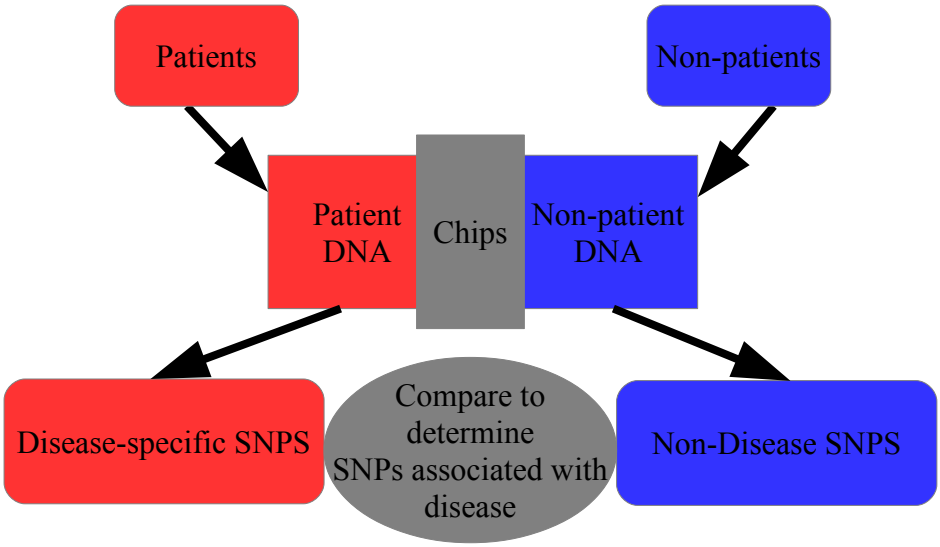
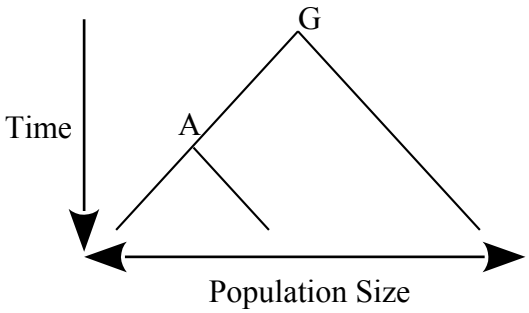
Identity By Descent Mapping for the Discovery of Novel Susceptibility Loci in Large Disease Cohorts
Lecturer: Dan Newburger Scribe: Rachel Mellon
March 3, 2015

- What is a **complex disease/complex condition**?
  - disease caused by both genetic factors and environmental factors
  - eg: identical twins have the same genetic makeup, but one might live a healthier lifestyle, causing different diseases to surface for each twin
  - cancer and heart disease are examples of complex conditions
  - eg: Crohn's Disease
    - 54% genetic (specific alleles, for example), 46% environmental (diet, for example)
- What is **heritability**?
  - the proportion of inter-individual differences (variance) in a trait that are the result of genetic factors
- **Association Studies** can be used to understand complex disease

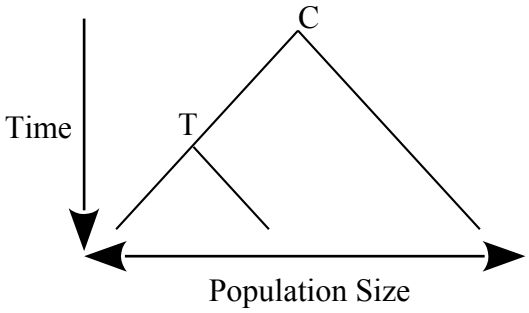


  - can help us **identify susceptibility variants** to different diseases, which can have many positive applications
    - novel biological insights → **clinical advances**
      - therapeutic targets, biomarkers, prevention
    - improved measures of individual aetiological processes → **personalized medicine**
      - diagnostics, prognostics, therapeutic optimization
  - could lead to a new standard of care!
  - by 2012, more than 140 studies with 10,000+ individuals participating
- **Missing Heritability**
  - for most complex diseases, we currently only have between 5% and 50% of the heritability of that disease explained
  - how can we determine the other 50%-95%?
  - GWAS (current method) is great for diseases where the associated allele is very common in the population but the disease itself is not
  - IBD Mapping can be used for set of problems with a missing variant: where there are low-frequency variants with an intermediate effect causing disease

        - GWAS won't work here because of indirect association
- Association Mapping Studies – Standard GWAS Significance Test
    - Example 1: allele does not confer disease
        - Case (patient) Alleles: 64 GG, 32 GA, 4 AA
        - Control (healthy) Alleles: 63 GG, 32 GA, 5 AA
        - Test for significance of G/A SNP: Chi-square $p \approx 0.90$



    - Example 2: allele **does** confer disease
        - Case (patient) Alleles: 36 CC, 48 CT, 16 TT
        - Control (healthy) Alleles: 49 CC, 42 CT, 9 TT
        - Test for significance of C/T SNP: Chi-square $p \approx 0.01$

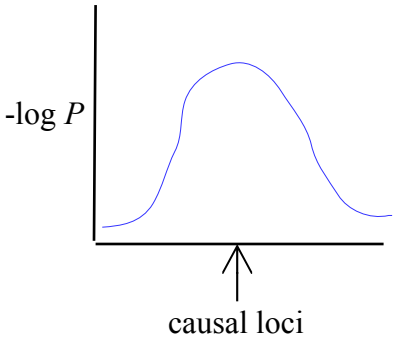


    - But it's not enough to say the C/T SNP causes the disease! Many times, the SNP that GWAS can identify as correlated with disease has a high co-occurance with the SNP that actually causes the genetic difference. These are called indirect associations.
        - **Linkage Disequilibrium** → two close together SNPs are more likely to be inherited in common
        - **Direct Association** → if a SNP is disease relevant, the GWAS method (chi-squared test) will identify it at should above in Example 2
        - **Indirect Association** → if there are multiple close together SNPs, it's possible that the one on the chip is irrelevant to the disease, but because of Linkage Disequilibrium can still help us identify the presence of that disease
    - Traditional GWAS doesn't work for recent/rare causal variants
        - If multiple rare variants of intermediate effect size exist at the same locus, can use IBD mapping instead
- What is **identity by descent (IBD)**?
    - Two corresponding regions in the chromosomes of two different individuals from the same ancestor
        - captures rare variant loci!
    - Run pairwise IBD detection across all cases/controls

- in the cases, there will be connected clusters that might not have been obvious in the initial selection
- Pairwise mapping statistic:

$$\frac{\text{(\# connection in cases)}}{\text{(\# cases choose 2)}} - \frac{\text{(\# connections in control)}}{\text{(\# controls choose 2)}}$$

- Permutation testing:



- Why isn't IBD Mapping used for everything?
  - Great accuracy!
  - Speed: can't analyze cohorts of size > 10,000, because running one analysis on a data set that size takes a ton of hours
  - Current IBD methods developed by grad students (>30,000 work)
    - Parente2
      - uses mathematical representation of relationships
      - uses sparse sliding windows
      - outputs score that determines related individuals and unrelated individuals
    - SpeeDB
      - super fast for filtering out obviously non-IBD regions by efficiently filtering candidate IBD
      - Example of 4 SNPs in the test region:

s1	AA	AA	compatible
s2	AA	AT	compatible
s3	AT	AT	compatible
s4	AA	TT	conflict

- **IBD Detection Pipeline** (Dan Newburger's research)
  - Training pipeline (begin with 1000 individuals):
    - Control Subset → Phase with HAPI-UR → correct missing data → Train Parente on haplotypes
  - Testing pipeline
    - Cohort minus training controls → **Filter IBD candidates using SpeeDB → Parente inference on genotypes** → Analyze!