CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 16

Genetics
Reading assignment

- Chapter 13
Gene association studies

- Goal: identify genes/markers associated with disease
- Example: BRCA1 – associated with risk of breast cancer
- Lots of hype on the news recently: companies promise to “sequence” your genome and tell you:
  - likely ancestry
  - risk for disease
- Examples:
  - www.23andme.com
  - www.decodeme.com
  - and many others
First...definitions

- **Genotype** – genetic composition of our genome
- **Phenotype** – observable consequence of genotype – e.g. skin/hair color, IQ, disease state, etc.
- We have two copies of each chromosome (homologous chromosomes), each received from one of the parents
- Each gene can, thus, have two forms (alleles), e.g. A1/A2
- Each gene may be associated with a phenotype
- Dominant gene – phenotype of A1/A2 is the same as phenotype of A1/A1
- Recessive gene - otherwise
More definitions

- Genotype A/A is called homozygous (both chromosomes have the same allele)
- Genotype A/B is called heterozygous (mother and father's chromosomes disagree)

Notes:

- Phenotypes not necessarily directly correlated with a single gene – polygenic traits
- Probability gene correlates with a phenotype – penetrance
- Link between genotype and phenotype can be qualitative (gene “form” matters) or quantitative (gene dosage matters)
Technology – what we measure?

- Definition of allele/genotype depends on what we can measure – constantly changing

- We are looking for things that differ within a population – polymorphic markers:
  - Restriction fragment length polymorphism (RFLP) – measures presence/absence of particular sites in the genome
  - Variable number tandem repeats (VNTR) – specific repeat elements that occur in different copy numbers
  - Single-nucleotide polymorphisms (SNPs) – single letter differences between chromosomes (>500,000 characterized)
  - Copy number variants (CNV) – genomic regions whose copy number differs between individuals
Allele frequencies

- Population genetics questions:
  - what alleles exist in a certain population?
  - what is the relative abundance of the alleles?
  - how “diverse” is a population?

- Given a locus (gene or genomic region), assume there are $K$ possible alleles in a population and allele $j$ occurs with frequency $p_j$

- How “uniform” is the locus in the population? Likelihood two random individuals have same allele

  \[
  \text{homzygosity } F = \sum_{i=1}^{K} p_i^2
  \]
Allele frequencies...

- Usually we focus on the differences:
  \[ H = 1 - F = 1 - \sum_{i=1}^{K} p_i^2 \]

- Interesting tidbit – most variation occurs within populations rather than between, e.g. two Africans are more different from each other than the average African is from the average Chinese (see book for details)

- However, allele frequencies can be used to infer population membership for an individual
Who am I?

- My alleles are $A_1$, $B_2$, $C_1$, $D_3$ (assume homozygous for clarity)

- Am I European or Asian?

- Need to know:
  \[
  p_{A_1}^{\text{Europe}}, p_{B_2}^{\text{Europe}}, p_{C_1}^{\text{Europe}}, p_{D_3}^{\text{Europe}}
  \]
  \[
  p_{A_1}^{\text{Asia}}, p_{B_2}^{\text{Asia}}, p_{C_1}^{\text{Asia}}, p_{D_3}^{\text{Asia}}
  \]

- \[p(\text{me, European}) = (p_{A_1}^{\text{Europe}})^2 \times (p_{B_2}^{\text{Europe}})^2 \times (p_{C_1}^{\text{Europe}})^2 \times (p_{D_3}^{\text{Europe}})^2\]

- Similarly for $p(\text{me, Asian})$

- If $p(\text{me, European}) > p(\text{me, Asian})$ I can infer that I have European ancestry
Who am I?

- Inferring ancestry as described is overly-simplistic
- Can do more fancy statistics
- However: any statistical approach is error prone – answer is associated with level of confidence, i.e. probability answer is wrong (remember P-values?)
- Beware of anyone who claims to infer your ancestry from genotype
- Beware of anyone who claims to infer disease susceptibility from genotype - need genetic/risk counselors not companies providing information for “entertainment purposes”
Recombination

- Genetic change not only caused by mutations
- Recombination – DNA “jumps” between homologous chromosomes due to cross-over events
Association studies

- The set of alleles on a same chromosome – haplotype

- If a particular allele of a gene is always associated with a phenotype (disease) – is this gene causing the disease?

- Most likely – gene is associated/nearby with the gene causing the disease (their alleles always appear on the same haplotype)

- Due to recombination a set of original haplotypes rapidly becomes broken apart

- How likely is it that two alleles remain on the same haplotype (are linked) during evolution?
Linkage analysis

- Preservation of linkage depends on distance between the genes and rate of recombination
- Given two genes (A, B) – how can we estimate whether recombination occurred between them?
- How likely is it that $A_1$ and $B_1$ are both on the same haplotype by chance?
  \[ p(A_1)p(B_1) \]
- How different is this from the observed ratios? - Linkage Disequilibrium
  \[
  D = p(A_1B_1) - p(A_1)p(B_1) \\
  D = p(A_2B_2) - p(A_2)p(B_2) \\
  D = p(A_1B_1)p(A_2B_2) - p(A_1B_2)p(A_2B_1)
  \]
Linkage analysis

- Linkage disequilibrium usually measured as ratio to maximum possible disequilibrium - $D/D_{\text{max}}$

\[
D_{\text{max}} = \min(p(A_2)p(B_1), p(A_1)p(B_2)) \text{ if } D > 0
\]

\[
D_{\text{max}} = \min(p(A_1)p(B_1), p(A_2)p(B_2)) \text{ if } D < 0
\]

- Another measure – Pearson's correlation coefficient

\[
r^2 = \frac{D^2}{p(A_1)p(A_2)p(B_1)p(B_2)}
\]
Additional resources

- www.hapmap.org
- www.1000genomes.org
- www.personalgenomes.org
Homework

- Prove the equalities on slide 13 (D = ...)
- Derive the formula for Dmax on slide 14 (problem 3.5 in book)

Due Tuesday Nov 4 – submit by E-mail to me and Mohammad