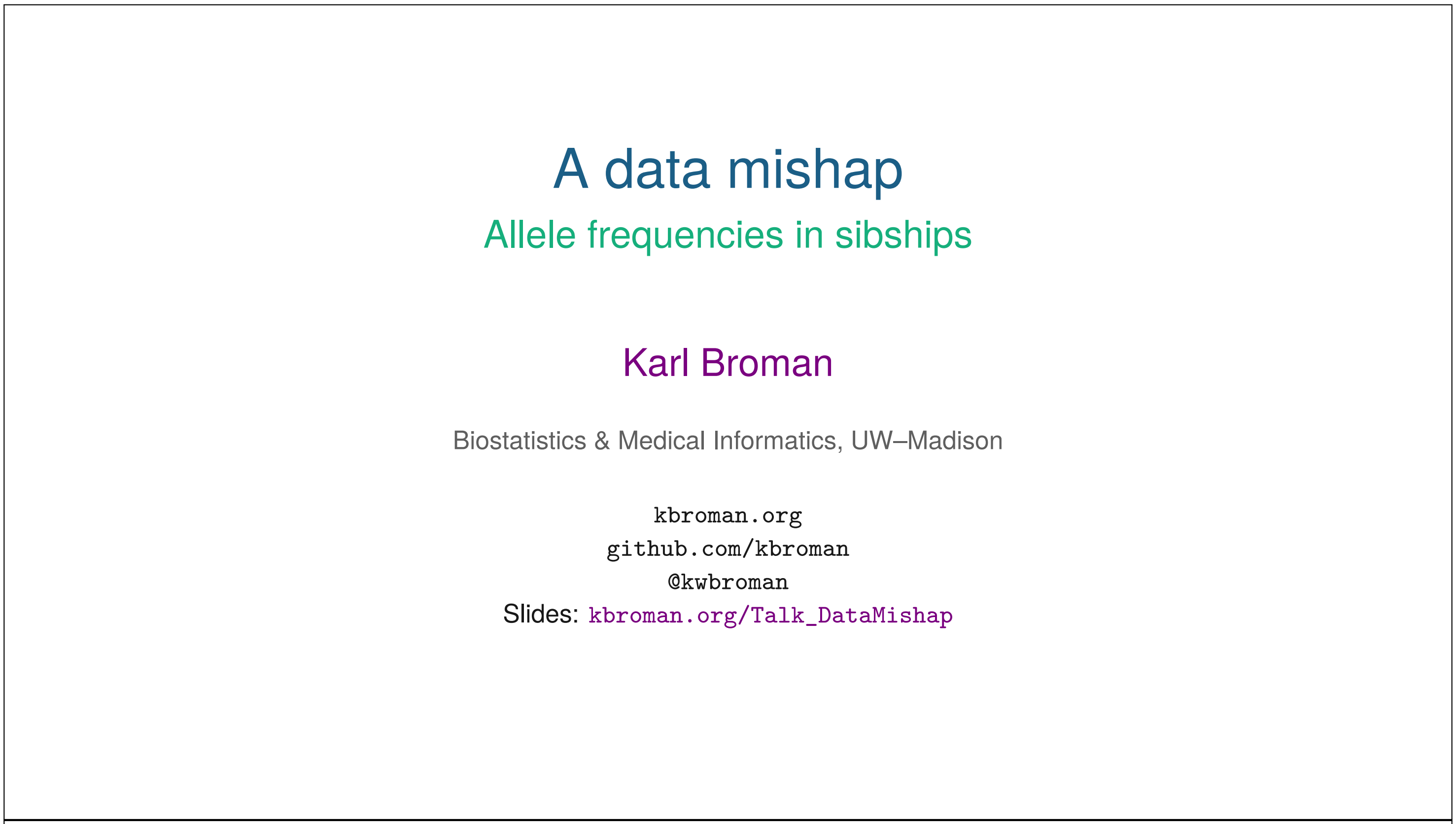


A data mishap

Allele frequencies in sibships

Karl Broman

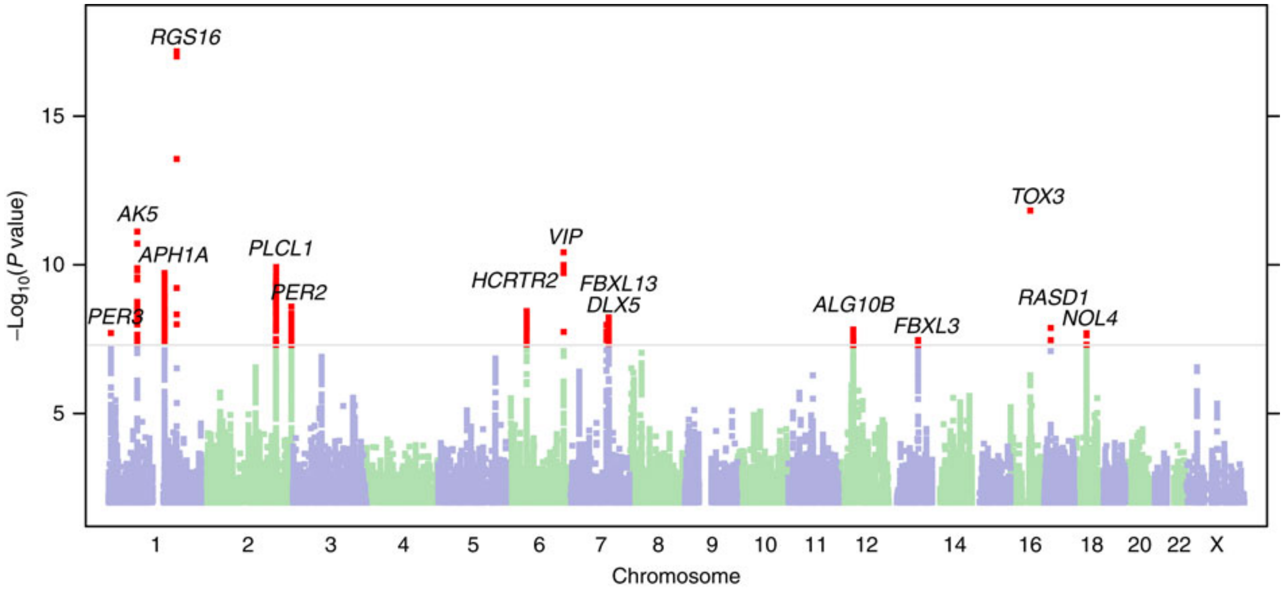
Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman

Slides: kbroman.org/Talk_DataMishap

These are slides for a 5-min talk about a data mishap, for a community night (https://www.littlemissdata.com/fdf/datamishapsnight) organized by Caitlin Hudon (@beeonaposy) and Laura Ellis (@LittleMissData).

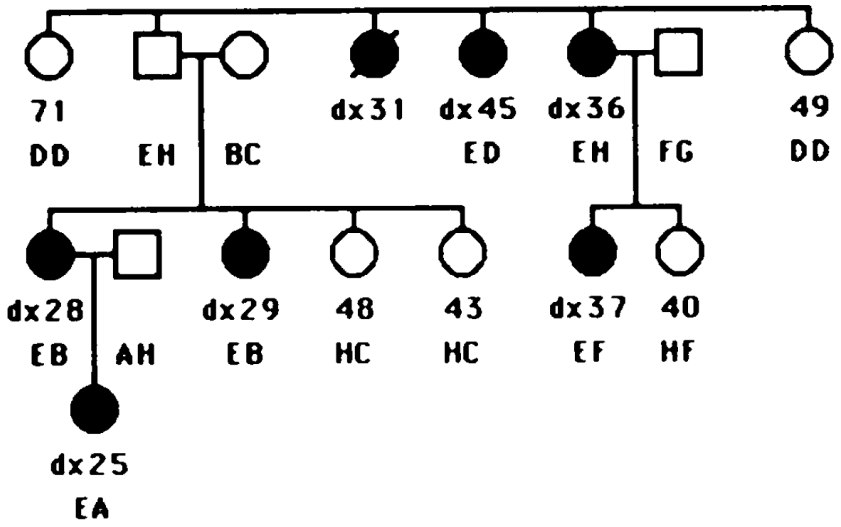
GWAS for “morning person”



Hu et al (2016) doi:10.1038/ncomms10448

Genome-wide association studies (GWAS) have been a revolution in human genetics. This figure is from a study of 23andMe participants who were asked whether they’re a morning person. This binary trait was associated with genotype at markers across the genome, immediately showing genes associated with the trait.

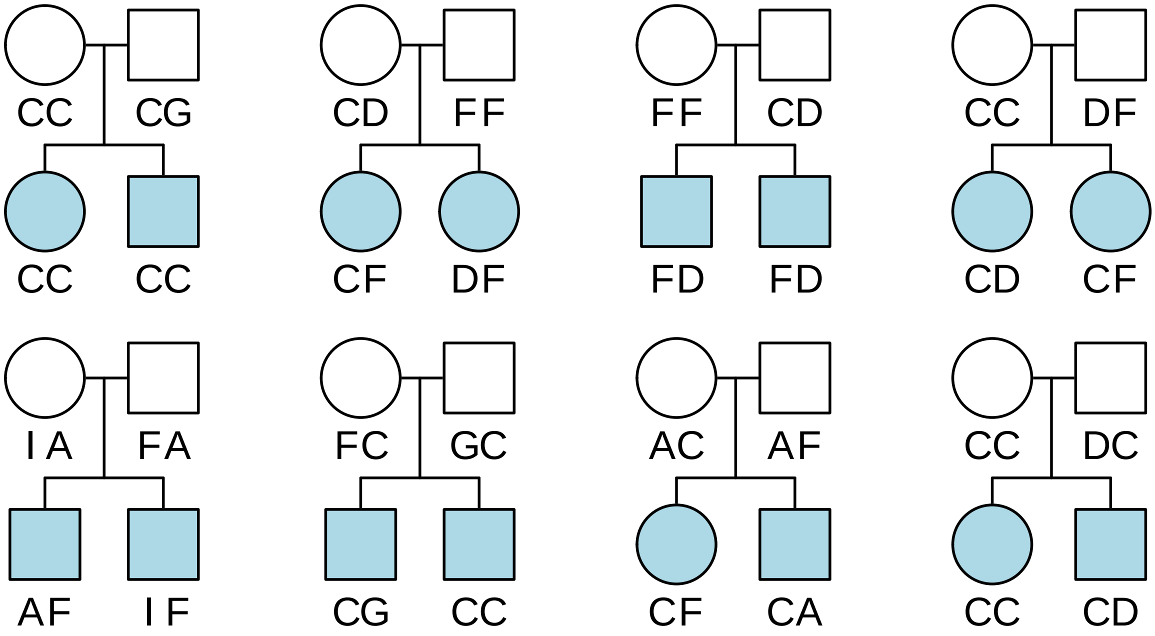
BRCA pedigree



Hall et al (1990) doi:10.1126/science.2270482

Back in the day, gene discovery involved the collection and analysis of large families. This is one of the families from the study that identified the BRCA1 gene. An important insight there was focusing on families with early-onset breast cancer.

Affected sib pairs



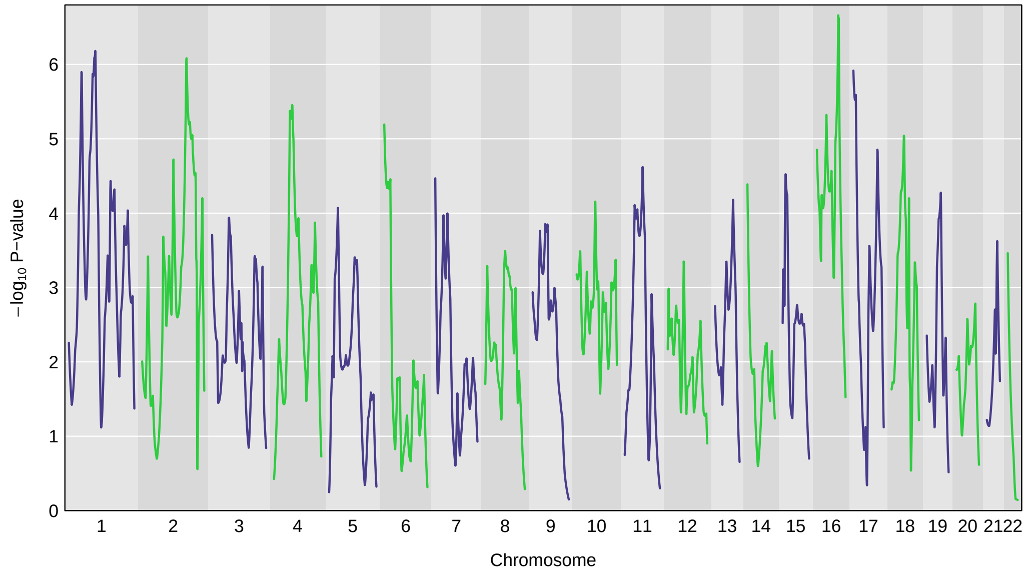
In-between, there was a period where we thought we could find disease genes by gathering a moderate number of affected sibling pairs. You look for regions where affected sibpairs had more similar genotypes than you would expect by chance.

Marshfield, Wisconsin

Marshfield

In 1998 I was a postdoc in a genetics lab in Marshfield, WI (2 1/2 hours drive north of Madison). My advisor hooked me up with an affected sibpair study on prostate cancer. I did the initial data cleaning and a basic analysis, hoping to wow the famous people involved with my prowess.

Prostate cancer genome scan



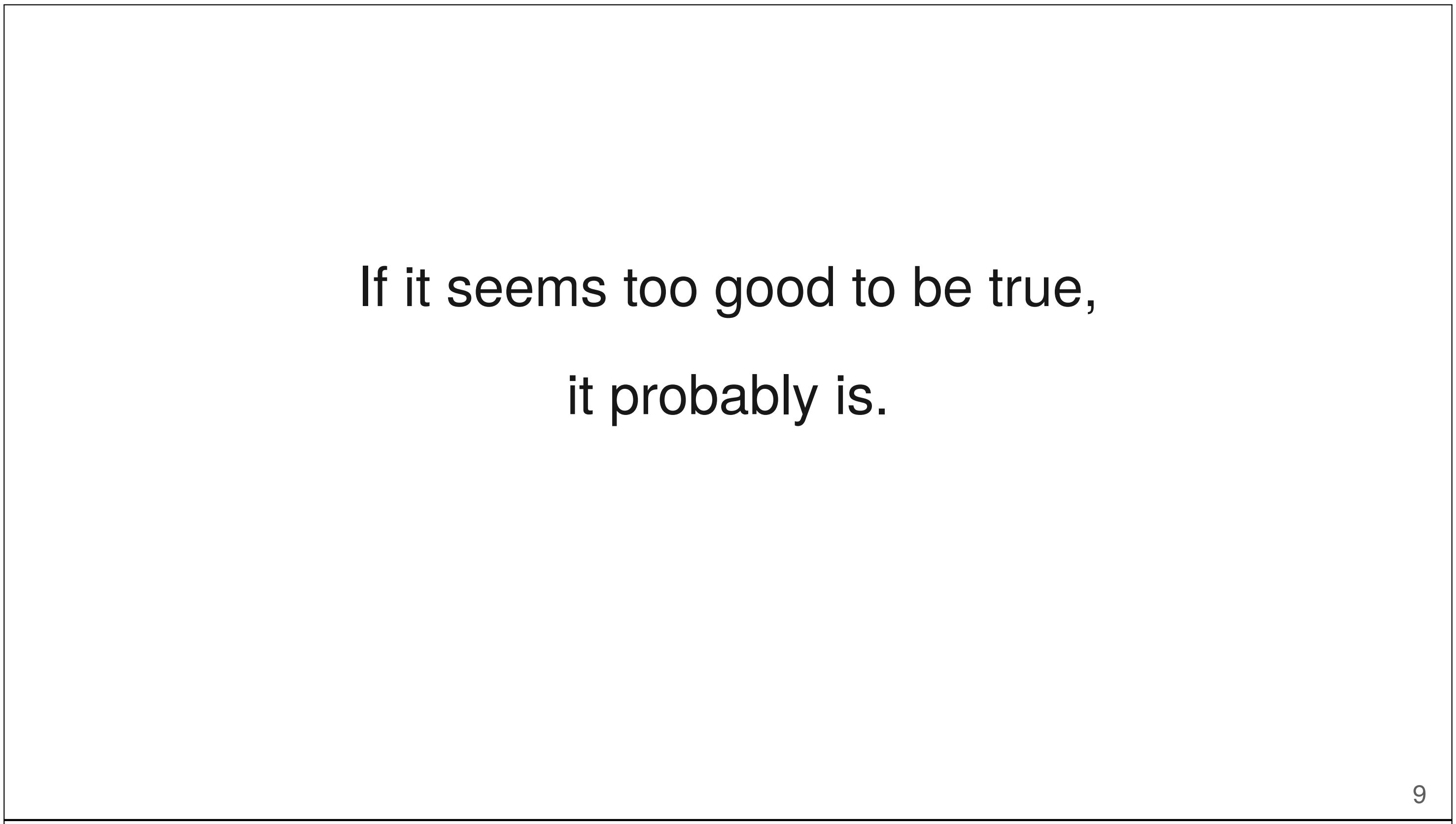
This plot (of $-\log_{10}$ p-values) is an approximation of my initial results. We're looking for values around 3, so these were super exciting to me: much higher association than I would have expected, and on many more chromosomes than I would have expected.

so happy

It was so awesome.

bit.ly/faxpic

I immediately faxed my results off to my collaborators. (That’s how we shared results with each other in 1998.)



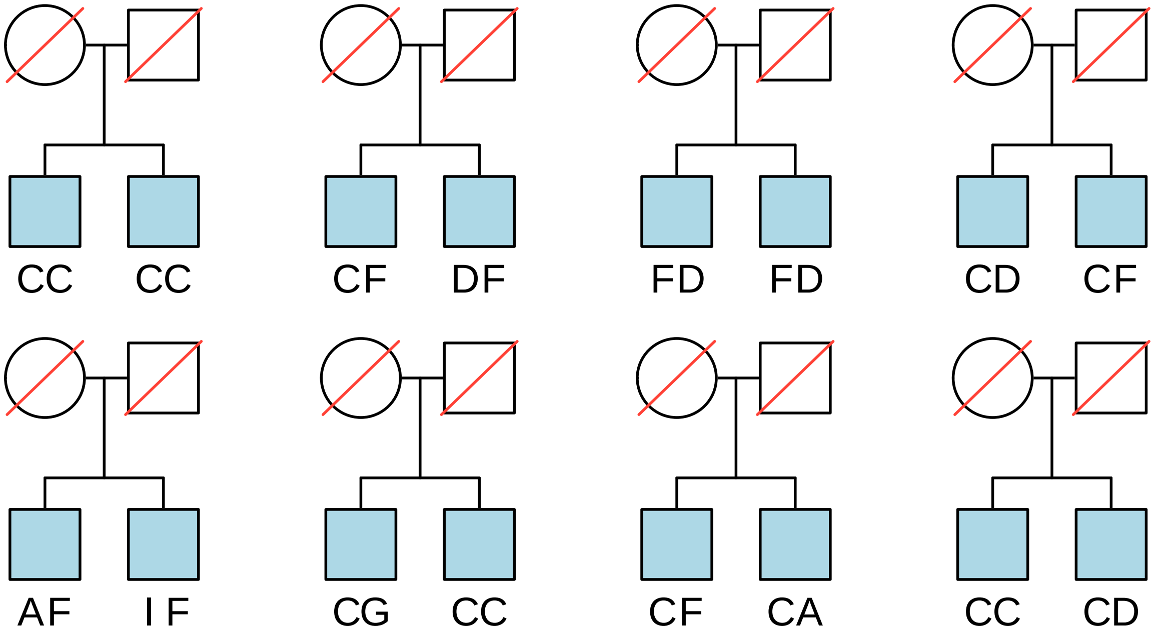
If it seems too good to be true,

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.

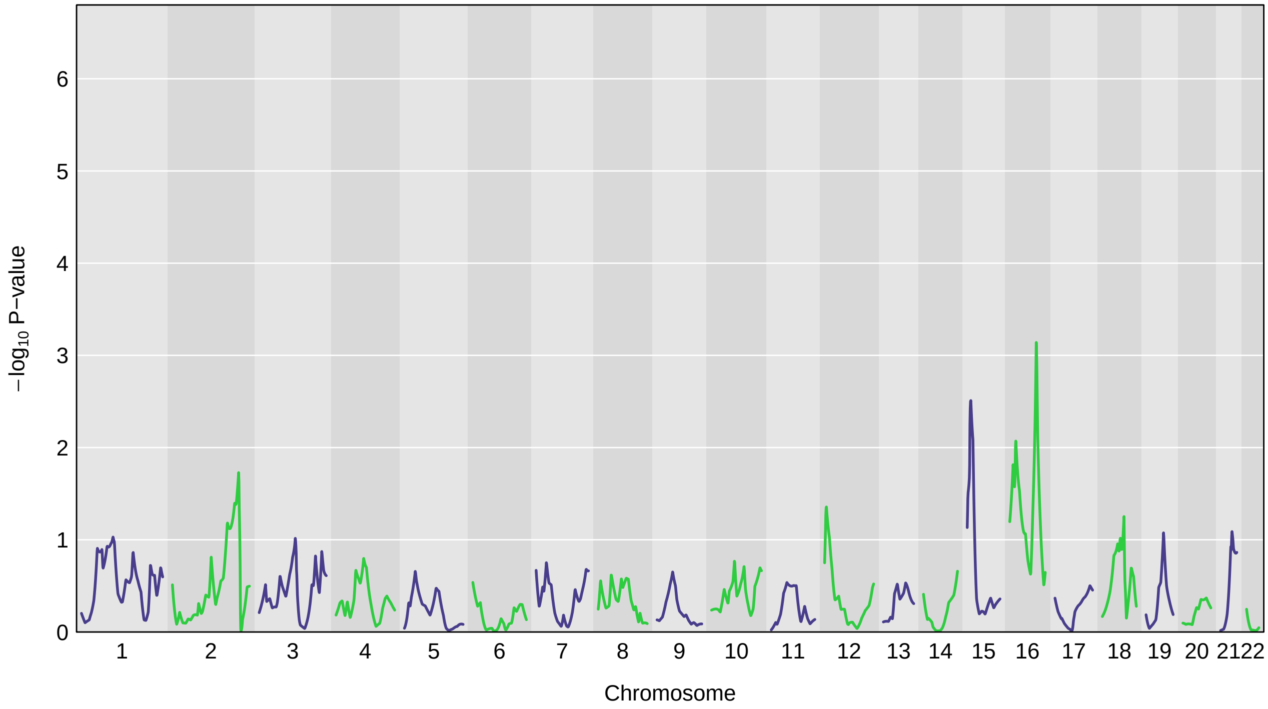
Prostate cancer pairs



In this prostate cancer study, the affected sibpairs are all old, and there’s essentially no data on the parents. In this case, our method for determining sharing is particularly sensitive to the allele frequencies.

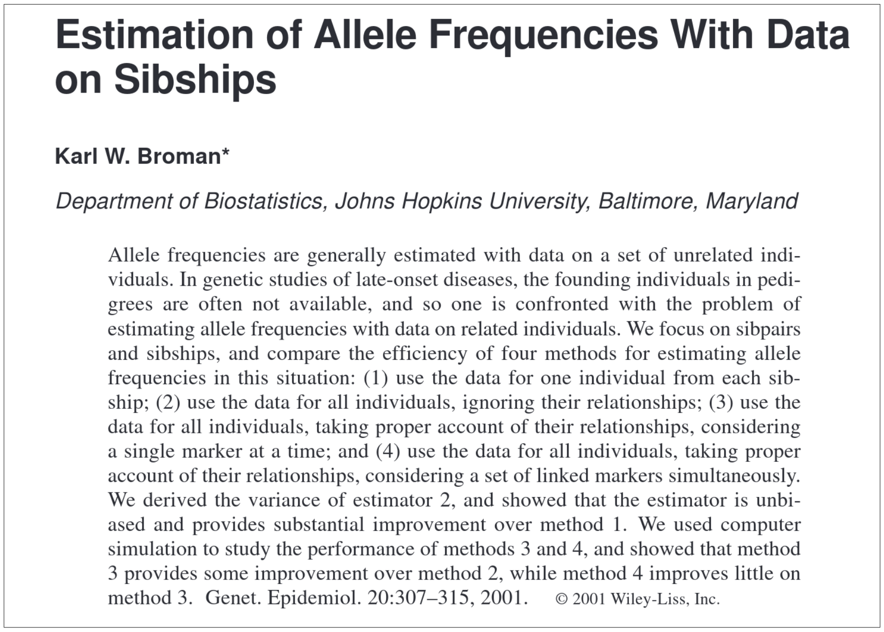
It’s not obvious how to estimate the allele frequencies, but also the simple approach I took had a bug that really through things off.

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.



Estimation of Allele Frequencies With Data on Sibships

Karl W. Broman*

Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland

Allele frequencies are generally estimated with data on a set of unrelated individuals. In genetic studies of late-onset diseases, the founding individuals in pedigrees are often not available, and so one is confronted with the problem of estimating allele frequencies with data on related individuals. We focus on sibpairs and sibships, and compare the efficiency of four methods for estimating allele frequencies in this situation: (1) use the data for one individual from each sibship; (2) use the data for all individuals, ignoring their relationships; (3) use the data for all individuals, taking proper account of their relationships, considering a single marker at a time; and (4) use the data for all individuals, taking proper account of their relationships, considering a set of linked markers simultaneously. We derived the variance of estimator 2, and showed that the estimator is unbiased and provides substantial improvement over method 1. We used computer simulation to study the performance of methods 3 and 4, and showed that method 3 provides some improvement over method 2, while method 4 improves little on method 3. Genet. Epidemiol. 20:307–315, 2001. © 2001 Wiley-Liss, Inc.

doi:10.1002/gepi.2

My collaborators were pretty nice about it. And I ended up writing a paper about the problem. That paper also had a major flaw, which is also interesting and instructive, but that's another story.

Slides: kbroman.org/Talk_DataMishap

kbroman.org

github.com/kbroman

@kwbroman