Relationships, Relatedness, and the Coancestry of Genome

Elizabeth Thompson
Department of Statistics
University of Washington.

MathAcrossCampus Talk
University of Washington, Mar 10, 2011
Mendelian segregation: Identity by descent

- **Mendel's first law** (1866): Each individual has two genome copies; one maternal, one paternal. At every location, to each offspring independently, a parent copies a random one of the two homologous genes (chunks of DNA) he/she has at that genome location.
- Genes are **identical by descent** (*ibd*) if they are copies of the same gene in a common ancestor.

Given I have blood type O, there is increased probability my cousin has blood type O, because there is positive probability we have ibd genes, and *ibd* genes should be of the same allelic type.

- **In a pedigree**: *ibd* is well-defined, relative to the founders.
**ibd in multiple diploid individuals: one locus**

- At every location, I share one copy \(ibd\) with my mother; i.e. 50%.
- At every location, there is 50% chance I share my paternal DNA \(ibd\) with my brother, and (independently) 50% chance I share my maternal DNA with him.

- So there is chance \((1/4,1/2,1/4)\) chance of sharing \((0,1,2)\) \(ibd\) with him. That is, overall (on average) 50% of my genome.

- The average proportions are the same, but the patterns different.

- In general, the 4 genes of 2 individuals can have 15 different \(ibd\) combinations — Bell #s; # distinct partitions of 4 labeled objects.

- In general, the 12 genes of 6 individuals can have 4,213,597 different \(ibd\) patterns, but very few of these can arise in a given pedigree relationship (even a complex one).
Inheritance of chromosome segments

- Chromosomes are inherited in large chunks, $\sim 10^8$bp or 100 Mbp (1 CMbp$=10^8$bp).
- In any meiosis, crossovers occur as a Poisson process along the chromosome.
- In any meiosis, the chance that the DNA at two positions derives from different parental chromosomes increases with distance along the chromosome.
- At large distances, this probability is $\approx 1/2$ — independent inheritance.

Each mat/pat genome of $3 \times 10^9$ bp ($\sim 3,000$ Mbp) is packaged into 22 chromosomes sized from 51 to 245 Mbp.
*ibd* in remote relatives; (K. P. Donnelly, 1983)

Relatives separated by \( m \) meioses.

\[
\Pr(\text{2 kids get same}) = 1/2
\]

\[
\Pr(\text{descendants share}) = 2 \times (1/2)^m
\]

\[
\Pr(\text{share any genome length } L \text{ CMbp}) = 1 - \exp(-(m - 1)L/2^{m-1})
\]

Length of *ibd* segment \( \sim m^{-1} \) CMbp.

\[
\text{Length of } ibd \text{ segment} \sim m^{-1} \text{ CMbp.}
\]

\[
\begin{array}{c|cccc}
\text{Meioses apart} & 5 & 10 & 15 & 20 \\
\text{Log10 prob sharing} & -3.0 & -2.0 & -1.0 & 0.0 \\
\end{array}
\]

\[
\begin{array}{c|cc}
\text{meioses apart} & m = 12 & m = 20 \\
\text{ibd at point} & 0.0005 & 2 \times 10^{-6} \\
\text{any } ibd \ (L = 30 \text{ CMbp}) & 0.148 & 0.001 \\
\text{length } ibd \text{ segment} & 8.5 \text{ Mbp} & 5 \text{ Mbp} \\
\end{array}
\]

- *ibd* segments are rare but not short.
Estimating *ibd* from genetic data

- There is a large amount of variation in our genomes: at about 1 in 1000 bp, there will be two different possible alleles. These are SNPs; *single nucleotide polymorphisms*.

- DNA chunks that are *ibd* from a recent common ancestor are the same allelic type for the SNPs in the chunk (with high probability).

- DNA that is not *ibd* will be of “independent” allelic type—basically, there will be differences at many SNPs.

- Each SNP alone gives almost no information, but *ibd* comes in chunks, with more and larger chunks in closer relatives.

- Modern genetic data enable us to detect chunks of DNA that are *ibd*, using these dense but individually uninformative SNPs.

- Different relationships give rise to different probabilities of *ibd* patterns. If we can detect the portions of genome *ibd* among individuals, we should be able to estimate relationships.
Why estimate relationships/relatedness?

- Forensic questions: identifying individuals from their relatives victims of natural or man-made disasters

- Legal questions: Identifying parents, children, siblings: paternity testing, adoptions, immigration cases.

- Medical Genetics: for example, sib pair studies. Validation of stated pedigree relationships. Sample swaps.

- Conservation Genetics: studying/managing breeding for severely endangered species: California condor, Przewalski horse, Caribbean iguanas

- Ecological Genetics: non-invasive sampling of hair or feces; gene flow, and reproductive success in natural populations, dispersal of seed, pollen, and juveniles perennial plants, armadillos, salmon
Przewalski Horses; mixed-up records

Only “true” wild horse:
- 66 chromosomes (vs 64)
- Captive-bred (13 founders)
  - 1927-1997
One was known Mongolian domestic; one a hybrid(?)
Askania Nova; main “pure” group, and one more recent (1953) founder.

Many uncertainties; horses mixed up. Wrong ones shipped.
- concerns as to validity of International Stud Book.
San Diego “pure” stallion (1985), led to establishing of two groups (“pure” / “mixed”) in USA, but he was not. etc. etc.
1992: genetic marker data used to resolve many pedigree errors.

Now reintroduced in China & Mongolia, but still threatened.
Meiosis indicators and independent random switches

- In meiosis (parent-offspring transmission of DNA) \( i \), let

\[
S_i(t) = \begin{cases} 
0 & \text{if parent's maternal genome is copied to offspring} \\
1 & \text{if parent's paternal genome is copied to offspring}
\end{cases}
\]

- Mendel’s first Law:
  1. \( \Pr(S_i(t) = 1) = \Pr(S_i(t) = 0) = \frac{1}{2} \),
  2. All meioses \( i \) are independent.

- Poisson process of random switches \( 0 \leftrightarrow 1 \) in \( S_i(t) \):

\[
\Pr(S_i(t + y) \neq S_i(t)) = \frac{1 - \exp(-2y)}{2}
\]

Here \( y \) is in units of \( 10^8 \) bp (CMbp).
(Note: \( \Pr(S_i(t + y) \neq S_i(t)) \uparrow y \) and \( \rightarrow 1/2 \) as \( y \rightarrow \infty \).)

- \( S_i(t) \) is a Markov process.

\[
\Pr(S_i(t_k) = 1 \mid S_i(t_j), \ j = 1, 2, \ldots, (k - 1)) = \Pr(S_i(t_k) = 1 \mid S_i(t_{k-1})).
\]

where \( t_1 < t_2 < \ldots < t_{k-1} < t_k \).
From $S_i$ to \textit{ibd}: Genome shared with my cousin

When will the cousins share DNA \textit{ibd}?

<table>
<thead>
<tr>
<th>$S_2 = S_4$</th>
<th>$S_1 = S_3$ (Chance 50%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

$S_5 = S_6$  $S_5 = S_6 = 1$  No way.

$Pr(S_2 = S_4) = Pr(S_1 = S_3) = 1/2$

Total prob of maternal \textit{ibd} is

$(1/4) \times ((1/2) + (1/4) + (1/4))$.

- On average, the cousins share 1/4 of their maternal genomes.
- The $S_i$ are Markov, but \textit{ibd} is not.
Double first cousins and quadruple-half-first-cousins

- Each shares 1/4 of her maternal and of her paternal genome *ibd* with the other individual (on average).

- For QHFC, each of the mom and dad of each individual is related to *both* the mom and the dad of the *other* individual, but mom is not related to dad.

- For DFC, probability of sharing maternal *and* paternal genome *ibd* with the other individual is \((1/4) \times (1/4) = 1/16\). For QHFC this is 1/32.
Two types of quadruple second cousins

- QHFC exist in animal populations (e.g., horses?), but not (often?) in human populations. Quadruple-2nd-cousins exist in small human populations.

- Not all quadruple second cousins are related the same: For the cyclic type, each of mom and dad of each individual is first cousin to both mom and dad of the other.

- Overall: each shares 1/8 maternal genome and 1/8 paternal genome. But for the exchange type: Prob share both genomes is 1/64. And for the cyclic type: Prob share both genomes is 1/128.
Offspring of the two types of quadruple second cousins

- On their maternal chromosomes, the kids share $(1/16)$ \textit{ibd} (on average). Are the two pairs “equally related”?
- Unless $y = 0$ or $y = \infty$, the probability of sharing \textit{ibd} at distance $y$ are different!
- The distribution of lengths of \textit{ibd} segments are different; on average smaller in the cyclic case.
- In principle, even these two relationships are distinguishable.
Not all first cousins have the same amount of \textit{ibd}

- Back to first cousins;
  Let $\rho = (1 - \exp(-2y))/2 = \Pr(S_i(t + y) \neq S_i(t))$, ($y$ in CMbp).
  \[
  \Pr(\text{ibd}(t + y)|\text{ibd}(t)) = (1 - \rho)^2(\rho^2 + (1 - \rho)^2) + \rho^2/2
  \]

- In a 200 Mbp chromosome, mean \textit{ibd} is 50 Mbp (25%), but almost 10% have no \textit{ibd} and 10% have over 100 Mbp.

- Genome-wide, maternal cousins share 25% of maternal genome, on average, but mean $\pm$ 2 stdev is 0.16 to 0.34 (1/6 to 1/3).
Estimating relationships??; back to Przewalski horses

- Because genomes are short (variance of $ibd$ is high), estimating relationships without specific hypotheses does not work well. The $ibd$ does not determine the pedigree relationship.

- If we have specific relationship hypotheses, then we can estimate:
  - Przewalski horses; switched foals/stallions/mares,
  - Human sib-pair studies; non-sibs easily detected.
  - Human genetic studies; switched samples etc.

- In endangered species and other genetic studies, we may be more interested in proportion of genome shared $ibd$ – the actual relatedness – than in the pedigree relationship.

- For example:
  - Human genome-wide association studies:
    - Individuals assumed unrelated: detect and eliminate close relatives.
  - Conservation genetics:
    - Example of the California condor.
Estimating actual relatedness; California Condors

Genetically; Three groups

1850-1950; Population declines 100,000 to 200 (?)
1975-84; Population highly endangered; eggs taken from wild.
1984-5: Population crash; 10 survivors into captivity; also 7 eggs
Condors live long, fly far, mate for life; how are these related ??
Topa-Topa in LA Zoo 20 years, maybe brother to AC5 –from wild
Who should be bred? Who released? Maintain the gene pool.
Now over 200 total: 100 in SD/LA, 100 fly (semi-)free.
Finding genes for traits, using *ibd*

- We can estimate actual realized levels of *ibd*.

- With a clear pedigree hypothesis, we can use estimated *ibd* to validate the hypothesis.

- Conversely, given a pedigree, can we find the *ibd* segments? And why would we want to?

- Related individuals having similar trait values are (more) likely to share genome *ibd* in regions where there are genes affecting the trait.

- By finding the regions where *ibd* is correlated with similarity in trait values, we can find where there are genes affecting the trait.

- We do not want to just consider pairs of individuals – there is much more information in looking jointly, and using the descent among pedigree members.
Specifying *ibd* in a pedigree; the *ibd* graph

FGL = founder genome label.

- Nodes are (unlabeled) distinct genomes.
- Edges are (labeled) observed individuals.
- Only *ibd* matters, not (labeled) founder origins (FGL), and no longer the pedigree once *ibd* is known/inferred from SNP data!
Changes in *ibd* graph along a chromosome

Switch in meiosis to K. Switch in meiosis to J.

- Crossover events change the nodes present in observed individuals, and hence the structure of the *ibd* graph. The edges are the same, but may connect different nodes. Nodes may appear/disappear. (Nodes labeled for convenience only.)

- Changes are few (on bp scale); recall in any 1 meiosis, crossovers occur at $\sim 10^8$ bp, or once per CMbp per meiosis.

- Components of the *ibd* graph tend to be small, when only current generation(s) observed for trait.
How to estimate the *ibd* graph

- Recall the independent meiosis processes $S_i$, Markov over SNP locations $j$. Now $S_{i,j} = 0/1$ as maternal/paternal DNA transmitted in meiosis $i$ at SNP $j$.

- Recall *ibd* is not Markov; but *ibd* at $j$ is a function of $S_{\bullet,j} = \{S_{i,j}\}$.

- The SNP data $Y_{\bullet,j}$ we see at SNP $j$, depend only on $S_{\bullet,j}$, and, in fact, only through the *ibd* graph $D_j$ at $j$.

This dependence structure provides way to simulate realizations of the $\{S_{i,j}\}$ over all $i$ and $j$, conditional on all the SNP data $Y_{\bullet,j}$ over all $j$. 
Framework for trait data analysis

- Given SNP data \( \{Y_{*,j}\} \), sample \( \{S_{i,j}\} \) (jointly) for SNPs \( j \).
- Store this one sample of \( \{D_j\} \).
- In any region, \( D_T \) for trait location(s) is as defined by \( D_j \), reduced to observed \( O_T \).
- Use trait model, to compute \( P(Y_T \mid D_T) \), for multiple trait locations/models/traits....

- Only \textit{ibd} graph connects the SNP and trait analyses.
Storing and indexing ibd graphs

- *ibd* graphs can be compactly stored; save only the change points.

- Among trait-observed individuals:
  - Many different $\{S_{i,j}\}$ give the same (unlabeled) *ibd* graph, $\mathcal{D}$.
  - Many realizations give the same $\mathcal{D}_j$ at some SNPs $j$.
  - Many *ibd* graphs, $\mathcal{D}_j$, are unchanged over many SNPs $j$.

- Trait analysis depends only on $\mathcal{D}$; compute $P(\mathbf{Y}_T | \mathcal{D}_T)$ or trait-data statistic only for distinct $\mathcal{D}_T$.

- IBDgraph software by Statistics students Hoyt and Lucas Koepke allows for efficient insertion, querying, equality testing, and set operations on the collection of *ibd*-graphs, at SNPs or over SNP ranges.

- The IBDgraph software takes only a few seconds to run, and can reduce trait likelihood computations by an order of magnitude or more.
In 1990s, our analyses failed…pedigree? markers? methods?

Details of the ancestral pedigree are surely wrong/biased. We want to use the \textit{ibd} information, but not the ancestral pedigree.

With modern data, we could infer \textit{ibd} within the three families and then also between families.
Combining information among pedigrees

- Advantages of remote relatives (distant cousins):
  - less shared environment than close relatives.
  - but some genetic (allelic) homogeneity among some families
due to remote (unknown) relationships between them
  - many different variants can mess up a functional gene.

- \textit{ibd} genome segments are few, not short (Donnelly, 1983).
  Recall the example of pair of individuals separated by \( k = 20 \) meioses, total genome length \( L = 30 \) in units of CMbp.
  - Prob share any genome \( \approx 1 - \exp\left(-\frac{(k - 1)L}{2^{k-1}}\right) \approx 10^{-3} \)
  - Expected length of shared segment is \( \approx \frac{1}{k-1} \) units \( \approx 5 \) Mbp.

- Even without knowing the pedigree, \textit{ibd} genome segments will show same allelic SNP types over the segment. Population variation insures non-\textit{ibd} will show differences, if segments are not too short.

- Segments of length down to 1 Mbp are easily(?) detected with modern SNP data, even without using pedigree information.
  (Statistics students: Chris Glazner and Marshall Brown.)
• Within families, crossovers change the gene \( \text{ibd} \) graph along a chromosome.

• There may be \( \text{ibd} \) between founders in a given family,

• ... and/or between founders of different families.

• Generally, such links will be few and sparse, but, with ascertainment, several families might share \( \text{ibd} \) at some points.

• Again components of these graphs are not large/complex.

• Again, the component graphs are slowly varying (on bp scale).
Conclusions: *ibd* is fun! (but also important)

- *ibd* is the basis of all genetic similarities among relatives.

- Because we are diploid, there are many possible proportions and combinations of genome shared, at a locus and across the genome,

- Inferred *ibd* can be used to estimate relationships, degree of relatedness, or proportions of genome shared, but for a given relationship, the proportion has high variance.

- Human (and animal) genomes are short; $3 \times 10^9$ bp is a lot of DNA, $3 \times 10^6$ SNPs is a lot of variation, but per-generation inheritance is in chunks of order $10^8$bp (1 CMBp).

- In analyzing the genetics of a trait only the *ibd* matters.

- *ibd* within and among pedigrees can be inferred from SNP data, stored compactly, and used in multiple trait data analyses.