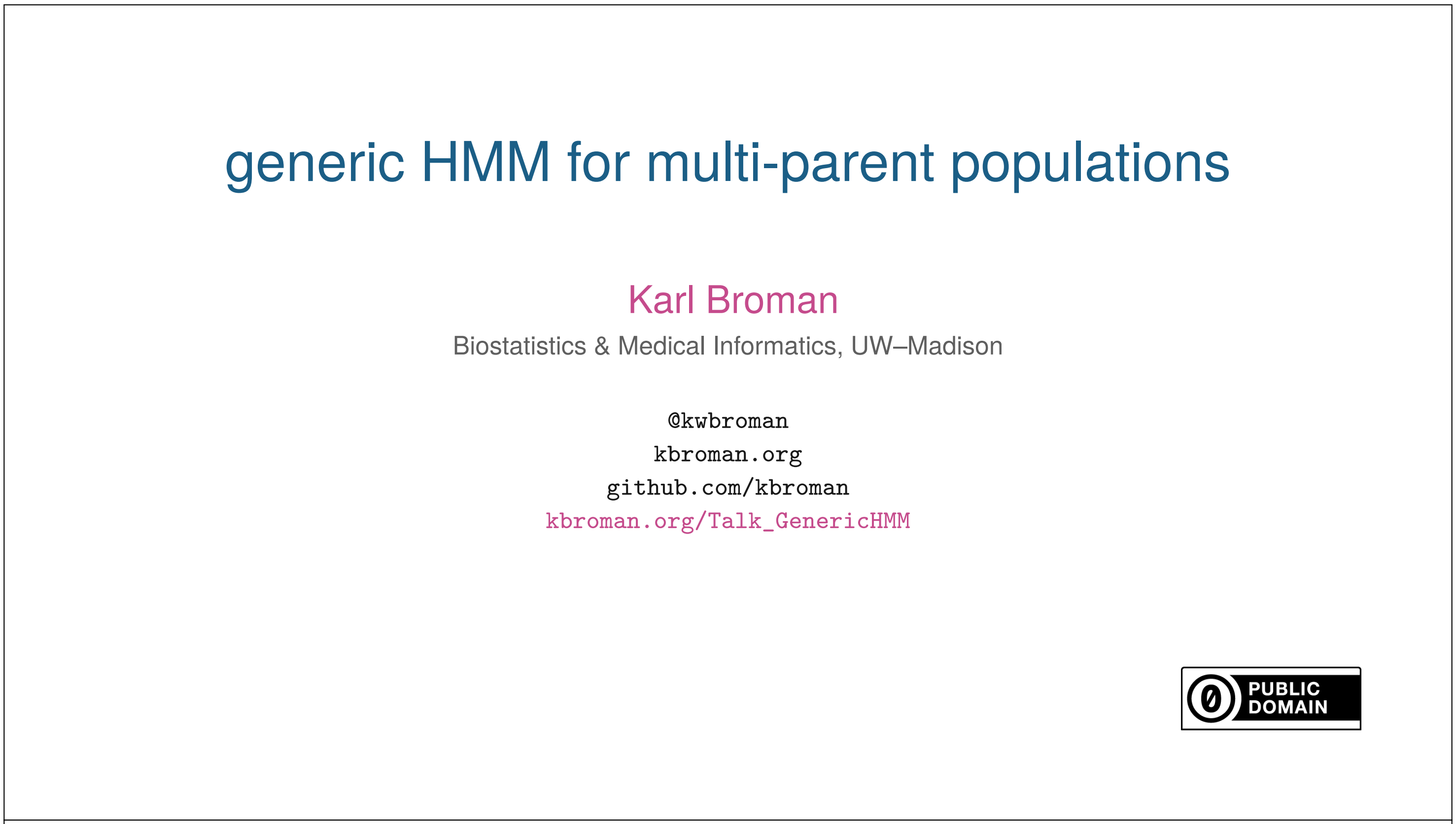


generic HMM for multi-parent populations

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

@kwbroman
kbroman.org
github.com/kbroman
kbroman.org/Talk_GenericHMM

These are slides for a talk for the CTC (www.complextrait.org/ctc2021/) on 1 Sept 2021.

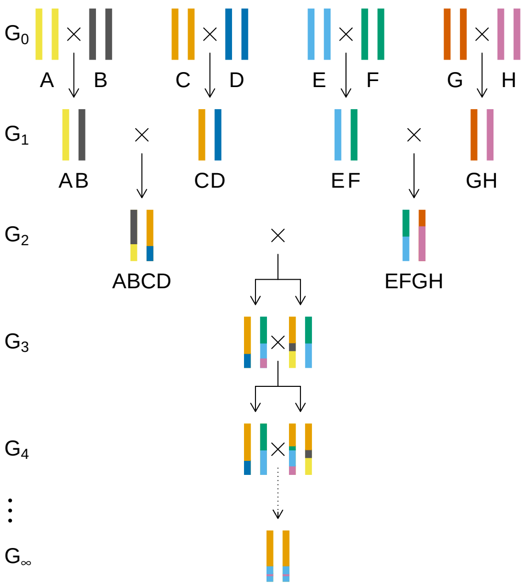
Slides: kbroman.org/Talk_GenericHMM/generic_hmm.pdf

Slides with notes: kbroman.org/Talk_GenericHMM/generic_hmm_notes.pdf

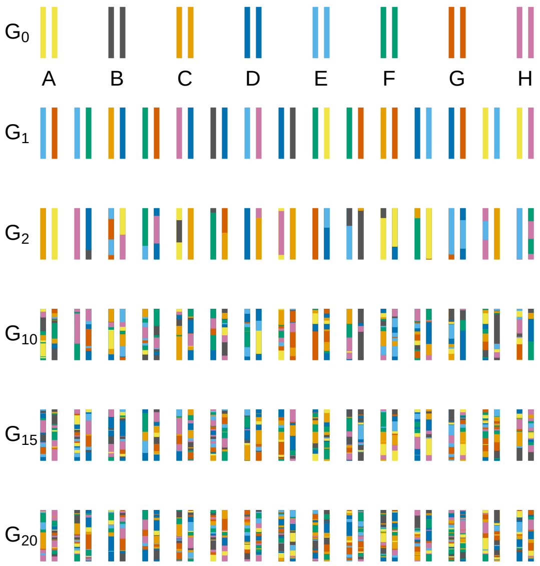
Source: github.com/kbroman/Talk_GenericHMM

Related paper on bioRxiv: doi.org/gswx

Recombinant Inbred Lines



Advanced Intercross Population

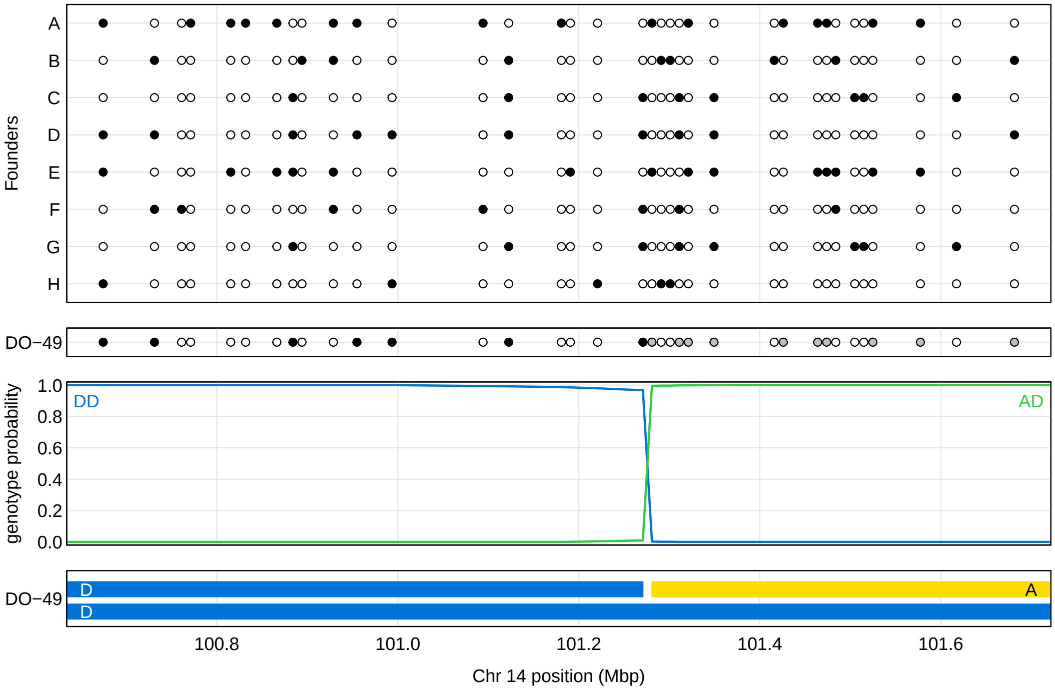


Multi-parent populations are experimental crosses starting from multiple inbred founder lines.

Major examples include the Collaborative Cross, a set of 8-way recombinant inbred lines in mouse, and Hetereogeneous Stock, which have been developed in both mice and rats and are advanced intercross populations derived from 8 founders. The Diversity Outbred mouse population is similar to HS. In plants, multi-parent recombinant inbred lines are called MAGIC lines (for multiparent advanced generation inter-cross).

The offspring chromosomes will be mosaics of the founder chromosomes. Multi-parent populations can be homozygous (like RIL) or heterozygous (like HS). The number of founders need not be 8.

Genome reconstruction

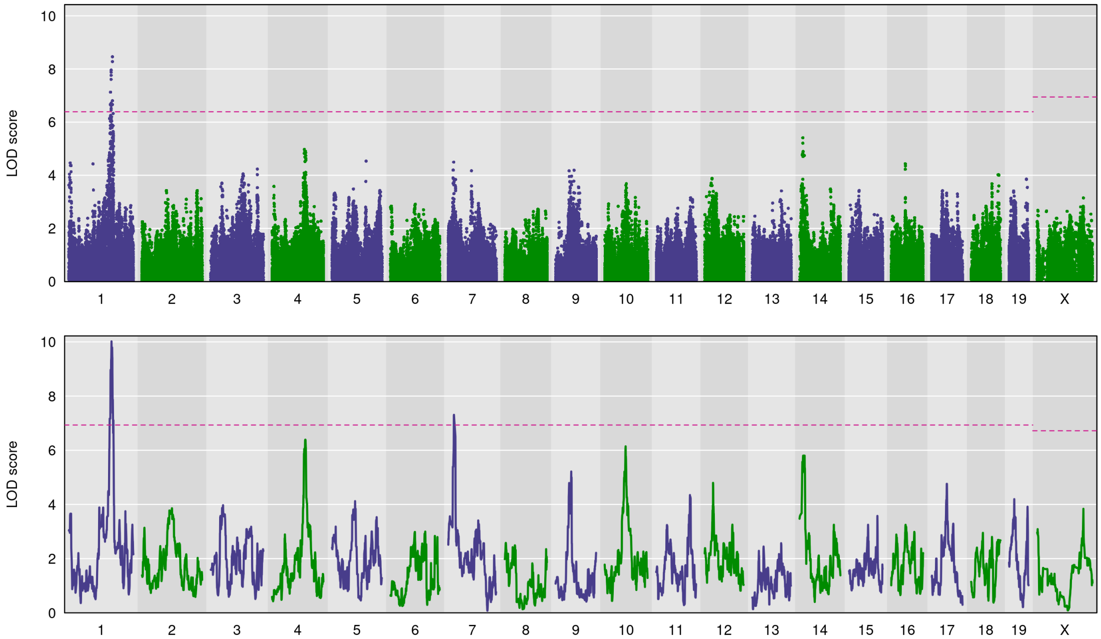


A key step in the analysis of multi-parent populations is genome reconstruction: using dense SNP genotypes in the founders and MPP offspring to infer the haplotypes across the genome.

Here we consider a 1 Mbp region on chromosome 14 in a single Diversity Outbred Mouse. Open and closed circles indicate AA and BB genotypes at SNPs. Gray circles indicate AB heterozygous genotypes. Using the SNP data along the chromosome, we can calculate the probability of each possible genotype at each position.

For this mouse, the left half of the interval looks to be homozygous DD, while the right half looks to be heterozygous AD.

QTL genome scan



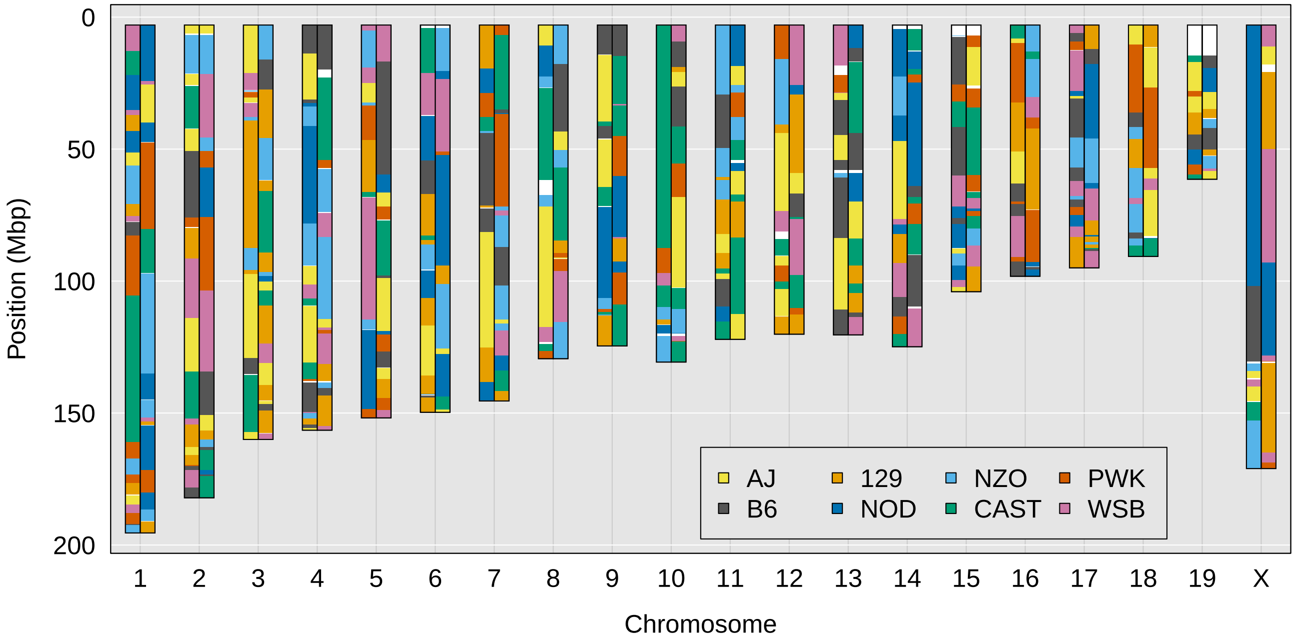
One could skip the whole genome reconstruction and just do QTL analysis at the SNPs, as is done in GWAS. If the genotyped SNPs include individual causal polymorphisms, this could be best.

But if there are multiple causal polymorphisms in a region QTL analysis with the inferred haplotypes may be more powerful. Moreover, if the founder strains have been sequenced, you can use the reconstructed genomes to get inferred genotypes at all polymorphisms in the founders. (Similar approaches were used in human GWAS, based on HapMap SNPs.)

Here, the single-SNP analysis shows significant evidence for a single QTL on chromosome 1. The haplotype analysis indicates evidence for a second QTL on chromosome 4.

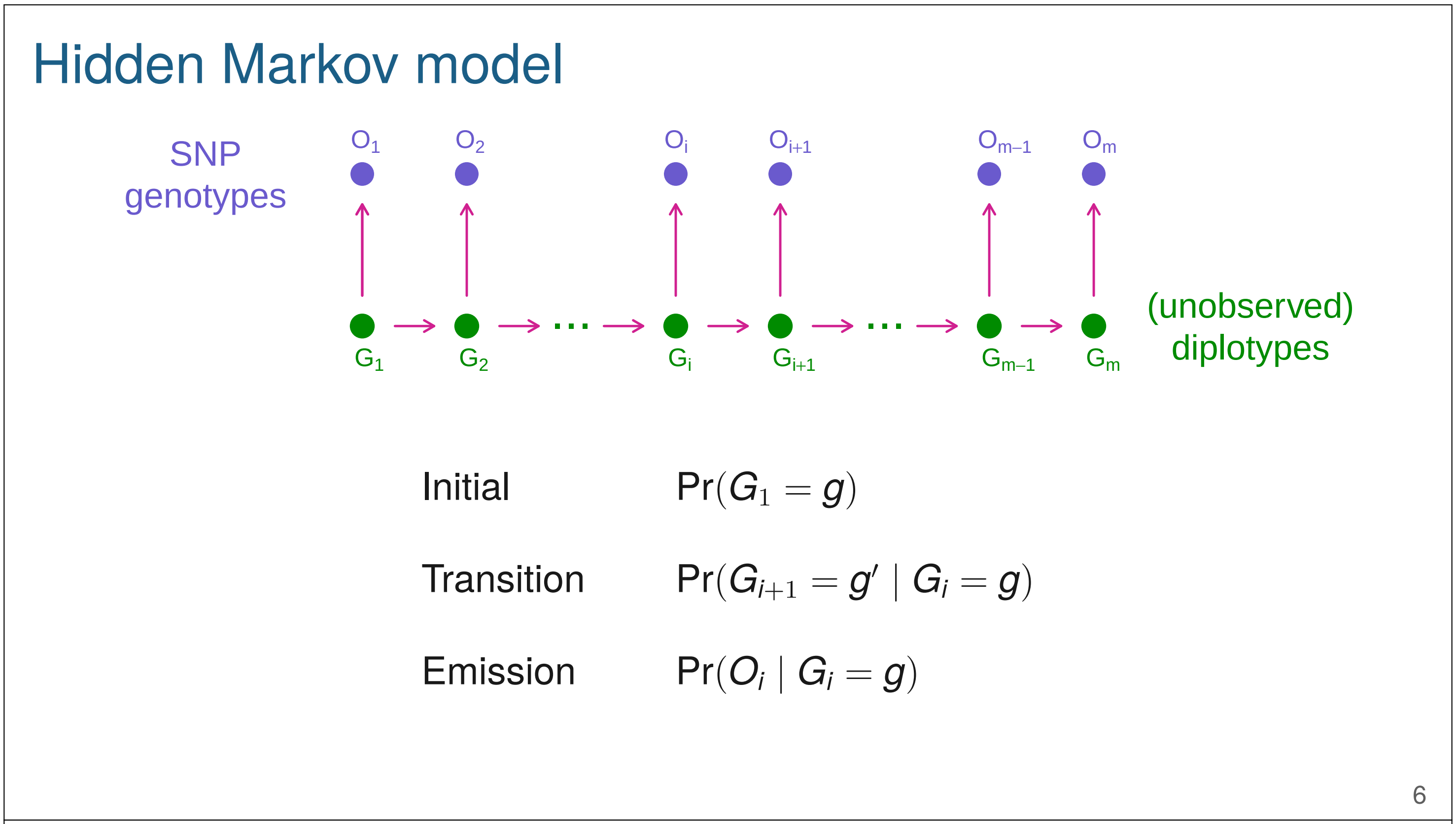
Beyond QTL mapping, genome reconstructions are useful in data diagnostics. For example, the estimated number of crossovers is useful when assessing sample quality.

DO genome



Here is the reconstructed genome of a Diversity Outbred mouse. (The white segments are undetermined.)

Our goal is to figure this out, using SNP genotypes on this mouse plus the 8 founder lines.



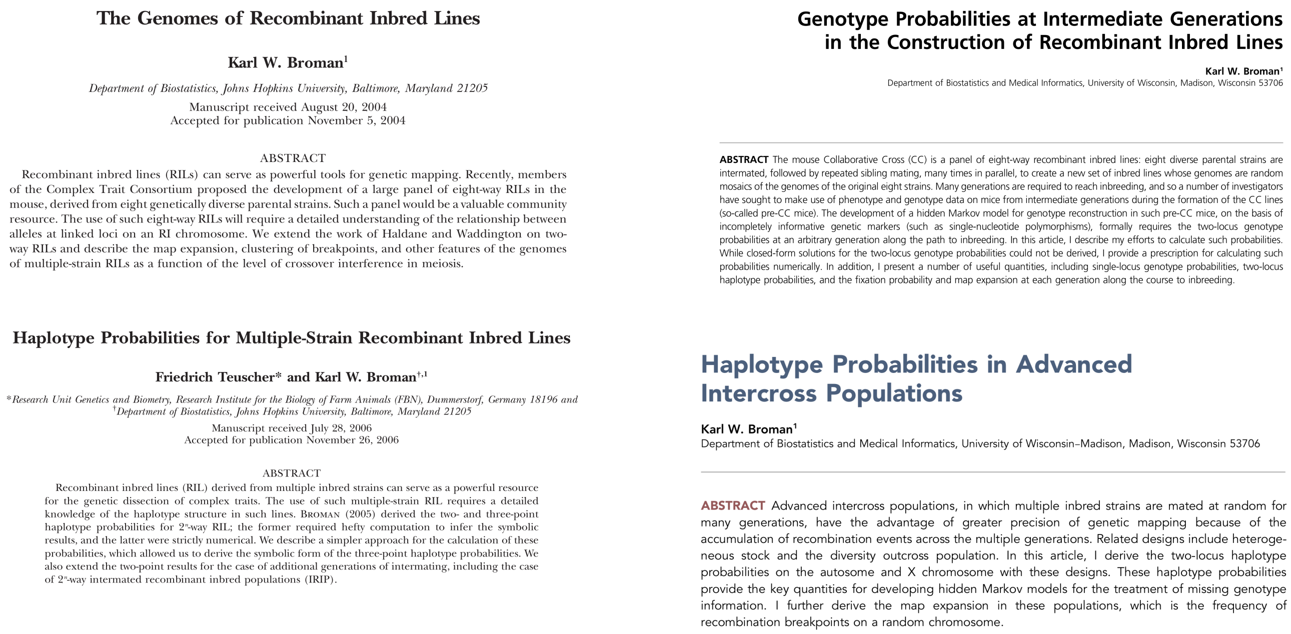
Hidden Markov model

Initial	$\Pr(G_1 = g)$
Transition	$\Pr(G_{i+1} = g' \mid G_i = g)$
Emission	$\Pr(O_i \mid G_i = g)$

The main approach for genome reconstruction is to use a hidden Markov model. The underlying diplotypes we're trying to determine follow a Markov chain $\{G_i\}$, but are unobserved. We observe SNP genotypes $\{O_i\}$, with an assumed conditional independence structure, where given G_i, O_i is conditionally independent of everything else.

Three sets of parameters govern the model: the initial and transition probabilities, which concern the pattern of underlying genotypes on the MPP chromosomes; and the emission probabilities, which relate the underlying genotypes to the observed SNP genotypes and largely concern a model for SNP genotyping errors.

Exact probabilities



The Genomes of Recombinant Inbred Lines

Karl W. Broman[1]

Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205

Manuscript received August 20, 2004
Accepted for publication November 5, 2004

ABSTRACT

Recombinant inbred lines (RILs) can serve as powerful tools for genetic mapping. Recently, members of the Complex Trait Consortium proposed the development of a large panel of eight-way RILs in the mouse, derived from eight genetically diverse parental strains. Such a panel would be a valuable community resource. The use of such eight-way RILs will require a detailed understanding of the relationship between alleles at linked loci on an RI chromosome. We extend the work of Haldane and Waddington on two-way RILs and describe the map expansion, clustering of breakpoints, and other features of the genomes of multiple-strain RILs as a function of the level of crossover interference in meiosis.

Haplotype Probabilities for Multiple-Strain Recombinant Inbred Lines

Friedrich Teuscher* and Karl W. Broman[†,1]

**Research Unit Genetics and Biometry, Research Institute for the Biology of Farm Animals (FBN), Dummerstorf, Germany 18196 and* [†]*Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205*

Manuscript received July 28, 2006
Accepted for publication November 26, 2006

ABSTRACT

Recombinant inbred lines (RIL) derived from multiple inbred strains can serve as a powerful resource for the genetic dissection of complex traits. The use of such multiple-strain RIL requires a detailed knowledge of the haplotype structure in such lines. Broman (2005) derived the two- and three-point haplotype probabilities for 2^n-way RIL; the former required hefty computation to infer the symbolic results, and the latter were strictly numerical. We describe a simpler approach for the calculation of these probabilities, which allowed us to derive the symbolic form of the three-point haplotype probabilities. We also extend the two-point results for the case of additional generations of intermating, including the case of 2^n-way intermated recombinant inbred populations (IRIP).

Genotype Probabilities at Intermediate Generations in the Construction of Recombinant Inbred Lines

Karl W. Broman[1]

Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53706

ABSTRACT The mouse Collaborative Cross (CC) is a panel of eight-way recombinant inbred lines: eight diverse parental strains are intermated, followed by repeated sibling mating, many times in parallel, to create a new set of inbred lines whose genomes are random mosaics of the genomes of the original eight strains. Many generations are required to reach inbreeding, and so a number of investigators have sought to make use of phenotype and genotype data on mice from intermediate generations during the formation of the CC lines (so-called pre-CC mice). The development of a hidden Markov model for genotype reconstruction in such pre-CC mice, on the basis of incompletely informative genetic markers (such as single-nucleotide polymorphisms), formally requires the two-locus genotype probabilities at an arbitrary generation along the path to inbreeding. In this article, I describe my efforts to calculate such probabilities. While closed-form solutions for the two-locus genotype probabilities could not be derived, I provide a prescription for calculating such probabilities numerically. In addition, I present a number of useful quantities, including single-locus genotype probabilities, two-locus haplotype probabilities, and the fixation probability and map expansion at each generation along the course to inbreeding.

Haplotype Probabilities in Advanced Intercross Populations

Karl W. Broman[1]

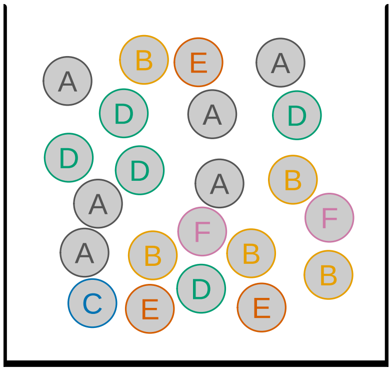
Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, Wisconsin 53706

ABSTRACT Advanced intercross populations, in which multiple inbred strains are mated at random for many generations, have the advantage of greater precision of genetic mapping because of the accumulation of recombination events across the multiple generations. Related designs include heterogeneous stock and the diversity outcross population. In this article, I derive the two-locus haplotype probabilities on the autosome and X chromosome with these designs. These haplotype probabilities provide the key quantities for developing hidden Markov models for the treatment of missing genotype information. I further derive the map expansion in these populations, which is the frequency of recombination breakpoints on a random chromosome.

I've spent quite a lot of time studying the pattern of genotypes on MPP chromosomes, first with a paper on multi-way recombinant inbred lines, but then following up with three further papers considering extra generations of outbreeding, the genotypes at intermediate generations, and the patterns in advanced intercross populations such as Diversity Outbred mice.

The mathematics is interesting but tedious. And is it necessary? It would be nice to have a generic approach that could be used generally.

Generic model



k founders in proportions $\{\alpha_i\}$

n generations of random mating

Random chromosome:

$$\pi_i = \alpha_i$$

$$t_{ij} = \alpha_j \left[1 - (1-r)^n\right] \quad \text{when } i \neq j$$

Map expansion:

$$n(1 - \textstyle\sum \alpha_i^2)$$

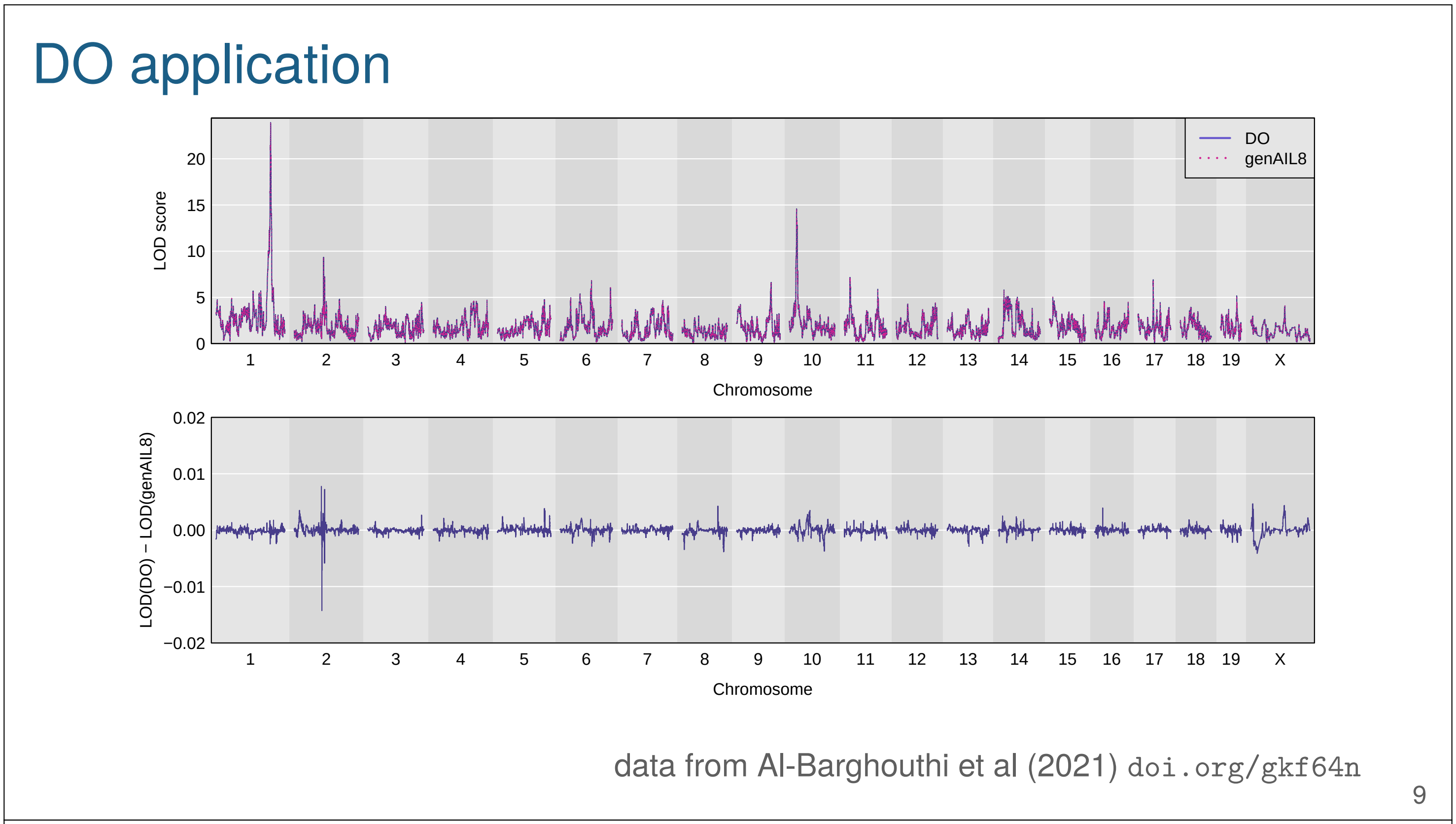
$$= n\left(\tfrac{k-1}{k}\right) \quad \text{if } \alpha_i \equiv 1/k$$

And that is what I propose here. Imagine a population of k founders in known (but not necessarily equal) proportions, and that a multi-parent population is formed by random mating for n discrete generations. In this case, we can calculate the transition probabilities exactly.

We could apply these equations more generally. We need just specify the proportions of the founders (which should be known from the design of the study) and the effective number of generations of random mating. The latter might be calibrated by considering the map expansion (the proportional increase in the number of recombination breakpoints, relative to a single meiosis). This could be approximated by computer simulation.

For a heterozygous population, like HS or the DO, we draw two random chromosomes. For a homozygyous population, like MAGIC lines or the Collaborative Cross, we can pretend that they are doubled haploids, with a single random chromosome like above.

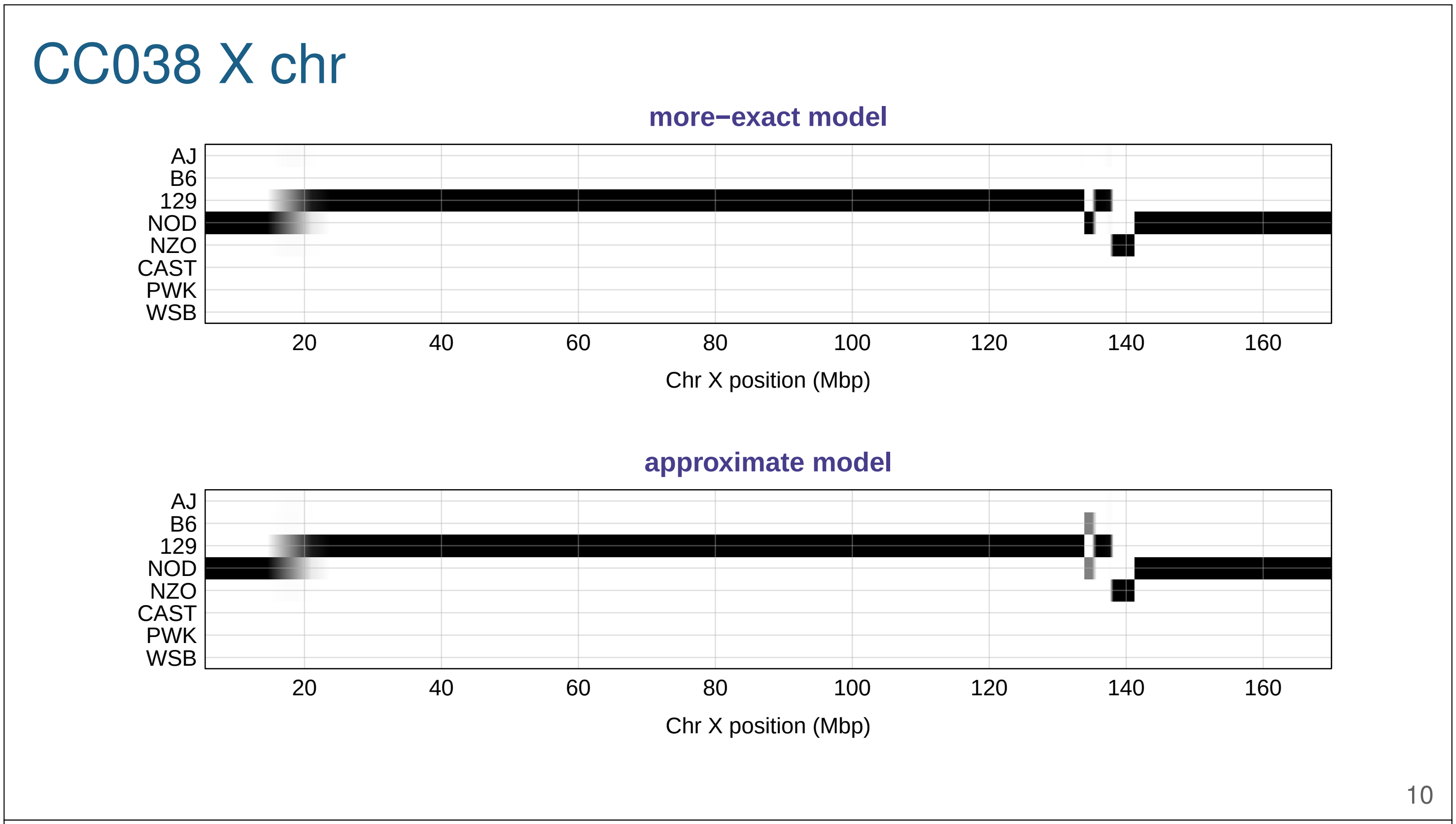
For the X chromosome, we can use the same equations, replacing n with $\left(\frac{2}{3}\right)n$, due to recombination only happening on the X chromosome in females.



DO application

data from Al-Barghouthi et al (2021) `doi.org/gkf64n`

If we apply our approach to data from Diversity Outbred mice, the results with the generic model proposed above are basically identical to the use of the more-exact model. For data from Al-Bargouthi et al (2021), this is the biggest difference seen: the LOD curves are not distinguishable, as the biggest difference is just 0.01.



CC038 X chr

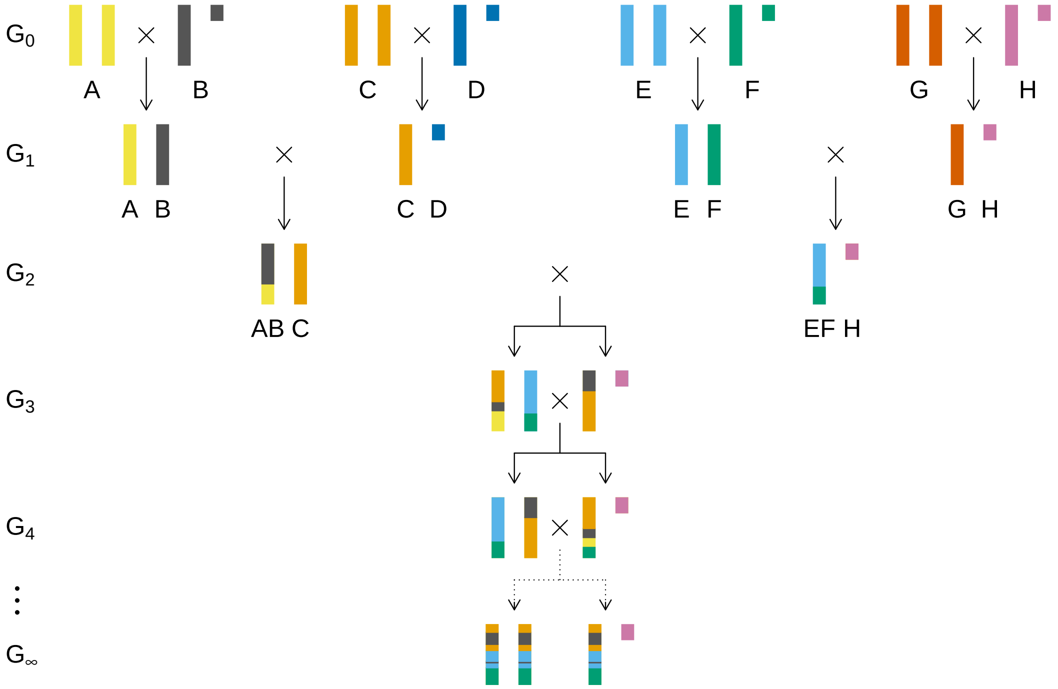
Reconstructions of the genomes of Collaborative Cross lines are identical for autosomes, but there are important differences for the X chromosome.

This slide shows the reconstruction of the X chromosome in Collaborative Cross line CC038, but the exact model (top panel) and by the approximate model (bottom panel).

The analysis differs in that the top model excludes three of the eight founders and weighs one of the other five more highly.

These results differ in a region around 135 Mbp, where in the bottom panel, B6 and NOD are assigned equally probability, as they are identical in the region, but the top panel was able to exclude B6.

X chr in CC



The X chromosome in the Collaborative Cross behaves different than autosomes. We list the crosses female × male; note that the Y chromosome comes from the H strain and the X chromosome comes from the five strains A, B, C, E, and F, with the average proportion from the C strain between twice that of the others.

This can be really useful information (provided that it is correct), particularly as the X chromosome shows reduced polymorphism compared to the autosomes. Many of the CC founders share large stretches of DNA on the X chromosome.

Summary

- Generic model for genome reconstruction in multi-parent populations
- Specific relative proportions of founders + effective number of generations of random mating
- Basic conclusion: HAPPY is effective
- bioRxiv manuscript: `doi.org/gswx`

It's always good to provide a summary.

Slides: kbroman.org/Talk_GenericHMM

bioRxiv manuscript: doi.org/gswx

kbroman.org

github.com/kbroman

@kwbroman

kbroman.org/qtl2

Here's where you can find me and these slides, as well as a preprint giving further details on the work.