Linkage Analysis using IBD distribution

Bo Peng† (bpeng@rice.edu)

† Department of Statistics, Rice University
Mar. 30, 2004
Outline

1. Identity by Descent
   - Definitions of IBS and IBD
   - Difference between IBS and IBD
Outline

1. Identity by Descent
   - Definitions of IBS and IBD
   - Difference between IBS and IBD

2. Distribution of IBD values
   - One locus, a simple case
   - Two loci with recombination
Outline

1. **Identity by Descent**
   - Definitions of IBS and IBD
   - Difference between IBS and IBD

2. **Distribution of IBD values**
   - One locus, a simple case
   - Two loci with recombination

3. **Linkage Analysis using IBD distribution**
   - Main ideas
   - Model details
   - Practical issues

4. **An example**
Identity-by-state (IBS): two alleles of the same form (i.e., having the same DNA sequence) are said to be IBS.

Problem: Exact sequence of alleles are sometimes unknown. Alleles might be grouped by observed properties.
**Identity-by-state (IBS):** two alleles of the same form (i.e., having the same DNA sequence) are said to be IBS.

Problem: Exact sequence of alleles are sometimes unknown. Alleles might be grouped by observed properties.

**Identity-by-descent (IBD):** two IBS alleles are descended from (and are therefore replicates of) the same ancestral allele.

Problem: Need pedigree information. Need highly polymorphic markers to determine IBD.
Difference between IBS and IBD, an example

Two siblings, no recombination
Difference between IBS and IBD, an example

Two siblings, no recombination

• IBS: A1, A2, B1
Difference between IBS and IBD, an example

Two siblings, no recombination

- IBS: A1, A2, B1
- IBD: only A1 (why?)
Distribution of IBD: A Simple Case

A pair of full siblings whose parents are non-inbred and genetically unrelated to each other. Let $D = \#\text{ibd}$

```
A1    x                        A2    x

A1     A2                     A3    A4
A1    x                        A2    x
```

A1   A2      A3   A4
A pair of full siblings whose parents are non-inbred and genetically unrelated to each other. Let $D = \#\text{ibd}$

W.o.l.o.g, assume $A_1, A_2$ are descended as illustrated. Then

$$P(D = 0) = P(A_3 \neq A_1 \text{ and } A_4 \neq A_2) = \frac{1}{4}$$
A pair of full siblings whose parents are non-inbred and genetically unrelated to each other. Let \( D = \#\text{ibd} \)

W.o.l.o.g, assume A1, A2 are descended as illustrated. Then

\[
P(D = 0) = P(A_3 \neq A_1 \text{ and } A_4 \neq A_2) = \frac{1}{4}
\]

\( D \) has distribution \( \text{Bin}(2, \frac{1}{2}) \). I.e.,

\[
P(D = 0) = \frac{1}{4} \quad P(D = 1) = \frac{1}{2} \quad P(D = 2) = \frac{1}{4}
\]
For two markers: $D_A$ and $D_B$ with recombination rate $\theta$ between $A$ and $B$.

- The marginal distribution of $D_A$ and $D_B$ are both $Bin\left(2, \frac{1}{2}\right)$
- Conditional distribution of $D_B$ given $D_A$ is dependent on $\theta$. 
Distribution of IBD: Two Loci

For two markers: $D_A$ and $D_B$ with recombination rate $\theta$ between $A$ and $B$.

- The marginal distribution of $D_A$ and $D_B$ are both $Bin(2, \frac{1}{2})$.
- Conditional distribution of $D_B$ given $D_A$ is dependent on $\theta$.

Some facts:

- The probability that $A$ and $B$ have the same IBD status for the gametes from one parent is $\phi = \theta^2 + (1 - \theta)^2$. (no recombination or two recombination.)
- The probability that $A$ and $B$ have the same IBD status for the gametes from both parents is $\phi^2$.
- Other cases are more complicated ...
Conditional distribution of $D_B$

Conditional distribution $P(D_B \mid D_A)$ ... used later

<table>
<thead>
<tr>
<th></th>
<th>$D_B = 0$</th>
<th>$D_B = 1$</th>
<th>$D_B = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_A = 0$</td>
<td>$\phi^2$</td>
<td>$2\phi (1 - \phi)$</td>
<td>$(1 - \phi)^2$</td>
</tr>
<tr>
<td>$D_A = 1$</td>
<td>$\phi (1 - \phi)$</td>
<td>$1 - 2\phi (1 - \phi)$</td>
<td>$\phi (1 - \phi)$</td>
</tr>
<tr>
<td>$D_A = 2$</td>
<td>$(1 - \phi)^2$</td>
<td>$2\phi (1 - \phi)$</td>
<td>$\phi^2$</td>
</tr>
</tbody>
</table>

For example,

\[
P(D_B = 0 \mid D_A = 1) = P\text{ (same ibd status)} \times P\text{ (different ibd status)} = \phi (1 - \phi)
\]

\[
P(D_B = 1 \mid D_A = 1) = 1 - P(D_B = 0 \mid D_A = 1) = 1 - 2\phi (1 - \phi)
\]
Some other facts

The covariance between $V_A$ and $V_B$ is

$$E(D_A D_B) = \sum E(D_A D_B | D_A) P(D_A)$$
$$= \frac{1}{2} \times (1 - 2\phi (1 - \phi) + 2\phi (1 - \phi)) + \frac{1}{4} \times 2 \left(2\phi (1 - \phi) + 2\phi^2\right)$$
$$= \frac{1}{2} + \phi$$

$Cov(D_A, D_B) = E(D_A D_B) - E(D_A) E(D_B) = \phi + \frac{1}{2} - 1 = \frac{2\phi - 1}{2}$

$Corr(D_A, D_B) = \frac{Cov(D_A, D_B)}{\sqrt{Var(D_A) \cdot Var(D_B)}} = 2\phi - 1$

Note that $Corr(D_A, D_B) = 0$ if $\theta = \frac{1}{2}$, $Corr(D_A, D_B) = 1$ if $\theta = 0$. These are intuitively true.
Linkage Analysis using IBD: Assumptions

Suppose that

- There is only a single disease susceptibility locus $A$. (This is the simplest model).
Linkage Analysis using IBD: Assumptions

Suppose that

- There is only a single disease susceptibility locus $A$. (This is the simplest model).

- Locus $A$ has two alleles $A_1$ and $A_2$ at population frequencies $p$ and $q = 1 - p$, respectively.
Suppose that

- There is only a single disease susceptibility locus $A$. (This is the simplest model).
- Locus $A$ has two alleles $A_1$ and $A_2$ at population frequencies $p$ and $q = 1 - p$, respectively.
- The penetrance vector for genotype $A_1A_1$, $A_1A_2$, $A_2A_2$ is $(f_1, f_2, f_3)$.
  1. for simple recessive disease, the vector is $(1, 0, 0)$ or $(0, 0, 1)$
  2. for simple dominant disease, the vector is $(1, 1, 0)$ or $(0, 1, 1)$
  3. for quasi-recessive or quasi-dominant disease, $0 \leq f_i \leq 1$. 
Linkage Analysis using IBD: Assumptions

Suppose that

- There is only a single disease susceptibility locus $A$. (This is the simplest model).

- Locus $A$ has two alleles $A_1$ and $A_2$ at population frequencies $p$ and $q = 1 - p$, respectively.

- The penetrance vector for genotype $A_1A_1$, $A_1A_2$, $A_2A_2$ is $(f_1, f_2, f_3)$.
  1. for simple recessive disease, the vector is $(1, 0, 0)$ or $(0, 0, 1)$
  2. for simple dominant disease, the vector is $(1, 1, 0)$ or $(0, 1, 1)$
  3. for quasi-recessive or quasi-dominant disease, $0 \leq f_i \leq 1$.

- If allele frequencies and penetrance vector are unknown (as in most of the cases), other quantities estimated from the sample population can be used.
Linkage Analysis using IBD: Ideas

- Data: the genotype at locus $B$ of affected sib-pairs and their parents. Let $X = 0, 1, 2$ be the number of affected siblings.
Linkage Analysis using IBD: Ideas

- Data: the genotype at locus $B$ of affected sib-pairs and their parents. Let $X = 0, 1, 2$ be the number of affected siblings.

- We can calculate $P(D_B \mid X)$ if we know allele frequencies and penetrance vector.

$$P(D_B \mid X) = \sum_{D_A} P(D_B \mid D_A) P(D_A \mid X)$$

If no linkage exists ($\theta = \frac{1}{2}$), $P(D_B \mid X) = P(D_B) = Bin(2, \frac{1}{2})$. 
Linkage Analysis using IBD: Ideas

- Data: the genotype at locus $B$ of affected sib-pairs and their parents. Let $X = 0, 1, 2$ be the number of affected siblings.

- We can calculate $P(D_B \mid X)$ if we know allele frequencies and penetrance vector.

  $P(D_B \mid X) = \sum_{D_A} P(D_B \mid D_A) P(D_A \mid X)$

  If no linkage exists ($\theta = \frac{1}{2}$), $P(D_B \mid X) = P(D_B) = Bin(2, \frac{1}{2})$.

- The deviation of $P(D_B \mid X)$ to $P(D_B)$ is a (monotone) function of $\theta$.

- The unconditioned distribution of $D_B$ or $D_A$, if interested, can be obtained by

  $P(D) = \sum_{X=0}^{2} P(D \mid X) P(X)$
How to calculate $P(D_A \mid X)$?

$P(D_B \mid X)$ can be evaluated from $P(D_B \mid D_A)$ and $P(D_A \mid X)$:

$$P(D_B \mid X) = \sum_{D_A=0}^{2} P(D_B \mid D_A) P(D_A \mid X)$$

Let $K$ denote mating types, then

$$P(D_A \mid X) = \frac{\sum_K P(D_A, X \mid K) P(K)}{P(X)} = \frac{\sum_K P(D_A, X \mid K) P(K)}{\sum_K P(X \mid K) P(K)}$$
How to calculate $P(D_A | X)$?

$P(D_B | X)$ can be evaluated from $P(D_B | D_A)$ and $P(D_A | X)$:

$$P(D_B | X) = \sum_{D_A=0}^{2} P(D_B | D_A) P(D_A | X)$$

Let $K$ denote mating types, then

$$P(D_A | X) = \frac{\sum_K P(D_A, X | K) P(K)}{P(X)} = \frac{\sum_K P(D_A, X | K) P(K)}{\sum_K P(X | K) P(K)}$$

For example, when $K = A_1A_1 \times A_1A_1$, $P(K) = p^4$, $P(X = 2 | K) = f_1^2$, etc

$$P(D_A, X = 2 | K) = P(D_A | K) P(X = 2 | K) = \begin{cases} 
\frac{1}{4} f_1^2 & D_A = 0 \\
\frac{1}{2} f_1^2 & D_A = 1 \\
\frac{1}{4} f_1^2 & D_A = 2 
\end{cases}$$
If $p, f_i$ are unknown: $K_P$, $V_A$ and $V_D$

Let $f_i$ (can also be explained as genotypic means), $p, q$ as defined before.

1. $V_G = p^2 f_1^2 + 2pq f_2^2 + q^2 f_3^2$: Total genetic variance
   $V_D = p^2 q^2 (f_1 - 2f_2 + f_3)^2$: Residual dominance variance
   $V_A = 2pq [p (f_2 - f_1) + q (f_3 - f_2)]^2$: Additive variance.

2. $K_P = p^2 f_1 + 2pq f_2 + q^2 f_3$: prevalence of the disease in the population
   $K_M = K_P + \frac{V_D + V_A}{K_P}$: prevalence of the disease in monozygotic twins
   $K_O = K_P + \frac{1}{2} V_A$: prevalence of the disease in offsprings
   $K_S = K_P + \frac{1}{2} V_A + \frac{1}{2} V_D$: prevalence of the disease in siblings.

$(K_P, K_M, K_O, K_S) \sim (K_P, V_D, V_A)$ has one less variable than $(p, f_1, f_2, f_3)$. However, these three variables are enough to determine the conditional distributions.
Decomposition of genetic variance

$V_G$ (total genetic variance) can be decomposed to $V_D$ (residual dominance variance) and $V_A$ (additive variance), using the following additive model

\[ f_1 = 2\mu_{A_1} + \mu_{A_1A_1} \]
\[ f_2 = \mu_{A_1} + \mu_{A_2} + \mu_{A_1A_2} \]
\[ f_3 = 2\mu_{A_2} + \mu_{A_2A_2} \]

where $\mu_{XX}$ are residual dominance deviations (supposed to be small), $\mu_{A_1}, \mu_{A_2}$ are effect of allele $A_1$ and $A_2$.

\[ V_D = p^2 \mu_{A_1A_1}^2 + 2pq\mu_{A_1A_2}^2 + q^2 \mu_{A_1A_2}^2 \]
\[ V_A = V_G - V_D \]
Decomposition of genetic variance: cont.

- $V_A$ measures the variance of breeding values and is the chief determinant of how much a population will respond if subject to selection. (?)
Decomposition of genetic variance: cont.

- $V_A$ measures the variance of breeding values and is the chief determinant of how much a population will respond if subject to selection. (?

- $V_D$ measures the variance contributed when the risk of heterozygotes is not exactly halfway between that of the two homozygotes.
Decomposition of genetic variance: cont.

- $V_A$ measures the variance of breeding values and is the chief determinant of how much a population will respond if subject to selection. (?)

- $V_D$ measures the variance contributed when the risk of heterozygotes is not exactly halfway between that of the two homozygotes.

- Exact additive assumption holds if $f_2 = \frac{1}{2} (f_1 + f_3)$
Decomposition of genetic variance: cont.

- $V_A$ measures the variance of breeding values and is the chief determinant of how much a population will respond if subject to selection. (?)

- $V_D$ measures the variance contributed when the risk of heterozygotes is not exactly halfway between that of the two homozygotes.

- Exact additive assumption holds if $f_2 = \frac{1}{2} (f_1 + f_3)$

- $\mu_{A_1}, \mu_{A_2}$ is estimated by minimizing $V_D$. $\mu_{A_1} = pf_1 + qf_2$, $\mu_{A_2} = pf_2 + qf_3$. 
Decomposition of genetic variance: cont.

- $V_A$ measures the variance of breeding values and is the chief determinant of how much a population will respond if subject to selection. (?

- $V_D$ measures the variance contributed when the risk of heterozygotes is not exactly halfway between that of the two homozygotes.

- Exact additive assumption holds if $f_2 = \frac{1}{2} (f_1 + f_3)$

- $\mu_{A_1}, \mu_{A_2}$ is estimated by minimizing $V_D$. $\mu_{A_1} = pf_1 + qf_2$, $\mu_{A_2} = pf_2 + qf_3$.

- $V_D$ and $V_A$ are usually estimated from sample data, using two-way analysis of variance.
Results

The distribution of IBD can be written directly as $V_D$, $V_A$ etc. For example, when $\theta = 0$

\[
P(D_B = 0 \mid X = 2) = \frac{1}{4} - \frac{\frac{1}{2}V_A + \frac{1}{4}V_D}{4 \left( K_p^2 + \frac{1}{2}V_A + \frac{1}{4}V_D \right)}
\]

\[
P(D_B = 1 \mid X = 2) = \frac{1}{2} - \frac{\frac{1}{2}V_D}{4 \left( K_p^2 + \frac{1}{2}V_A + \frac{1}{4}V_D \right)}
\]

\[
P(D_B = 2 \mid X = 2) = \frac{1}{4} + \frac{\frac{1}{2}V_A + \frac{3}{4}V_D}{4 \left( K_p^2 + \frac{1}{2}V_A + \frac{1}{4}V_D \right)}
\]

Note

1. $V_A$, $V_D$ are positive so the distribution of $P(D_B \mid X)$ is skewed to $D_B = 2$ compared to that of $P(D_B)$.

2. We can estimate $P(D_B)$ from sample. By comparing $P(D_B)$ and $P(D_B \mid X)$ for different $\theta$, we can estimate $\theta$. 
An example

Juvenile diabetes mellious (1975)

- Marker: HLA complex. A region on the short arm of chromosome 6 that contains many genes involved in the immune system. HLA stands for "human lymphocyte antigen." These markers are highly polymorphic.
An example

Juvenile diabetes mellious (1975)

- Marker: HLA complex. A region on the short arm of chromosome 6 that contains many genes involved in the immune system. HLA stands for "human lymphocyte antigen." These markers are highly polymorphic.

- 17 sibships, 15 of which contain exactly two affected sibs.
An example

Juvenile diabetes mellious (1975)

- Marker: HLA complex. A region on the short arm of chromosome 6 that contains many genes involved in the immune system. HLA stands for "human lymphocyte antigen." These markers are highly polymorphic.

- 17 sibships, 15 of which contain exactly two affected sibs.

- Of these 15 independent sibpairs, 10 had an IBD score of 2, four had an IBD score of 1 and one pair had an IBD score of 0.
Analysis and conclusion

- Null hypothesis: $\theta = \frac{1}{2}$. Disease susceptibility locus is not close to HLA complex.
Analysis and conclusion

- Null hypothesis: $\theta = \frac{1}{2}$. Disease susceptibility locus is not close to HLA complex.

- $K_p = 0.0058$, $V_A = 3.848 \times 10^{-4}$, $V_D = 4.269 \times 10^{-4}$ are estimated from other paper/data.
Analysis and conclusion

- Null hypothesis: $\theta = \frac{1}{2}$. Disease susceptibility locus is not close to HLA complex.

- $K_p = 0.0058$, $V_A = 3.848 \times 10^{-4}$, $V_D = 4.269 \times 10^{-4}$ are estimated from other paper/data.

- Multiple sets of $(p, f_1, f_2, f_3)$ can be estimated based on different assumptions. They all lead to the same IBD distribution.
Analysis and conclusion

- Null hypothesis: $\theta = \frac{1}{2}$. Disease susceptibility locus is not close to HLA complex.

- $K_p = 0.0058$, $V_A = 3.848 \times 10^{-4}$, $V_D = 4.269 \times 10^{-4}$ are estimated from other paper/data.

- Multiple sets of $(p, f_1, f_2, f_3)$ can be estimated based on different assumptions. They all lead to the same IBD distribution.

- The expected percentage of affected sib pairs ($X = 2$) with IBD = 0, 1, and 2 are found to be 2.5%, 34% and 63.5%. (Observed percentages are: 66.7%, 26.7% and 6.7%).
Analysis and conclusion

• Null hypothesis: $\theta = \frac{1}{2}$. Disease susceptibility locus is not close to HLA complex.

• $K_p = 0.0058$, $V_A = 3.848 \times 10^{-4}$, $V_D = 4.269 \times 10^{-4}$ are estimated from other paper/data.

• Multiple sets of $(p, f_1, f_2, f_3)$ can be estimated based on different assumptions. They all lead to the same IBD distribution.

• The expected percentage of affected sib pairs ($X = 2$) with IBD = 0, 1, and 2 are found to be 2.5%, 34% and 63.5%. (Observed percentages are: 66.7%, 26.7% and 6.7%).

• Conclusion: strong linkage.
Analysis and conclusion

- Null hypothesis: $\theta = \frac{1}{2}$. Disease susceptibility locus is not close to HLA complex.

- $K_p = 0.0058$, $V_A = 3.848 \times 10^{-4}$, $V_D = 4.269 \times 10^{-4}$ are estimated from other paper/data.

- Multiple sets of $(p, f_1, f_2, f_3)$ can be estimated based on different assumptions. They all lead to the same IBD distribution.

- The expected percentage of affected sib pairs ($X = 2$) with IBD = 0, 1, and 2 are found to be 2.5%, 34% and 63.5%. (Observed percentages are: 66.7%, 26.7% and 6.7%).

- Conclusion: strong linkage.

- Test of significance is available. (David’s talk.)
About this presentation

This presentation uses

- Lyx
- Tikbibtex
- Foiltex
- Ppower4
References


*go back to front, ibd, distribution of ibd, model*