# Computer simulations

## The genomes of recombinant inbred lines

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

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

This is a second case study, whose central lesson concerns the value of computer simulation, but which also concerns sort of a deep dive into a crazy probability problem that started out as a computer simulation but turned into a combination of simulation, numeric calculations, and analytic derivations (partly through symbolic, computational algebra).

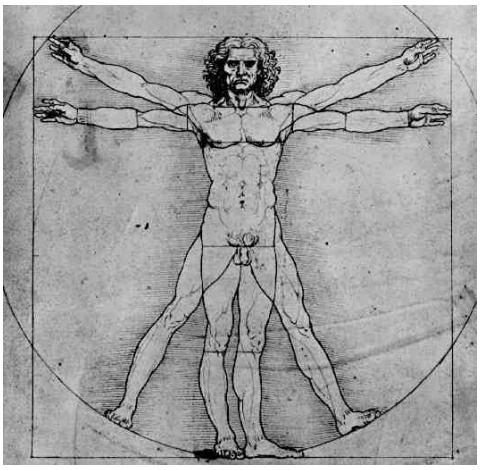

My own research is in genetics, and I focus large on mouse genetics.

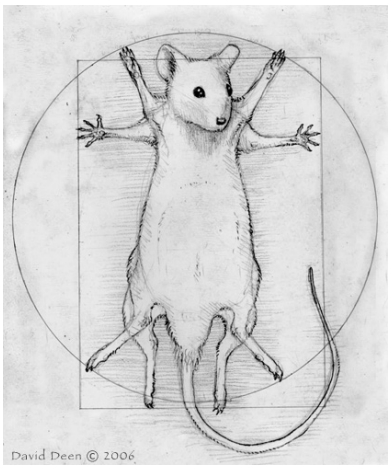
Mice have a lot of advantages for biomedical research. One advantage is that you can generate inbred lines. That is, by repeated brother-sister matings, you can create a line that is entirely homozygous.

The mice in this picture are members of an inbred strain, and so all are genetically identical. And mice like humans have two copies of every chromosome (one from their father and one from their mother), and because of the inbreeding those two copies are identical. So if you take a male and female from this strain and mate them, you get the same thing back.

Thus, the mice you work with today are the same as the mice you'll work two years from now. And you can completely eliminate genetic variation from your studies. Or you can introduce genetic variation in exactly the way you want.

In particular, you can make very large crosses. A particular female mouse can only give so many offspring, but because you have unlimited copies of her, you can expand any cross to be as large as you would like.





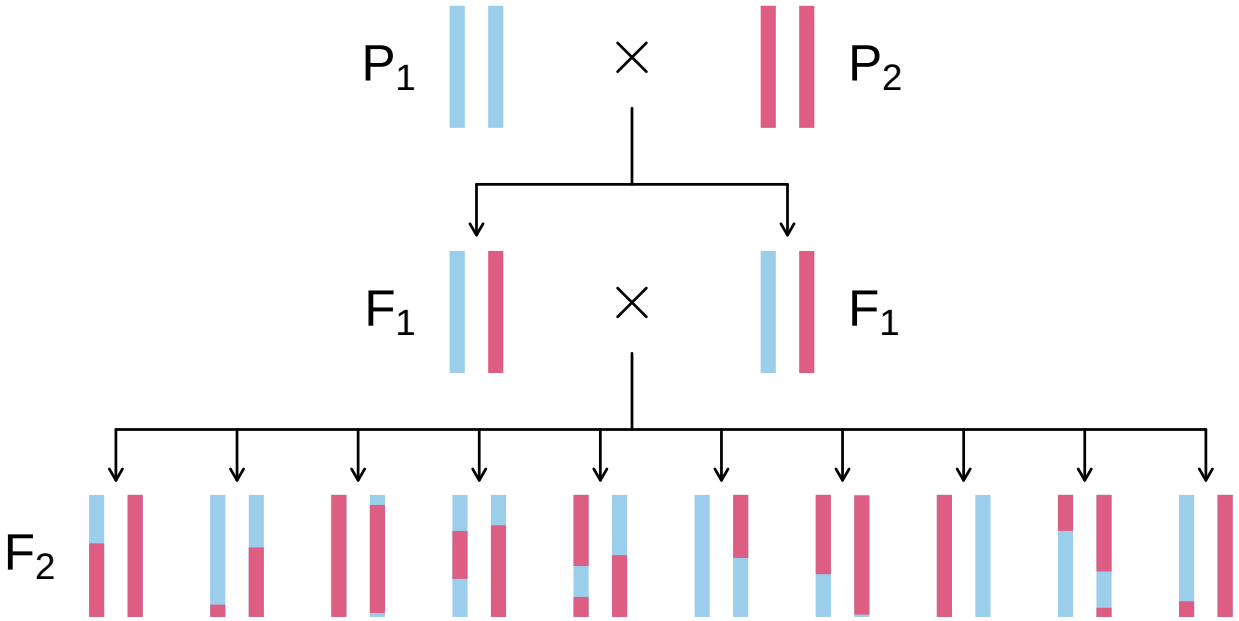

daviddeen.com



Of course, we're not really interested in curing mouse disease. Well, some of my collaborators are interested in mice specifically, but most studies of mice are carried out in order to learn about humans. And mice have a remarkable amount in common with humans.

I like to think of mice as models for humans in three ways. First, the particular genes involved in a trait (say blood pressure) in the mouse could also be involved in the corresponding human trait. Second, the genetic architecture underlying a trait in the mouse (and the pathways and such) can tell us something about the genetic architecture of human traits. Finally, the methods we develop for gene mapping in mice might carry over to direct human studies.

# Intercross





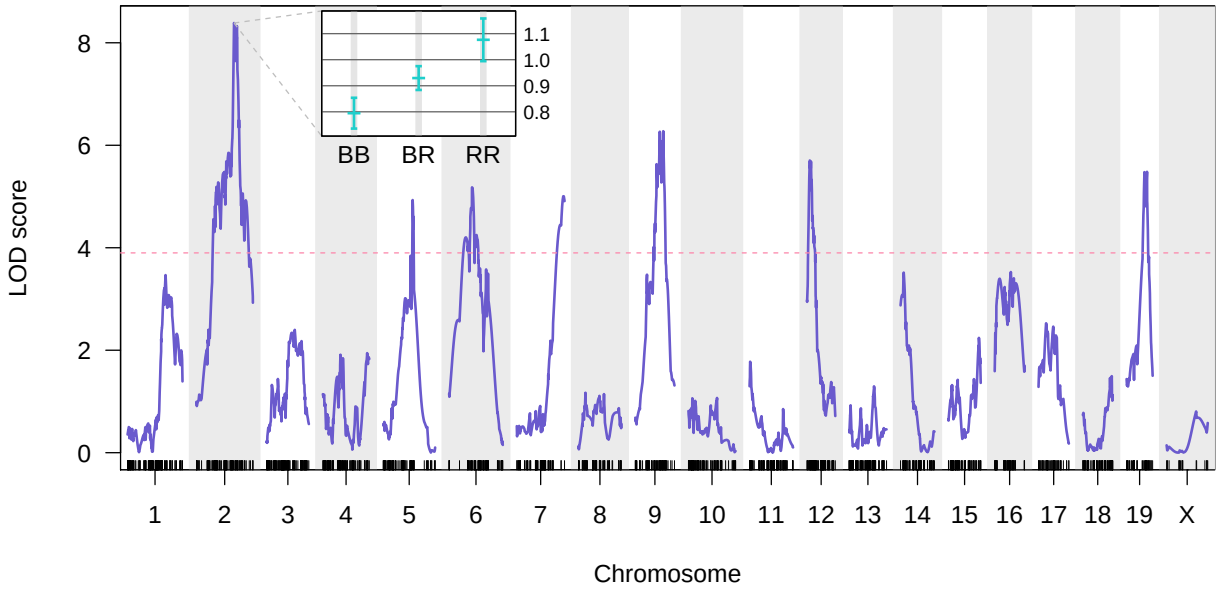

To learn about the genetics of a trait, we can perform one of a series of crosses; the most common is the intercross. You take two strains that differ in the trait of interest. Say the blue strain has low blood pressure and the pink strain has high blood pressure. You cross them and get the $F_1$ hybrid, which will get a copy of each chromosome. Then you cross $F_1$ siblings to get intercross progeny.

The intercross mice will get a chromosome from each parent that could be the blue or pink chromosome intact but will more generally be a mosaic of the two chromosomes as a result of recombination at meiosis (the process of cell division that gives rise to sperm and egg cells).

We would gather a bunch of intercross mice, measure the blood pressure of each (which could be the hardest part of this project), and then measure genotype along the chromosomes. We would then look for regions of the genome where genotype (whether a mouse was homoyzogous blue, heterozygous, or homozygous pink) is associated with the phenotype (blood pressure).

Regions of the genome that affect a quantitative trait like blood pressure are called "quantitative trait loci" (QTL).

# QTL mapping



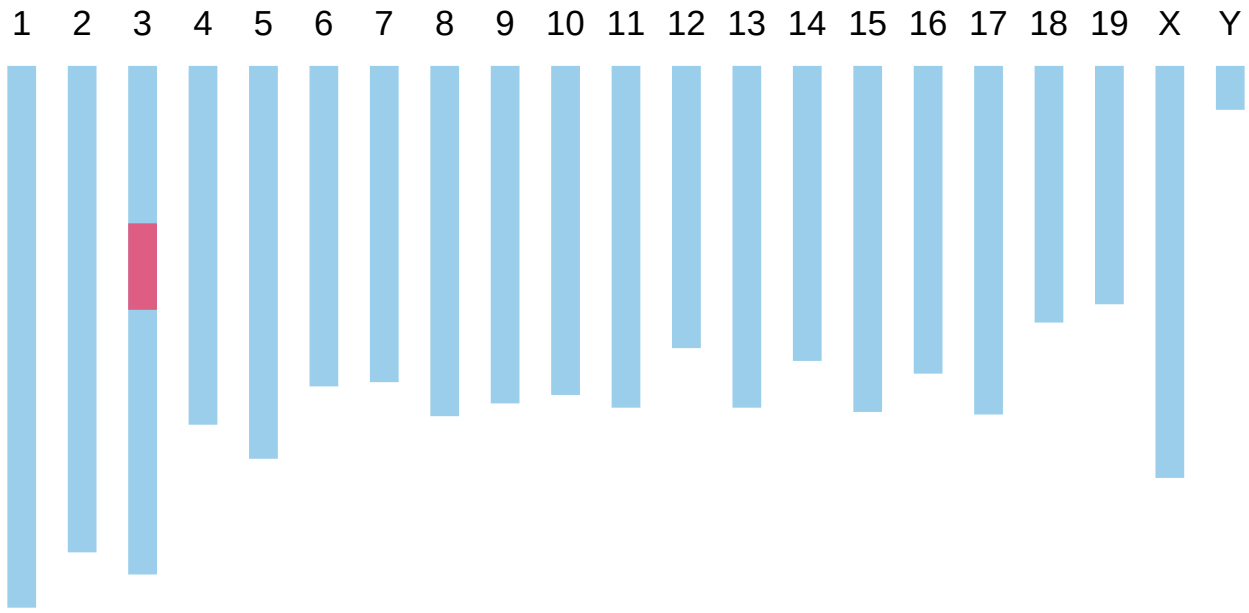



So we would then scan along the genome, at each position measuring the association between genotype and phenotype and plotting some test statistic.

Effectively, we perform an ANOVA at each genomic position. (For historical reasons, we generally plot a different test statistic, the LOD score, which is a $\log_{10}$ likelihood ratio comparing the hypothesis of genotype/phenotype association to the null hypothesis of no association.)

Where the test statistic is large, there is some association between genotype and phenotype. So there will be a difference among the average trait values for the different genotypes, as seen here on chromosome 2, which is clearly not blood pressure (but rather, I think, log insulin level).

# Congenic line





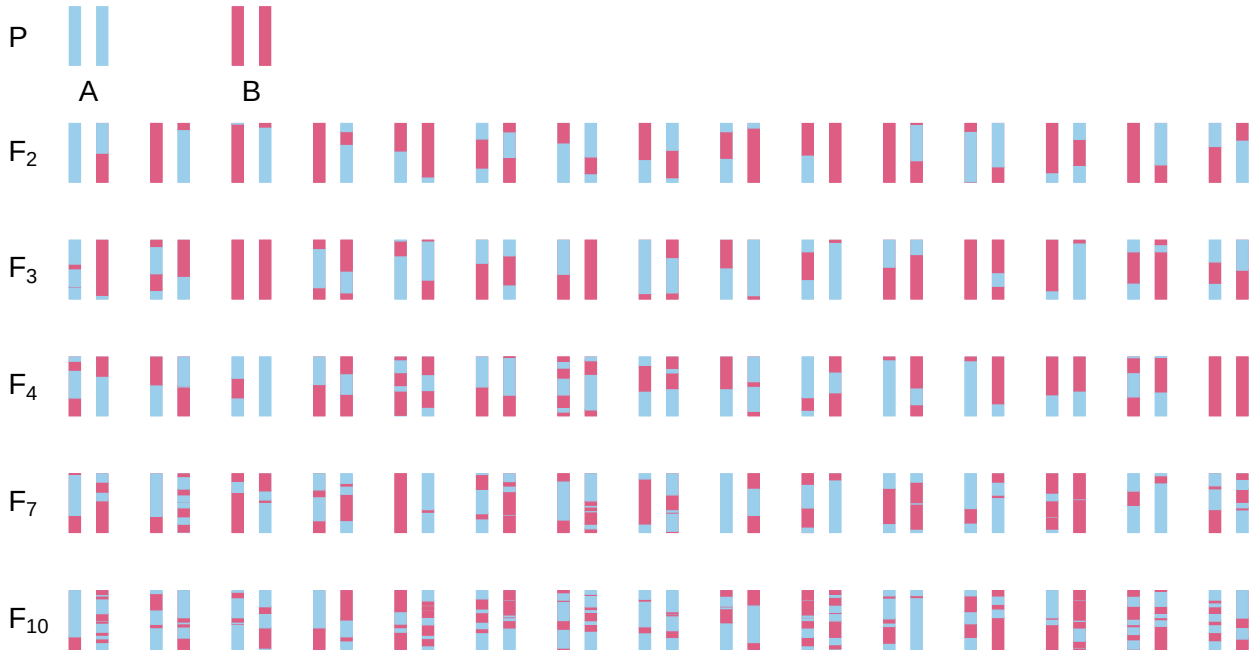

When we find a QTL, our goal is then to try to identify the gene underneath that location. The traditional approach is to form what's called a congenic line, where you take the blue strain and insert a chunk of the pink strain into it. If this new line shows a difference in phenotype from the blue strain, you have shown that something in that region is affecting the trait.

You'd then do a series of crosses to break up that region into smaller and smaller pieces until you are able to pinpoint the particular gene that is responsible for the effect.

This is very time consuming and expensive, and it is often not successful. QTL are usually mapped to very large intervals, like half a chromosome, and it takes a huge effort to fine map down to the gene level.

So there's been a lot of effort to devise ways to speed up the process.

# Advanced intercross lines

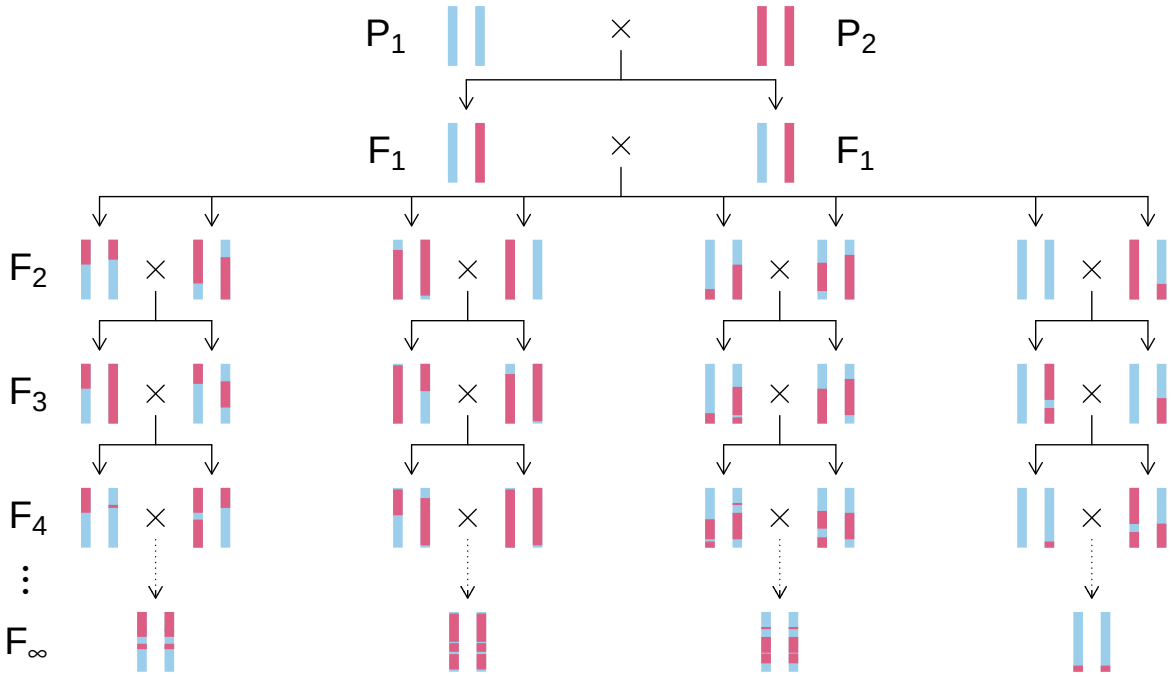





One approach is to do repeated generations of crosses to break up the chromosomes into smaller pieces. These are called advanced intercross lines. After 10 generations, you've broken up the chromosomes into much smaller pieces, and so the correlation among genotypes is much reduced and as a consequence your mapping precision is much higher.

But this is really time consuming; you can only do about four generations of mice per year, so 10 generations is 2.5 years.

# Recombinant inbred lines



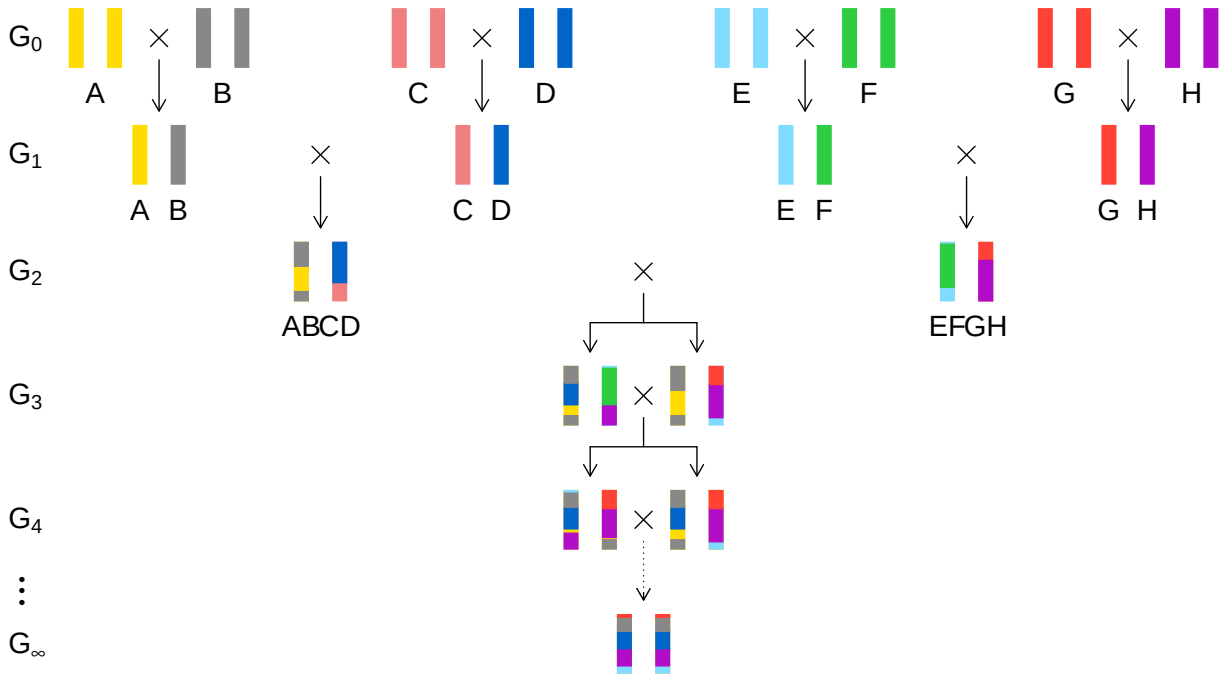



Another approach is to create what are called recombinant inbred lines. You start with an intercross and then do repeated sibling mating to generate a new inbred line that has bits of each of the founder lines. Because of the multiple generations of crossing, you end up with more dense crossovers that occured in different generations. You do this multiple times in parallel to create a panel of lines that you can use for mapping. (Note that many plant species, you can do the same thing but using selfing rather than sib-mating: repeatedly crosses of a given plant to itself. The progress to inbreeding is much more rapid in that case.)

This is also a lot of work, but the advantage here is that the final lines become a permanent resource. Once you've developed the lines, they can be shared among labs to study many different phenotypes. And you only need to genotype each line one time. There are a number of different panels of mouse RIL, but until recently they have been rather limited (like 10-15 lines from a given pair of founders). The BXD lines have been expanded considerably recently; there are like 80 or so BXD lines.

But in addition to the disadvantage of the cost and time to develop such a panel, because they involve just two strains they are of interest only for traits that show a difference between those two particular founders. For example, the BXD panel may not be useful for your trait of interest.

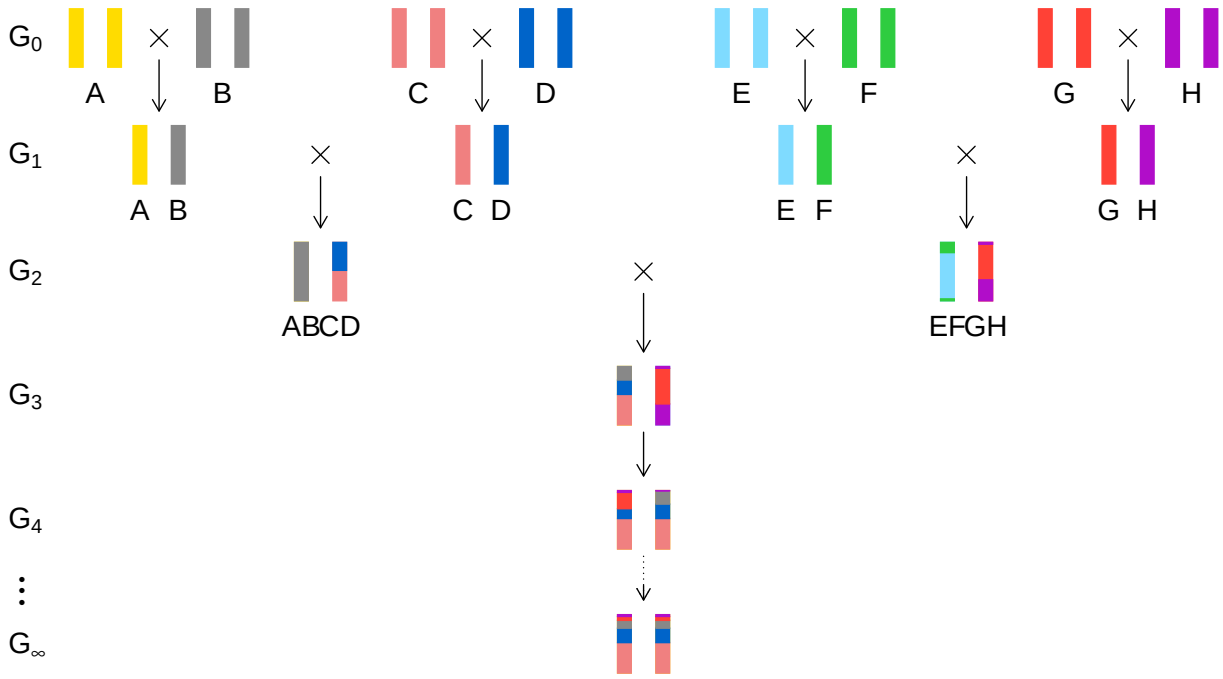
# Collaborative Cross






In 2000, a group of mouse geneticists came up with the idea of the Collaborative Cross: to create a panel of recombinant inbred lines formed from a set of eight founders. The idea is you'd have the mapping resolution and other advantages of RILs, and since you were starting with a diverse set of lines, this single panel would be useful for a variety of purposes.

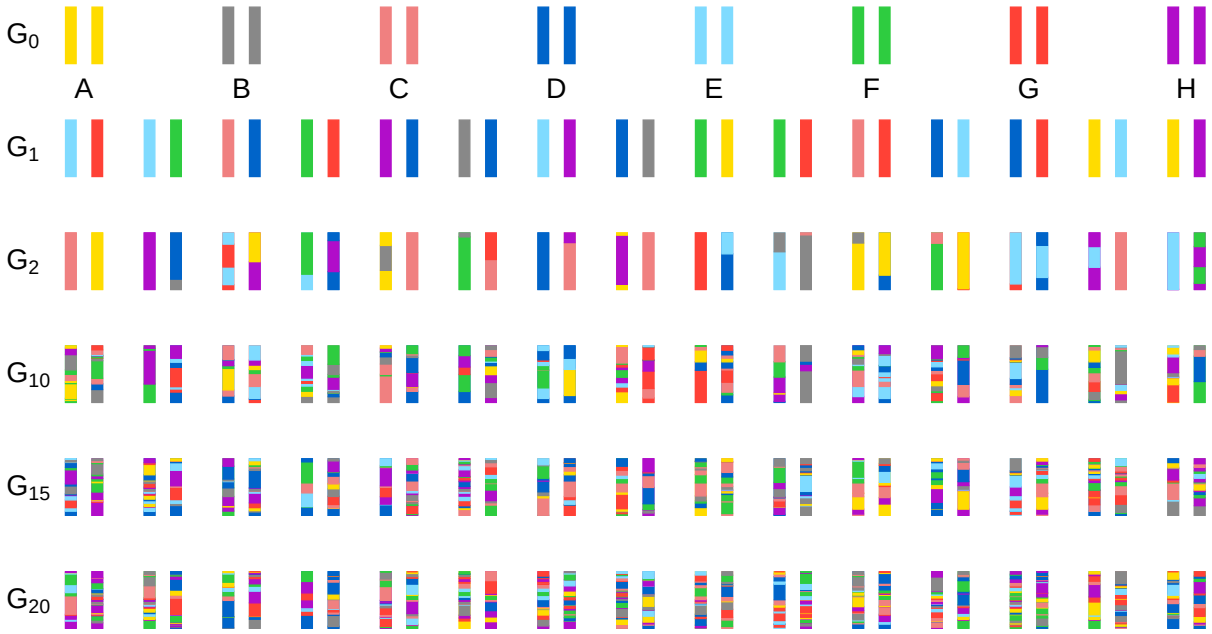
# MAGIC






This idea has really taken off in plants, in which case they are called MAGIC lines. RILs by selfing from 8 or maybe even 16 or 32 founders. There are MAGIC populations in Arabidopsis, wheat, maize, rice, and other plant species.

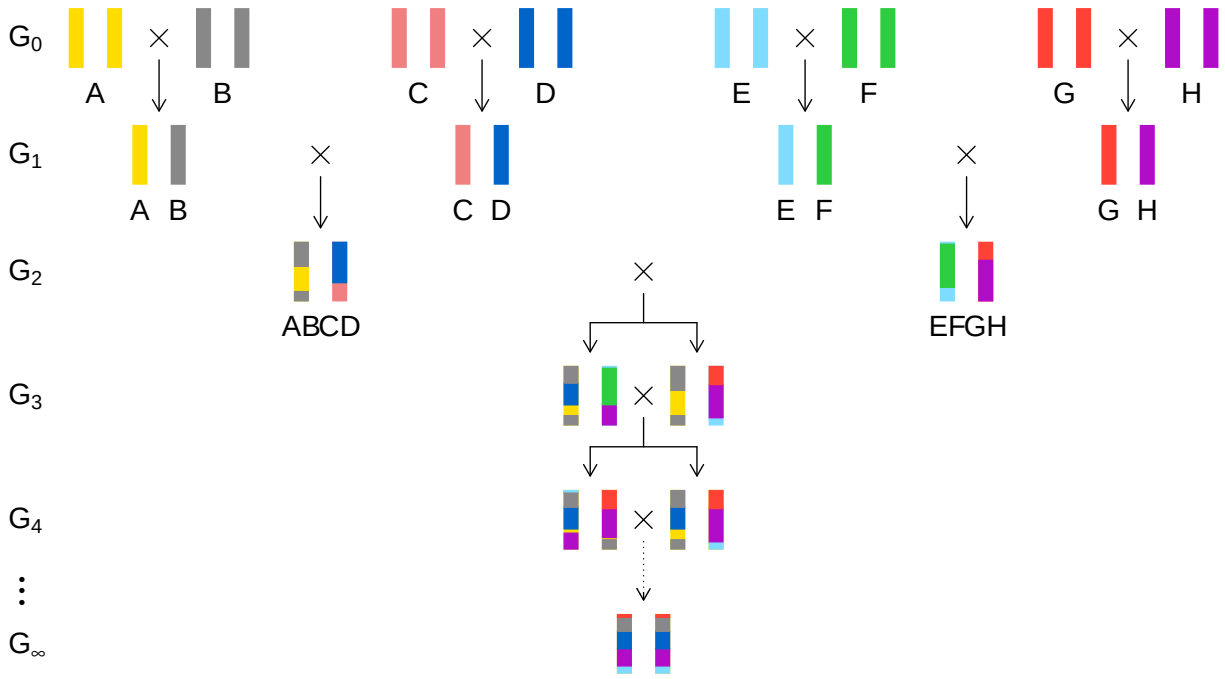
# Heterogeneous stock






The advanced intercross idea can also be extended to use more than two strains. An advanced intercross with 8 founders is sometimes called heterogeneous stock.

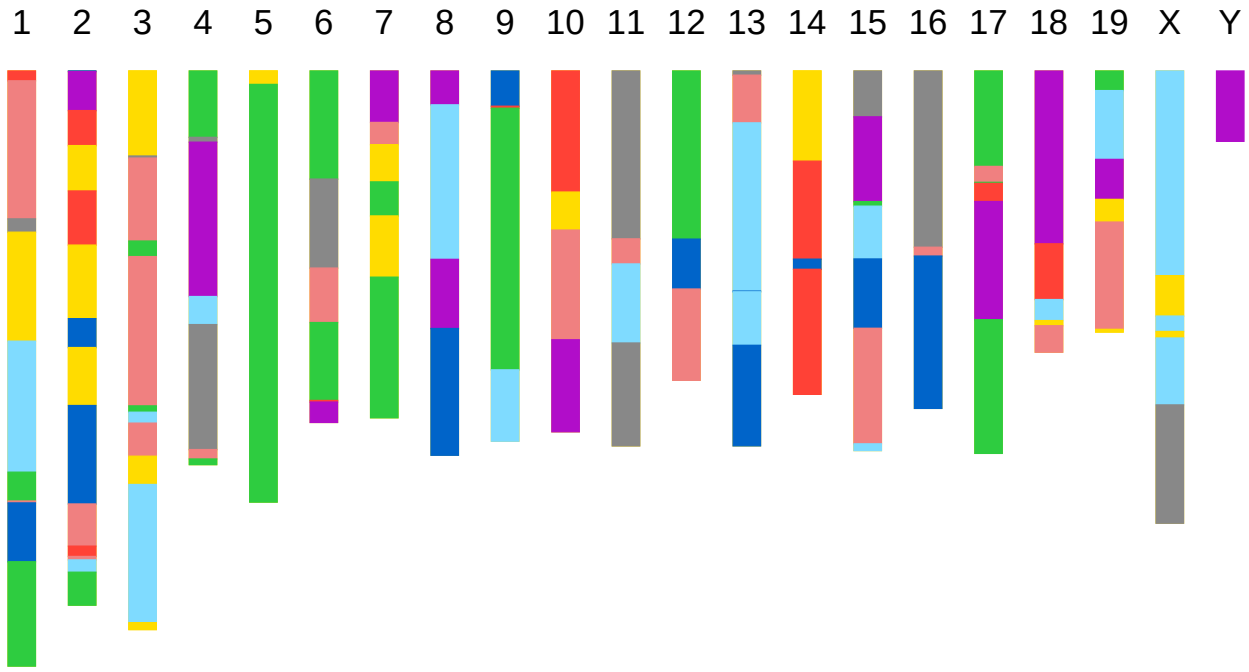
# Collaborative Cross






Back to these 8-way recombinant inbred lines, the Collaborative Cross: I became fascinated with the genotype pattern on the fixed, recombinant chromosome.

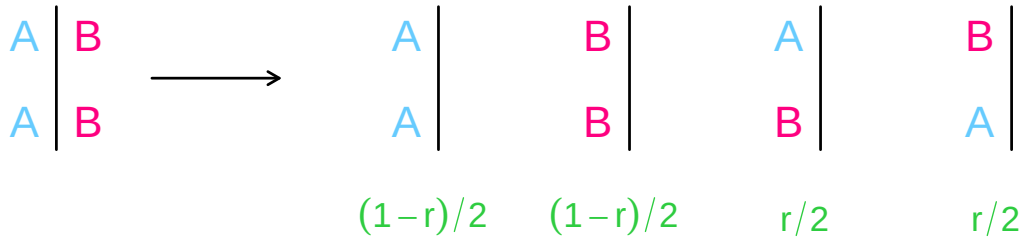
# CC genome




13

Here's a picture of a Collaborative Cross genome. At any one position, it's got one of the eight colors. What can we say about the rate of exchanges along the fixed chromosome, and are the switches equally likely among the eight colors? Can we calculate the transition matrix along the chromosome?

# Recombination fraction



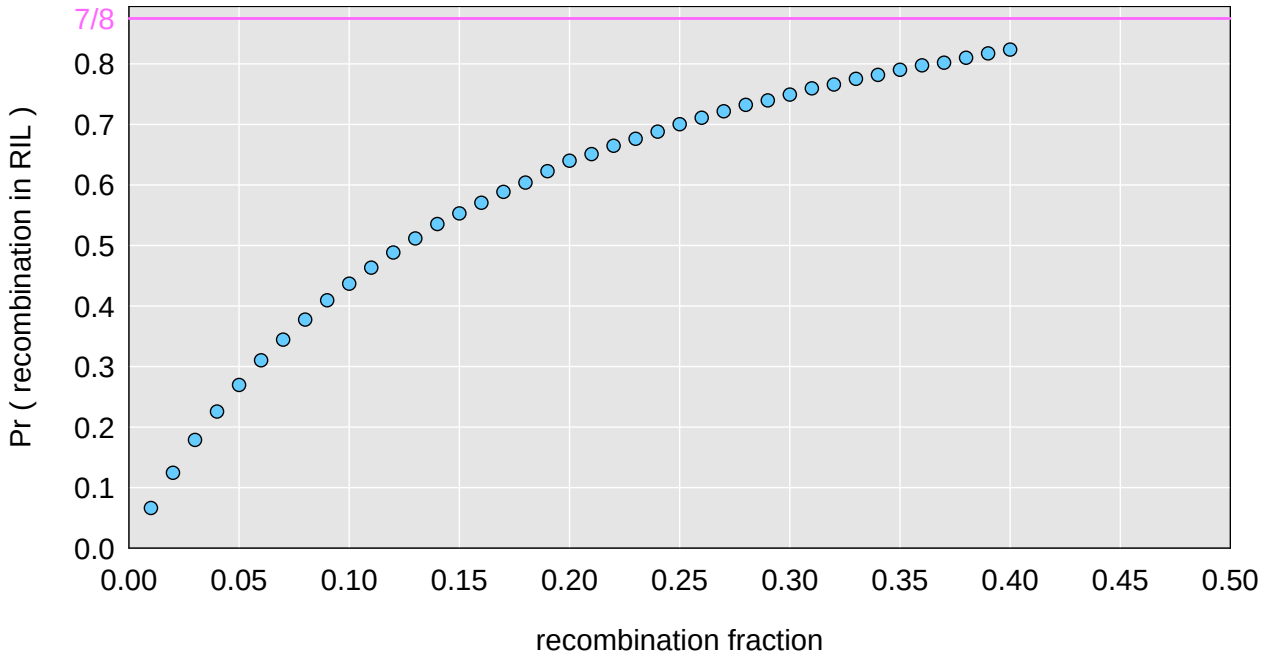

r is the "recombination fraction"



A key quantity to consider is the recombination fraction for an interval. If a parent has alleles A on one chromosome and B on the other, the sperm or egg cell produced at meiosis could have a parental type chromosomes, or it could have a so-called “recombinant” chromosome, with A at one locus and B at the other. The chance that meiosis produces a recombinant chromosome is called the recombination fraction. It depends on the distance between the two positions.

Now the question is, for an interval that has recombination fraction $r$, what is the chance that the fixed Collaborative Cross chromosome will have the same allele, vs switch from one allele two another? What’s the analogous quantity to the recombination fraction for the fixed CC chromosome?

# Simulation results






I had this idea one night, and thought, "I'll just simulate it." So I started some simulations running: simulating the 8-way cross followed by repeated sibling mating until fixation, using intervals with different recombination fractions.

The next morning I had this curve: the relationship between recombination fraction at meiosis (on the x-axis) and the chance that the CC chromosome is fixed for alternate alleles (on the y-axis). A nice smooth curve, and we can be confident that it hits 7/8 when the recombination fraction is 1/2. (If the two positions are recombining completely at random, then you're picking two of the eight alleles for the two loci on the fixed chromosome, equally likely, and so there's a 7/8 chance that you'll pick different alleles.

# Haldane & Waddington 1931

## INBREEDING AND LINKAGE*

J. B. S. HALDANE AND C. H. WADDINGTON
*John Innes Horticultural Institution, London, England*

Received August 9, 1930

TABLE OF CONTENTS





Well, I happened to know about this famous old paper that does this sort of calculation analytically for the two-way recombinant inbred line case. They cover sibling mating as well as selfing, and both autosomes and the X chromosome.

# Result for selfing

Then $c_n + \lambda d_n \equiv c_n + \frac{1}{4}(1 - 2x)d_n + \frac{1}{2}\lambda(1 - 2x)d_n$

$$\therefore \lambda = \frac{1 - 2x}{2 + 4x}.$$

Then since $d_\infty = 0$, and $c_1 = 0$, $d_1 = 2$,

$$c_\infty = c_\infty + \lambda d_\infty = c_1 + \lambda d_1 = \frac{1 - 2x}{1 + 2x}.$$

Put $y = D_\infty$ (the final proportion of crossover zygotes)

$$\therefore C_\infty + D_\infty = 1, C_\infty - D_\infty = c_\infty \therefore y = \tfrac{1}{2}(1 - c_\infty).$$

$$\therefore y = \frac{2x}{1 + 2x}. \qquad (1.3)$$



The paper is pretty hard going, but they derive some simple and now well-known equations, such as this one for selfing. If the recombination fraction (at meiosis) between two loci is $x$, then the chance the two-way RIL by selfing will be fixed for alternate alleles is $2x/(1 + 2x)$.

# Result for sib-mating

Omitting some rather tedious algebra, the solution of these equations is:

$$\zeta = \frac{q}{2-3q}, \quad \theta = \frac{2q}{2-3q}, \quad \kappa = \frac{1}{2-3q},$$
$$\lambda = \frac{1-2q}{2-3q}, \quad \mu = \frac{1-2q}{2-3q}, \quad \nu = \frac{2q}{2-3q}$$

as may easily be verified.

$$\therefore \; c_\infty = c_n + 2e_n + \frac{1}{1+6x}[(1-2x)(d_n + 2f_n + 2j_n + \tfrac{1}{2}k_n) + 2g_n + 4x(h_n + i_n)] \qquad (3.4)$$

and $y = \frac{1}{2}(1-c_\infty)$.

In the case considered, $d_0 = 1, \therefore c_\infty = \zeta d_0 = 1-2x/1+6x$. Hence the proportion of crossover zygotes, $y = 4x/1+6x$ (3.5).

Omitting some rather tedious algebra, the solution of these equations is:

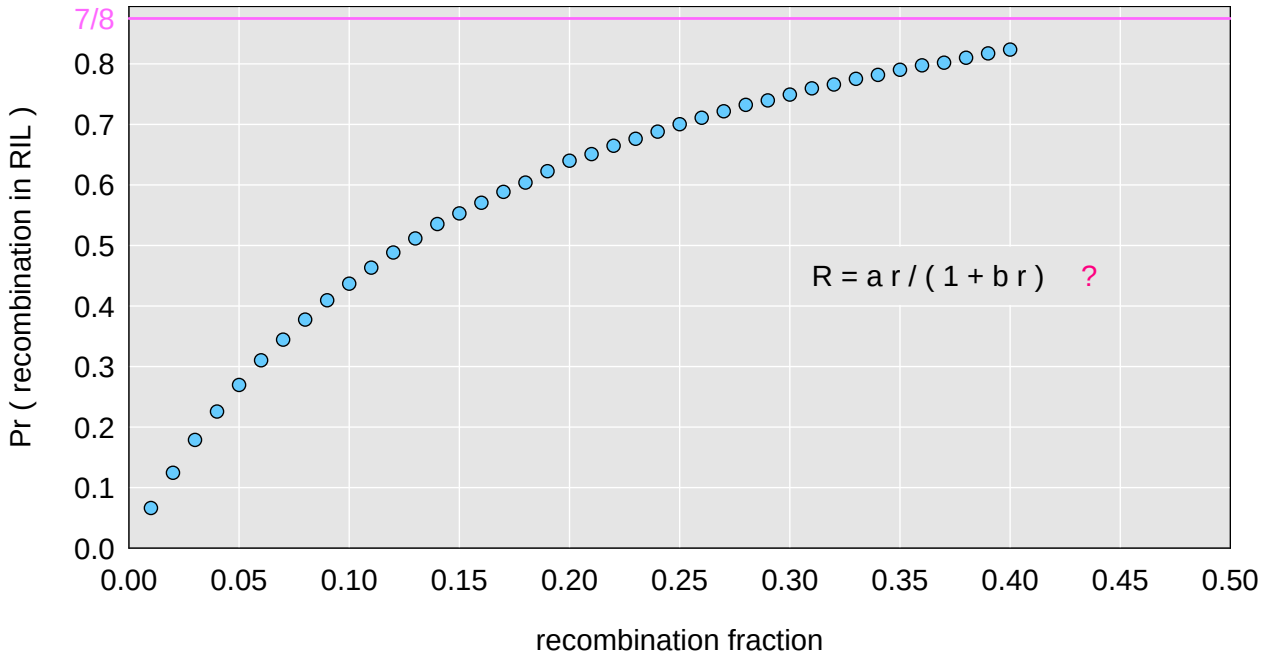
$$\zeta = \frac{q}{2-3q}, \quad \theta = \frac{2q}{2-3q}, \quad \kappa = \frac{1}{2-3q},$$
$$1-2q \qquad 1-2q \qquad 2q$$



For sib-mating, the result is $4x/(1+6x)$. That is, if the recombination fraction between two loci is $x$, the chance that two-way RIL by sibling mating will be fixed for alternate alleles is $4x/(1+6x)$.

I particularly like that they say, "Omitting some rather tedious algebra." If you've had a look at this paper, you may be surprised to then learn that there was tedious algebra omitted.

# Simulation results






So going back to my simulation results, I figured "well, for 2-ways by selfing it's $2x/(1+2x)$, and for 2-ways by sib-mating it's $4x/(1+6x)$, so maybe for 8-ways by sib-mating it's like $ax/(1+bx)$ for some $a$ and $b$.

# Non-linear regression

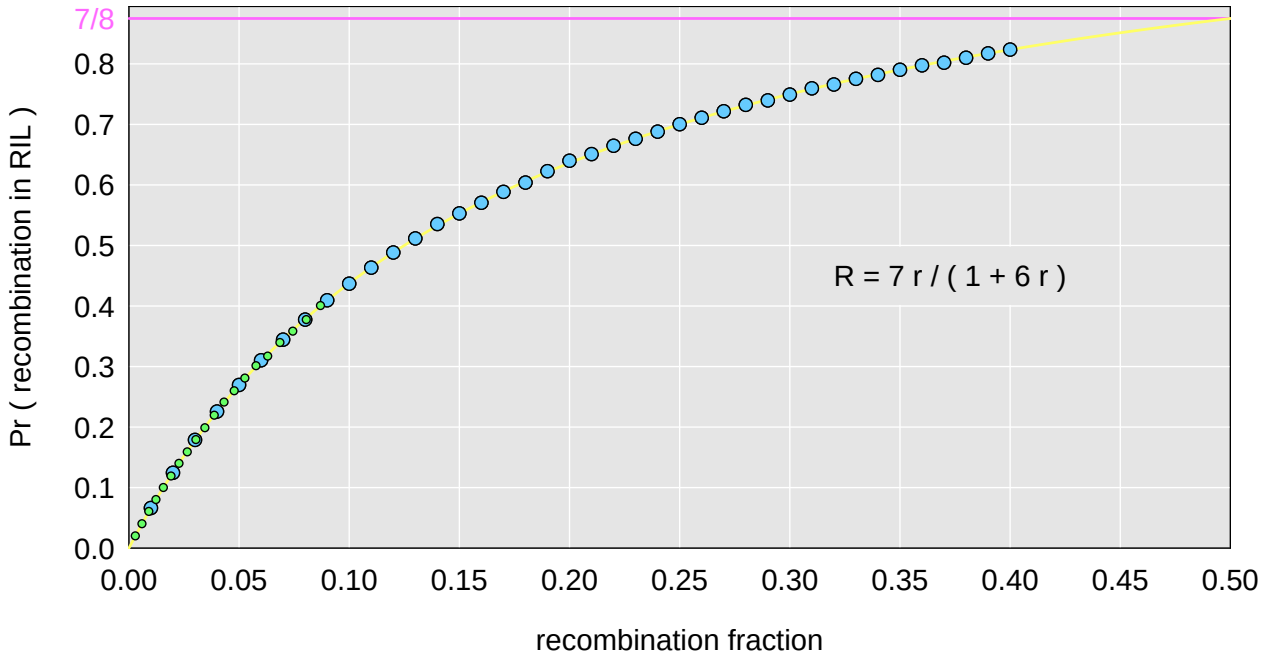
```
out <- nls( R ~ a*r/(1 + b*r),
            data = data.frame(r=r, R=R),
            start = list(a=4, b=6))
summary(out)
```

```
                                        More data
  Estimate  Std. Error          Estimate  Std. Error
a    7.016       0.011       a    7.003       0.008
b    6.023       0.016       b    6.005       0.012
```



How can I figure out $a$ and $b$? Non-linear regression, using `nls()` in R. It was immediately clear that the answer was $7r/(1+6r)$ for 8-way RIL by sib-mating. I did some more simulations to get more precise coefficient estimates, but the answer was clear.

# Simulation results






Add that curve to my simulated data and it fits perfectly. A triumph of computer simulation. Fancy algebra accomplished with no detailed knowledge whatsover. Certainly no tedious algebra.

Okay, but not entirely satisfying. And by this point I was hooked: how to prove this equation?

# Markov chain

- Sequence of random variables $\{X_0, X_1, X_2, \dots\}$ satisfying

$$\Pr(X_{n+1} \mid X_0, X_1, \dots, X_n) = \Pr(X_{n+1} \mid X_n)$$

- Transition probabilities $P_{ij} = \Pr(X_{n+1} = j \mid X_n = i)$
- Here, $X_n$ = "parental type" at generation n.
- We are interested in absorption probabilities

$$\pi_j = \Pr(X_n \to j \mid X_0)$$



So they key thing here is that if you let $X_n$ denote the "parental type" meaning the pattern of alleles at two positions each of the two individuals' two chromosomes at generation $n$, then the $X$'s form a Markov chain. That is, if I tell you the pattern at a given generation, that's all you need to know to determine what will happen at the next generation; the history of how we got there doesn't matter.

The key parameters for the Markov chain are the transition probabilities. That is, if I tell you the parental type at generation $n$, what are the probabilities for each possible type at generation $n + 1$?

The key thing we're trying to determine are so-called absorption probabilities. What is the chance that the Markov chain will be absorbed into a particular fixation state (from which it won't be able to exit), given the starting state (which is defined by the cross design).

# Absorption probabilities

Consider the case of absorption into the state $\begin{array}{c|c} A & A \\ A & A \end{array}$

(write this AA|AA)

Let $h_i$ = probability, starting at i, of being absorbed into AA|AA.

Then $h_{AA|AA} = 1$ and $h_{AB|AB} = 0$.

Condition on the first step: $h_i = \sum_k P_{ik} h_k$

For selfing, this gives a system of 3 linear equations.

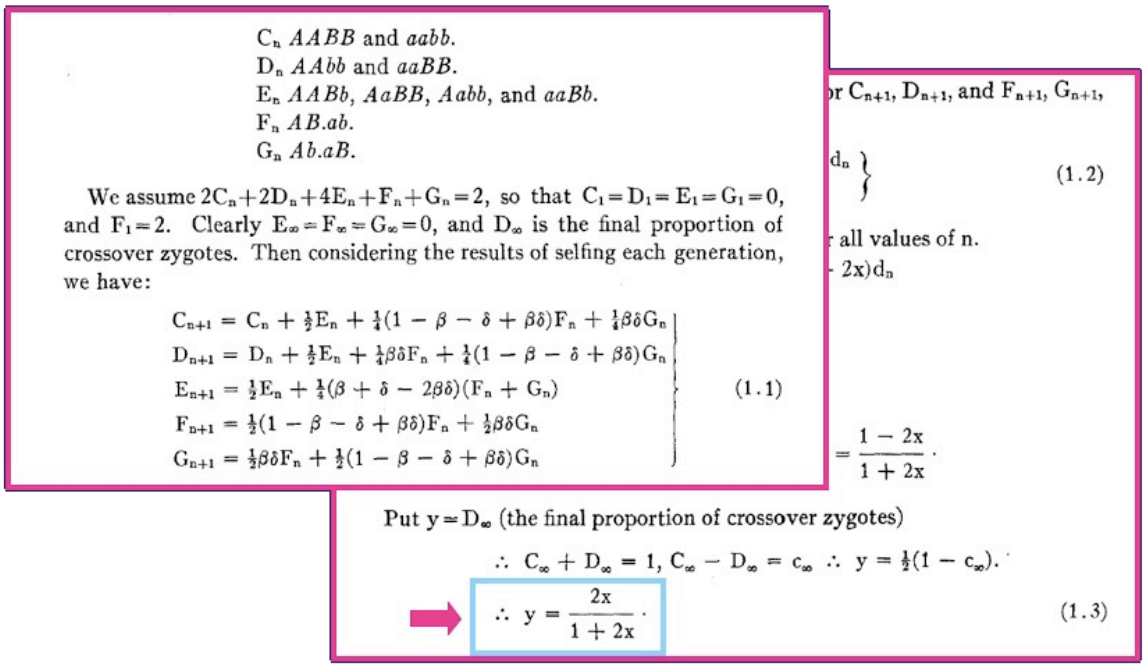


A key trick with Markov chains is to condition on the first step. Let's focus on a particular absorbing state, say A's at both loci on both chromosomes. And let $h_i$ is the probability that, starting at state $i$, you'll be absorved into that particular $AA|AA$ state. Then $h_i$ is start at $i$, jump to state $k$, and from there be absorbed into the $AA|AA$ state.

So that gives us a set of linear equations that we can solve, since the transition matrix $P$ is known.

For 2-way RIL by selfing, you end up with a system of three linear equations. There are five distinct states, but two are absorbing and so the $h$'s for them are just 0 or 1, so there are three unknowns.

# Equations for selfing



$C_n$ *AABB* and *aabb*.
$D_n$ *AAbb* and *aaBB*.
$E_n$ *AABb*, *AaBB*, *Aabb*, and *aaBb*.
$F_n$ *AB.ab*.
$G_n$ *Ab.aB*.

We assume $2C_n+2D_n+4E_n+F_n+G_n=2$, so that $C_1=D_1=E_1=G_1=0$, and $F_1=2$. Clearly $E_\infty=F_\infty=G_\infty=0$, and $D_\infty$ is the final proportion of crossover zygotes. Then considering the results of selfing each generation, we have:

$$\left.\begin{aligned}
C_{n+1} &= C_n + \tfrac{1}{2}E_n + \tfrac{1}{4}(1-\beta-\delta+\beta\delta)F_n + \tfrac{1}{4}\beta\delta G_n \\
D_{n+1} &= D_n + \tfrac{1}{2}E_n + \tfrac{1}{4}\beta\delta F_n + \tfrac{1}{4}(1-\beta-\delta+\beta\delta)G_n \\
E_{n+1} &= \tfrac{1}{2}E_n + \tfrac{1}{4}(\beta+\delta-2\beta\delta)(F_n+G_n) \\
F_{n+1} &= \tfrac{1}{2}(1-\beta-\delta+\beta\delta)F_n + \tfrac{1}{2}\beta\delta G_n \\
G_{n+1} &= \tfrac{1}{2}\beta\delta F_n + \tfrac{1}{2}(1-\beta-\delta+\beta\delta)G_n
\end{aligned}\right\} \quad (1.1)$$

…r $C_{n+1}$, $D_{n+1}$, and $F_{n+1}$, $G_{n+1}$,

$d_n$ } (1.2)

… all values of n.
… $2x)d_n$

$= \dfrac{1-2x}{1+2x}$.

Put $y=D_\infty$ (the final proportion of crossover zygotes)

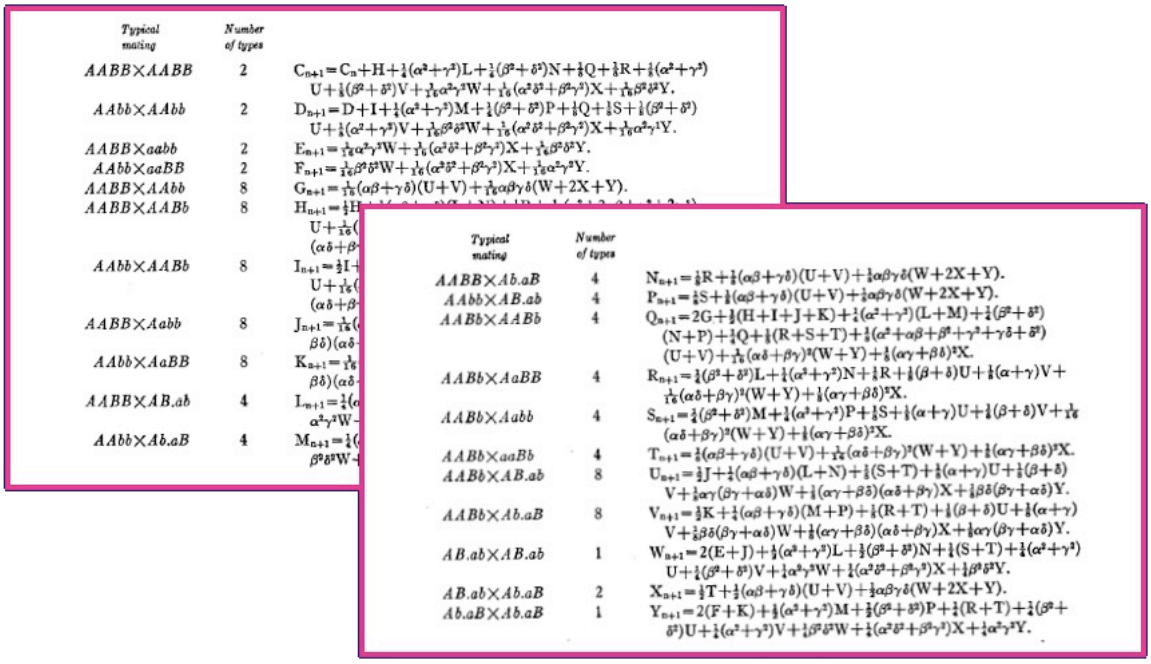
$\therefore\ C_\infty + D_\infty = 1,\ C_\infty - D_\infty = c_\infty\ \therefore\ y = \tfrac{1}{2}(1-c_\infty).$

$$\therefore\ y = \frac{2x}{1+2x}. \quad (1.3)$$



Here's the 2-way selfing business from the Haldane and Waddington paper. They are showing the 5×5 transition matrix and they go one to solve the set of related equations to get to that $2x/(1+2x)$ solution.

# Equations for sib-mating



| Typical mating | Number of types | |
|---|---|---|
| $AABB\times AABB$ | 2 | $C_{n+1}=C_n+H+\frac{1}{4}(\alpha^2+\gamma^2)L+\frac{1}{4}(\beta^2+\delta^2)N+\frac{1}{8}Q+\frac{1}{8}R+\frac{1}{4}(\alpha^2+\gamma^2)$ $U+\frac{1}{8}(\beta^2+\delta^2)V+\frac{1}{16}\alpha^2\gamma^2W+\frac{1}{16}(\alpha^2\delta^2+\beta^2\gamma^2)X+\frac{1}{16}\beta^2\delta^2Y.$ |
| $AAbb\times AAbb$ | 2 | $D_{n+1}=D+I+\frac{1}{4}(\alpha^2+\gamma^2)M+\frac{1}{4}(\beta^2+\delta^2)P+\frac{1}{8}Q+\frac{1}{8}S+\frac{1}{8}(\beta^2+\delta^2)$ $U+\frac{1}{8}(\alpha^2+\gamma^2)V+\frac{1}{16}\beta^2\delta^2W+\frac{1}{16}(\alpha^2\delta^2+\beta^2\gamma^2)X+\frac{1}{16}\alpha^2\gamma^2Y.$ |
| $AABB\times aabb$ | 2 | $E_{n+1}=\frac{1}{16}\alpha^2\gamma^2W+\frac{1}{16}(\alpha^2\delta^2+\beta^2\gamma^2)X+\frac{1}{16}\beta^2\delta^2Y.$ |
| $AAbb\times aaBB$ | 2 | $F_{n+1}=\frac{1}{16}\beta^2\delta^2W+\frac{1}{16}(\alpha^2\delta^2+\beta^2\gamma^2)X+\frac{1}{16}\alpha^2\gamma^2Y.$ |
| $AABB\times AAbb$ | 8 | $G_{n+1}=\frac{1}{16}(\alpha\beta+\gamma\delta)(U+V)+\frac{1}{16}\alpha\beta\gamma\delta(W+2X+Y).$ |
| $AABB\times AABb$ | 8 | $H_{n+1}=\frac{1}{2}H$ $U+\frac{1}{16}($ $(\alpha\delta+\beta$ |
| $AAbb\times AABb$ | 8 | $I_{n+1}=\frac{3}{4}I+$ $U+\frac{1}{16}($ $(\alpha\delta+\beta$ |
| $AABB\times Aabb$ | 8 | $J_{n+1}=\frac{1}{16}($ $\beta\delta)(\alpha\delta$ |
| $AAbb\times AaBB$ | 8 | $K_{n+1}=\frac{1}{16}$ $\beta\delta)(\alpha\delta$ |
| $AABB\times AB.ab$ | 4 | $L_{n+1}=\frac{1}{4}(\alpha$ $\alpha^2\gamma^2W-$ |
| $AAbb\times Ab.aB$ | 4 | $M_{n+1}=\frac{1}{4}($ $\beta^2\delta^2W+$ |

| Typical mating | Number of types | |
|---|---|---|
| $AABB\times Ab.aB$ | 4 | $N_{n+1}=\frac{1}{8}R+\frac{1}{8}(\alpha\beta+\gamma\delta)(U+V)+\frac{1}{8}\alpha\beta\gamma\delta(W+2X+Y).$ |
| $AAbb\times AB.ab$ | 4 | $P_{n+1}=\frac{1}{8}S+\frac{1}{8}(\alpha\beta+\gamma\delta)(U+V)+\frac{1}{8}\alpha\beta\gamma\delta(W+2X+Y).$ |
| $AABb\times AABb$ | 4 | $Q_{n+1}=2G+\frac{1}{2}(H+I+J+K)+\frac{1}{4}(\alpha^2+\gamma^2)(L+M)+\frac{1}{4}(\beta^2+\delta^2)$ $(N+P)+\frac{1}{4}Q+\frac{1}{8}(R+S+T)+\frac{1}{8}(\alpha^2+\alpha\beta+\beta^2+\gamma^2+\gamma\delta+\delta^2)$ $(U+V)+\frac{1}{16}(\alpha\delta+\beta\gamma)^2(W+Y)+\frac{1}{8}(\alpha\gamma+\beta\delta)^2X.$ |
| $AABb\times AaBB$ | 4 | $R_{n+1}=\frac{1}{4}(\beta^2+\delta^2)L+\frac{1}{4}(\alpha^2+\gamma^2)N+\frac{1}{8}R+\frac{1}{8}(\beta+\delta)U+\frac{1}{8}(\alpha+\gamma)V+$ $\frac{1}{16}(\alpha\delta+\beta\gamma)^2(W+Y)+\frac{1}{8}(\alpha\gamma+\beta\delta)^2X.$ |
| $AABb\times Aabb$ | 4 | $S_{n+1}=\frac{1}{4}(\beta^2+\delta^2)M+\frac{1}{4}(\alpha^2+\gamma^2)P+\frac{1}{8}S+\frac{1}{8}(\alpha+\gamma)U+\frac{1}{8}(\beta+\delta)V+\frac{1}{16}$ $(\alpha\delta+\beta\gamma)^2(W+Y)+\frac{1}{8}(\alpha\gamma+\beta\delta)^2X.$ |
| $AABb\times aaBb$ | 4 | $T_{n+1}=\frac{1}{2}(\alpha\beta+\gamma\delta)(U+V)+\frac{1}{32}(\alpha\delta+\beta\gamma)^2(W+Y)+\frac{1}{8}(\alpha\gamma+\beta\delta)^2X.$ |
| $AABb\times AB.ab$ | 8 | $U_{n+1}=\frac{1}{8}J+\frac{1}{8}(\alpha\beta+\gamma\delta)(L+N)+\frac{1}{8}(S+T)+\frac{1}{8}(\alpha+\gamma)U+\frac{1}{8}(\beta+\delta)$ $V+\frac{1}{8}\alpha\gamma(\beta\gamma+\alpha\delta)W+\frac{1}{8}(\alpha\gamma+\beta\delta)(\alpha\delta+\beta\gamma)X+\frac{1}{8}\beta\delta(\beta\gamma+\alpha\delta)Y.$ |
| $AABb\times Ab.aB$ | 8 | $V_{n+1}=\frac{1}{8}K+\frac{1}{8}(\alpha\beta+\gamma\delta)(M+P)+\frac{1}{8}(R+T)+\frac{1}{8}(\beta+\delta)U+\frac{1}{8}(\alpha+\gamma)$ $V+\frac{1}{8}\beta\delta(\beta\gamma+\alpha\delta)W+\frac{1}{8}(\alpha\gamma+\beta\delta)(\alpha\delta+\beta\gamma)X+\frac{1}{8}\alpha\gamma(\beta\gamma+\alpha\delta)Y.$ |
| $AB.ab\times AB.ab$ | 1 | $W_{n+1}=2(E+J)+\frac{1}{2}(\alpha^2+\gamma^2)L+\frac{1}{2}(\beta^2+\delta^2)N+\frac{1}{4}(S+T)+\frac{1}{4}(\alpha^2+\gamma^2)$ $U+\frac{1}{4}(\beta^2+\delta^2)V+\frac{1}{4}\alpha^2\gamma^2W+\frac{1}{4}(\alpha^2\delta^2+\beta^2\gamma^2)X+\frac{1}{4}\beta^2\delta^2Y.$ |
| $AB.ab\times Ab.aB$ | 2 | $X_{n+1}=\frac{1}{2}T+\frac{1}{2}(\alpha\beta+\gamma\delta)(U+V)+\frac{1}{2}\alpha\beta\gamma\delta(W+2X+Y).$ |
| $Ab.aB\times Ab.aB$ | 1 | $Y_{n+1}=2(F+K)+\frac{1}{2}(\alpha^2+\gamma^2)M+\frac{1}{2}(\beta^2+\delta^2)P+\frac{1}{4}(R+T)+\frac{1}{4}(\beta^2+$ $\delta^2)U+\frac{1}{4}(\alpha^2+\gamma^2)V+\frac{1}{4}\beta^2\delta^2W+\frac{1}{4}(\alpha^2\delta^2+\beta^2\gamma^2)X+\frac{1}{4}\alpha^2\gamma^2Y.$ |



The case of sibling mating is somewhat more complicated. There are 22 distinct states to keep track of, and they're showing here the transition matrix.

# Result for sib-mating

Omitting some rather tedious algebra, the solution of these equations is:

$$\zeta = \frac{q}{2-3q}, \quad \theta = \frac{2q}{2-3q}, \quad \kappa = \frac{1}{2-3q},$$
$$\lambda = \frac{1-2q}{2-3q}, \quad \mu = \frac{1-2q}{2-3q}, \quad \nu = \frac{2q}{2-3q}$$

as may easily be verified.

$$\therefore \; c_\infty = c_n + 2e_n + \frac{1}{1+6x}[(1-2x)(d_n + 2f_n + 2j_n + \tfrac{1}{2}k_n) + 2g_n + 4x(h_n + i_n)] \qquad (3.4)$$

and $y = \frac{1}{2}(1-c_\infty)$.

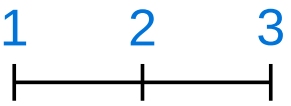
In the case considered, $d_0 = 1, \therefore c_\infty = \zeta d_0 = 1-2x/1+6x$. Hence the proportion of crossover zygotes, $y = 4x/1+6x$ (3.5).



Solving a system of equations that's a bit smaller but still rather similar to those, they get to the $4x/(1+6x)$ solution.

Following this approach, I was able to prove my $7r/(1+6r)$ equation, but by this time I had become obsessed with the situation for three points.

# 3-point coincidence




- $r_{ij}$ = recombination fraction for interval (i, j)
  Assume $r_{12} = r_{23} = r$.
- Coincidence = c = Pr(double recombinant)/$r^2$
  = Pr(rec'n in 23 | rec'n in 12)/Pr(rec'n in 23)
- No interference = 1
  Positive interference $< 1$
  Negative interference $> 1$
- Generally c is a function of r



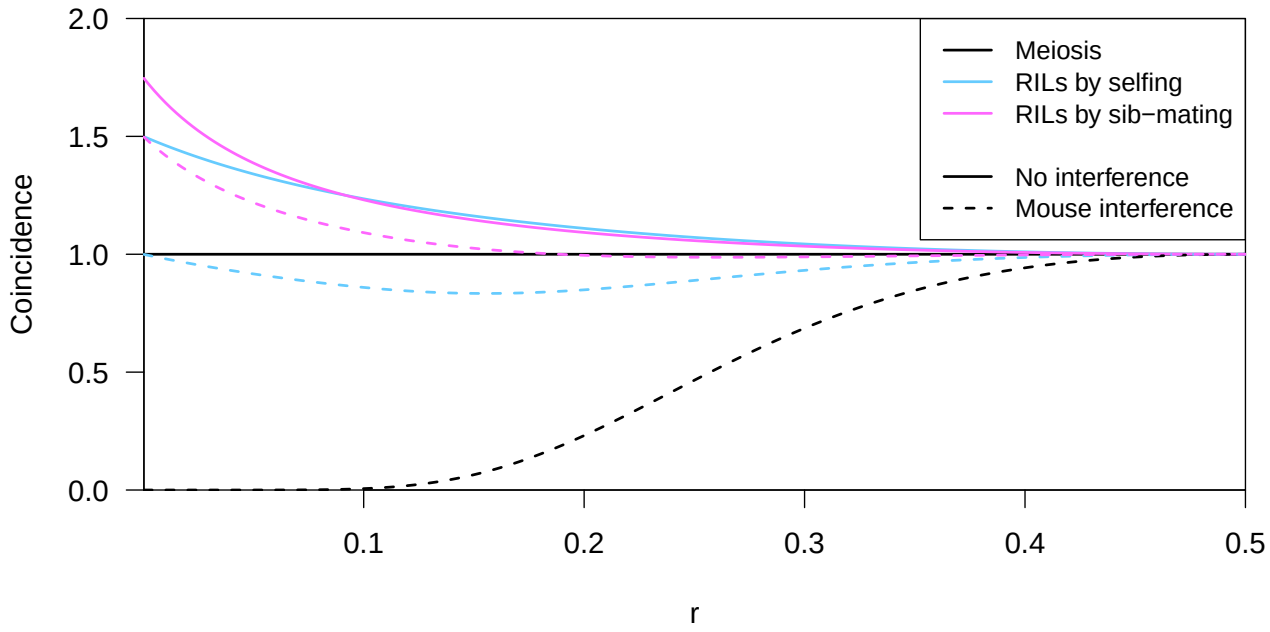
I mentioned in the introductory lecture for this course about crossover interference: the tendency of these meiotic exchanges to not occur too close together. One measure of interference is to consider three points along a chromosome and calculate the so-called 3-point coincidence, which you can view as the probability of exchange in both intervals divided by what would be expected by chance.

Alternatively, and my preferred definition, is to think about recombination in the second interval, and ask how is the chance of a recombination event in the second interval affected by the presence of a recombination event in the first interval: the conditional probability divided by the prior probability.

The coincidence being 1 indicates no interference, $< 1$ indicates "positive" interference (crossovers tend not to be close together), and $> 1$ indicates "negative interference" (crossovers tend to be clustered).

The coincidence is generally a function of the size of the intervals, measured by the recombination fraction, *r*.

# Coincidence






It turns out that, for 2-way recombinant inbred lines, the equivalent measure to coincidence for the fixed RIL chromosome can be derived directly with little effort, as it just depends on the recombination fraction for the three intervals (the two disjoint intervals plus the large interval that is the union of the two). The Haldane and Waddington paper has a section on this.

The solid curves here are for the case of no crossover interference, with two-way RIL by selfing or sib-mating. In both cases, the curves are entirely above 1, indicating clustering of breakpoints on the fixed interval.

Using a model for meiosis in the mouse, which exhibits strong positive crossover interference (the black dashed curve), we end up with the dashed curves for the coincidence-type quantity for the fixed chromosomes. For 2-way RILs by sib-mating, it is almost entirely above 1. This indicates that even if the crossovers at meiosis tend not to be close together, the breakpoints on the fixed RIL chromosome will tend to be clustered. Haldane and Waddington derived this, but didn't really dwell on it.

# Coincidence in 8-way RILs

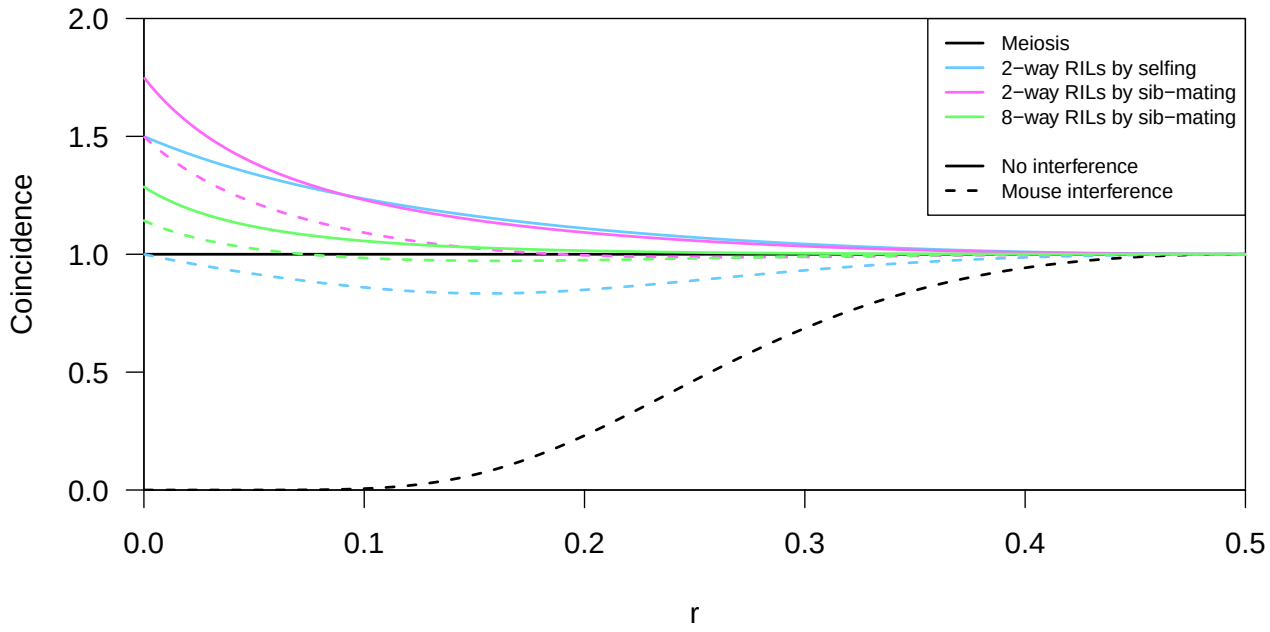
- The trick that allowed us to get the coincidence for 2-way RILs doesn't work for 8-way RILs.
- It's sufficient to consider 4-way RILs.
- Calculations for 3 points in 4-way RILs is still astoundingly complex.
  - 2 points in 2-way RILs by sib mating:
    55 parental types $\rightarrow$ 22 states by symmetry
  - 3 points in 4-way RILs by sib mating:
    2,164,240 parental types $\rightarrow$ 137,488 states by symmetry
- Even counting the states was difficult.



Now I'm interested in derived this quantity for 8-way RILs by sibling mating. But it turned out that the trick used for 2-way RILs doesn't work for the 8-way case. However, it turns out it's sufficient to consider just 4-way RILs, because of the bottleneck with two parents each with two chromosomes.

Still, it was difficult just to keep track of the possible states when 3 points on 4 chromosomes are considered.

# Coincidence






Nevertheless, with some very hefty numerical calculations (not simulations, or algebra), I was able to derive the green equations here, which are the quantities analogous to the 3-point coincidence but for the fixed chromosome in 8-way RIL by sib-mating.

Note that even with strong crossover interference, the coincidence is mostly above 1, again indicating clustering of breakpoints on the RIL chromosome. But really it's pretty close to 1, as for no interference.

# The formula

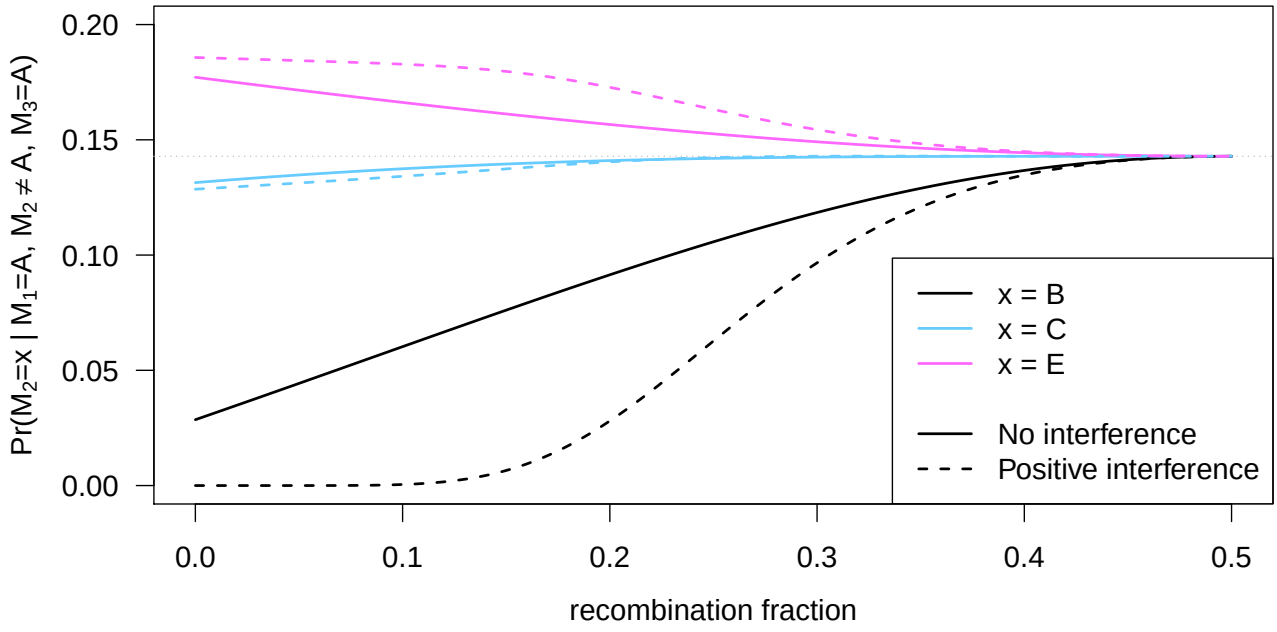
$$C = \frac{(1+6r)[280+1208r-848r^2+5c(7-28r-368r^2+344r^3)-2c^2(49-324r+452r^2)r^2-16c^3(1-2r)r^4]}{49(1+12r-12cr^2)[5+10r-4(2+c)r^2+8cr^3]}$$



A few years later a collaborator showed me another trick to help in this work, and we were able to derive the exact formula for the 3-point coincidence curve.

# 3-point symmetry

$$\Pr(M_2 = x \mid M_1 = A, M_2 \neq A, M_3 = A)$$

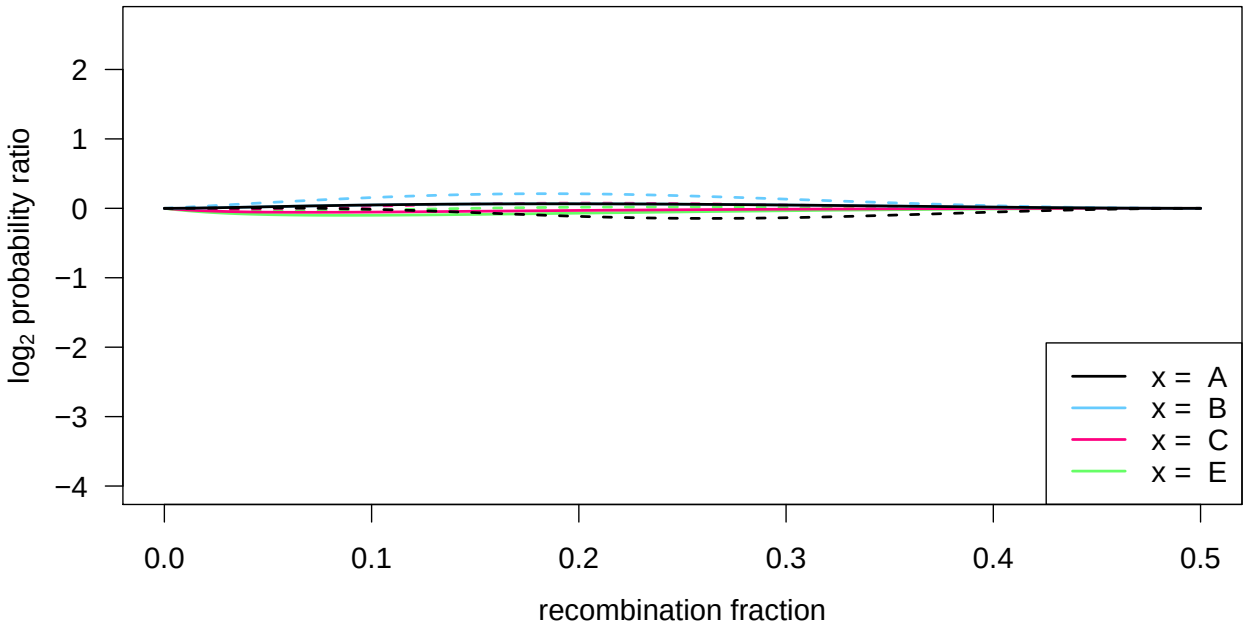





There are some other interesting things you can look at here. First, concerning a sort of 3-point symmetry. If you observe A at the outer two of a set of three points and know that you had a double-exchange and were not A in the middle, what is the chance for each of the other seven alleles in the middle? In the cross A×B×C..., it turns out that you're unlikely to be B, because you'd need a double-crossover in that first meiosis, and you're more likely to be E,F,G, or H (which are all equally likely).

# Markov property

$$\log_2 \left\{ \frac{\Pr(M_3{=}A \mid M_2{=}A, M_1{=}x)}{\Pr(M_3{=}A \mid M_2{=}A)} \right\}$$

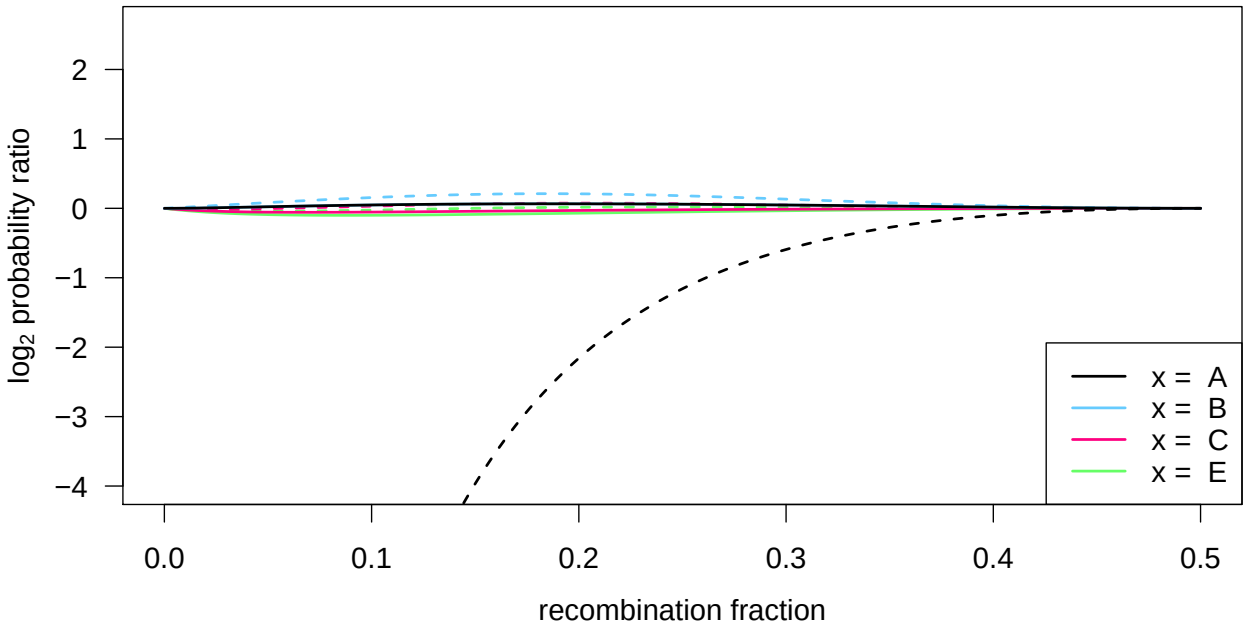





You can also investigate the Markov property for the fixed 8-way RIL chromosome. If I tell you that you're A at point 2, how much does knowledge of point 1 tell you about the probabilities of what you end up being at point 3? Not very much in this case, with a non-exchange.

# Markov property

$$\log_2 \left\{ \frac{\Pr(M_3{=}A \mid M_2{=}B, M_1{=}x)}{\Pr(M_3{=}A \mid M_2{=}B)} \right\}$$

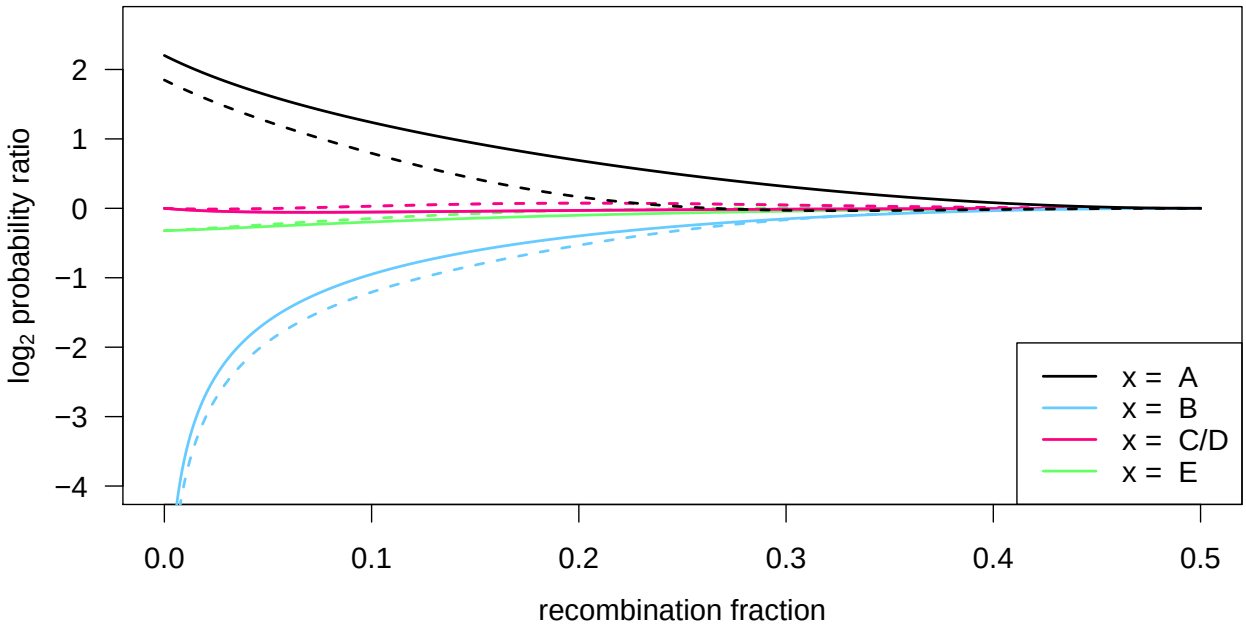





But how about if I ask about switching from B to A? Does knowledge of point 1 change your understanding of that? Well, if you were previously A and crossover interference is strong, you're very likely to not switch back to A, but otherwise it doesn't really matter.

# Markov property

$$\log_2 \left\{ \frac{\Pr(M_3{=}A \mid M_2{=}C, M_1{=}x)}{\Pr(M_3{=}A \mid M_2{=}C)} \right\}$$

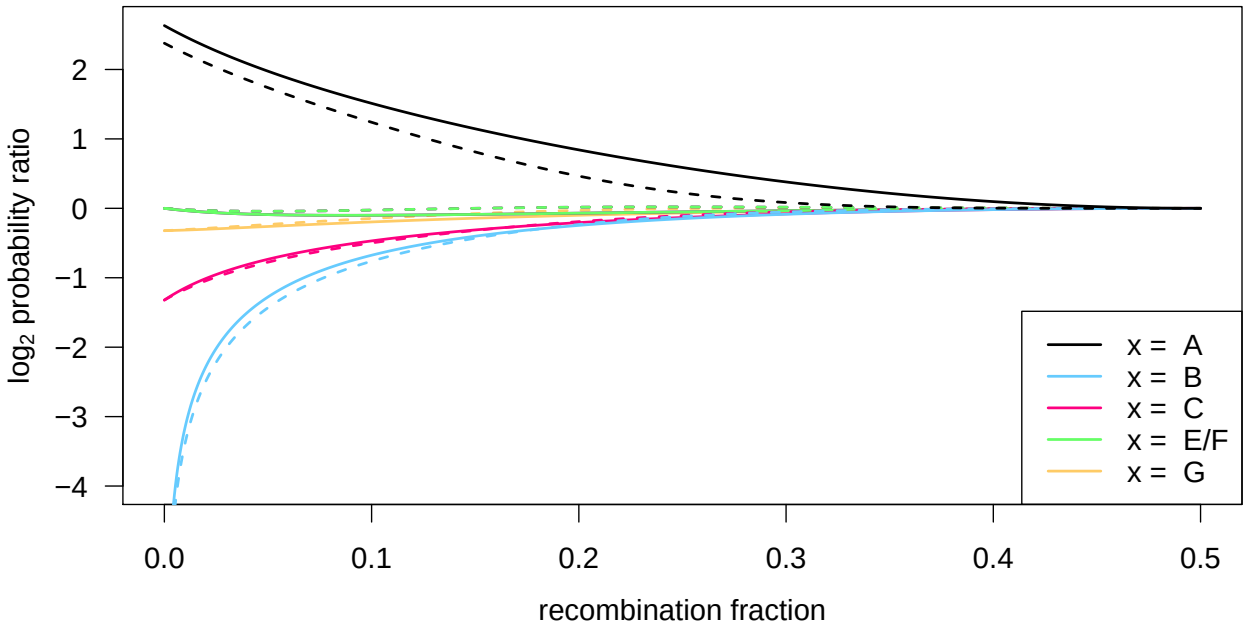





But if I ask about switching from C to A, knowledge of the previous state can matter a lot. You're somehow more likely to switch from back to A from C, if you were previously A.

# Markov property

$$\log_2 \left\{ \frac{\Pr(M_3{=}A \mid M_2{=}E, M_1{=}x)}{\Pr(M_3{=}A \mid M_2{=}E)} \right\}$$






The same goes for the last case: switching from E to A is more likely if you'd previously been A. This is saying that you'll tend to have chunks of E stuck into the middle of a longer stretch of A. Which I did notice in simulations.

# Whole genome simulations

- 2-way selfing, 2-way sib-mating, 8-way sib-mating
- Mouse-like genome, 1665 cM
- Strong positive crossover interference
- Inbreed to complex fixation
- 10,000 simulation replicates

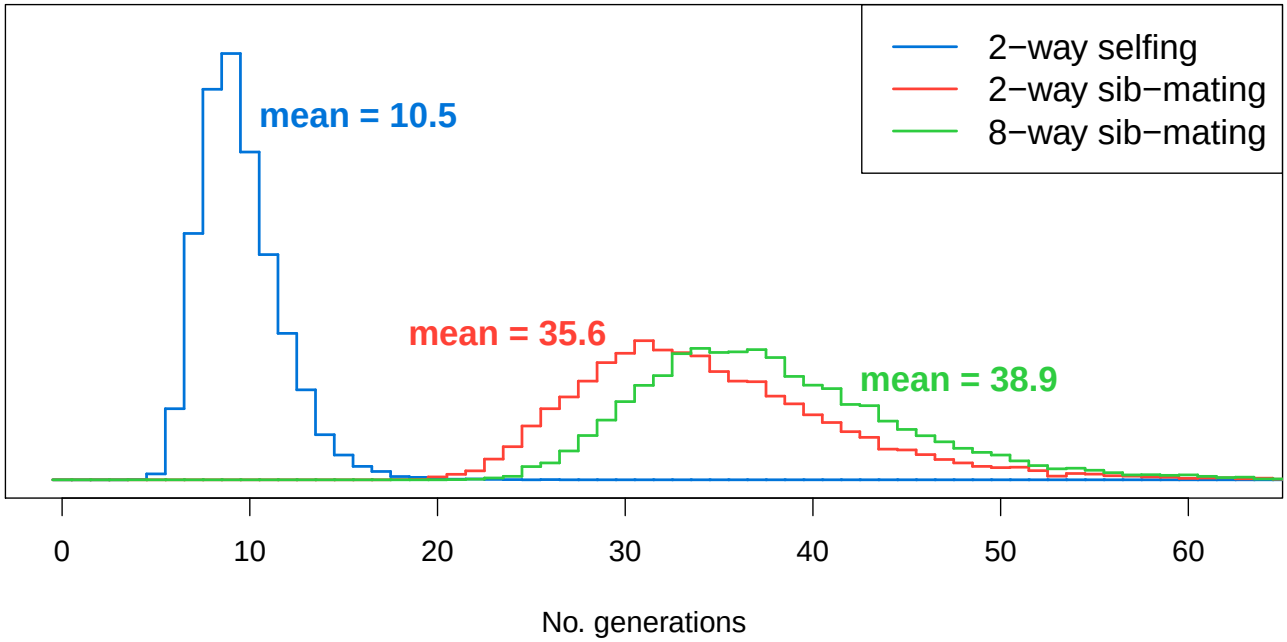


I'd started out trying to sell you on the value of computer simulation, but now I've turned completely to numerical and then exact symbolic calculations. Simulations have the advantage of being easy to get started and get some quick solutions, but the results can be both imprecise (due to simulation error) and overly specific (due to difficulty of covering very general conditions).

But let's get back to another way in which simulations have a big advantage over exact calculations: the ability to consider complete, realistic cases. In our exact calculations, we could look at two or three points along a chromosome. But there are a lot of interesting questions that can only be answered by considering the whole genome at once.

So here we simulate recombinant inbred lines for a mouse-like genome with 20 chromosomes and a total genome length of 1665 cM.

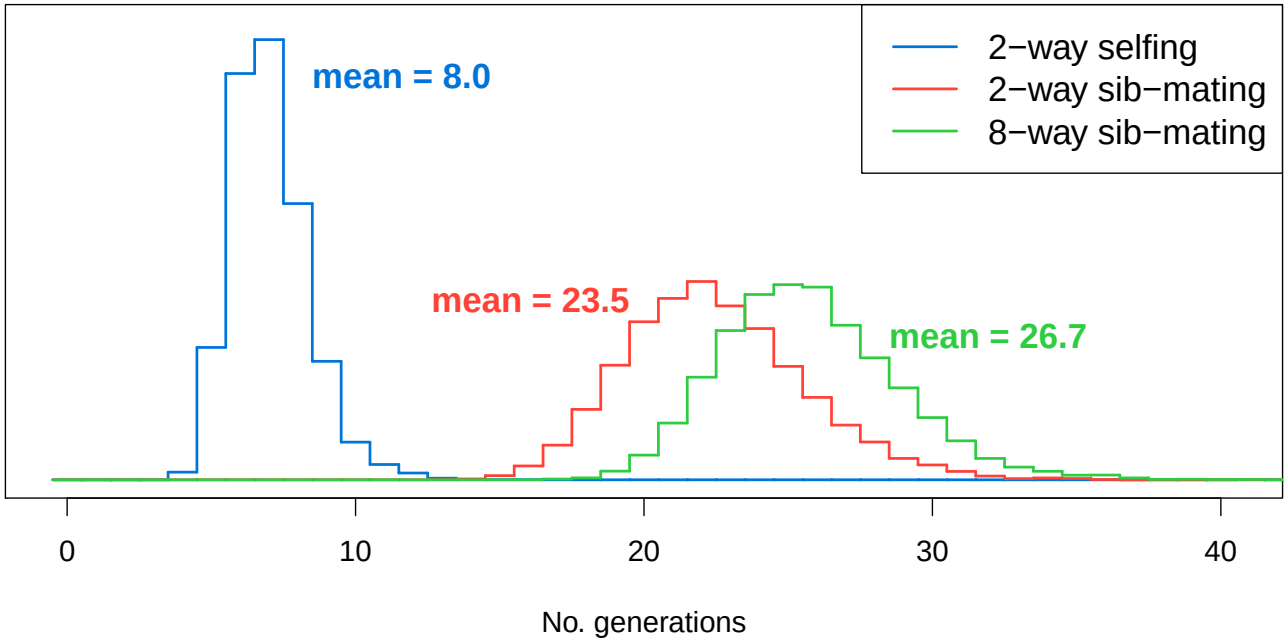
# No. generations to fixation






The first question: how many generations does it take to get complete inbreeding. We usually think of 6 generations for 2-way RIL by selfing and 20 generations for 2-way RIL by sibling mating, but here were seeing like 10 generations for selfing and 36 generations for sib-mating. 8-way RIL by sibmating take an extra 3 generations, but note that this includes the two extra generations to just get the 8 genomes together.

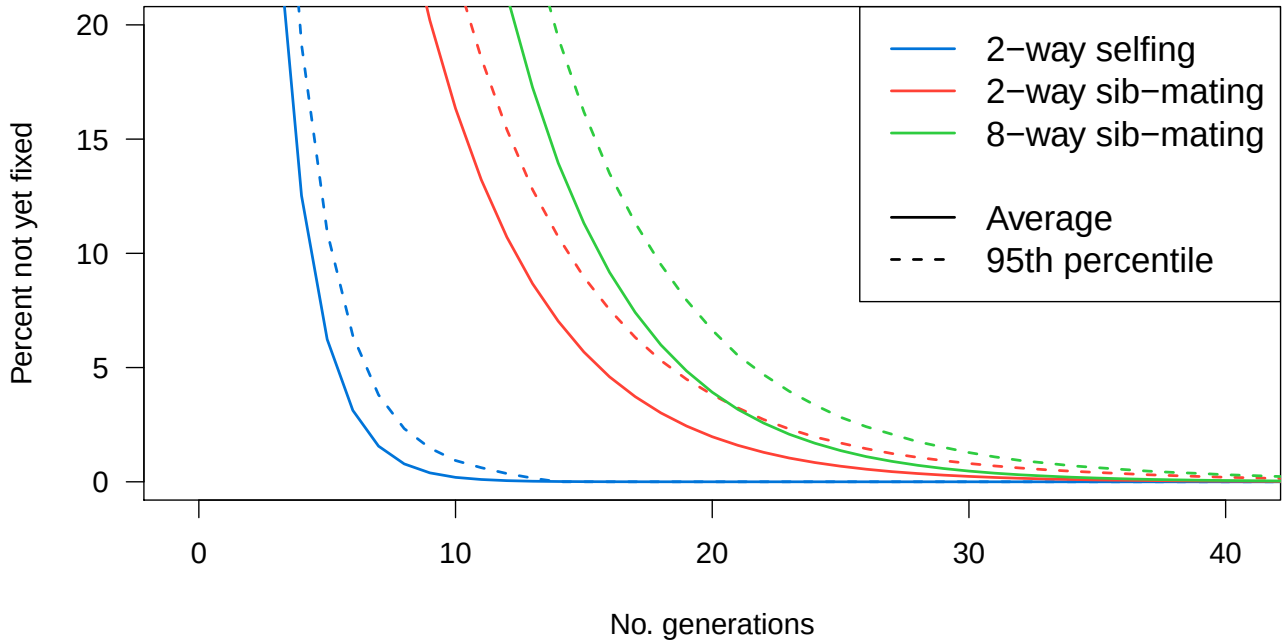
# No. generations to 99% fixation






If we loosen the requirement of complete fixation and just look to see how long it takes to 99% of the genome fixed, we cut off quite a bit, particularly from the sib-mating cases. But still note that while on average it takes 27 generations to get 99% fixation in 8-way RIL by sib-mating, there is a lot of variation; you could could be done in 20 or it could take 32 or so.

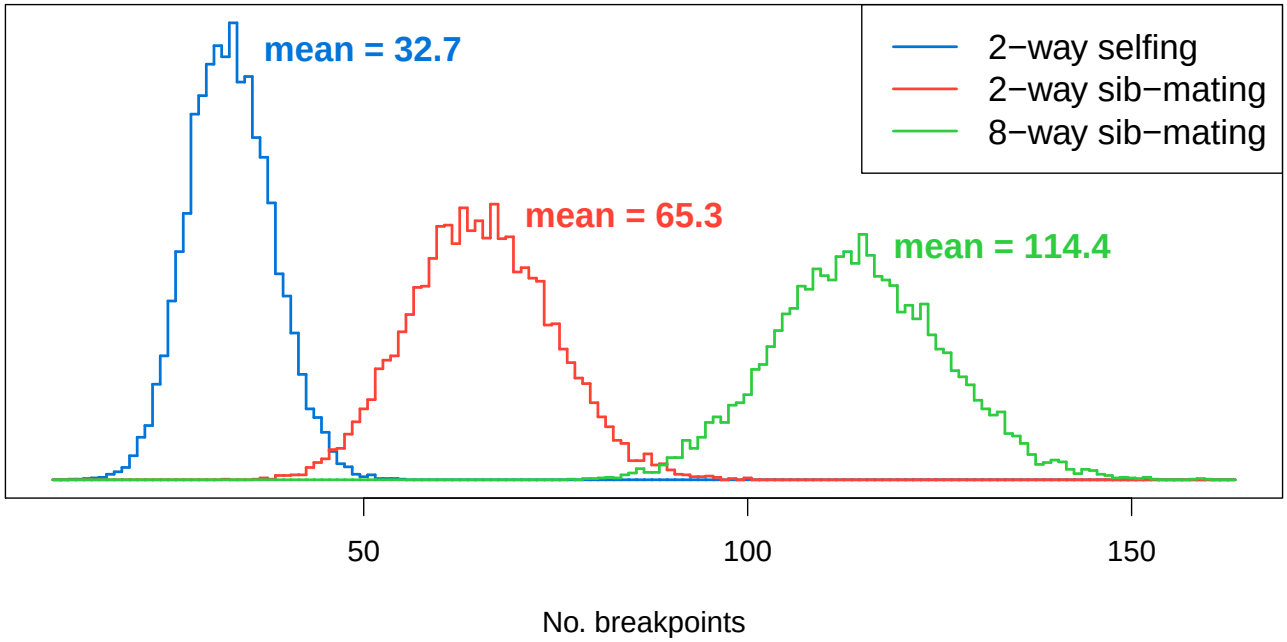
# Percent genome not fixed






What percent of the genome is not fixed? Here we show the average and 95th percentiles be generation; impossible to get this analytically.

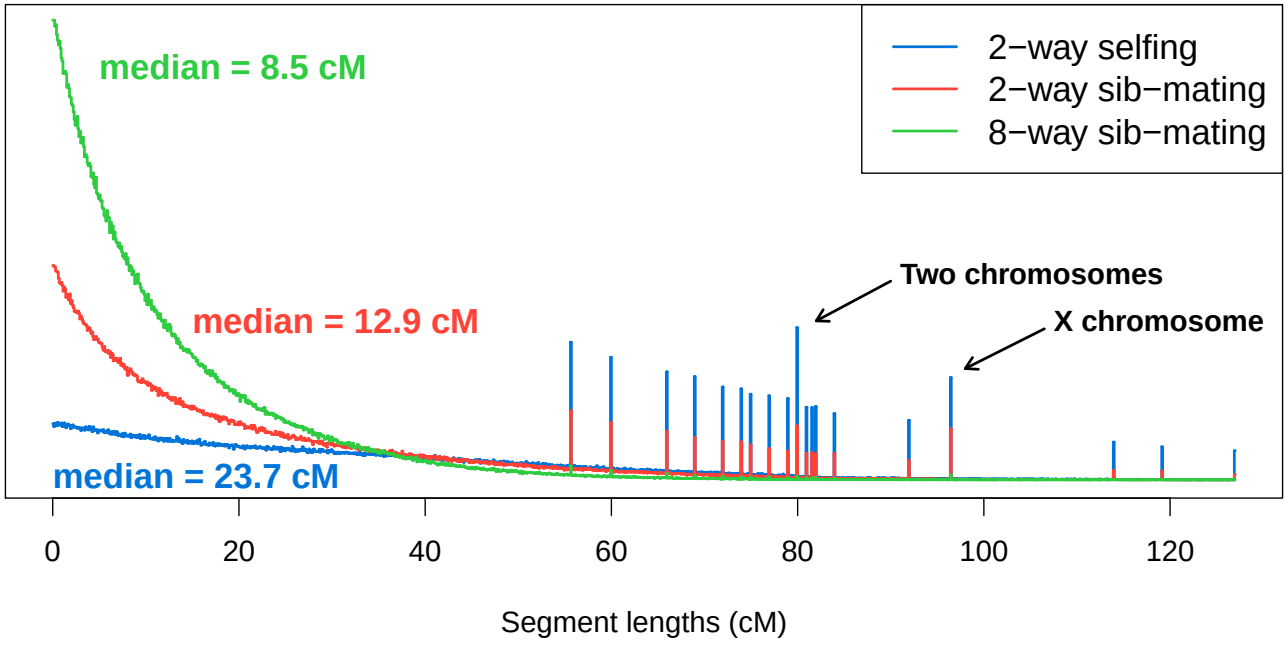
# No. breakpoints






How many crossovers are there, across the genome?

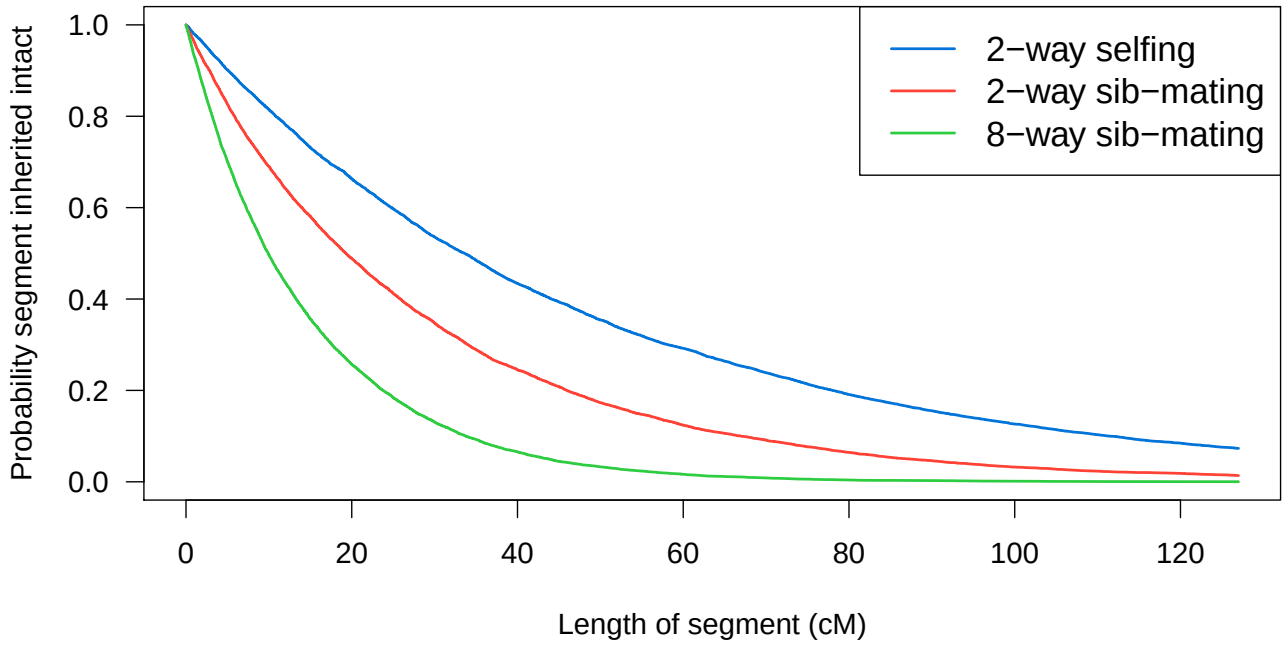
# Segment lengths






How big are the segments between crossovers? Note all of the spikes at whole chromosomes that were inherited intact.

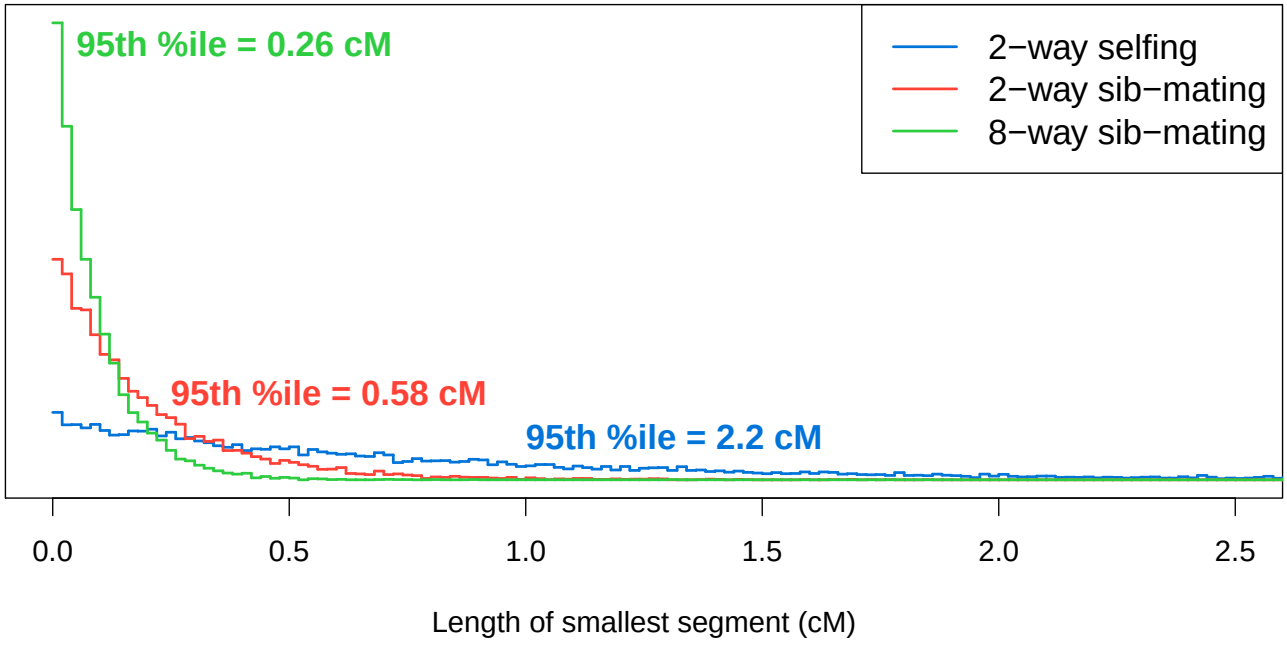
# Probability a segment is inherited intact






How likely is it that a particular segment is inherited intact, as a function of the length of the segment?

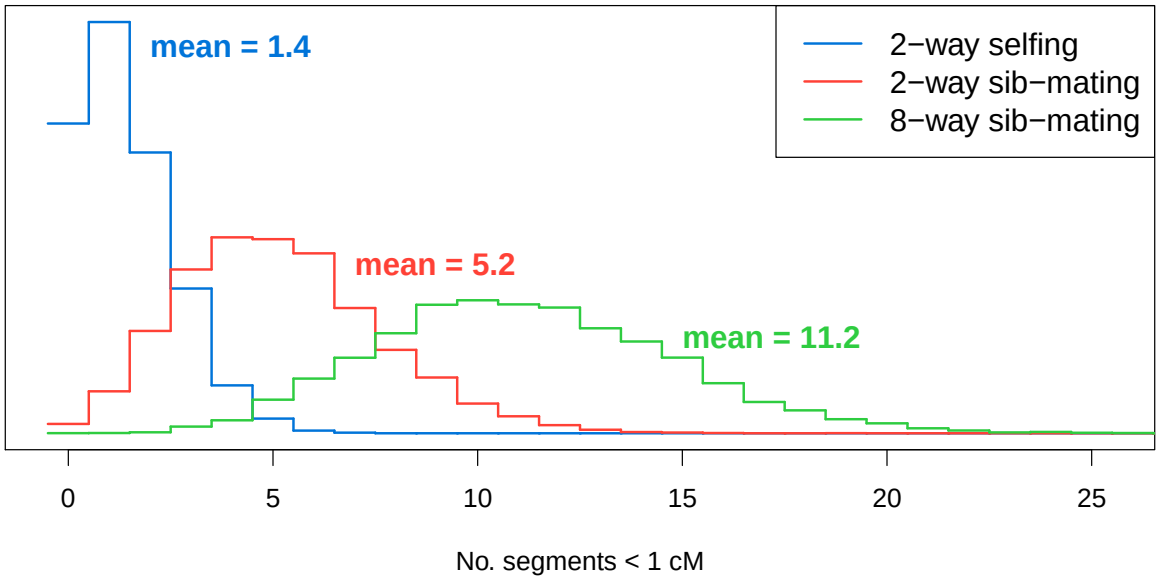
# Length of smallest segment






What is the length of the smallest segment between crossovers, genome-wide?

# No. segments $<$ 1 cM

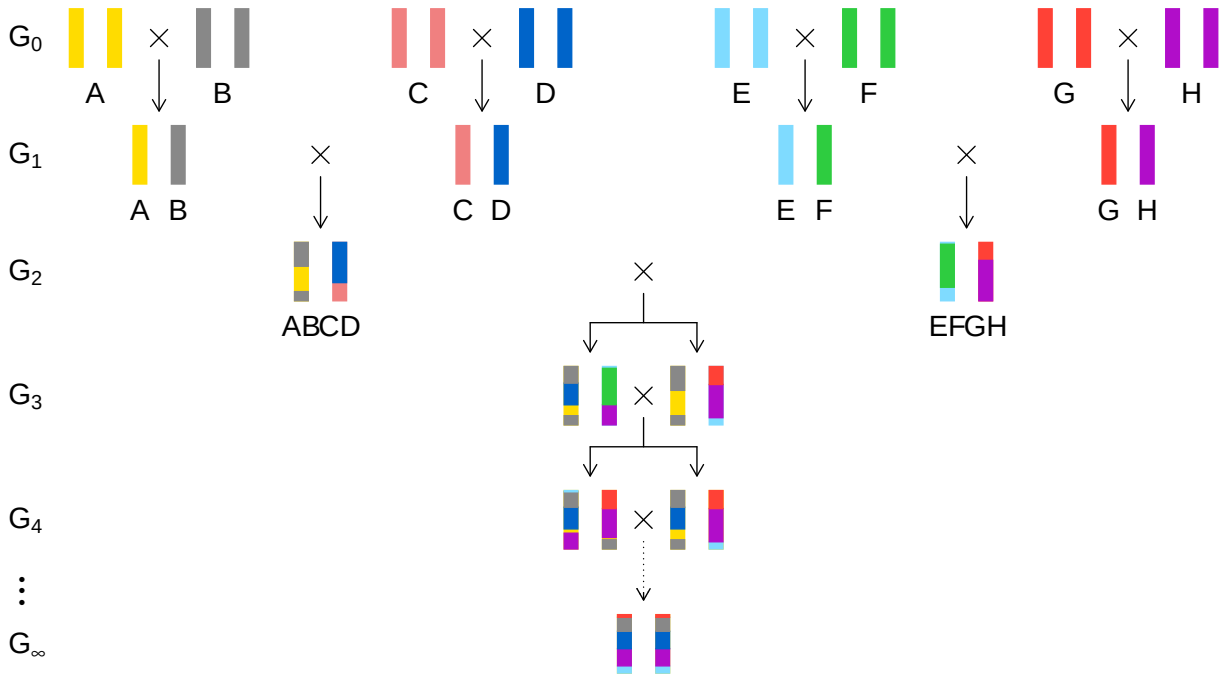





How many inter-breakpoint segments are $<$ 1 cM? An average of 11 in an 8-way RIL genome.

Easy to answer by simulation; impossible to get at any other way.

# Collaborative Cross






Finish this work on the Collaborative Cross genomes, I'd thought I was done with this sort of applied probability. But then the first lines started to be developed, and they were wanting to genotype and phenotype them at intermediate generations, prior to full inbreeding. Which led to the question: what do these lines look like at the intermediate generations?

# The PreCC

What happens at $G_2F_k$?

$\Pr(g_1 = i)$ as a function of k

$\Pr(g_1 = i, g_2 = j)$ as a function of k and the recombination fraction



We've covered the fully fixed chromosome case, but how about after *k* generations of sibling mating? What are the genotype frequencies at a fixed point? What are the transition probabilities from one position to another?

# Crazy table

**Table 4 Two-locus haplotype probabilities at generation $F_k$ in the formation of four-way RIL by sibling mating**

| Chr. | Individual | Prototype | No. states | Probability of each |
|---|---|---|---|---|
| A | Random | *AA* | 4 | $\frac{1}{4(1+6r)}-\left[\frac{6r^2-7r-3rs}{4(1+6r)s}\right]\left(\frac{1-2r+s}{4}\right)^k+\left[\frac{6r^2-7r+3rs}{4(1+6r)s}\right]\left(\frac{1-2r-s}{4}\right)^k$ |
| | | *AB* | 4 | $\frac{r}{2(1+6r)}+\left[\frac{10r^2-r-rs}{4(1+6r)s}\right]\left(\frac{1-2r+s}{4}\right)^k-\left[\frac{10r^2-r+rs}{4(1+6r)s}\right]\left(\frac{1-2r-s}{4}\right)^k$ |
| | | *AC* | 8 | $\frac{r}{2(1+6r)}-\left[\frac{2r^2+3r+rs}{4(1+6r)s}\right]\left(\frac{1-2r+s}{4}\right)^k+\left[\frac{2r^2+3r-rs}{4(1+6r)s}\right]\left(\frac{1-2r-s}{4}\right)^k$ |
| X | Female | *AA* | 2 | $\frac{1}{3(1+4r)}+\frac{1}{6(1+r)}\left(-\frac{1}{2}\right)^k-\left[\frac{4r^3-(4r^2+3r)t+3r^2-5r}{4(4r^2+5r+1)t}\right]\left(\frac{1-r+t}{4}\right)^k+\left[\frac{4r^3+(4r^2+3r)t+3r^2-5r}{4(4r^2+5r+1)t}\right]\left(\frac{1-r-t}{4}\right)^k$ |
| | | *AB* | 2 | $\frac{2r}{3(1+4r)}+\frac{r}{3(1+r)}\left(-\frac{1}{2}\right)^k+\left[\frac{2r^3+6r^2-(2r^2+r)t}{2(4r^2+5r+1)t}\right]\left(\frac{1-r+t}{4}\right)^k-\left[\frac{2r^3+6r^2+(2r^2+r)t}{2(4r^2+5r+1)t}\right]\left(\frac{1-r-t}{4}\right)^k$ |
| | | *AC* | 4 | $\frac{2r}{3(1+4r)}-\frac{r}{6(1+r)}\left(-\frac{1}{2}\right)^k-\left[\frac{9r^2+5r+rt}{4(4r^2+5r+1)t}\right]\left(\frac{1-r+t}{4}\right)^k+\left[\frac{9r^2+5r-rt}{4(4r^2+5r+1)t}\right]\left(\frac{1-r-t}{4}\right)^k$ |
| | | *CC* | 1 | $\frac{1}{3(1+4r)}-\frac{1}{3(1+r)}\left(-\frac{1}{2}\right)^k+\left[\frac{9r^2+5r+rt}{2(4r^2+5r+1)t}\right]\left(\frac{1-r+t}{4}\right)^k-\left[\frac{9r^2+5r-rt}{2(4r^2+5r+1)t}\right]\left(\frac{1-r-t}{4}\right)^k$ |
| X | Male | *AA* | 2 | $\frac{1}{3(1+4r)}-\frac{1}{3(1+r)}\left(-\frac{1}{2}\right)^k+\left[\frac{r^3-(8r^3+r^2-3r)t-10r^2+5r}{2(4r^4-35r^3-29r^2+15r+5)}\right]\left(\frac{1-r+t}{4}\right)^k+\left[\frac{r^3+(8r^3+r^2-3r)t-10r^2+5r}{2(4r^4-35r^3-29r^2+15r+5)}\right]\left(\frac{1-r-t}{4}\right)^k$ |
| | | *AB* | 2 | $\frac{2r}{3(1+4r)}-\frac{2r}{3(1+r)}\left(-\frac{1}{2}\right)^k+\left[\frac{r^4+(5r^3-r)t-10r^3+5r^2}{4r^4-35r^3-29r^2+15r+5}\right]\left(\frac{1-r+t}{4}\right)^k+\left[\frac{r^4-(5r^3-r)t-10r^3+5r^2}{4r^4-35r^3-29r^2+15r+5}\right]\left(\frac{1-r-t}{4}\right)^k$ |
| | | *AC* | 4 | $\frac{2r}{3(1+4r)}+\frac{r}{3(1+r)}\left(-\frac{1}{2}\right)^k-\left[\frac{2r^4+(2r^3-r^2+r)t-19r^3+5r}{2(4r^4-35r^3-29r^2+15r+5)}\right]\left(\frac{1-r+t}{4}\right)^k-\left[\frac{2r^4-(2r^3-r^2+r)t-19r^3+5r}{2(4r^4-35r^3-29r^2+15r+5)}\right]\left(\frac{1-r-t}{4}\right)^k$ |
| | | *CC* | 1 | $\frac{1}{3(1+4r)}+\frac{2}{3(1+r)}\left(-\frac{1}{2}\right)^k+\left[\frac{2r^4+(2r^3-r^2+r)t-19r^3+5r}{4r^4-35r^3-29r^2+15r+5}\right]\left(\frac{1-r+t}{4}\right)^k+\left[\frac{2r^4-(2r^3-r^2+r)t-19r^3+5r}{4r^4-35r^3-29r^2+15r+5}\right]\left(\frac{1-r-t}{4}\right)^k$ |

$s=\sqrt{4r^2-12r+5}$ and $t=\sqrt{r^2-10r+5}$; the autosomal haplotype probabilities are valid for $r<\frac{1}{2}$.



So I was pulled back into this sort of investigation again, and while I wasn't able to solve all of the things I wanted, I was able to characterize the haplotype probabilities at these intermediate generations, which led to yet another rather crazy paper with this especially crazy table of results. (Just try and re-type this table, copy editor; I dare you!)

## Lesson

Computer simulations are hugely valuable.



So I'd lost track of my the central theme several times in this, but the key lesson to draw from this is: computer simulations are hugely valuable for data scientists.

# Uses of simulations

- Study probabilities
- Estimate power/sample size
- Evaluate performance of a method
- Evaluate sensitivity/robustness of a method



Computer simulations have lots of uses: to study probabilities (as here), but also to estimate power or sample size, to evaluate how statistical methods perform, and to evaluate the sensitivity or robustness of a method.

Are you're concerned about whether your method will work well if the data have errors or biases? Well, simulate data like you fear, and see try it out!

# Relative advantages?

- Simulations
- Numerical calculations
- Analytic calculations



In this work on recombinant inbred lines, I've used a combination of simulations, numerical calculations and analytic calculations.

Analytic calculations can be the most satisfying, and can give you generally useful closed-form solutions. But you usually have to make a lot of simplifying assumptions, and it can be really hard work.

Numerical calculations can have precision that approaches exact results, and can at times be easier to achieve and applicable in more complex cases.

But computer simulations, while being potentially noisy and having to repeat everything for a variety of different cases, are often easy to get going and can allow consideration of fully realistic situations and to investigate complex outcomes.

When it doubt, simulate!

# References


- Haldane & Waddington (1931) Inbreeding and Linkage. Genetics 16:357–374
- Broman KW (2005) The genomes of recombinant inbred lines. Genetics 169:1133–1146
- Teuscher & Broman (2007) Haplotype probabilities for multiple-strain recombinant inbred lines. Genetics 175:1267–1274
- Broman KW (2012) Genotype probabilities at intermediate generations in the construction of recombinant inbred lines. Genetics 190:403–412
- Broman KW (2012) Haplotype probabilities in advanced intercross populations. G3 2:199–202




Here are a bunch of references for the work I've presented.