# general HMM for multi-parent populations

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

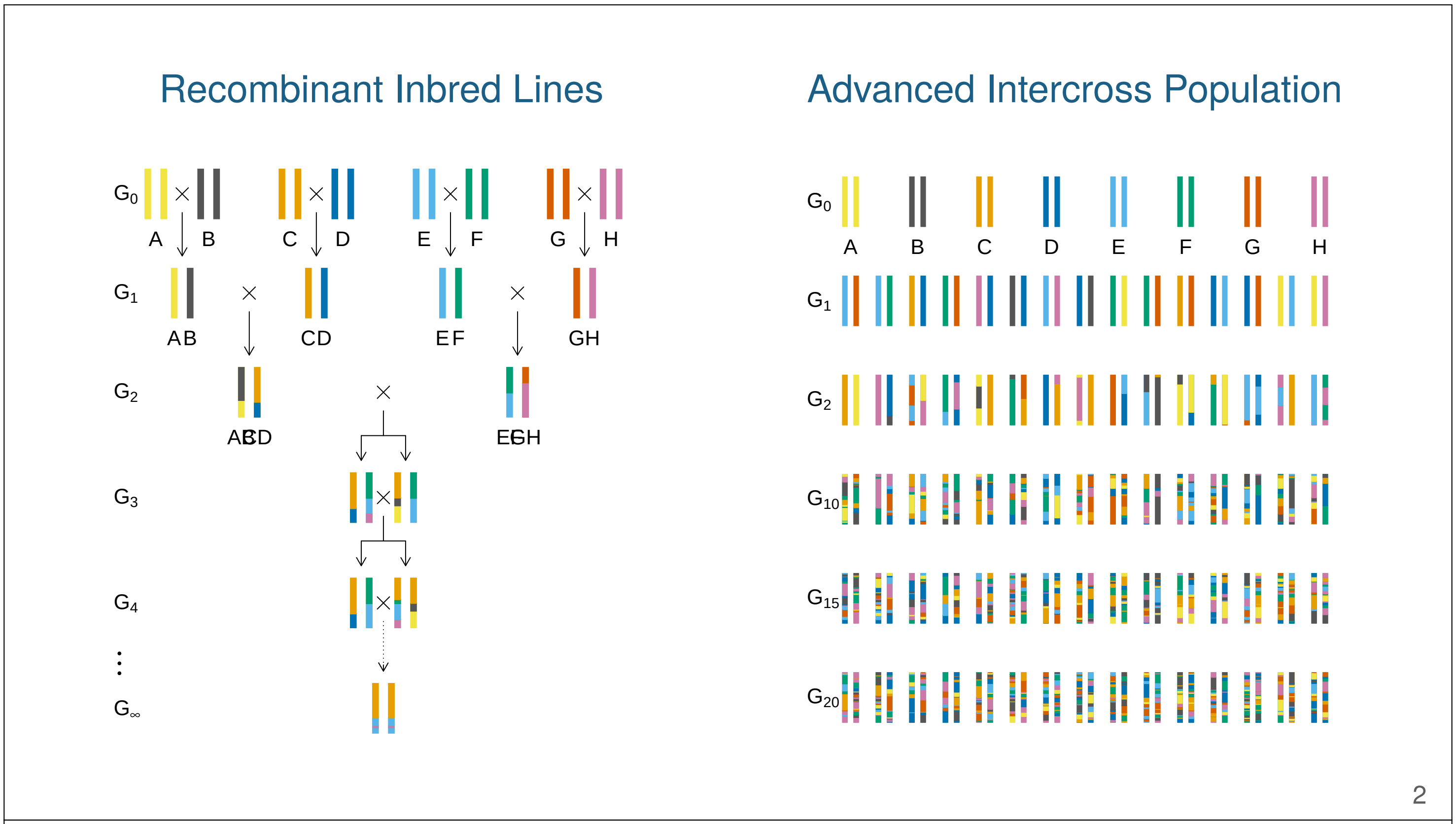
`@kwbroman`
`kbroman.org`
`github.com/kbroman`
`kbroman.org/Talk_GeneralHMM`

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

Slides: `kbroman.org/Talk_GeneralHMM/general_hmm.pdf`

Slides with notes: `kbroman.org/Talk_GeneralHMM/general_hmm_notes.pdf`

Source: `github.com/kbroman/Talk_GeneralHMM`



## Recombinant Inbred Lines

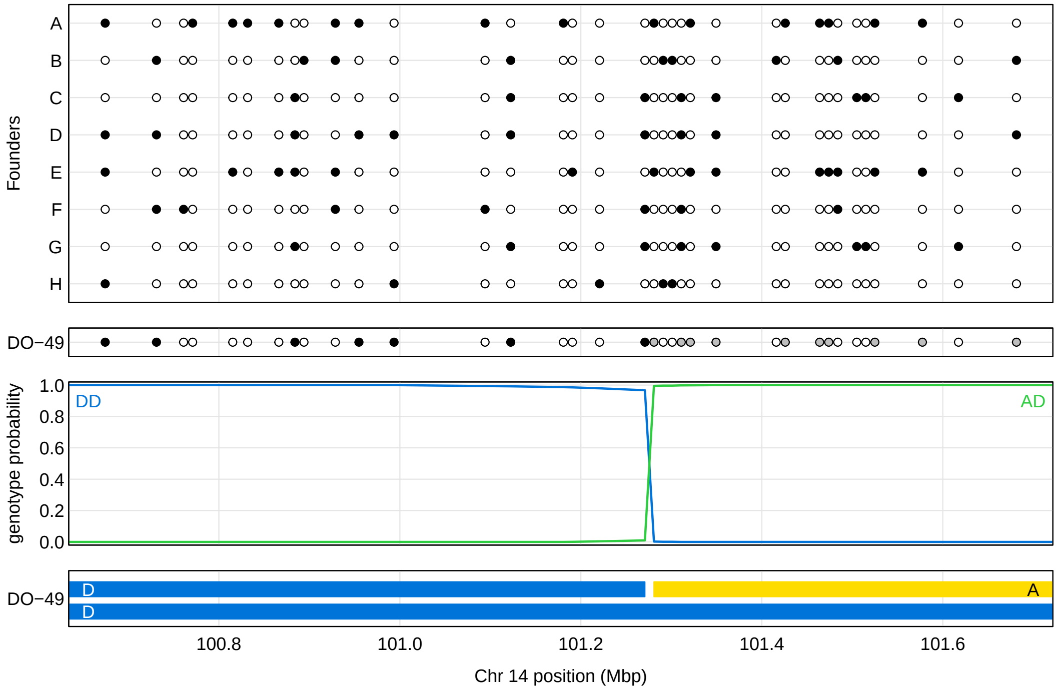
## Advanced Intercross Population

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

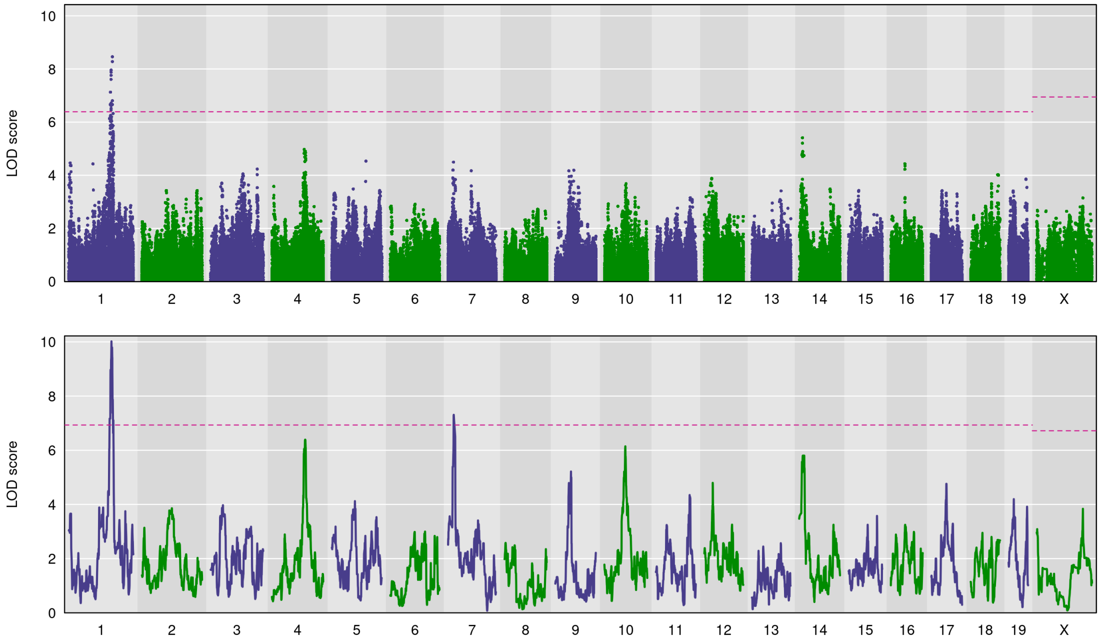


A key step in the analysis of multi-parent populations is genome reconstruction: using dense SNP genotypes in the founders and MPP offspring to infer the haplotypes across the genome.

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



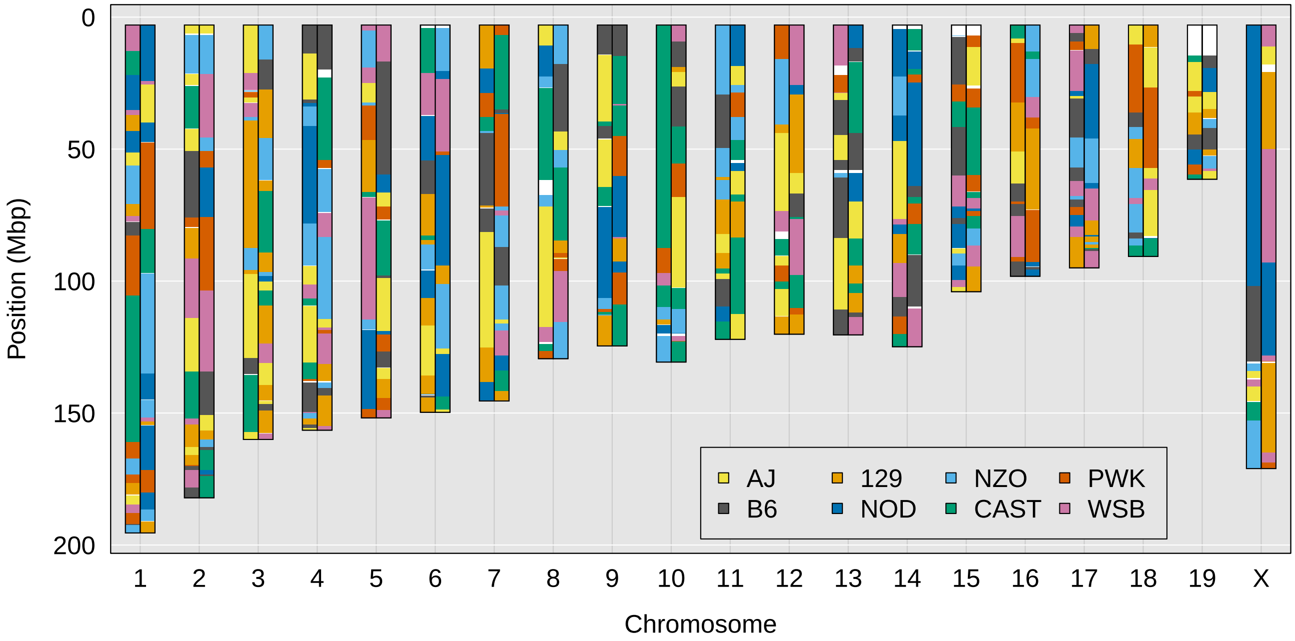
One could skip the whole genome reconstruction and just do QTL analysis at the SNPs, as is done in GWAS. If the genotyped SNPs include individual causal polymorphisms, this could be best.

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

## Exact probabilities

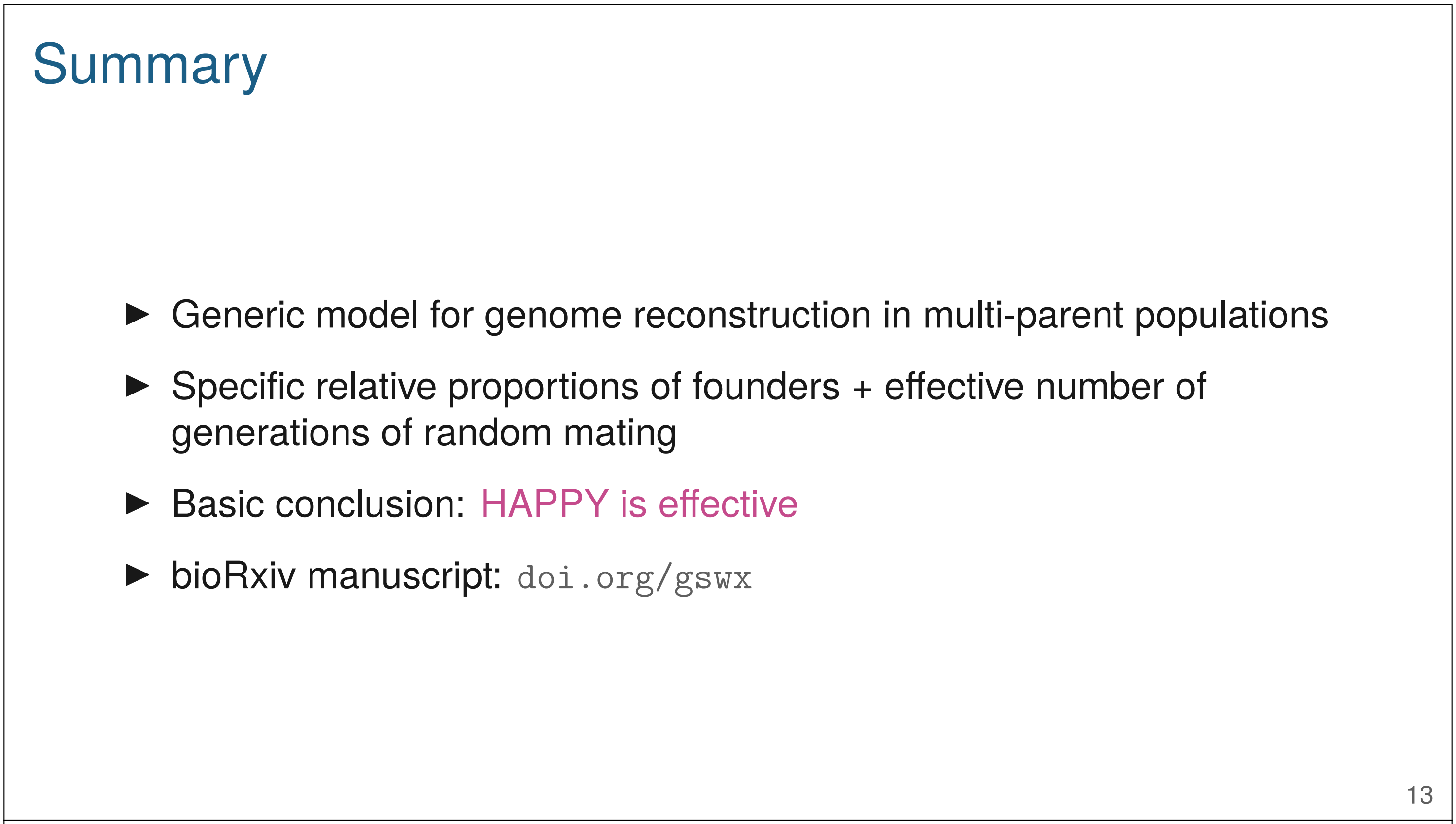
## Generic model

# Genome expansion

# DO application

# X chr in CC

# X chr reconstruction



# 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_GeneralHMM

bioRxiv manuscript: doi.org/gswx

kbroman.org

github.com/kbroman

@kwbroman

kbroman.org/qtl2

Here's where you can find me and these slides.