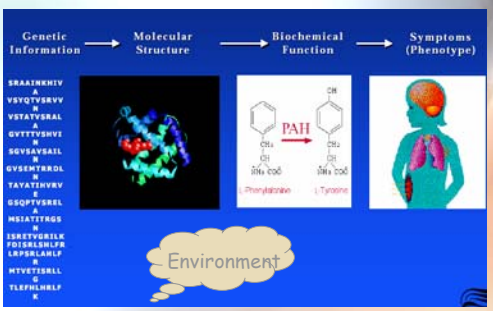
Basics in Genetics Analysis

Heping Zhang

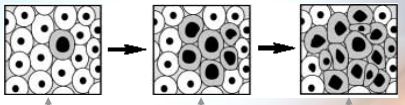
Genetics and Diseases



Dr. Doug Brutlag Lecture Scribing "central paradigm" //www.s-star.org/ 2

Diseases Progression

How does the Breast Cancer grows and spread?



A malignant tumor within the breast

Malignant tumors suppress normal ones

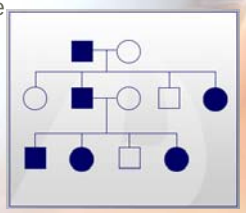
Cancer cells spread

Known and Probable Risk Factors

- Being a woman
- Getting older
- Having a personal history of BC or ovarian cancer
- Having a family history of breast cancer
- Having a previous biopsy showing carcinoma in situ
- Having your first period before age 12
- Starting menopause after age 55
- Never having children
- Having your first child after age 30
- Having a mutation in the BRCA1 or BRCA2 genes
- Drinking more than 1 alcoholic drink per day
- Being overweight after menopause or gaining weight as an adult.

Genetic Epidemiology

- How is a disease transmitted in families? (Inheritance patterns).
- What is the recurrence risk for relatives?
- Mendelian disorders
  - Autosomal or X-linked
  - Dominant or recessive



Yale University Center of Human Genetics 5

Terminology

- Marker: a known DNA sequence that can be identified by a simple assay

  e.g., D1S80, D4S43, D16S126
- Allele: a viable DNA coding that occupies a given locus (position) on a chromosome

  e.g., A, a;
- Genotype: the observed alleles at a genetic locus for an individual

  e.g., AA, Aa, aa;
  - Homozygous: AA, aa
  - Heterozygous: Aa
- Phenotype: the expression of a particular genotype
  - Continuous: blood pressure
  - Dichotomous: Cancer, Hypertension

Mendel's Laws

- **First Law**
  Segregation of Characteristics: the sex cell of a plant or animal may contain one factor (allele) for different traits but not both factors needed to express the traits.
- **Second Law**
  Independent Assortment: For two characteristics, the genes are inherited independently.
- **Third Law**
  Dominants and Recessives: each inherited characteristic is determined by two heredity factors/genes, one from each parent which determine whether a gene will be dominant or recessive.

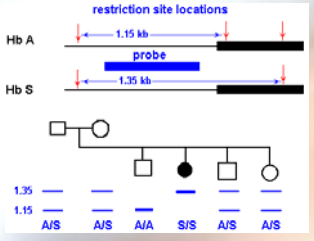
DNA Polymorphism and Human Variation

DNA Polymorphism

- **Restriction Fragment Length Polymorphism**
- **Variable Number of Tandem Repeats**
  - **Minisatellites**
  - **Microsatellites**
- **Single Nucleotide Polymorphism**
  - **Single-base substitutions**
  - **Single-base insertion/deletions**

RFLP

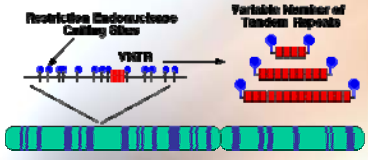
- **A technique discovered in 1975 in which organisms may be differentiated by analysis of patterns derived from cleavage of their DNA**
- **Only two alleles: present or not present**



http://www.people.virginia.edu/~rjh9u/hbsrflp.html

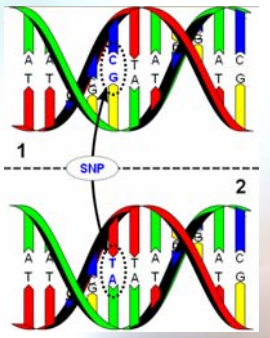
VNTR

- Successively repeating blocks of oligonucleotide of variable lengths (Nakamura, et al. 1987)
- Two families of VNTR
  - Minisatellite
    - Sequences of 11-16 bp repeated 1000 times.
    - Highly repetitive and dispersed into the genome
  - Microsatellite (Short Tandem Repeat)
    - Short sequences of 100-200 bp given by the repetition of 1-6 bp sequences



http://www.people.virginia.edu/~rjh9u/vntr1.html

SNP



A DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual)

http://en.wikipedia.org/wiki/Single_nucleotide_polymorphism

Characteristics of SNP

- The most common genetic polymorphism
- Distributed throughout genome with high density
- More stable and easy to assay
- Major cause of genetic diversity among different (normal) individuals
- Facilitates large scale genetic association studies as genetic markers.
- Most of SNPs neither change protein synthesis nor cause disease directly
  - Serve as landmarks: may be physically close to the mutation site on the chromosome
  - Shared among groups of people with common characteristics

Approaches: Linkage and Association

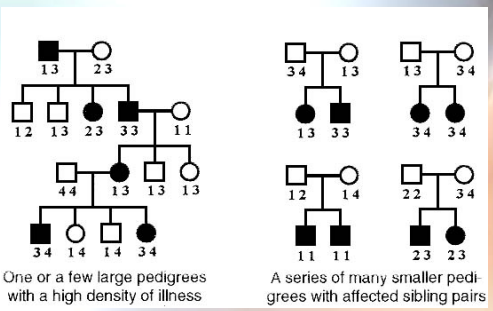
- Linkage studies use individual families where members are affected and attempt to demonstrate linkage between the occurrence of the disease and genetic markers (creates associations within families, but not among unrelated people)
- Association studies are based on populations and attempt to show an association between a particular allele and susceptibility to disease (a statistical statement about the co-occurrence of alleles and phenotypes)

Linkage Studies

- Goal

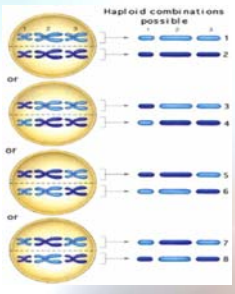
  To obtain a crude chromosomal location of the gene or genes associated with a phenotype of interest, e.g. a genetic disease or an important quantitative traits.
- Co-inheritance of Disease and Marker Genes in Families
  - Mendel's 2nd Law states that disease genes and genetic markers are inherited independently.
  - However, for a marker in close proximity to a disease locus, their genes may go together in family pedigrees (creating genetic linkage), only occasionally interrupted by "crossing-over.

Pedigree



Recombination

- Random segregation of chromatids



2 x 23 human chromosomes

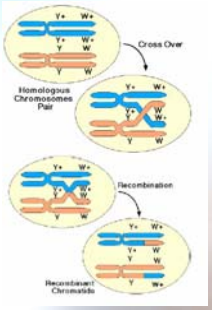
=>

2^{23} possible haploid combinations

Genes on different chromosomes recombine with probability $\theta = .5$

Recombination

- Crossover between homologous pairs of chromosomes



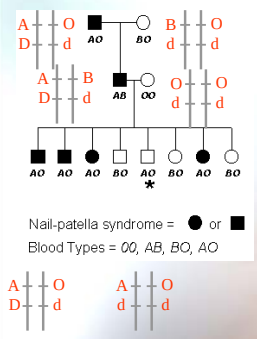
Genes on the same chromosome recombine with probability θ depending on their distance and location on the chromosome

LOD (Log Ratio of Odds) Scores

- Tests a genetic model which states that the two markers (or a disease locus and a genetic marker) are linked with a recombination fraction of θ and which requires that parameters are specified:
  - Dominant or recessive
  - Degree of penetrance
  - Allele frequencies
- LOD score :

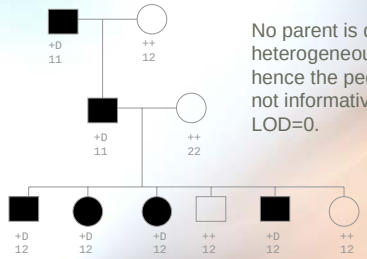
$$\log \left\{ \frac{\text{Likelihood of two markers are linked}}{\text{Likelihood of two markers are unlinked}} \right\}$$

Linkage Distance between Genes



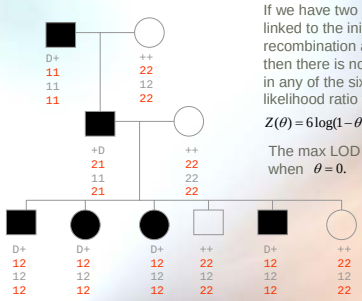
Recombination Frequency	Lod Score
0.050	0.951
0.100	1.088
0.125	1.099
0.150	1.090
0.200	1.031
0.250	0.932

Model-based Multipoint Linkage Analysis



No parent is doubly heterogeneous and hence the pedigree is not informative. LOD=0.

Model-based Multipoint Linkage Analysis



If we have two markers tightly linked to the initial marker (no recombination among them), then there is no recombination in any of the six offspring. The likelihood ratio for it is

$$Z(\theta) = 6\log(1-\theta) - 6\log(0.5)$$

The max LOD =1.8, reached when $\theta = 0$.

Model-free Linkage Analysis

- Identity by Descent (IBD): Two individuals are said to share a gene identity by descent if one gene is a directly-inherited copy of the other or if they are both directly-inherited copies of a common ancestral gene.
- Identify by State (IBS): Two individuals are said to share an allele identity by state if they both have that allele (not necessarily inherited from the same ancestral).
- Idea of model-free linkage analysis
  - Test the allele sharing for each affected relative pair (or larger group of affected relatives) against the expected sharing in the absence of linkage.
  - The usual measure of allele sharing is IBD

Sib Pair Studies

- For each sib pair, count the number of alleles shared IBD.
- Average the counts over all sib pairs.
- Estimate the expectation and variable under the null hypothesis of no linkage.
- Standardize the statistic and compare it with the standard normal distribution.
- Perform a one-sided test.

Linkage Equilibrium

Risch and Merikangas (1996 Science) pointed out that for genes of modest effect, strategies employing linkage disequilibrium (LD) would be more powerful than traditional linkage analysis.

State of random association between alleles at different markers

Locus 1 A--- p_1
a--- p_2

Locus 2 B --- q_1
b --- q_2

Gamete	freq	LE
AB	p_{11}	p_1q_1
Ab	p_{12}	p_1q_2
aB	p_{21}	p_2q_1
ab	p_{22}	p_2q_2

Linkage Disequilibrium (LD)

- Genes which are not in linkage equilibrium
- Genes which occur more often than expected from the law of independent assortment
- Non-random association in a population of alleles at two closely linked loci based on having a common ancestor

How Does LD Occur

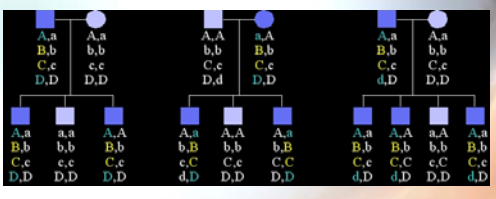
- Alleles that are closely linked will be commonly inherited
- But, in time, disequilibrium will disappear due to recombination (i.e. Allele frequencies will equalize). If two alleles 1 Mb apart are in disequilibrium, then in 70 generations the disequilibrium will decay by 50%
- Any new mutation (allele) will occur on a specific chromosome and the mutated allele will be associated with the alleles present at all loci on that chromosome
- With more meiotic events, recombination between loci causes decay of LD and the alleles return to equilibrium
- The decay takes longer for alleles closely linked due to less chance of recombination

How Is LD Measured?

D, D_{max} and D'

- $D = f_{AB}-f_Af_B = f_{ab}f_{AB} - f_{Ab}f_{aB}$
  $D = 0$ *LE;* $D < 0$ or > 0 *LD*
- $D_{max} = \min(f_Af_b, f_af_B)$, when $D > 0$
  $= \max(-f_Af_B, -f_af_b)$, when $D < 0$
- $D'=D/D_{max}$
  $0 \le |D'| \le 1$
  D' is the measure of the strength of LD

Linkage vs. LD

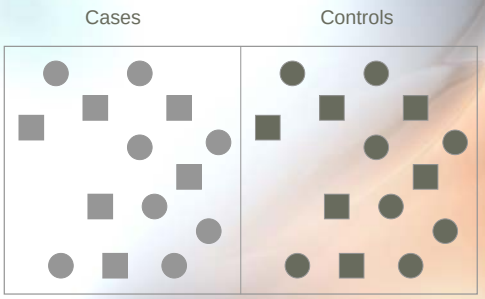


All loci are "linked" to the unobserved disease allele within each of the 3 families, but only alleles "B" and "C" are in LD (associated) with the disease allele across families.

Linkage vs. LD

1. Linkage focuses on a locus, LD on an allele
2. Linkage is resulted from recombination events in the last 2-3 generations, LD from much earlier, ancestral recombination events
3. Linkage measures co-segregation in a pedigree, LD in a population (essentially a huge huge pedigree)
4. Linkage is usually detected for markers reasonable close to the disease gene (1cM) , LD for markers even closer (0.01-0.02 cM)

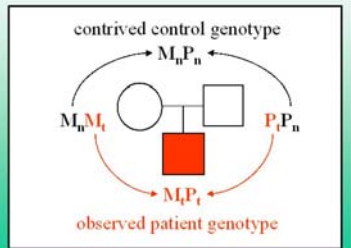
Association Studies



Haplotype Relative Risk (HRR)

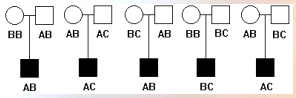
The "haplotype relative risk" approach

Falk CT, Rubinstein P (1987) Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* 51: 227-233.



Transmission Disequilibrium Test (TDT)

- A sample of cases is collected along with their parents, all of whom are genotyped for a marker
- The heterozygous parents are examined
- By chance, there is a 50% chance of them transmitting either allele of the marker to their affected offspring.
- If one marker allele is associated with the disease then this allele will be transmitted on more than 50% of occasions.



The A allele is transmitted to affected offspring four times out of five.

TDT-McNemar Test

Combinations of Transmitted and Nontransmitted Marker Alleles A and a among 2*n* Parents of *n* Affected Children

	Nontransmitted		
Transmitted	A	a	Total
A	n_{11}	n_{12}	$n_{11}+n_{12}$
a	n_{21}	n_{22}	$n_{21}+n_{22}$
Total	$n_{11}+n_{21}$	$n_{12}+n_{22}$	$2n$

$$c^2 = (n_{12} - n_{21})^2 / (n_{12} + n_{21})$$

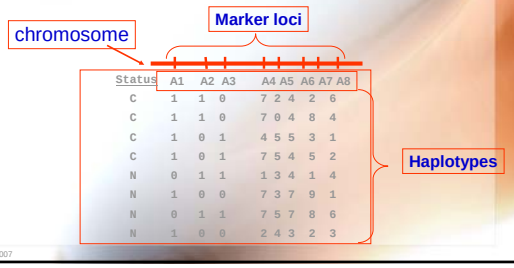
How Is TDT Used?

- Check the results of an association study

  Confirm whether a parent heterozygous for an associated and a non-associated allele transmits the associated allele more often to affected offspring
- Starts with couples with more than one affected offspring
  - Select parent that is heterozygous for marker M1
  - Test compares number of such parents who transmit the M1 allele to their affected offspring versus transmitting other allele
  - Fundamentally a test of association and linkage

Haplotype

- An ordered list of alleles of multiple linked loci on a single chromosome
  - A set of alleles from closely linked loci, usually inherited as a unit.
  - Diploid individuals have two haplotypes, a haplotype inherited from the father and a haplotype inherited from the mother
  - Haplotypes are often used to determine parentage.



Haplotypes vs. SNPs

- **ADVANTAGES**
  - Haplotypes are more informative
  - Haplotypes may enhance the power for LD analysis
  - Haplotypes can be used to study the evolutionary relationship of SNPs
- **DISADVANTAGE**
  - May not be completely resolved in the absence of family data or experimentation

Haplotype Block

A block is a set of *s* consecutive SNPs, which, although in theory could generate as many as 2^s different haplotypes, in fact shows markedly fewer in our sample of *n*, perhaps as few as *s*+1. In this case, there will be a subset of SNPs in the block whose alleles in our sample essentially determine those of the remaining SNPs in the block. These have been called haplotype tags. Outside the block, much more distinct haplotypes exist.

Haplotype Block

Daly et al (2001) "it became evident that the region could be largely decomposed into discrete haplotype blocks, each with a striking lack of diversity

Daly et al., Nat. Genet., 2001

SNP Haplotype Reconstruction

- **With a random sample of multilocus genotypes at a set of SNPs, we can attempt**
  - **To estimate the frequencies of all possible haplotypes**

Subject 1	AA	Bb	Cc
Subject 2	Aa	Bb	cc
Subject 3	AA	BB	cc
Subject 4	aa	BB	Cc
Subject 5	Aa	Bb	CC

A	B	C
A	b	c
	or	
A	B	c
A	b	C

Desired Results

Individual phase

Subject 1	A	B	C
	A	b	c
Subject 2	A	b	c
	a	B	c
Subject 3	A	B	c
	A	B	c
Subject 4	a	B	C
	A	B	c
Subject 5	A	b	C
	a	B	C
	.		
	.		
	.		

Haplotype frequencies

Haplotype 1	ABC	0.1
Haplotype 2	ABc	0.3
Haplotype 3	AbC	0.1
Haplotype 4	Abc	0.2
Haplotype 5	aBC	0.2
Haplotype 6	aBc	0.1
Haplotype 7	abC	0.0
Haplotype 8	abc	0.0

SNP Haplotype Reconstruction

- **With a random sample of multilocus genotypes at a set of SNPs, we can attempt**
  - **To estimate the frequencies of all possible haplotypes**

Subject 1	AA	Bb	Cc
Subject 2	Aa	Bb	cc
Subject 3	AA	BB	cc
Subject 4	aa	BB	Cc
Subject 5	Aa	Bb	CC

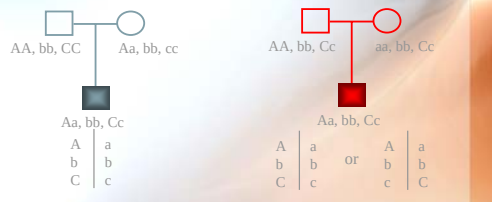
A	B	C
A	b	c
	or	
A	B	c
A	b	C

  - **To infer the haplotypes of all individuals**

Subject 1	(1,4) (2,3)
Subject 2	(2,8) (4,6)
Subject 3	(2,2)
Subject 4	(5,6)
Subject 5	(1,7) (3,5)

1,4 or 2,3

Using Parental Information



Issues and Unsolved Challenges

- How to deal with missing or ambiguous genotypes?
- How to handle large number of loci simultaneously?
- How to accelerate convergence and avoid local-mode?

References

- **Genetic Recombination**

  Dr. Craig Woodworth, Genetic Recombination in Eukaryotes, Lecture Notes, (www.clarkson.edu/class/by214/powerpoint)
- **Biochemistry**

  Mary Campbell, Saunders Press
- **Molecular Biology of the Cell**

  Alberts et.al., Garland Press
- **On-line resources**

  http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html

  Harvard MCB educational link (http://golgi.harvard.edu/BioLinks/EduRes.html)