Model misspecification

Estimating allele frequencies in sibships

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

This lecture concerns an old problem in human genetics. The main topic is about estimating allele frequencies with data on correlated individuals (siblings), but really about a case study in recognizing and diagnosing problems, evaluating the relative quality of different estimates, and on trade-offs between extracting full information vs just getting reasonable answers.

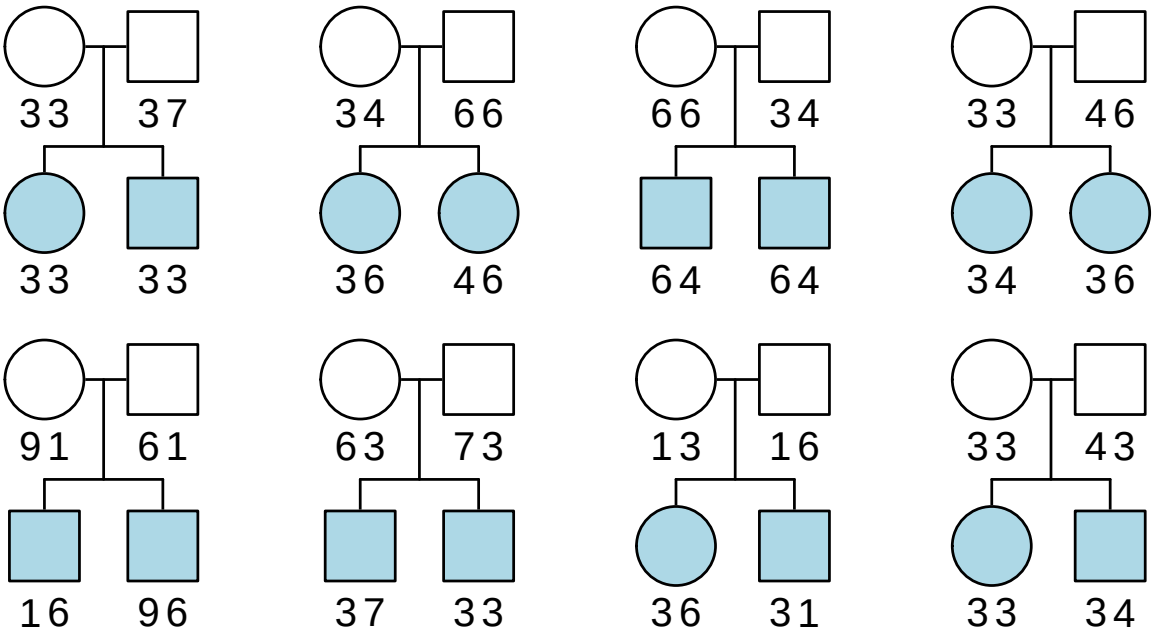
Mapping disease genes

Look for genomic regions where

individuals with similar phenotypes
also have similar genotypes

Our interest here is in mapping disease genes in humans. Mapping disease genes is, in essence, about finding genomic regions where individuals with similar phenotypes also have similar genotypes.

Affected sib pairs



In the late 1990s, one of ways we tried to identify disease genes was with affected sibling pair studies. You gather a bunch of pairs of siblings who were both affected with a disease, and then get genotype data for them, and look for genomic regions where the affected sibpairs had more similar genotypes than you would expect by chance.

IBS vs IBD

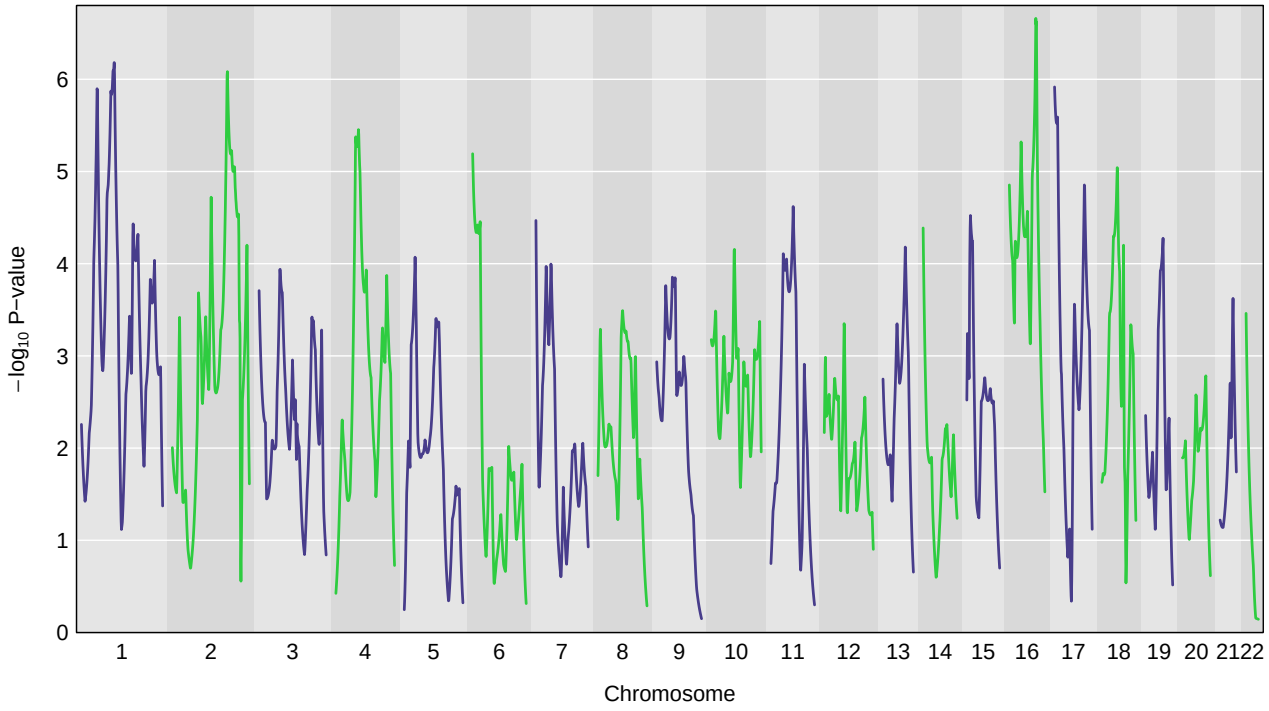
IBS = identical by state
= same allele number

IBD = identical by descent
= copies of the same ancestral allele

non-inbred sibs are IBD = 0, 1, 2
with probability = 1/4, 1/2, 1/4

In measuring genetic similarity at a locus, it's valuable to distinguish between alleles being "identical by state" (meaning they just look the same) and "identical by descent" (meaning that they are copies of an ancestral allele). Non-inbred siblings will have IBD status 0, 1, or 2, with probability 1/4, 1/2, 1/4, respectively.

Prostate cancer genome scan



In an affected sib-pair study, we'll scan across the genome; at each point will seek to estimate the proportion of alleles shared IBD between affected siblings and compare that to what is expected for siblings.

The first such study I was involved in was of prostate cancer, and consisted of maybe 150 affected sibling pairs. This plot (of $-\log_{10}$ p-values) is an approximation of my initial results. We're looking for values around 3, so these were super exciting to me. I distinctly remember faxing these results to my collaborators, thinking "I am so awesome. I will conquer all diseases."

Lesson

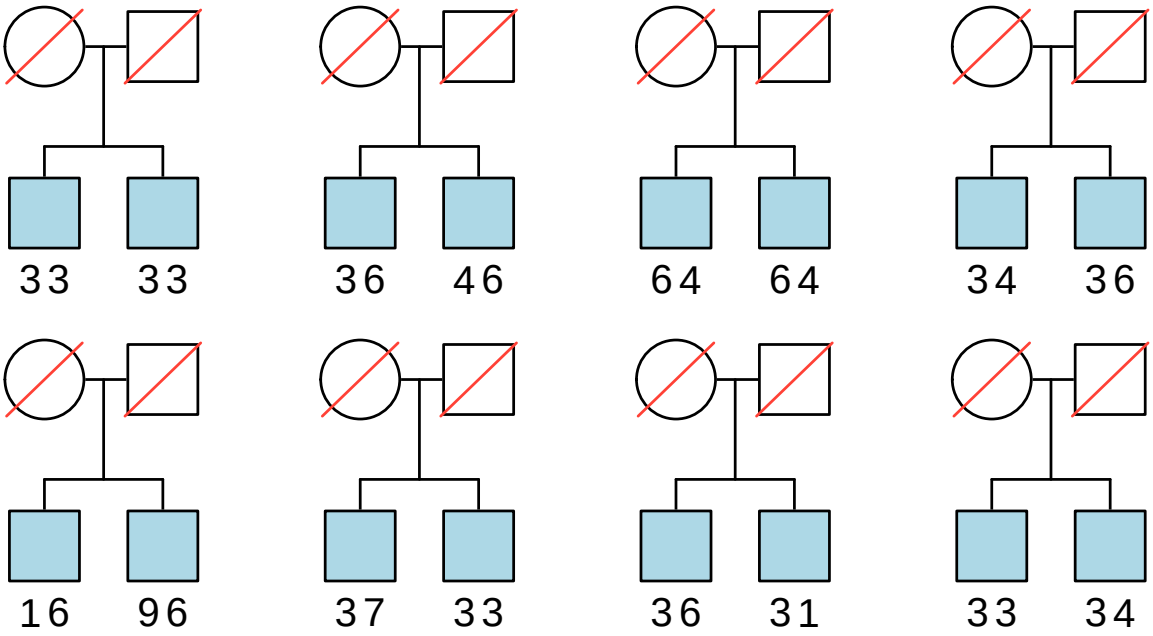
If it seems too good to be true,

it probably is.

But as soon as I sent that fax, I was like, "Huh. Those results seem too good to be true."

It turns out that I'd messed up the allele frequencies and so the results were all messed up.

Prostate cancer pairs

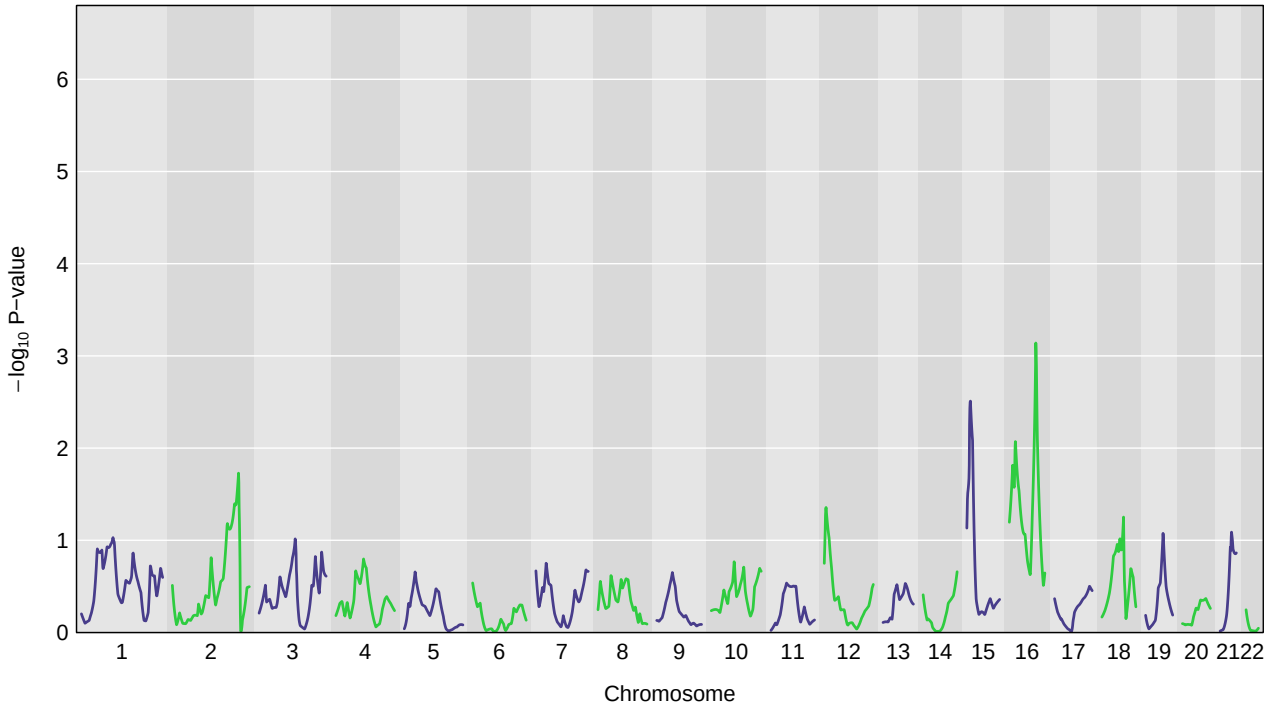


In this prostate cancer study, the affected sibpairs are all old, and there's essentially no data on the parents. In this case, determining the number of alleles shared IBD turns out to be particularly sensitive to the allele frequencies.

For example, if they're both 3/3 and 3 is relatively rare, that it's very likely that they're IBD=2. But if 3 is quite common, then they could reasonably be IBD=1 or 0.

If they both share a 6 allele, is that IBD=1, or IBD=0? If 6 is rare, you'd lean towards IBD=1, but if 6 is common, it could be either.

Prostate cancer genome scan – corrected



The unusually strong results I got were entirely due to a mistake in the code that estimated the allele frequencies. If I use more reasonable estimates, this is what I get. There's maybe evidence for a disease locus on chr 16 and possibly also 15, but the evidence isn't very strong.

And this is sort of what we'd expect given the size of this study. We're hoping to find some evidence of a disease gene, but we're not going to see the whole genome lighting up.

Estimating allele frequencies

Usually, you would use the founders in the pedigrees.
(assumed unrelated)

What if you only have sibships?

And just how should we estimate the allele frequencies? The normal way is to use the founders founders in the pedigrees (here, the parents), who we would generally assume to be unrelated. But in this case the parents are entirely absent. We just have data on the siblings themselves.

Estimating allele frequencies with sibpairs

Method 1: Use a random sibling from each

$$\text{var}(\hat{p}^{(1)}) = \frac{p(1-p)}{2n}$$

Method 2: Use everyone, ignoring relationships

$$\text{var}(\hat{p}^{(2)}) = (3/4)\left(\frac{p(1-p)}{2n}\right)$$

relative efficiency = 4/3 = 1.33
(best possible = 1.5)

Well, one way is to just use a single sibling from each pair. Then you have a set of unrelated individuals. Alternatively, you could use the data on all of the siblings. But then you are potentially overcounting some alleles.

Which one is the better estimate? Is it better to omit some data so that what's left is a set of pure, independent counts? Or is it better to use all of the data and ignoring the over-counting of alleles.

Both methods give unbiased estimates. And the standard error for the first estimate is that of a binomial proportion with sample size $2n$, where n is the number of sibling pairs, since each individual gives you data on 2 alleles.

I'll show in a moment that the second estimate has variance that is $3/4$ that of the first estimate, so a relative efficiency of $4/3$. The best possible estimate you can get with sibling pair data has a relative efficiency of $3/2$, since two siblings are on average giving you data on three distinct alleles.

My favorite equations

$$\begin{aligned} \mathrm{E}(X) &= \mathrm{E}[\mathrm{E}(X|Z)] \\ \mathrm{var}(X) &= \mathrm{E}[\mathrm{var}(X|Z)] + \mathrm{var}[\mathrm{E}(X|Z)] \\ \mathrm{cov}(X,Y) &= \mathrm{E}[\mathrm{cov}(X,Y|Z)] + \mathrm{cov}[\mathrm{E}(X|Z), \mathrm{E}(Y|Z)] \end{aligned}$$

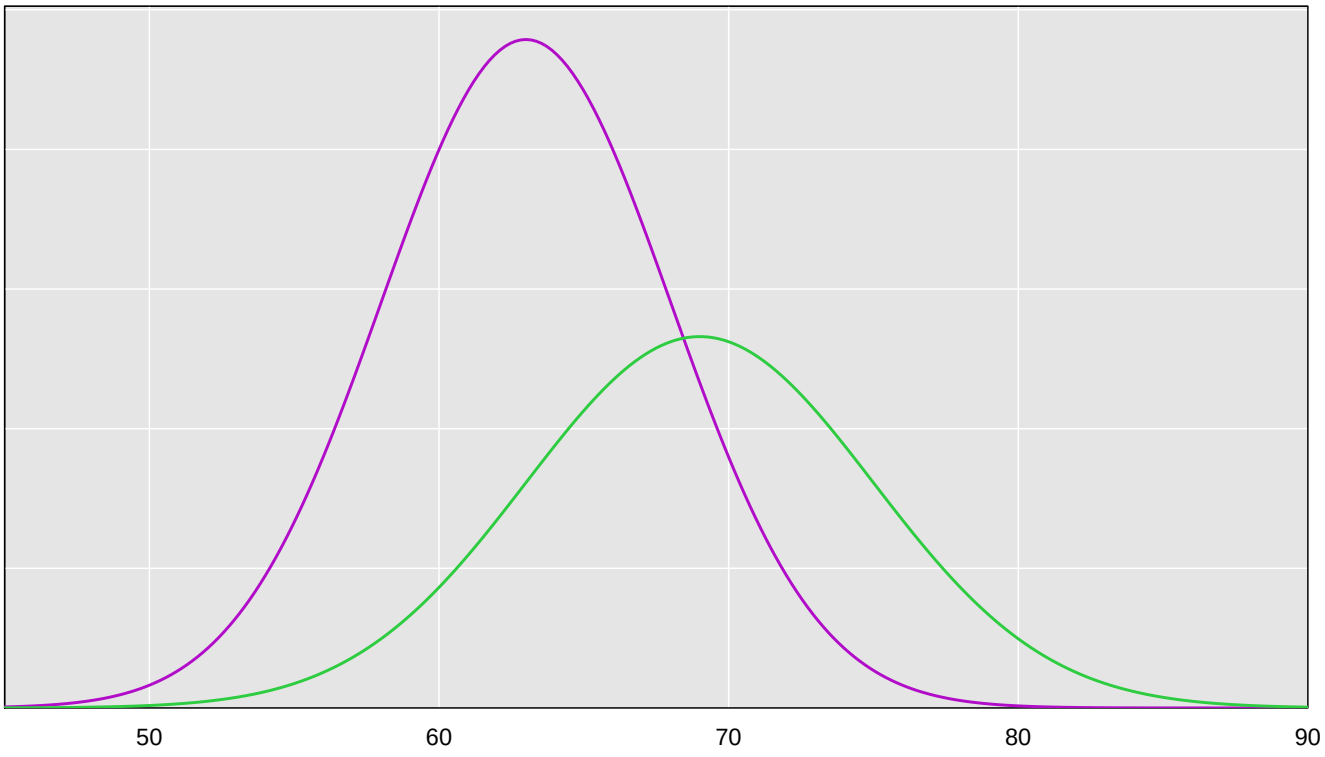
Everything is a mixture

How do we determine the variance of that second estimate? Well it turns out that we'll be using some of my favorite equations, particularly that the variance of a random variable can be expressed as the mean conditional variance plus the variance of the conditional mean.

Another fave

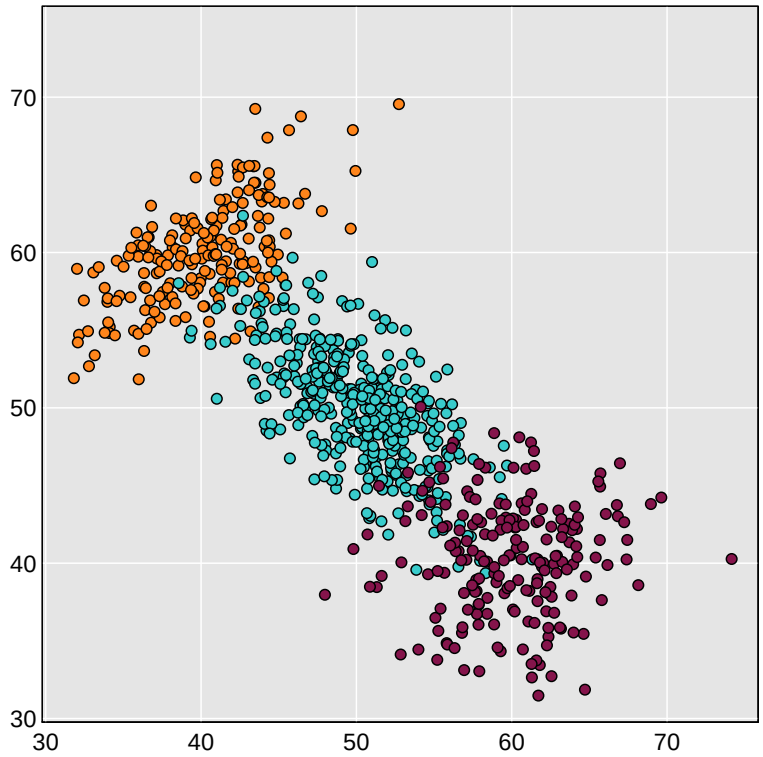
$$\begin{aligned} \mathrm{cov}(X, aY + bZ) &= a\,\mathrm{cov}(X, Y) + b\,\mathrm{cov}(X, Z) \\ \\ \text{thus}\quad \mathrm{var}(X + Y) &= \mathrm{cov}(X + Y, X + Y) \\ &= \mathrm{cov}(X + Y, X) + \mathrm{cov}(X + Y, Y) \\ &= \mathrm{cov}(X, X) + \mathrm{cov}(X, Y) + \mathrm{cov}(Y, X) + \mathrm{cov}(Y, Y) \\ &= \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X, Y) \end{aligned}$$

Here's another of my favorite equations, which is what you need in order to remember that the variance of the sum is the sum of the variances plus twice the covariance.



13

Regarding those initial equations, one way to think about them is of data that are mixtures. If there are two types of individuals, each following a different normal distribution, then the overall average is the average of the two individual averages, and the overall variance is the average of the variances plus the variance of the averages.



For the covariance, think of a pair of correlated variables where again the data are a mixture of a set of underlying bivariate distributions. The overall covariance is the average of the individual covariances plus the covariance of the averages.

Back to that SE

Let X_i = number of 1 alleles in sib *i*.

We want $\text{var}(X_1 + X_2)$

$$\begin{aligned}
\text{And so really } \text{cov}(X_1, X_2) &= \text{E}[\text{cov}(X_1, X_2|\text{IBD})] + \text{cov}[E(X_1|\text{IBD}), E(X_2|\text{IBD})] \\
&= \text{E}[\text{cov}(X_1, X_2|\text{IBD})] \\
&= \sum_{k=0}^{2} \text{cov}(X_1, X_2|\text{IBD} = k)\, \Pr(\text{IBD} = k)
\end{aligned}$$

Now back to that SE, we can focus on a particular allele, say allele 1. And then we count the number of 1 alleles in each sibling. We want the variance of the sum, which means we're going to want to covariance of the counts for the two siblings. And it turns out to be easiest to look at this in terms of the conditional covariance, given the number of alleles that the two siblings share IBD.

$\text{E}(X_i|\text{IBD})$ doesn't depend on the IBD status, so this ends up just being a number, and so the covariance of the two conditional means is 0, so that term drops out.

Also

$$\begin{aligned} \mathrm{cov}(X, Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \\ \mathrm{cov}(X, Y|Z) &= \mathrm{E}(XY|Z) - \mathrm{E}(X|Z)\mathrm{E}(Y|Z) \end{aligned}$$

Similar to the common expansion of the variance, we can write the covariance as the expected product minus the product of the two expected values, and this works whether you're conditioning on some other variable or not.

TABLE IV. Joint Distribution of the Numbers of 1 Alleles Carried by Two Individuals, Given the Number of Alleles They Share IBD

IBD	X_1, X_2	$\Pr(X_1, X_2 \mid \text{IBD})$
0	0,0	$(1-p)^4$
	0,1	$2p(1-p)^3$
	1,0	$2p(1-p)^3$
	1,1	$4p^2(1-p)^2$
	0,2	$p^2(1-p)^2$
	2,0	$p^2(1-p)^2$
	1,2	$2p^3(1-p)$
	2,1	$2p^3(1-p)$
	2,2	p^4
1	0,0	$(1-p)^3$
	0,1	$p(1-p)^2$
	1,0	$p(1-p)^2$
	1,1	$p(1-p)$
	1,2	$p^2(1-p)$
	2,1	$p^2(1-p)$
	2,2	p^3
2	0,0	$(1-p)^2$
	1,1	$2p(1-p)$
	2,2	p^2

Broman (2001) doi:10.1002/gepi.2

The basis of the calculation is figuring out the conditional distribution of (X_1, X_2) given IBD status, shown here. Remember X_i is the number of 1 alleles in the ith sib.

Consider the case IBD=2. Here the two sibs have the same genotypes, and so $X_1 = X_2$. The frequencies are according to the Hardy-Weinberg proportions: draw 2 alleles with replacement from a vat with p 1's,

The case IBD=0 is similarly easy. Here the two sibs have independent genotypes, and so the probabilities are according to the case that you draw 4 alleles with replacement from a vat with p 1's.

IBD=1 is like drawing three alleles with replacement from a vat of p 1's. In the $X_1 = 1, X_2 = 1$ case you then need to sum over the two possibilities: that the allele in common between the sibs is a 1 or that it's not.

Lessons

- Omitting data is usually bad
- Crudely ignoring correlations can be good
  - You might even be able to figure out the SE

The lesson that you learn here are that omitting data are ususally (making for a worse estimate), and that just ignoring correlations in the data may still give you an okay estimate. And in many cases, you might be able to figure out the SE of that estimate that ignores the correlations.

From what I've seen, this is true quite generally. If you find yourself thinking, "Should I omit some data, so what's left is all independent?" you should hesitate and think back to this situation.

I should also emphasize here that the analytic calculations I've shown are not strictly necessary. We could have just as well proceeded by computer simulation. We could simulate siblings' genotypes and then drive the two estimates and repeat many times, and we'd immediately get to the answers about unbiasedness and the 4/3 relative efficiency.

Method 3

Account for relationships in the estimate

Missing data = IBD status for a sib pair at a marker

Use EM algorithm:

E step: estimate IBD status given allele frequencies
M step: estimate allele frequencies given IBD status

Now we can actually estimate the allele frequencies accounting for the relationships between the siblings. The key idea is that if we need which alleles were IBD, we could prevent the overcounting we get when ignoring the relationships.

This naturally leads to yet another instance of an EM algorithm. In the E step, we estimated IBD status given current estimates of allele frequencies. In the M step, we re-estimate the allele frequencies, given the IBD status.

Method 4

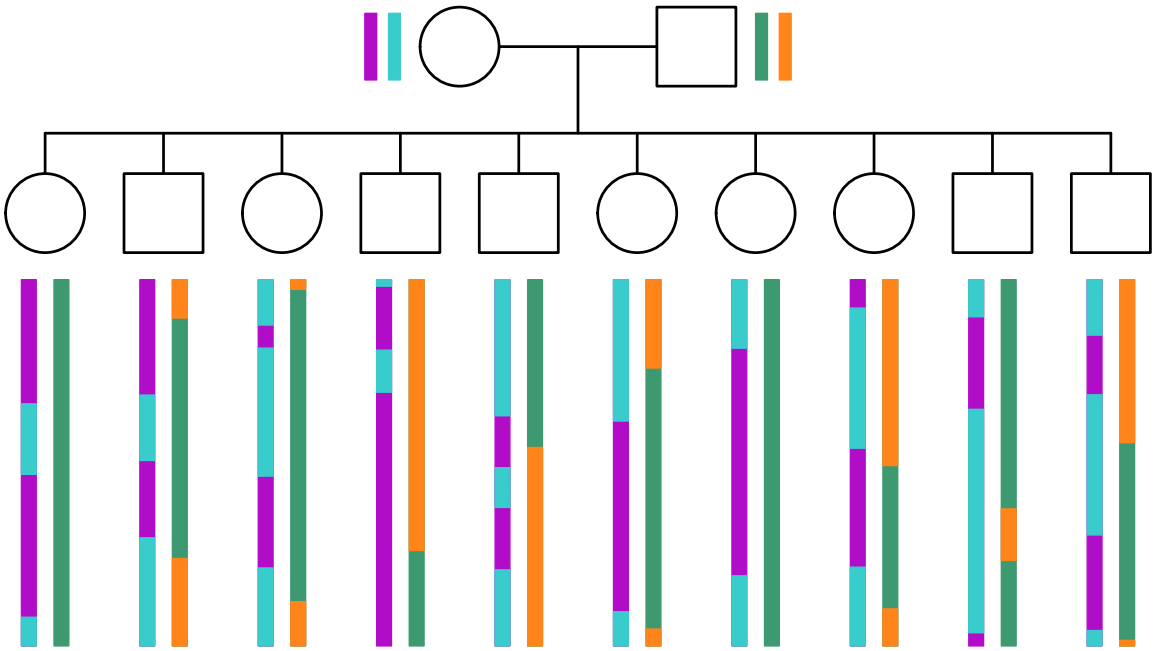
Make use of the multiple markers on a chromosome

- Markers along chromosome give improved info about IBD status
- Again, an EM algorithm:
  - Estimate IBD along chromosome given allele frequencies
  - Re-estimate allele frequencies using IBD information

But if that method were to work, we could do even better by taking account of surrounding markers to get improved estimates of IBD status which should then give us improved estimates of allele frequencies.

So again we have an EM algorithm, but at the E step we use all of the markers on a chromosome to infer IBD status; the M step will remain the same as before.

Siblings' chromosomes



To illustrate the use of surrounding markers, here are the chromosomes for a set of siblings in a large sibship. We've colored the parents' chromosomes as purple, blue, green, and orange. The chromosome from the mom is purple and blue; the chromosome from the data is green and orange. Two siblings share an allele IBD at a particular position if they got the same color. You can see that those large patches of color should mean that surrounding markers can be used to help determine IBD status.

Average relative efficiency

	Allele frequency			
Method	0.05	0.10	0.15	0.20
1		1.00		
2		1.33		
3	1.46	1.45	1.44	1.43
4	1.48	1.46	1.45	1.44

Broman (2001) doi:10.1002/gepi.2

We can't assess the relative merits of methods 3 and 4 analytically, but we can carry out a simulation to see how well they work. This table shows the relative efficiency of each method, relative to the first method where we just use one sibling per pair. Methods 3 and 4 improve a bit on method 2, and actually approach the best possible, a relative efficiency of 1.5. Methods 3 and 4 work somewhat better for less-frequent alleles, as it is easier to establish IBD in that case.

Method 3

	Allele frequency			
het	0.05	0.10	0.15	0.20
0.7	1.45	1.44	1.43	1.42
0.8	1.46	1.45	1.44	1.43
0.9	1.48	1.47	1.47	1.45

Broman (2001) doi:10.1002/gepi.2

Looking more closely at method 3, we note that its performance also depends on the amount of information at the marker, measured by the heterozygosity (the chance that an individual is heterozygous), as when heterozygosity is large, it's easier to infer IBD status.

Method 4

	Allele frequency			
d (cM)	0.05	0.10	0.15	0.20
0.1	1.50	1.49	1.48	1.48
1	1.49	1.48	1.47	1.46
5	1.48	1.46	1.44	1.44
10	1.47	1.45	1.43	1.42
(method 3)	1.46	1.45	1.44	1.43

Broman (2001) doi:10.1002/gepi.2

Looking more closely at method 4, we note that its performance improves when the marker density increases. (d is the distance between markers, in cM.) With very dense markers, it achieves the ideal relative efficiency of 1.5.

Summary

Method	Progr. time	CPU time	Rel. Eff.
1	2 min	1 msec	1.00
2	2 min	1 msec	1.33
3	1 morning	2 msec	1.45
4	1 afternoon	2.5 sec	1.46

Here's a different sort of summary of the results that includes the time to implement the methods in software, the time to actually obtain the estimates, and the average relative efficiency.

Methods 3 and 4 are considerably more work to program. That you can implement method 3 in the morning and method 4 in the afternoon is perhaps overly optimistic; especially the latter. But it's maybe on that order. And while method 3 is slower than methods 1 and 2, it's still pretty close to instantaneous. Method 4, on the other hand, is like 1000 times slower.

This is a commonly observed trade-off in data science: more complex methods can give improved estimates, but they take much longer to implement as well as execute. In this particular context, method 2 is probably sufficient, but method 3 gives sufficiently improved estimates that it's probably worth it. Method 4 is clearly not worth it and was really just considered here because it's sort of cute.

(It's good to remind you again here that when you think of programming time, it is good to think about the different contributions to that: formulating the problem, actually writing the software, then debugging the software, and finally running it.)

One last thing

- Turns out, I made a mistake in Method 3
  Mary Sara McPeek (U Chicago) spotted it
- Fixed problem, re-ran simulations, and...

  the correct MLE was worse than my mistaken estimate

A funny thing: it turns out that I messed up my formulation of method 4, and Mary Sara McPeek spotted it and pointed it out. I re-ran the simulations with the correction, and the true MLE actually performed worse than my mistaken version.