Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

L#8:(Mar-30-2010)
Genome Wide Association Studies

Outline

1. Genetic Data
   - Linkage Disequilibrium (LD)

LD and HWE

- Two important concepts to explore
- (1) **LD**: Linkage Disequilibrium
- (2) **HWE**: Hardy-Weinberg Equilibrium
- *Concepts regarding the genetic component of the data — No connection to traits.*

LD and HWE

- Both LD and HWE are measures of allelic association....
- LD measures the associations among sites along the genome
- HWE measures the association at a single site between pair of homologous chromosmes.

Linkage Disequilibrium (LD)

- Association between two adjacent variant sites become lost over time as recombination events occur in the region separating them. Asymptotically the genomes will go to linkage equilibrium, making all the sites acting independently.
- If the sites are all independent then only the “causal variant” site will contribute to the “probability raising” and will not have any “screening off” due to some other confounding (correlated) sites.
- However, in general the variant sites are not yet in linkage equilibrium and there exist strong dependence among the sites.

- **The SNP sites that are usually analyzed in GWAS could be within genes, but may not be functional.** That is, these SNP sites may not directly cause the disease.
- Usually "tag SNPs" that are analyzed are selected to represent the haplotypes occurring within a haplotype block–they are non-functional but closely associated to functional/causal SNPs.
- These sites are likely to be *associated with* disease because they are in LD (**Linkage Disequilibrium**) with the *functional variant*.
- LD is measured in terms of two closely related measures: D' and r^2.
- These measures are very closely related to Pearson's χ^2-statistics.

LD: D'

- Consider the distribution of alleles for n individuals across two sites: Assume that the two sites are *independent of each other* – **in Linkage Equilibrium**.
- *The presence an allele at one site does not influence the particular allele observed at the second site.*
- Assume: At site 1 the alleles are A and a, with population frequencies p_A and p_a, respectively. At site 2 the alleles are B and b, with population frequencies p_B and p_b, respectively.

	Site 2 B	b	
Site A	$n_{11} = Np_Ap_B$	$n_{12} = Np_Ap_b$	$n_{1.} = Np_A$
1 a	$n_{21} = Np_ap_B$	$n_{22} = Np_ap_b$	$n_{2.} = Np_a$
	$n_{.1} = Np_B$	$n_{.2} = Np_b$	$N = 2n$

LD: D'

- If sites 1 and 2 are in fact associated with one another, then the observed counts will deviate from the numbers shown in the earlier table.
- Represent the deviation by a single scalar D.
- $H_0 : D = 0$ corresponds to the null hypothesis that the two sites are independent (in LE: Linkage Equilibrium).

	Site 2		
	B	b	
Site A	$n_{11} = N(p_A p_B + D)$	$n_{12} = N(p_A p_b - D)$	$n_{1.}$
1 a	$n_{21} = N(p_a p_B - D)$	$n_{22} = N(p_a p_b + D)$	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$N = 2n$

Estimating D'

- D can be expressed in terms of the joint probability of A and B and the product of the individual allele probabilities:

$$D = p_{AB} - p_A p_B.$$

- Note that we can estimate D as

$$\widehat{D} = \widehat{p_{AB}} - \widehat{p_A}\widehat{p_B} = \widehat{p_{AB}} - (n_{1\cdot}/N)(n_{\cdot 1}/N).$$

- $\widehat{p_{AB}}$ has to be estimated by an MLE estimator

Estimating D'

- Let $\theta = (p_{AB}, p_{Ab}, p_{aB}, p_{ab})$ be estimated from the genotype counts from two biallelic loci

$$\begin{aligned}\log L(\theta | n_{11}, \ldots, n_{33}) \\ \propto \quad & (2n_{11} + n_{12} + n_{21}) \log p_{AB} + (2n_{13} + n_{12} + n_{23}) \log p_{Ab} \\ & + (2n_{31} + n_{21} + n_{32}) \log p_{aB} + (2n_{33} + n_{32} + n_{23}) \log p_{ab} \\ & + n_{22} \log(p_{AB} p_{ab} + p_{Ab} p_{aB}).\end{aligned}$$

		Site 2		
		BB	*Bb*	*bb*
Site 1	*AA*	n_{11}	n_{12}	n_{13}
	Aa	n_{21}	n_{22}	n_{23}
	aa	n_{31}	n_{32}	n_{33}

- One can estimate D by first substituting $p_A p_B + D$ for p_{AB}, $p_A p_b - D$ for p_{Ab}, etc. and solve the maximization problem for $\widehat{D}$ using numerical optimization.
- Alternatively, write $p_{Ab} = p_A - p_{AB}$, $p_{aB} = p_B - p_{AB}$, and $p_{ab} = 1 - p_A - p_B - p_{AB}$, and estimate p_{AB}. Solve for $\widehat{D} = \widehat{p_{AB}} - \widehat{p_A}\widehat{p_B}$.
- A rescaled value of D, given by D' is used for a measure of LD:

$$D' = \frac{|D|}{D_{\max}},$$

where $D_{\max}$ bounds D from above:

$$D_{\max} = \begin{cases} \min(p_A p_b, p_a p_B), & \text{if } D > 0; \\ \min(p_A p_B, p_a p_b), & \text{otherwise.} \end{cases}$$

- Note that $0 \leq D' \leq 1$
- If D' is close to 1, then the two sites are assumed to be in "complete LD." The sites are in the same haplotype block.
- If D' is close to 0, then the two sites are assumed to be independent — with a recombination hot-spot separating them. The sites belong to two distinct adjacent haplotype blocks.

The quantity r^2

- The quantity r^2, measuring LD, is based on Pearson's χ^2-statistic for the test of no association.
- Consider an $r \times c$ contingency table corresponding to the counts of individuals with two bi-allelic sites: Site 1: A, a and site 2: B, b.

$$\chi^2_1 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where $i = 1, \ldots, r$; $j = 1, \ldots, c$; and O_{ij} and E_{ij} are respective observed and expected cell counts for the i,jth cell of an $r \times c$ table.

- r^2 is defined as

$$r^2 = \chi^2_1 / N.$$

Relation between D' and r^2

$$(O_{ij} - E_{ij})^2 = (ND)^2.$$

Thus

$$\begin{aligned}
\chi_1^2 &= \sum_{ij} \frac{(ND)^2}{E_{ij}} \\
&= (ND)^2 \left(\frac{1}{Np_Ap_B} + \frac{1}{Np_Ap_b} + \frac{1}{Np_ap_B} + \frac{1}{Np_ap_b} \right) \\
&= ND^2 \left(\frac{p_ap_b + p_ap_B + p_Ap_b + p_Ap_B}{p_Ap_Bp_ap_b} \right) \\
&= \frac{ND^2}{p_Ap_Bp_ap_b}
\end{aligned}$$

Relation between D' and r^2

- In summary,
$$r^2 = \chi_1^2/N = \frac{D^2}{p_A p_b p_a p_b}.$$
- Thus r^2 is simply D^2, further adjusted by the marginal probabilities.
- r^2 is usually preferred, because of its straightforward relationship with the χ^2 statistics and the null hypothesis H_0 that the two sites are independent.

Caveat

- With the currently available technology (e.g., genotype sequencing), haplotypes are not observed — so the cell counts in the contingency tables are inferred.
- The estimation process (MLE or EM), introduces further errors into the r^2 measures — making it highly unreliable.
- Additionally, Pearson's χ^2-test assumes independent observations — which may be violated in the absence of HWE (Hardy-Weinberg Equilibrium). Note that the contingency table includes two observations per person

Summary

- D' and r^2 are both *measures* of linkage disequilibrium between loci; they estimate the amount of association between two sites.
- Conclusions from these must be drawn with caution — as they depend on certain implicit assumptions that are often violated.
- The Key problem: **Haplotype Phasing Problem**

LD Blocks

- Determine whether a group of adjacent loci are in LD.
- A measure of LD across a region (comprising multiple SNPs) is the average of all pairwise measures of D'

$$\bar{D}' = \frac{1}{n_L} \sum_{i,j \in L} D'_{ij},$$

where

- L is a set of loci within a region of interest
- D'_{ij} is the measure of LD between loci i and j for $i, j \in L$
- n_L is the number of ways of choosing two loci from the set L (i.e., $\binom{|L|}{2}$)
- the summation is over all such pairs of loci

LD Blocks

- Through characterization of regions of high average LDs, a genome (i.e., human's) can be partitioned in to *LD Blocks*.
- These blocks are separated by (recombination) *hotspots* – regions in which recombination events might have occurred with very high frequencies (and likely to happen in the future).
- In general, *alleles tend to be more correlated within an LD block* than across...

SNP tagging

- Once regions of high LD are identified, we will aim to determine the smallest subset of SNPs that characterizes the variability in the region — this process is called *SNP tagging* and the selected SNPs are called **Tag SNPs**.
- Example: Consider two SNPs (i and j) that are in perfect LD so that $D'_{i,j} = 1$. Genotyping both SNPs are unnecessary as their relationship is *deterministic* — knowledge of the genotype of one SNP completely defines the genotype of the second and there is no need to sequence both loci.
- Few (say 3 - 5) well-defined tag SNPs capture a substantial majority of the genetic variability within an LD block.

TAG SNPs

- Note: in general, tag SNPs are correlated with the true disease causing variant – but are not typically functional themselves.
- LD blocks differ substantially across race and ethnicity groups: It's shorter in Black/non-Hispanics than White and Hispanics.
- African population has much more genetic variability. It is older with many more recombination events than the European population.
- *A tag SNP may capture information on the true disease-causing variant in one racial group, but not another.*
- Thus in any GWAS, understanding population substructures and its effect on measures of LD is CRUCIAL!!!!

LD and Population Stratification

- **Population Stratification**: Presence of multiple subgroups (of sub-populations) among which there is minimal mating and gene-flow.
- *Ignoring population stratification in a sample could lead to confounding conclusions.*
- Population ad mixtures pose additional problems.
- Simpson's paradox – Yule-Simpson Effect. This paradox occurs in the presence of a confounding variable that is not properly accounted for in the analysis.

Hardy-Weinberg Equilibrium (HWE)

- HWE denotes independence of alleles at a single site between two homologous chromosomes.
- For instance, consider the simple case of biallelic SNP with genotypes AA, Aa and aa.
- HWE implies that the probability of an allele occurring on one homolog does not affect which allele will be present on the second homolog:

$$p_{AA} = p_A^2, p_{Aa} = p_{aA} = p_A p_a, \text{ and } p_{aa} = p_a^2,$$

where

$$p_A + p_a = 1.$$

Violation of HWE

- Tests of HWE include Pearson's χ^2-test and Fisher's exact test.
- When more than 20% of the expected counts are less than five, Fisher's exact test is recommended. The χ^2-test is computationally efficient but relies on asymptotic theories.
- The test s are based on the 2×2 table of genotypes at a single locus, as shown below:

	Homolog 2 A	a	
Homolog A	n_{11}	n_{12}	$n_{1.}$
1 a	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

χ^2-test

- Note: n_{11} and n_{22} are the counts for major and minor homozygous individuals: *AA* and *aa*, respectively.
- The two heterozygous genotypes are indistinguishable: One can only observe $n^*_{12} = n_{12} + n_{21}$.
- The expected values, corresponding to observations $O_{11} = n_{11}$, $O_{12} = n^*_{12}$ and $O_{22} = n_{22}$ are

$$E_{11} = np_A^2, E_{12} = 2np_A(1-p_A), \text{ and } E_{22} = n(1-p_A)^2.$$

- The χ^2-statistic:

$$\chi^2 = \sum_{i=1,2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_1.$$

- An estimate for p_A is

$$\widehat{p_A} = (2n_{11} + n^*_{12})/(2n).$$

HWE

- A statisticaly significant test of HWE suggests that *the SNP under investigation is in* **Hardy-Weinberg Disequilibrium** *(HWD).*
- HWD is usually assumed to be resulting from self-seleting mates: *non-random mating*.
- Deviation from HWE may also indicate non-neutral evolution: *positive or negative selection*
- Question: What is the relationship between HWE and population substructure?

HWE and Population Substructure

- Note:
  1. HWE implies constant allele frequencies over generations.
  2. HWE is violated in the presence of population admixtures.
  3. HWE is violated in the presence of population stratification.
- These observations and the corresponding statistical tests allow one to understand the population substructures and use them to correct the causal analysis of GWAS.

Allele Frequencies over Generations

- The genotype of a parent (at a single biallelic locus):

$$pr(AA) = p_A^2, pr(Aa) = 2p_A q_A, \text{ and } pr(aa) = q_A^2,$$

  where $q_A = p_a = (1 - p_A)$.
- The inheritance pattern. The conditional probability that an offspring inherits allele y, given that the parent has genotype X is $pr(y|X)$.

$$pr(A|AA) = 1, pr(A|Aa) = 1/2, \text{ and } pr(A|aa) = 0.$$

- Thus the population frequency of the allele A in the next generation is given by

$$\begin{aligned} pr(A) &= pr(A|AA)pr(AA) + pr(A|Aa)pr(Aa) + pr(A|aa)pr(aa) \\ &= p_A^2 + p_A q_A + 0 = p_A. \end{aligned}$$

Population Admixtures

- Population Admixtures occur as a result of matings between two populations for which alele frequencies differ.
- Assume that the two populations have two different frequencies for the allele A: p_{1A} and p_{2A}.
- Then the offsprings resulting from random matings of the two populations (assuming infinite populations sizes) will have frequencies:

$$\begin{aligned} pr(AA) &= p_{1A}p_{2A}, pr(Aa) = p_{1A}q_{2A} + p_{2A}q_{1A}, \text{ and} \\ pr(aa) &= q_{1A}q_{2A}. \end{aligned}$$

Note: $q_{iA} = 1 - p_{iA}$, $i = 1, 2$.

Population Stratification

- Population stratification is the combination of populations in which breeding occurs within but not between sub-populations.
- Within each sub-populations, we may have HWE (since the observed counts are as expected under random mating).
- Assume population 1 has allele frequency: $pr(A) = p_{1A}$ and population 2: $pr(A) = p_{2A}$. Assume that the two populations are of equal size, but $p_{1A} \ll p_{2A}$. The combined frequency is $p_A = (p_{1A} + p_{2A})/2$, but

$$\begin{aligned} pr(AA) &= (p_{1A}^2 + p_{2A}^2)/2 \approx p_{2A}^2/2, \text{ but} \\ p_A^2 &= (p_{1A} + p_{2A})^2/4 \approx p_{2A}^2/4. \end{aligned}$$

Hardy's Law

- Mendelian genetics: it was not then known how it could cause continuous characteristics. Udny Yule (1902) argued against Mendelism because he thought that dominant alleles would increase in the population.
- The American William E. Castle (1903) showed that without selection, the genotype frequencies would remain stable. Karl Pearson (1903) found one equilibrium position with values of $p = q = 1/2$.
- Reginald Punnett introduced the problem to G. H. Hardy, a British mathematician... who found biologists' use of mathematics as "very simple."
- The principle was known as Hardy's law in the English-speaking world until 1943, when Curt Stern pointed out that it had first been formulated independently in 1908 by the German physician Wilhelm Weinberg.

Hardy's Letter

- "To the Editor of Science:
- "I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making...
- "Suppose that Aa is a pair of Mendelian characters, A being dominant, and that in any given generation the number of pure dominants (AA), heterozygotes (Aa), and pure recessives (aa) are as $p : 2q : r$.

Hardy's Letter

- "Finally, suppose that the numbers are fairly large, so that mating may be regarded as random, that the sexes are evenly distributed among the three varieties, and that all are equally fertile. A little mathematics of the multiplication-table type is enough to show that in the next generation the numbers will be as $(p+q)^2 : 2(p+q)(q+r) : (q+r)^2$, or as $p_1 : 2q_1 : r_1$, say.
- "The interesting question is — in what circumstances will this distribution be the same as that in the generation before? It is easy to see that the condition for this is $q^2 = pr$. And since $q_1^2 = p_1 r_1$, whatever the values of p, q, and r may be, the distribution will in any case continue unchanged after the second generation."

Heterozygote advantage

- HWE may be violated under selection pressure: E.g., *heterozygote advantage*, or *heterotic balancing selection*. ... An individual who is heterozygous at a particular gene locus has a greater fitness than a homozygous individual.
- Example: Sickle cell anemia... a hereditary disease that damages red blood cells.
- Sickle cell anemia is caused by the inheritance of a variant hemoglobin gene (HgbS) from both parents. In these individuals hemoglobin (protein in red blood cells that carries oxygen to the tissues) is extremely sensitive to oxygen deprivation causing short life expectancy.

Heterozygote advantage

- However, a person who inherits the sickle cell gene from one parent and a normal hemoglobin gene (HgbA) from the other parent (a carrier of the sickle cell trait) has a normal life expectancy. The heterozygote is resistant to the malarial parasite which kills a large number of people each year.

Heterozygote advantage

- HgbS, which in the homozygous state causes sickle-cell anemia, is distributed throughout sub-Saharan Africa, the Mediterranean, the Middle East, and parts of India; the frequency of the carrier state ranges from 5 to over 40 percent.
- HgbE, the most common structural hemoglobin in the world population, is confined to the eastern regions of the Indian subcontinent, Myanmar, and Southeast Asia. Its frequency varies; carrier rates of over 60 percent of the population occur in eastern Thailand and parts of Cambodia.

[End of Lecture #8]