

Association Mapping part II

- Degrees of freedom
- Defining a population for an association study
- TDT and other "built-in control" methods
- HapMap project

Degrees of freedom

- I recommend the following essay by Gerard E. Dallal:
- <http://www.tufts.edu/~gdallal/dof.htm>
- What follows is my attempt to summarize it

Degrees of freedom

- Our data consist of a collection of observations
- We will want to estimating both parameters and variability
- Each observation provides a degree of freedom
- Some of those are “used up” by estimating parameters, such as the mean
- Left-over degrees of freedom are available to estimate variability

Mendelian genetics example

- Color in carnations is semi-dominant: RR red, RW pink, WW white
- We observe 43 red, 47 pink and 10 white flowered plants: 3 df
- Null hypothesis: Hardy-Weinberg $p^2 + 2pq + q^2$
- To use this null hypothesis we'll need to estimate p , which costs 1 df
- We'll also need to count our total number of plants, which costs 1 df
- One df is left, so we can do a test for variability (how far from H-W are we?)

Mendelian genetics example

- Color in a different flower is dominant: RR and RW red, WW white
- We observe 90 red and 10 white flowered plants: 2 df
- Null hypothesis: Hardy-Weinberg $p^2 + 2pq + q^2$
- To use this null hypothesis we'll need to estimate p , which costs 1 df
- We'll also need to count our total number of plants, which costs 1
- Nothing is left to do a test of variability!
- If you try to do the test, you will see that χ^2 is always 0, no matter what the data are; all the available information has already been used up

Defining a population

- In 1989 my thesis work relied on published HLA frequency tables
- European, Asian, African—enough resolution?
- For many mapping studies, no
- “African” is particularly problematic as African lineages have not been severely bottlenecked

Forensic populations

- Similar issues arise in DNA forensics
- $P(D|suspect)/P(D|population)$
- Problem case:
 - Suspect is Hispanic
 - Perpetrator is also Hispanic
 - “European” population data used in equation above
 - False conviction possible due to allele sharing between Hispanic suspect and perpetrator
 - Hispanic frequencies would be better, but....
 - Are Cuban Hispanics the same as Mexican or Brazilian Hispanics?

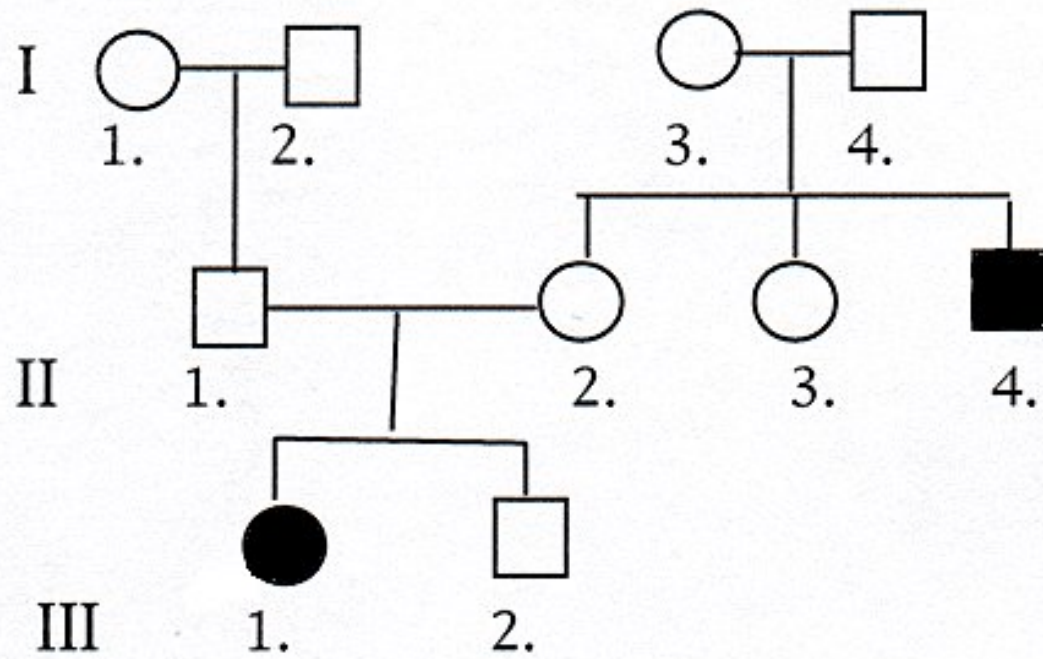
Family-based controls

- Some methods use family data to do association (not linkage) studies
- The idea is that the same individuals provide both “case” and “control” data
- This guarantees that both come from the same population

Transmission distortion test (TDT)

- Developed by Warren Ewens and associates in the 1990's
- Uses a trio: affected child and two parents
- For each allele, count transmissions/chances to transmit
- Compare to an expectation of 50%

Sample pedigree for TDT



TDT example

In a large study (Lie et al. 2000) of children with IDDM:

Haplotype	T	NT	%T	χ^2
0301-0302	342	65	84	189
0501-0201	247	78	76	88
0301-0303	15	13	54	0.1
0102-0502	3	3	50	0
0401-0402	33	34	49	0
0101-0501	47	115	29	29
0301-0301	16	48	25	16
0201-0201	15	56	21	23
0103-0603	12	74	13	32
0501-0301	7	47	13	30
0101-0503	0	12	0	12
0102-0602	0	164	0	164

Pros and cons of TDT

- Pros:
 - No issues with ethnicity of controls vs. cases
 - Sometimes detects association when family linkage studies fail
 - Nuclear family sufficient
- Cons:
 - Requires family studies
 - Requires LD
 - Large number of families needed
 - Information from complex pedigrees not fully used

HapMap

- Recombination hotspots tend to create “blocks” of LD
- Idea: within a block, you just need a few SNPs to know what haplotype is present
- HapMap project set out to find sets of SNPs sufficient to identify haplotypes
- In theory, this should give you full mapping accuracy without typing millions of SNPs
- Just type HapMap’s selected SNPs and infer the haplotype

HapMap Project

- Four populations:
 - CEPH (people of northern and western European ancestry living in Utah)
 - Yoruba from Ibadan, Nigeria
 - Han Chinese from Beijing
 - Japanese from Tokyo
- This is not exactly a worldwide exhaustive sample!

Idea behind HapMap SNPs

- Only common (frequency of at least 5%) SNPs
- Minimum SNPs necessary to distinguish common haplotypes
- Meant for mapping diseases caused by common variants

Arguments for HapMap strategy

- Most remaining mapping problems involve common, multifactorial, complex diseases
- Many researchers believe these are common-variant diseases
- Typing fewer SNPs is definitely cheaper

Arguments against HapMap strategy

- Common diseases might be caused by multiple rare variants, not common ones
- Haplotypes in HapMap may not be comprehensive—especially in Africa
- Full genome sequencing may eventually be the best strategy
- The actual disease-causing SNP is always the best marker