

SNPs, Human Variation and Linkage Disequilibrium

Bio 5488, Spring 2009
Monday 3/23/09, N.L. Saccone

Human DNA sequence variation

- SNPs are single nucleotide polymorphisms

Chr 1: A A C C A T A T C ... C G A T T ...

Chr 2: A A C C A T A T C ... C A A T T ...

Chr 3: A A C C G T A T C ... C G A T T ...

- Provide biallelic markers
- Coding SNPs may directly affect protein products of genes
- Non-coding SNPs still may affect gene regulation or expression

Some questions that SNPs can help us answer:

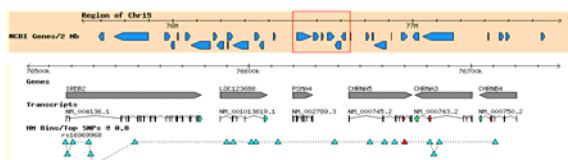
- Which genetic loci influence risk for common human diseases? (Disease gene mapping studies)
- Which genetic loci influence efficacy/safety of drug therapies? (Pharmacogenetics)
- Population genetics questions
 - evidence of selection
 - identification of recombination hotspots
- SNPs are the basis of current GWA studies

Genome Wide Association Studies (GWASes)

- Have gone very large-scale in a short time
 - Macular Degeneration (Klein et al., Science 2005) :116K SNPs in 96 cases, 50 controls
 - Type II Diabetes (3 papers in Science 2007)
 - E.g. Saxena et al.: 386,000 SNPs in 1464 cases, 1467 controls
 - Now: commercial chips covering 1M SNPs (and also copy-number variation)
- Still, some of the challenges have remained....

Genome Wide Association Studies (GWASes)

- Example from NICSNP (Bierut et al.; S. Saccone et al., Hum Mol Genet 2007): association between nicotine dependence and a non-synonymous coding SNP rs16969968 in CHRNA5
- Note: LD means several other potential "causative" variants are correlated with this SNPs



LD matters both in the design and in the interpretation of a GWAS.

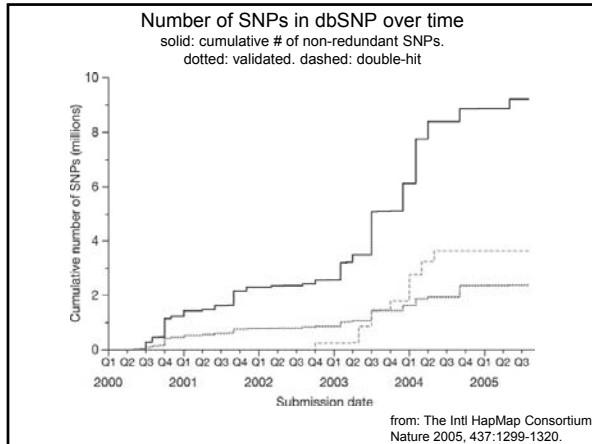
Human DNA sequence variation

Advantages of SNPs as markers:

- Common in the human genome
- More stable than microsatellites
- Genotyping methods are relatively "error-free"
- Genotyping is becoming cheaper and cheaper

Limitations:

- Biallelic (vs. very polymorphic microsatellite markers)
- Not all SNPs are "known" – direct sequencing also needed
- "Known" (or "characterized") SNPs tend to be "common" - some important variation will be "rare"



Mapping disease genes

- Some quick background
 - Linkage**
 - quantify co-segregation of trait and genotype in families

LOD score traditionally used to measure statistical evidence for linkage

- Association**
 - Common design: case-control sample, analyzed for allele frequency differences

Association in a case-control sample

Let $N = N_{\text{case}} + N_{\text{control}}$ (2N observations of alleles)

Most basic test for biallelic markers: compare allele frequencies in cases vs controls in a 2x2 table.

	case	ctrl	
A ₁	N ₁₁	N ₁₂	Chi-square with n-1 df (n = # of alleles)
A ₂	N ₂₁	N ₂₂	
			2N

$$\chi^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$

Association in a case-control sample

Alternatives: logistic regression

Let P = probability of being a case.

$$\text{Log}(P/(1-P)) = a_0 + (a_1x_1 + \dots + a_nx_n) + b_1G$$

x_i are covariates (e.g. gender, age)
 G represents genotype (0, 1 or 2 copies of a specified allele) (corresponds to a log-additive, that is, multiplicative model).

Statistical test: determine the improvement in fit when the genotype term is added. (Likelihood ratio chi-square).

Human DNA sequence variation

Note: the above test should work great if the marker you genotyped is actually the disease locus.

What if the marker is just "nearby" or "correlated" with the disease locus?

This is where the concept of "linkage disequilibrium" (LD) comes in.

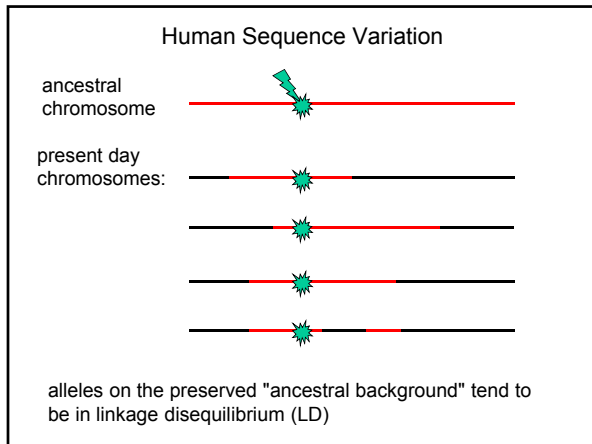
The International Hap Map Project

- goal: determine the common patterns of DNA sequence variation (LD among SNPs)
- one motivation: to help determine redundancy among SNPs for more efficient disease mapping and pharmacogenetics studies

Human DNA sequence variation

How to measure/describe "patterns" of DNA sequence variation?

How to use these patterns to find disease genes that affect phenotypes?



Linkage Disequilibrium (LD) involves haplotype frequencies.

Focus on pair-wise LD, SNP markers

Genotypes do not necessarily determine haplotypes:
Consider 2-locus genotype $A_1 A_2 B_1 B_2$.

Two possible phases :

Linkage Disequilibrium (LD) involves haplotype frequencies

Focus on pair-wise LD, SNP markers

Genotypes do not necessarily determine haplotypes:
Consider 2-locus genotype $A_1 A_2 B_1 B_2$.

Two possible phases :

A_1		A_2		A_1		A_2
B_1		B_2		B_2		B_1

Linkage Disequilibrium

Linkage Disequilibrium (LD), aka allelic association:

For two loci A and B:
LD is said to exist when alleles at A and B tend to co-occur on haplotypes in proportions different than would be expected under statistical independence.

Linkage Disequilibrium

Example: Consider 2 SNPs:

SNP 1: A 50% C 50%

SNP 2: A 50% G 50%

	snp1	snp2	expected freq
4 possible haplotypes:	A	A	$0.5 * 0.5$
	A	G	$0.5 * 0.5$
	C	A	$0.5 * 0.5$
	C	G	$0.5 * 0.5$

But perhaps in your sample you observe only the following:
 $A A C C A T A T C \dots C G A T T \dots$
 and
 $A A C C C T A T C \dots C A A T T \dots$

Linkage Disequilibrium

- How to formally measure LD between alleles at 2 loci?

To measure LD between alleles at 2 biallelic loci

Locus A Locus B
A₁, A₂ B₁, B₂

Given 2N haplotypes:
Haplotype freq for A_iB_j is

$$h_{ij} = \frac{n_{ij}}{2N}$$

Compare h_{ij} to the frequency expected under no association:

$$p_{A_1}p_{B_1} = \left(\frac{n_{11} + n_{12}}{2N}\right) \left(\frac{n_{11} + n_{21}}{2N}\right)$$

Define the disequilibrium coefficient: $D = h_{11} - p_{A_1}p_{B_1}$

	B ₁	B ₂	
A ₁	n ₁₁	n ₁₂	
A ₂	n ₂₁	n ₂₂	
			2N

Notes:

1. $D = h_{11} - p_{A_1}p_{B_1} = h_{22} - p_{A_2}p_{B_2}$
2. Choice of allele labeling may affect sign but not absolute value of D.

3. Range of D:
[max(-p_{A1}p_{B1}, -p_{A2}p_{B2}), min(p_{A1}p_{B2}, p_{A2}p_{B1})]

To see this:

Case 1: D < 0.

Then the 'most negative' D can be is the max of 0 - p_{A1}p_{B1} or 0 - p_{A2}p_{B2}.

Case 2: D > 0.

Relabel, and then use Case 1 argument.

	B ₁	B ₂	
A ₁	n ₁₁	n ₁₂	
A ₂	n ₂₁	n ₂₂	
			2N

Common LD measures

Disequilibrium coefficient:

$$D = h_{11} - p_{A_1}p_{B_1}$$

Normalized disequilibrium coefficient:

$$D' = D / |D|_{\max}, \text{ where}$$

$$|D|_{\max} = \begin{cases} \min(p_{A_1}p_{B_2}, p_{A_2}p_{B_1}) & \text{if } D > 0 \\ \min(p_{A_1}p_{B_1}, p_{A_2}p_{B_2}) & \text{if } D < 0 \end{cases}$$

Range of D' is [-1, 1]

Correlation coefficient:

$$r^2 = D^2 / (p_{A_1}p_{A_2}p_{B_1}p_{B_2})$$

Measuring LD

To measure significance: χ^2 (1 df): $\sum \frac{(obs - exp)^2}{exp}$

Example:

	B ₁	B ₂	
A ₁	50	0	50
A ₂	0	50	50
	50	50	100

$$\chi^2 = \frac{(50 - 25)^2}{25} + \frac{(0 - 25)^2}{25} + \frac{(0 - 25)^2}{25} + \frac{(50 - 25)^2}{25}$$

Measuring LD: D', r², chi-square significance

D' can equal 1 even though the p-value is not significant

Example 1:

	B ₁	B ₂	
A ₁	1	0	1
A ₂	49	50	99
	50	50	100

$$D' = (.01 - (.01)(.5)) / (.005) = 1$$

$$\chi^2 = 1.01, \text{ p-value} = 0.31.$$

$$r^2 = 0.01$$

Example 2: same haplotype freqs as before, but larger sample

	B ₁	B ₂	
A ₁	100	0	100
A ₂	4900	5000	9900
	5000	5000	10000

$$D' = 1$$

$$r^2 = \text{same as above}$$

$$\chi^2 = 50 + 50 + 0.5050 + 0.5050 = 101.01$$

$$\text{p-value is significant}$$

When calculating significance, sample size matters.

Measuring LD: D', r², chi-square significance

D' can equal 1 even though the p-value is not significant

Example 1:

(from last slide)

	B ₁	B ₂	
A ₁	1	0	1
A ₂	49	50	99
	50	50	100

$$D' = (.01 - (.01)(.5)) / (.005) = 1$$

$$\chi^2 = 1.01, \text{ p-value} = 0.31.$$

$$r^2 \text{ is small (0.01)}$$

Example 3: NOT the same haplotype freqs as before

	B ₁	B ₂	
A ₁	1	0	1
A ₂	4999	5000	9999
	5000	5000	10000

$$D' = 1$$

$$r^2 = 0.0001$$

$$\chi^2 = 1.0001$$

$$\text{p-value} = 0.32$$

Note r² is small in both cases.

LD measures

One more important example:

	B ₁	B ₂	
A ₁	10	0	10
	0	90	90
A ₂	10	90	100

$$D' = (.1 - (.1)(.1)) / (.09) = 1$$

$$\chi^2 = 100, \text{ p-value} \sim 0.0$$

$$r^2 = 1$$

	B ₁	B ₂	
A ₁	0	10	10
	10	80	90
A ₂	10	90	100

$$D' = (0 - (.1)(.1)) / (.01) = -1$$

$$\chi^2 = 1.23, \text{ p-value} = 0.27$$

$$r^2 = 0.012$$

LD measures

|D'| is 1 when the alleles of the two markers are as correlated as they can be, given the allele frequencies of the co-occurring alleles.

The range of r² depends on the marker allele frequencies.

r² equals 1 if and only if 1) the MAFs at the two loci match and 2) the minor alleles always co-occur

D' : useful for identifying regions of reduced recombination.
r² : useful for identifying markers that are good predictors of allelic status at other markers.

LD measures

r² (between a marker and disease locus) is closely related to the statistical power to detect disease association:

Suppose, in a sample of size N, the power to detect disease association, when the disease locus is genotyped, is X%. Then to achieve the same power if a marker is genotyped, the sample size must be increased to N/r².

(Pritchard and Przeworski, 2001)

Power = probability of correctly rejecting H₀.

Implications of this N/r² formula

1. Can *quantify* the degree to which genotyping only 'common' SNPs limits success: will be able to detect only disease loci satisfying the CDCV hypothesis

Ex/ Suppose all genotyped SNPs satisfy MAF ≥ 0.1

What is the maximum r² that a SNP having MAF = 0.1 can have with SNPs having these MAFs:

MAF	max possible r ²	required sample size
0.1	1	N (e.g. 1000)
0.05		
0.02		

Implications of this N/r² formula

MAFs	max possible r ²
0.1, 0.05	0.4737

	B ₁	B ₂	
A ₁	0.05(2N)	0.05(2N)	0.1(2N)
	0	0.9(2N)	0.9(2N)
A ₂	0.05(2N)	0.95(2N)	2N

$$D = 0.05 - 0.1 \cdot 0.05 = 0.045$$

$$r^2 = D^2 / (0.1 \cdot 0.9 \cdot 0.05 \cdot 0.95)$$

$$= 0.002025 / 0.004275 = 0.47368$$

Implications of this N/r² formula

1. Can *quantify* the degree to which genotyping only 'common' SNPs limits success to disease loci satisfying the CDCV hypothesis

Ex/ Suppose all genotyped SNPs satisfy MAF ≥ 0.1

What is the maximum r² that a SNP having MAF = 0.1 can have with SNPs having these MAFs:

MAF	max possible r ²	required sample size
0.1	1	N (e.g. 1000)
0.05	0.4737	2.11*N (e.g. 2111)
0.02	0.1837	5.44*N (e.g. 5444)

Thus knowledge of LD patterns is important for disease gene mapping.

Note: tight linkage between two loci will tend to maintain linkage disequilibrium.

Decay of linkage disequilibrium

After k generations, disequilibrium decays according to

$$D_k = (1 - \theta)^k D_0$$

where θ = the recombination fraction (assuming random mating).

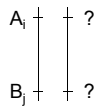
$$h_{11(1)} = (1-\theta) h_{11(0)} + \theta p_{A1}p_{B1}$$

so at generation 1,

$$D = h_{11} - p_{A1}p_{B1} = (1-\theta) (h_{11(0)} - p_{A1}p_{B1})$$

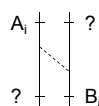
claim: $h_{ij(1)} = (1-\theta) h_{ij(0)} + \theta p_{Ai}p_{Bj}$,

nonrecombinant



$$(1-\theta) h_{ij(0)}$$

recombinant



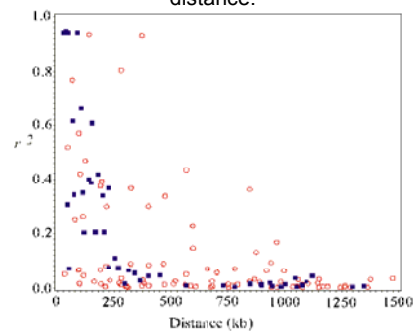
$$\theta p_{Ai}p_{Bj}$$

so at generation 1,

$$D = h_{11} - p_{A1}p_{B1} = (1-\theta) (h_{11(0)} - p_{1q1}) = (1-\theta) D_0$$

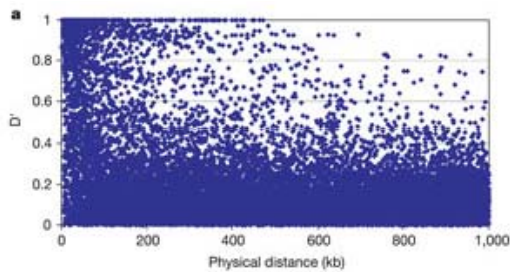
after k generations, get: $D_k = (1 - \theta)^k D_0$

LD is not a simple monotonic function of physical distance:



From Taillon-Miller et al., 2000 (O=Xq25, ■=Xq28)

LD is not necessarily a monotonic function of distance
Dawson et al., Nature 2002



Linkage Disequilibrium

• Potential sources of LD :

1. Linkage between loci
2. Random drift
3. Founder effect
4. Mutation
5. Selection
6. Population admixture / stratification

Linkage Disequilibrium

- Potential sources of LD :
 1. Linkage between loci
 2. Random drift
 3. **Founder effect**
 4. Mutation
 5. Selection
 6. Population admixture / stratification

Linkage Disequilibrium

Suppose have loci A, B, C in that order.
 Due to founder effect, suppose sample only 4 haplotypes out of the 8 possible:

A B C
 1 1 1
 1 2 1
 2 1 2
 2 2 2

Note:
 A and B are in equilibrium
 B and C are in equilibrium
 A and C are in complete disequilibrium

Disequilibrium not necessarily related to distance!

Linkage Disequilibrium

- Potential sources of LD :
 1. Linkage between loci
 2. Random drift
 3. Founder effect
 4. Mutation
 5. Selection
 6. **Population admixture / stratification**

An example of spurious association due to admixture/stratification:

population 1			population 2		
9	1	10	25	25	50
81	9	90	25	25	50
90	10	100	50	50	100

chi-square = 0

chi-square = 0

combined

34	26	60
106	34	140
140	60	200

chi-square = 7.26
 p-value = 0.007

SUMMARY

State-of-the-art design for complex disease gene mapping: SNP-based GWAS.

SNP-based GWASes rely in great part on marker-marker LD patterns.

Power to detect a association between disease and locus depends on LD (N/r^2 formula)

LD arises from a combination of many different factors.

LD is not a simple monotonic function of physical distance between markers.