Statistical Analysis

Goncalo Abecasis, Joel Hirschhorn,
Suzanne Leal & Dan MacArthur

Commonality of Complex and Mendelian Traits

- For both complex traits and Mendelian traits
  - Statistical evidence should be used in gene identification
    - Findings should be replicated
  - Although genes can be implicated
    - Difficult to know with certainty the causality of very rare variants
      - e.g. variants which are only observed only once or within a single family

Complex Traits

- Rare variants (e.g. MAF< 0.01) will have effects from large to approaching odd ratios=1.0
  - Large sample sizes are important for detecting associations with gene regions
    - International consortia for generating and sharing data across many traits as well as matched controls are necessary
    - Additionally very large publicly available cohorts of sequenced control individuals will be valuable for many different studies

Complex traits

- For very rare variants (e.g. MAF≤0.0005), even with large samples size not possible to test for associations with individual variants
  - For common rare variant (e.g. MAF>0.0005) can test for associations with individual rare variants
- Can use rare variant association tests, which test for associations with rare variants in aggregate across a region, e.g. gene
  - Even if association with the region is replicated
    - It does not tell if individual very rare variants are associated or not

Complex traits

- Caution should be used in that association results may be confounded by population substructure/admixture
  - Not clear if methods used to control for population substructure/admixture are adequate for associations studies of rare variants

Complex Traits

- For genes regions and higher frequency rare variants false positive can be avoided/reduced
  - By avoiding multiple testing which is not controlled for
  - Demand initial findings meet statistical rigor
    - Significance levels still need to be determined
- Necessary that findings are replicated in an independent sample

Complex Traits

- Although we can provide strong statistical evidence for regions, genes and higher frequency rare variants
  - Statistical evidence cannot be used to determine causality of very rare variants by testing individual variants
    - e.g. singletons or variants seen only in one family

Mendelian Traits

- Need to have more evidence than a single variant in an affected individual and variant not observed in databases
  - A small family segregating a rare variant is also not sufficient evidence
- Being able to establish linkage to a genetic region either in a few large families or multiple small families can provide statistical evidence that a region is involved in disease etiology
  - Can read the old literature to learn exactly how to perform parametric linkage analysis

Mendelian Traits

- Next generation sequencing, in a subset of family members, can be used to find variants within the implicated genetic region
- Having multiple families with variants, either identical or different, in the same gene provides evidence of involvement of the gene in disease etiology
  - Additionally these variants should be absent or only in very low frequencies in controls
  - Statistical tests can be performed to show there is differences in the frequencies in cases and controls
- Can provide strong evidence that a gene is involved in disease etiology

Mendelian Traits

- It is still not possible to say that variants are causal if they are only found in one family
  - Even if a family is large and the variant segregates with disease etiology
    - If within the linkage interval would expect even non-causal variants to segregate with disease etiology
      - Variant will be in linkage disequilibrium with the causal variant

Mendelian Traits

- If families are not available can also test for associations for Mendelian traits using the same rare variant aggregate association tests which were developed for complex traits
  - However can be problematic for diseases with locus heterogeneity or very reduced penetrance
    - Very large sample sizes are necessary

Very Rare Variants

- Seeing variant at higher frequencies in controls than cases rules out causality of variant
- However likewise if a variant is not seen in controls it is not evidence of causality
  - Due to recent population growth there are many extremely rare variants which in some cases are private
    - Therefore even if a variant not present in a very large ethnically matched control data set it is not evidence that a variant is causal

de Novo Variants

- Although non-synonymous de Novo mutations occur only once in every ~2 exomes
  - Some of these variants will fall simply by chance in genes for which a “story” can be built
    - Therefore an event being de Novo is not evidence that it is causal

Experimental Support

- Experimental support for a variant's effect on gene function can be seen as complementary to, and not a replacement for, strong direct statistical support for phenotype association
  - Functionality does not prove causality

Statistical Tests for Very Rare Variants

- However a class of variants only seen in cases but not in controls could provide additional evidence of causality and could be held to statistical standards
  - But will not work in all cases
- Test criteria have to be formed a priori and not based upon data observation
  - i.e. Not forming a statistical test to fit the particular example

Discussion Questions

- What are some of the caveats of testing for rare associations in complex traits?
- What is convincing evidence that a gene or a variant is involved in disease etiology?
- What type of novel application of statistical tests could be developed to provide evidence that very rare variants are involved in disease etiology?