(\mu) = logit(\mu)$, $\alpha = \{\alpha_1, \ldots, \alpha_q\}$ and $\beta = \{\beta_1, \ldots, \beta_p\}$ are the regression coefficients of the covariates and allele counts.
The score statistic of the marginal model for variant j is:

$$S_j = \sum_{i=1}^{n} G_{ij}(y_i - \hat{\mu}_i)$$

where $\hat{\mu}_i$ is the estimated mean under $H_0 : \beta = 0$ and is obtained by $h(\mu_i) = \alpha_0 + \alpha' X_i$

The Simple Burden Test

An aggregation test which collapses information for multiple variants. It is counting the number of variants per subject in region m

$$C_i = \sum_{j=1}^{m} w_j G_{ij}$$

in which w_j is some threshold indicator or weight. This is identical to testing $H_0 : \beta = 0$ in $h(\mu_i) = \alpha_0 + \alpha' X_i + \beta C_i$ the score statistic is then:

$$Q_{burden} = (\sum_{j=1}^{m} w_j S_j)^2$$

Flavours and Assumptions

- can accommodate different genetic models
- MAF threshold
- different weights are possible:
  - weight functions ($w_j = beta(MAF_j, a_1, a_2)$)
  - bioinformatic information
- there are some other tests outside of this regression framework:
  - Hoteling's Test (different categories of variants)
  - Wilcoxon-Rank test
- Assumptions:
  - all variants in region m are causal and have the same direction
  - violation leads to substantial loss of power

Adaptive Burden Test

- Addresses strong assumptions in the burden test.
- robust in the presence of null-variants
- allow for both damaging and protective effects
- most are based on two-step approaches
- a number of different methods:
  - aSum test
  - EREC
- Variable Threshold (VT) Test
  - selects the optimal MAF threshold for the burden test
- need fewer assumptions
- computational expensive

Variance Component Tests

Most common tests are C-Alpha, SSU, and SKAT. The SKAT test statistic is

$$Q_{SKAT} = \sum_{j=1}^{m} w_j^2 S_j^2$$

since SKAT uses S_j^2 instead of S_j it is robust to groupings of protective and damaging effects.

- more powerful if a region has **many non-causal** or **both protective and damaging variants**
- reduced power if a region has a higher proportion of causal variants with the same direction

So what to use???

- Depends on the unknown underlying genetic architecture
- Gene dependent
- Disease dependent

Omnibus Tests to the rescue

Omnibus Tests

- naive approach of taking the lowest p-value results in type I error inflation
- Fisher p-value combination:

$$Q_{fisher} = -2\log(p_{SKAT}) - 2\log(p_{burden})$$

- adaptive combination (SKAT-O):

$$Q_{opt} = (1-\ell)Q_{SKAT} + \ell Q_{burden}, 0 \leq \ell \leq 1$$

- ℓ is estimated with the minimum p-value of Q_{opt} over a grid of ℓs
- Disadvantages:
  - less powerful if one of the tests assumptions is largely true

What software to use?

Software

- KGGseq
- RvTests
- PlinkSeq
- R packages
- etc.

Which to choose?

- Doesn't really matter
- Chose one you understand!

KGGSeq

- HKU developed
- contains nearly everything, from annotation, QC to gene/pathway based testing
- command line, and GUI interface
- only requires vcf file and ped file

RvTests

- contains a large amount of different gene/pathway based tests
- Meta-analysis module
- annotations have to be done outside
- popular

P-Value Fishing

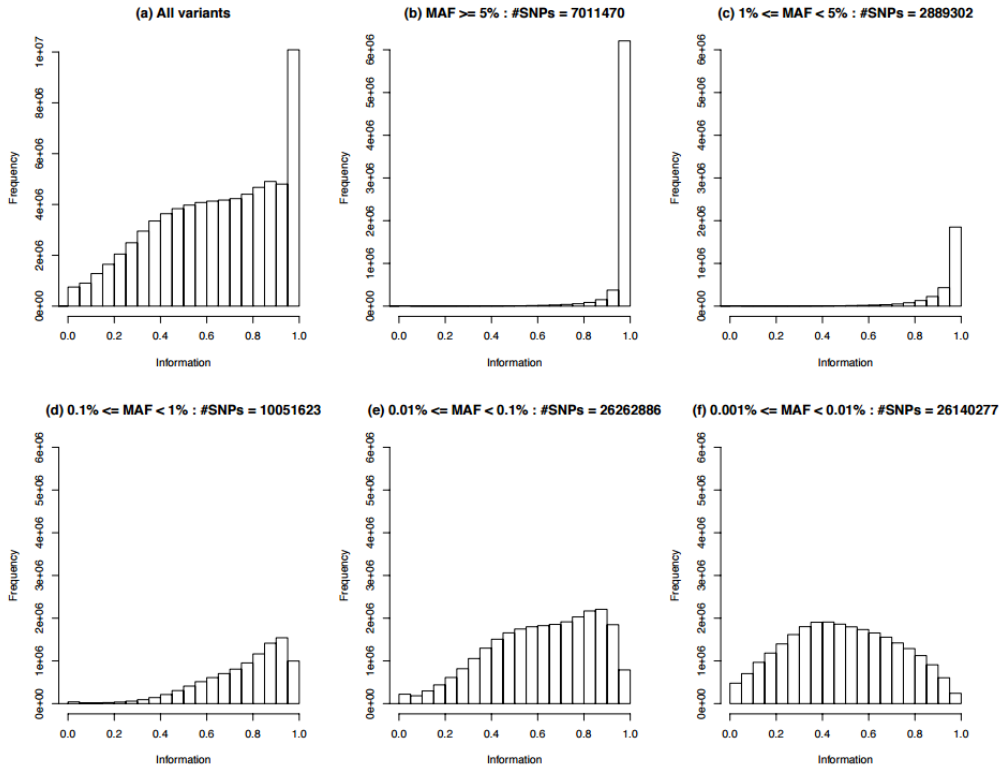
- there are a lot of different analysis options
- Be open about what you have done
- include non-significant findings
- DO NOT JUST SEARCH UNTIL YOU FIND SOMETHING
  - State your variant filtering methods, statistical tests, gene grouping, pathways
- adjust your α level if necessary
- Replications!!!

Table of Contents

Population Stratification Adjustments

- PCA and mixed models assume a smooth distribution of MAF over geographical space [3]
- rare variants are often sharply localized $\rightarrow$ PCA & LMM might fail
- targeted and exome sequencing contains limited information
- possible solutions:
  - use PCA to guide matching
  - PCA on rare variants is not more effective than on all or common variants
  - one can use off-target reads to improve PCA in targeted/exome sequencing
- aim for a homogeneous population
- tests behave differently given population differences

UKBioBank Example



https://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf

Imputation of rare variants

- Thought not possible just a few year ago
- still harbours some problems
  - no or less singeltons
  - only MAF bands can be considered
- larger sample size might solve some of the problems

Table of Contents

Conclusion

- array-genotype and exome sequencing while remain important until cost of WGS drops
- improve power by using publicly available ancestry-matched controls
- use a single platform for cases and controls (same aligner, same everything)
- challenges for WGS remain:
  - limited variant information to prioritize
  - limited information for variant grouping
- use an Omnibus tests when architecture is not known
- family-based association studies are an alternative

References I

Marouli, E. *et al.*
Rare and low-frequency coding variants alter human adult height.
Nature **542**, 186–190 (2017).
URL http://www.nature.com/doifinder/10.1038/nature21039.

Lee, S., Abecasis, G., Boehnke, M. & Lin, X.
Rare-Variant Association Analysis: Study Designs and Statistical Tests.
The American Journal of Human Genetics **95**, 5–23 (2014).
URL http://linkinghub.elsevier.com/retrieve/pii/S0002929714002717.

References II

Wang, C. *et al.*
Ancestry estimation and control of population stratification for sequence-based association studies.
Nature Genetics **46**, 409–415 (2014).
URL
http://www.nature.com/doifinder/10.1038/ng.2924.