Computational Workflows for Genome-Wide Association Study: II. Rare Variants

Sorin Istrail

Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

November 3, 2015

Outline

Rare Alleles in GWAS

The problem of rare allele is crucial for the GWAS analysis.

- The frequency and penetrance of causal alleles affect the statistical power to detect these alleles. Power increases with increased frequency and increased penetrance.
- A new measure, a single parameter: **the amount of variation in phenotype that can be "explained" by the genetic variant in question** might better capture this dependency.
- With such a measure, **rare highly penetrant alleles** and **common low penetrant alleles** are now on an equal footing.

Rare Alleles in GWAS

- The power to detect such an allele, therefore, depends on what is ultimately the most relevant measure of a genetic variant's contribution: *the proportion of phenotypic variance in the population that is explained by a particular variant.*
- This means, however, in rough terms, that rare variants of modest effects will be difficult to detect by any method because they can only explain a trivial fraction of the variation in a trait.

HapMap Designed for Common SNPs not rare SNPs

More reasons that rare alleles might be difficult to detect by association are:

- *even if they would have strong effect* because they are less represented in SNPs databases
- tag SNPs approaches are currently designed to tag common SNPs (usually with frequencies $> 5\%$)
- current state of the art population genetic models indicate that evolutionarily most rare allele with frequencies $< 5\%$ are likely to be recent mutations; old alleles tend to either disappear or become common; for them there will be less time for recombination and mutation - so their haplotype where they occur may not be disrupted. **Rare variants are expected to be on a single long haplotype**

Explosive Human Population Growth: Excess of Rare Genetic Variants

Seminal paper by Alon Keinan and Andy Clark, *"Recent explosive human population growth has results in an excess of Rare genetic variants"* SCIENCE, May 2012

- In the last 10,000 years the human population had an explosion in numbers: from a few million people 10,000 years ago to 7 billion today
- Three orders of magnitude expansion within about 400 generations (one generation about 25 years)
- Therefore a huge number of rare variants are relatively recent

Very hard to distinguish Rare SNPs from sequencing errors

- It is now possible to sequence large number of individuals
- such an approach introduces a new scale to the problem of **false positives** among newly identified variants such as SNPs.
- The ability to easily distinguish SNPs present only once in the sample (singletons) from sequencing errors decreases as the sample size increases.

Very hard to distinguish Rare SNPs from sequencing errors

- Although our sequencing technologies accuracy are improving, sequencing rrors will always be part of the picture.
- A sequencing error rate of 1 in a 10,000 bases (the *world standard for genomic sequence fidelity* is known as Bermuda Standards Schmutz et al in Nature (2004). This standard stated that genome assembly finished sequence should contain less than one error per 10,000 DNA bases (99.99% accuracy).

ery hard to distinguish Rare SNPs from sequencing errors

- In a sample of 10,000 individuals each base pair will have two errors on the average across the sample and the majority of monomorphic sites will appear polymorphic (most often as a singleton or doubleton; i.e., with the rare allele present in one or two copies in the sample.
- But if we now start filtering or correcting data it will lead to *missing many rare variants* because they are not observed as reliably.
- Any analysis of large sample sizes must account for the uncertainty inherent in sequencing by considering the variant calls probabilistically, and
- secondary validation or rare variants by an alternate sequencing procedure is essential.