MAGIC design

and other topics

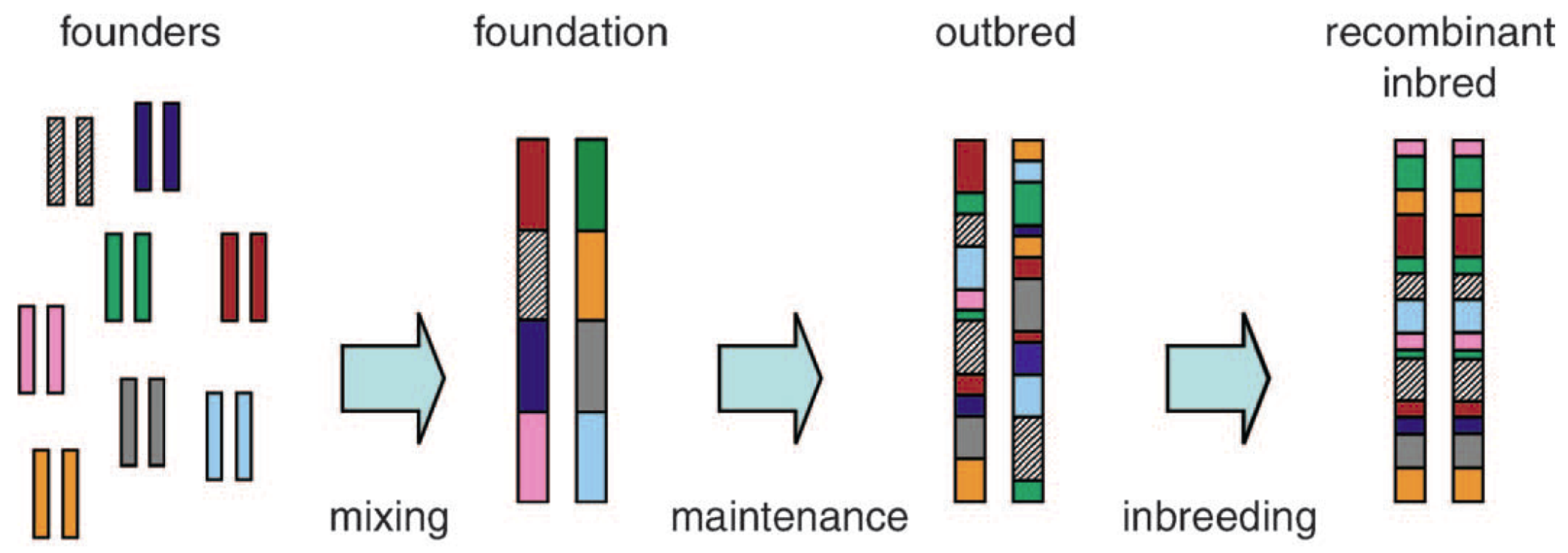
Karl Broman

Biostatistics & Medical Informatics
University of Wisconsin – Madison

`biostat.wisc.edu/ kbroman`
`github.com/kbroman`
`kbroman.wordpress.com`
`@kwbroman`

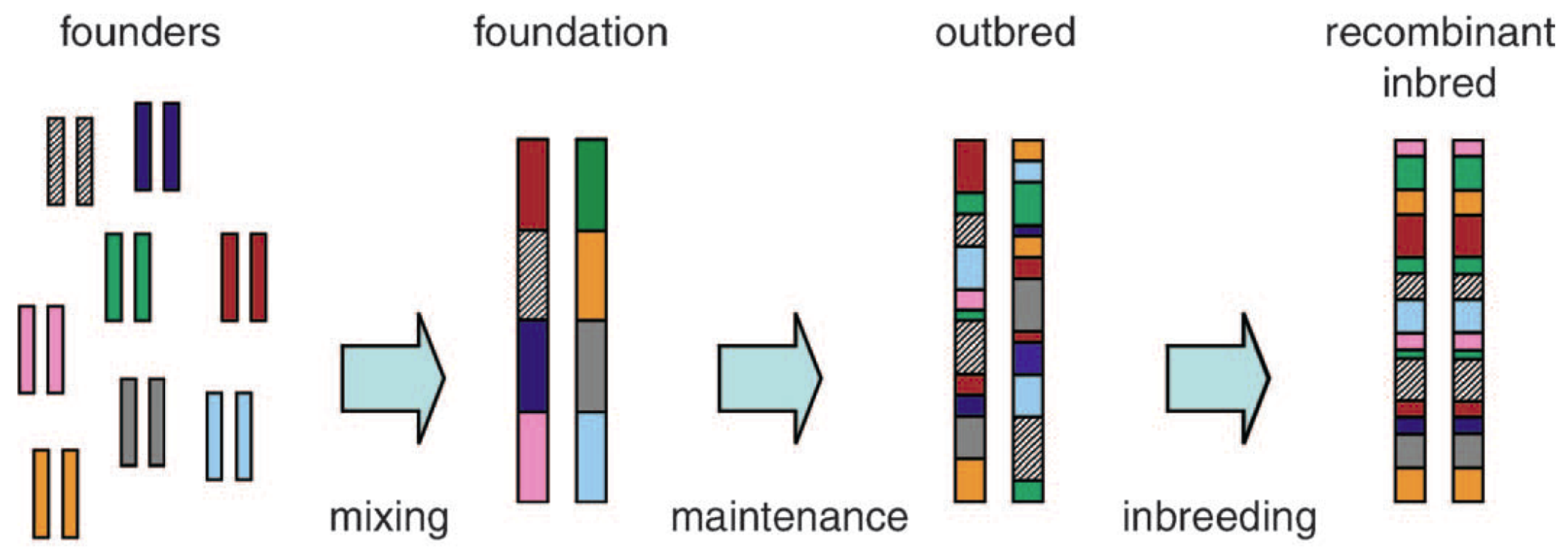
CC founders

MAGIC lines



Valdar et al., Genetics 172:1783, 2006

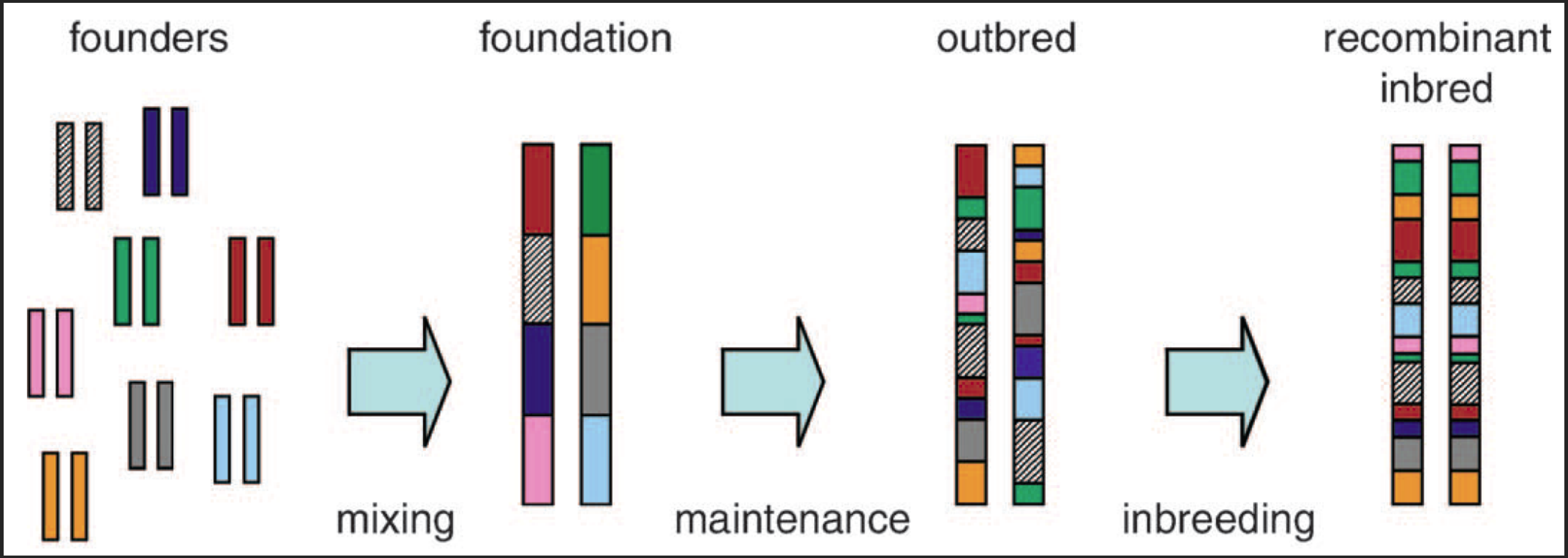
MAGIC lines



combine mix fix

Valdar et al., Genetics 172:1783, 2006

MAGIC lines

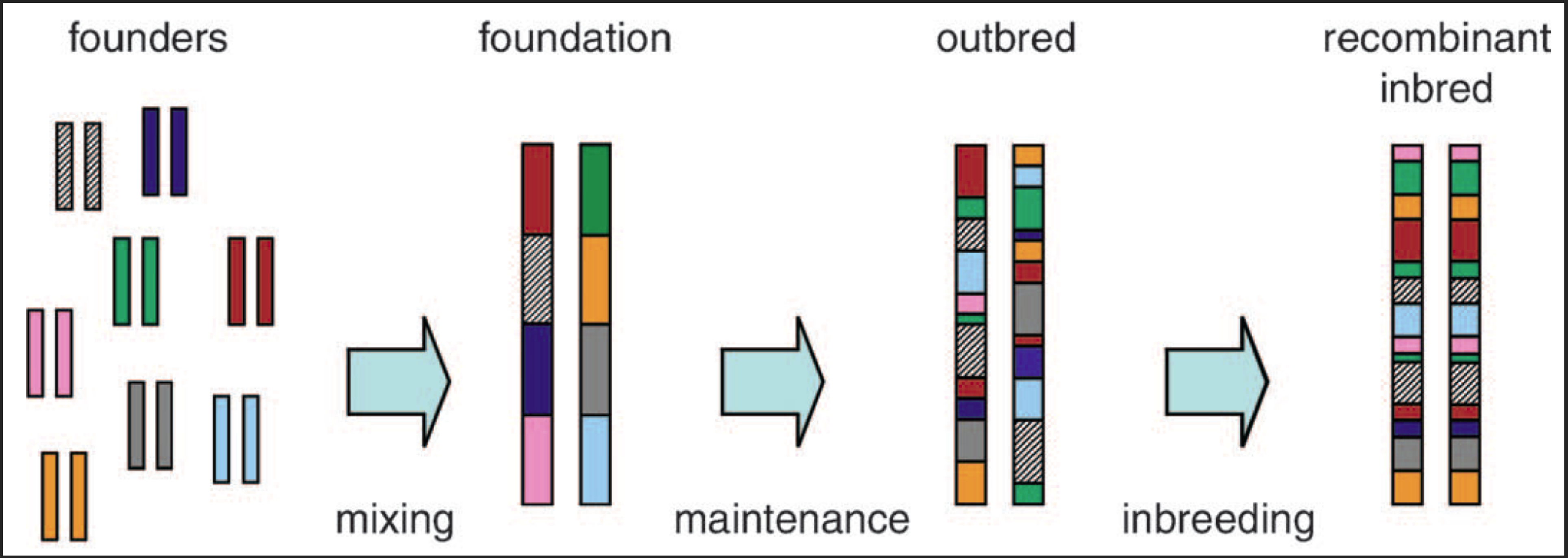


combine mix fix

How many?

Valdar et al., Genetics 172:1783, 2006

MAGIC lines



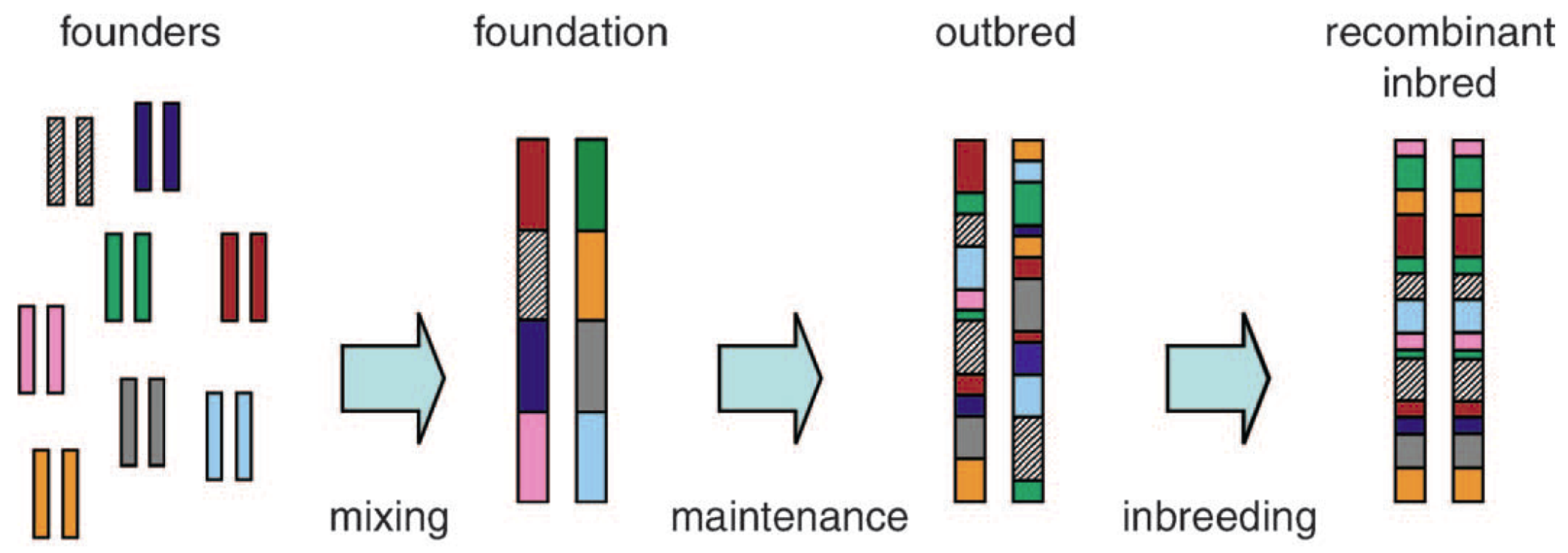
combine mix fix

How many?

Which?

Valdar et al., Genetics 172:1783, 2006

MAGIC lines



combine

mix

fix

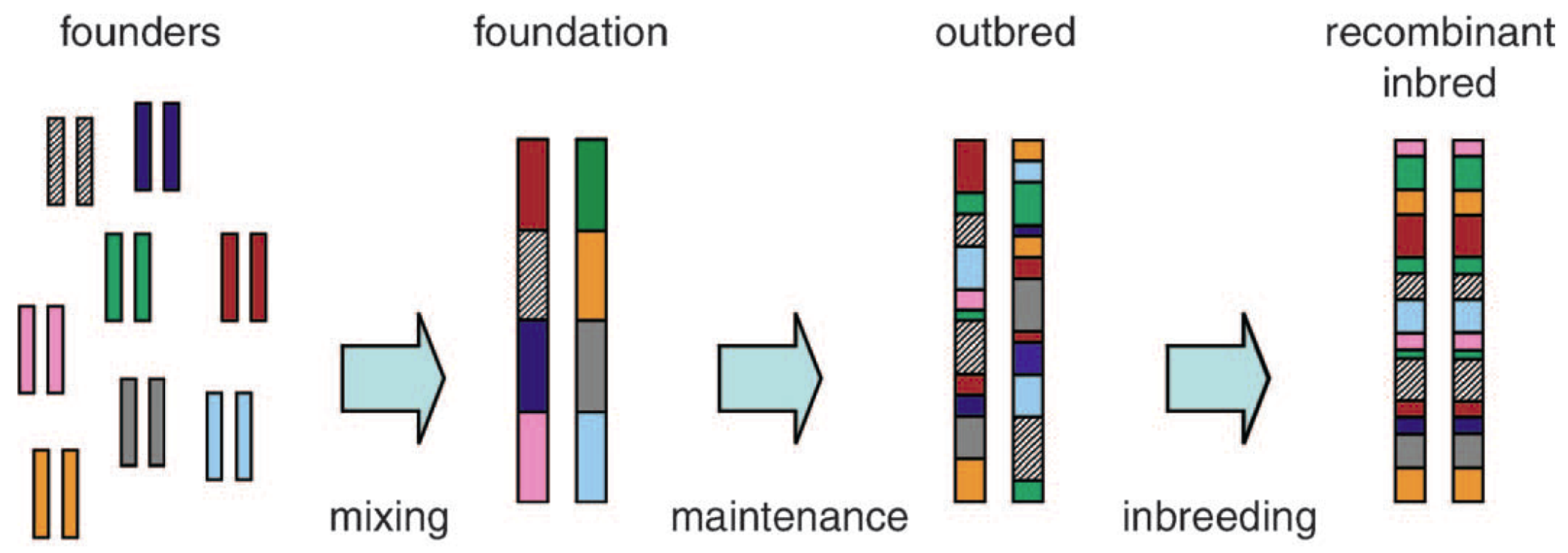
How many?

Which?

How long?

Valdar et al., Genetics 172:1783, 2006

MAGIC lines



combine

mix

fix

How many?

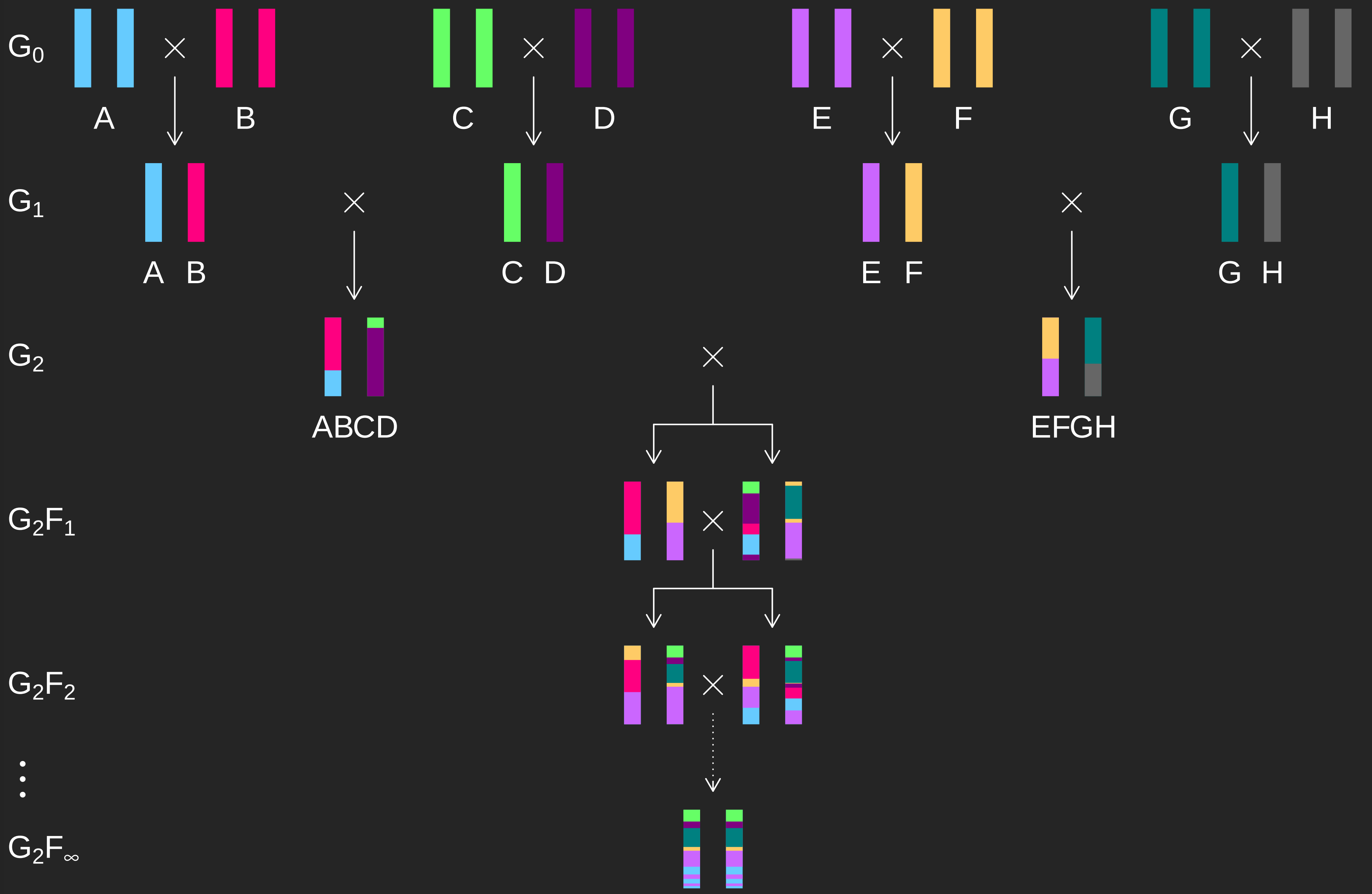
How long?

How?

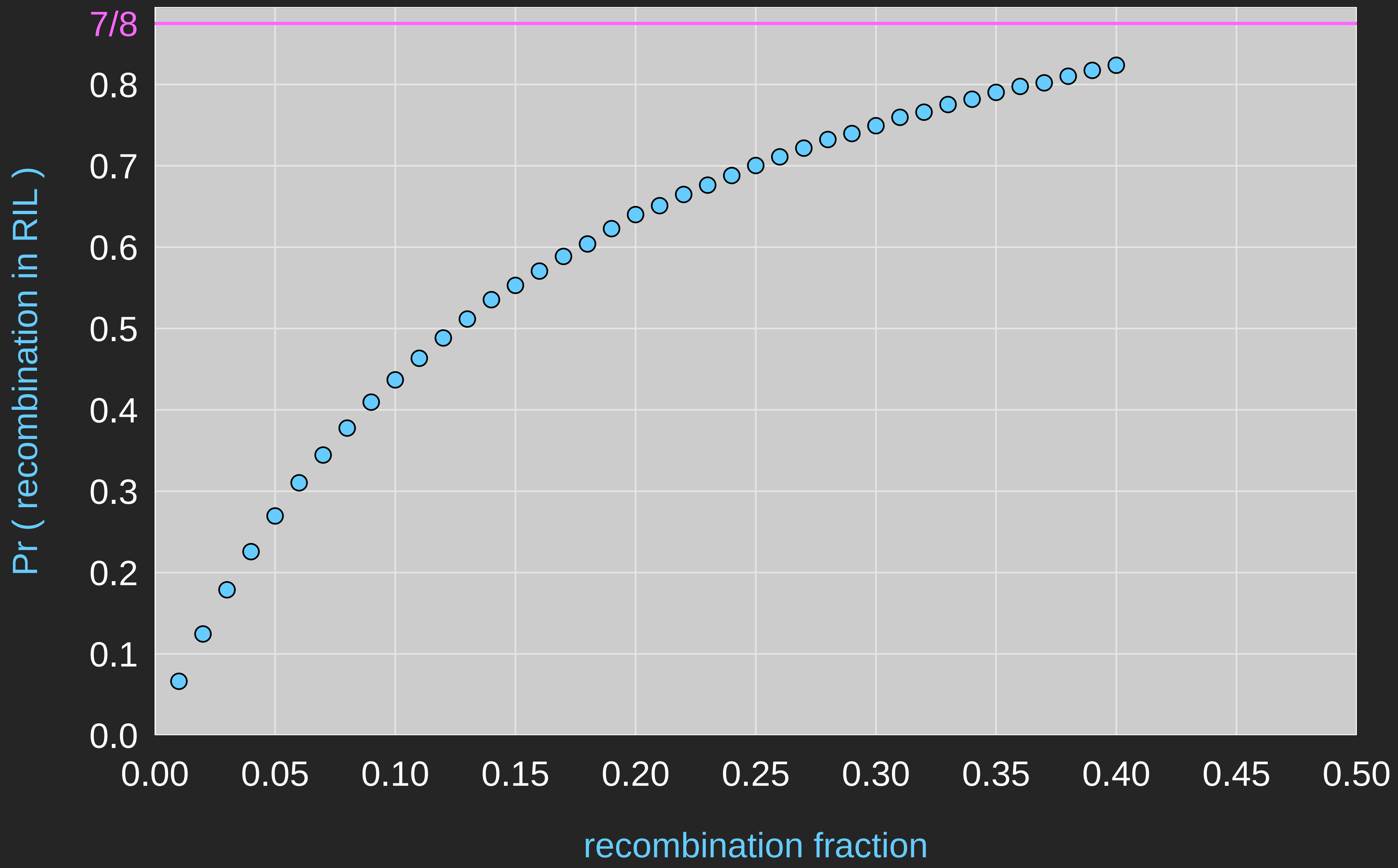
Which?

Valdar et al., Genetics 172:1783, 2006

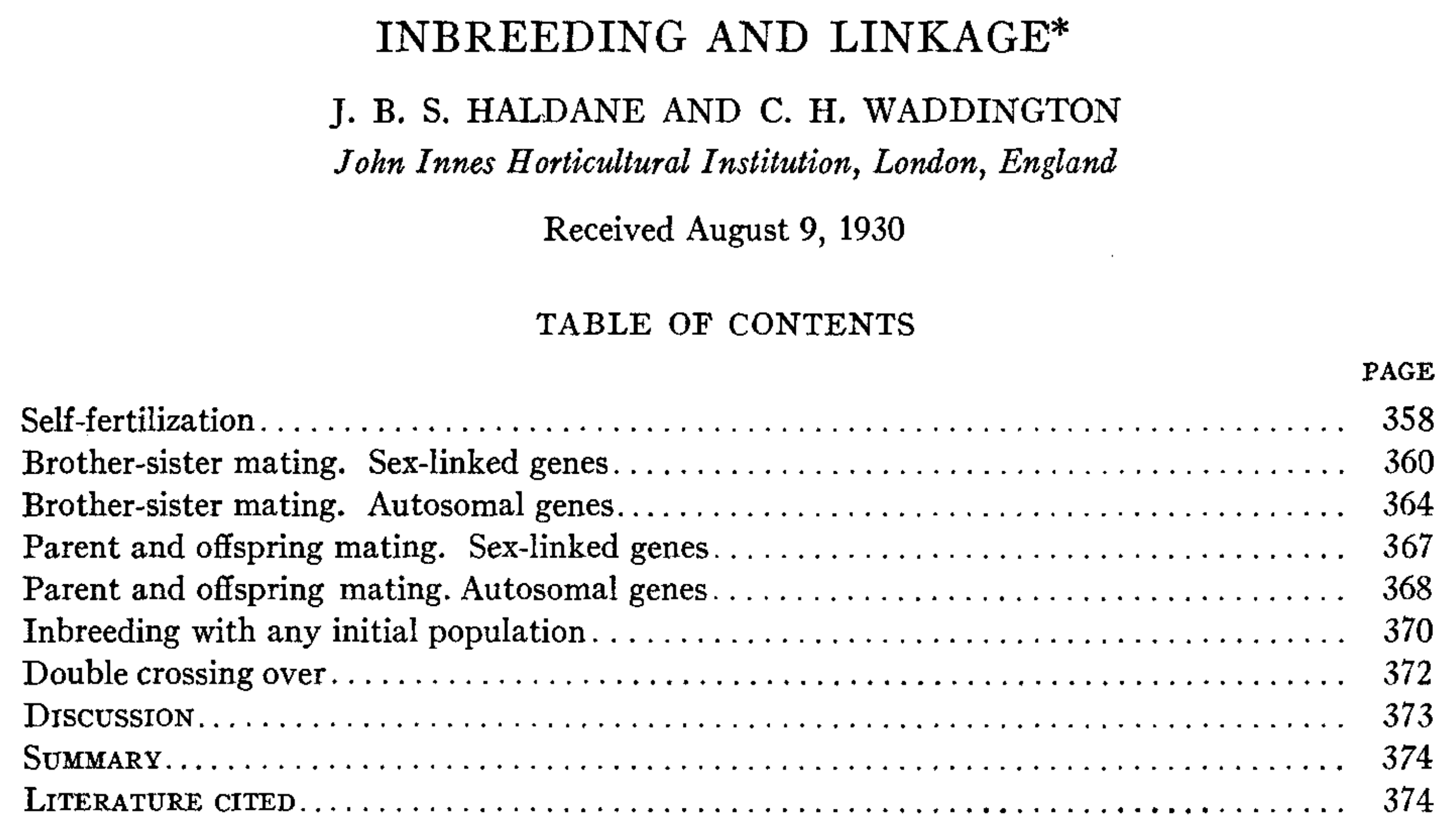
But first . . .



Simulation results



Haldane & Waddington 1931



INBREEDING AND LINKAGE*

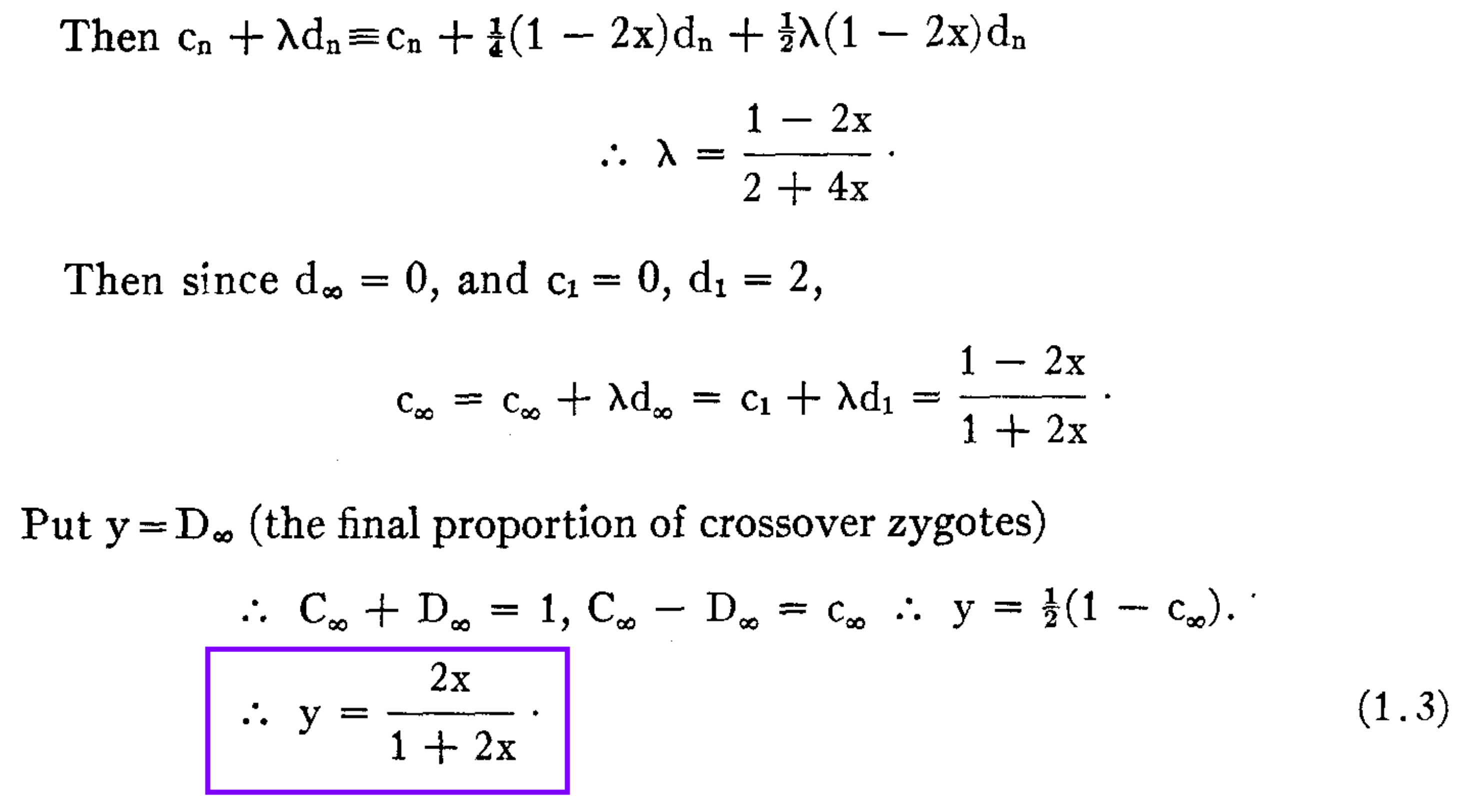
J. B. S. HALDANE AND C. H. WADDINGTON
John Innes Horticultural Institution, London, England

Received August 9, 1930

TABLE OF CONTENTS

Haldane & Waddington, Genetics 16:357, 1931

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 $c_1 = 0$, $d_1 = 2$,

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

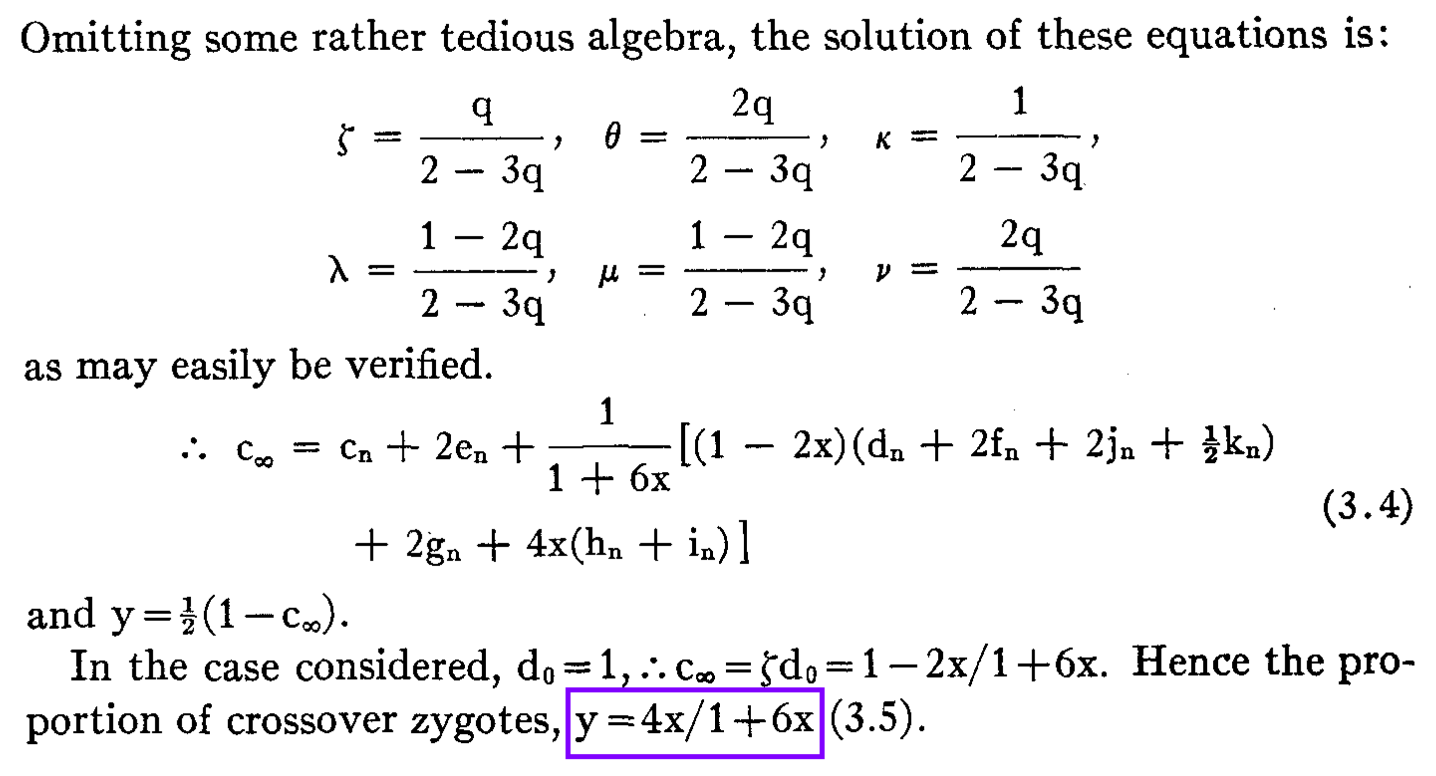
Put $y = D_\infty$ (the final proportion of crossover zygotes)

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

$$\boxed{\therefore y = \frac{2x}{1 + 2x}.} \tag{1.3}$$

Haldane & Waddington, Genetics 16:357, 1931

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.

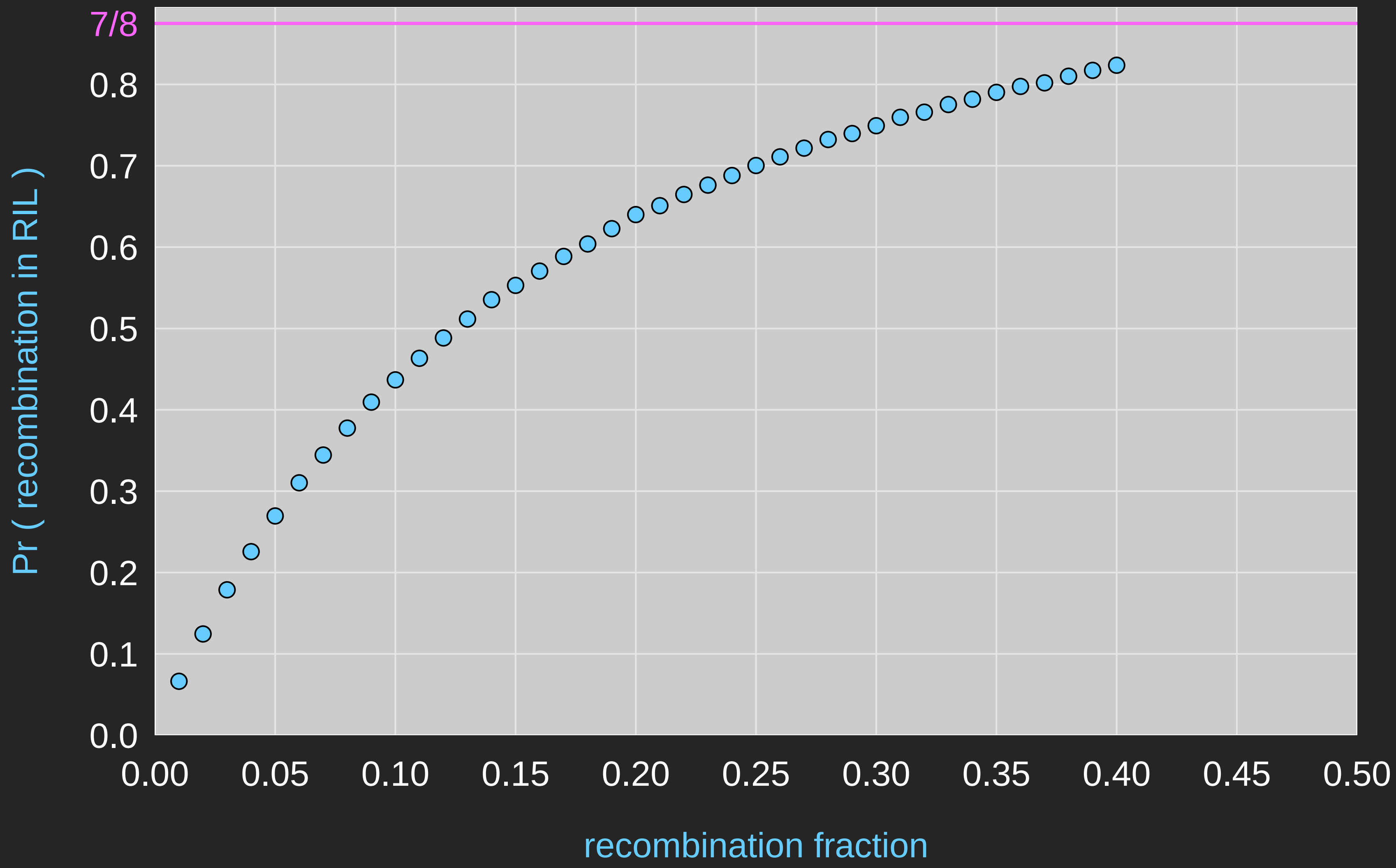
$$\therefore \; c_\infty = c_n + 2e_n + \frac{1}{1+6x}\left[(1-2x)(d_n + 2f_n + 2j_n + \tfrac{1}{2}k_n) + 2g_n + 4x(h_n + i_n)\right] \qquad (3.4)$$

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

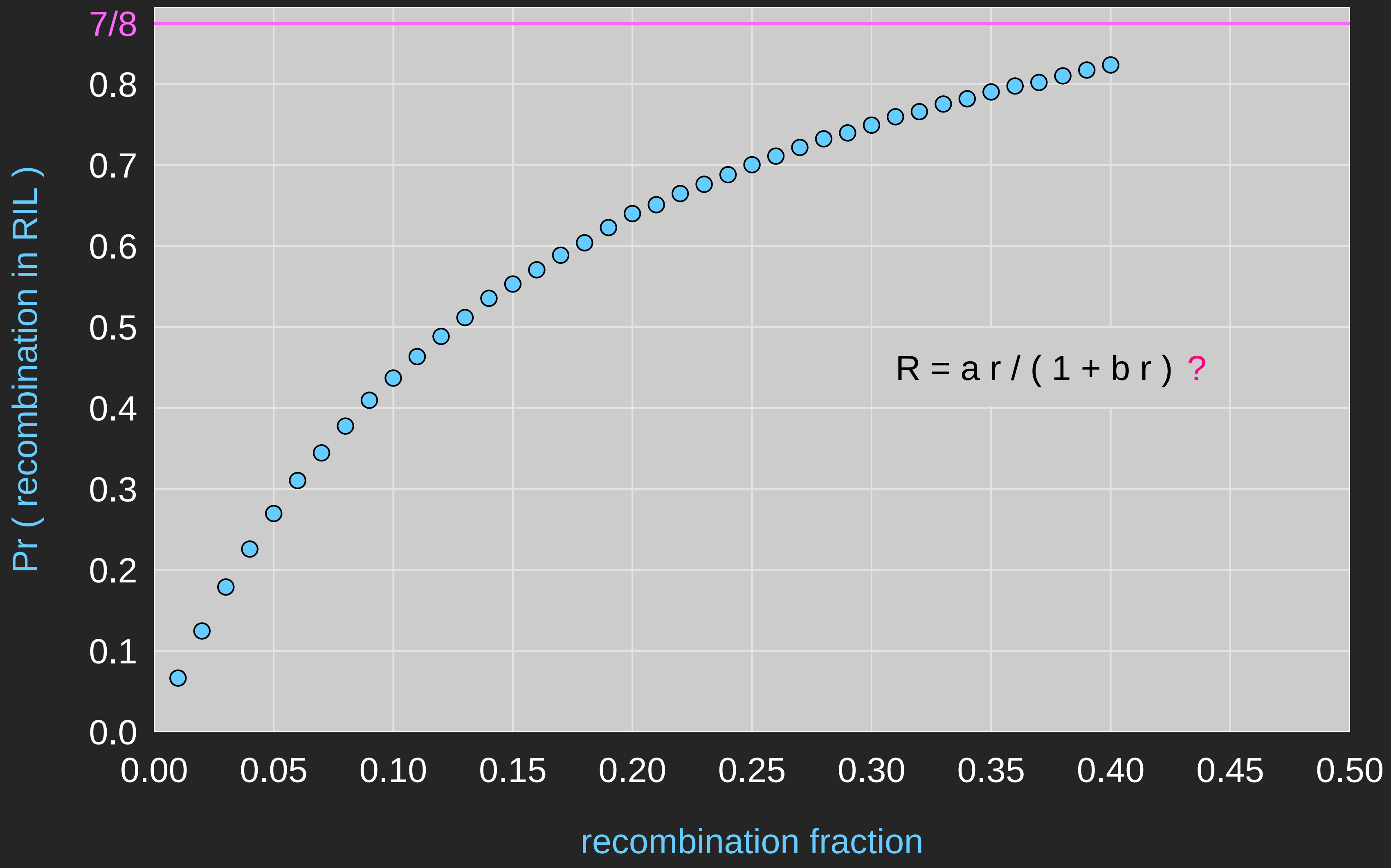
In the case considered, $d_0 = 1$, $\therefore c_\infty = \zeta d_0 = 1 - 2x/1+6x$. Hence the proportion of crossover zygotes, $\boxed{y = 4x/1+6x}$ (3.5).

Haldane & Waddington, Genetics 16:357, 1931

Simulation results



Simulation results



$R = a\,r / (1 + b\,r)$?

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)
```

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)
```

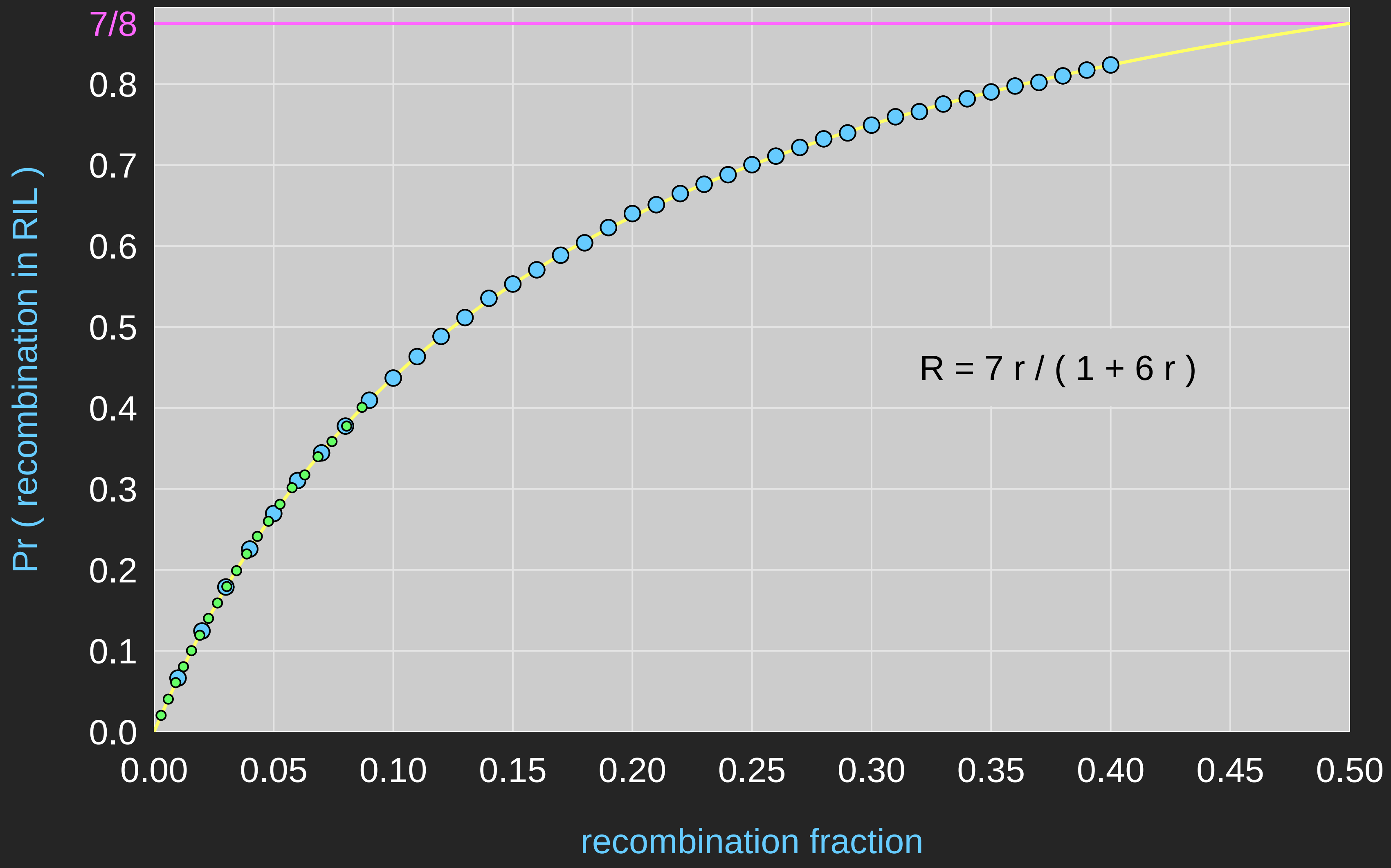
```
  Estimate  Std. Error
a    7.016       0.011
b    6.023       0.016
```

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
```

Simulation results



3-point coincidence

- r_{ij} = recombination fraction for interval (i, j)

  Assume $r_{12} = r_{23} = r$.

- Coincidence = $c = \Pr(\text{double recombinant})/r^2$

  $= \Pr(\text{rec'n in 23} \mid \text{rec'n in 12})/\Pr(\text{rec'n in 12})$

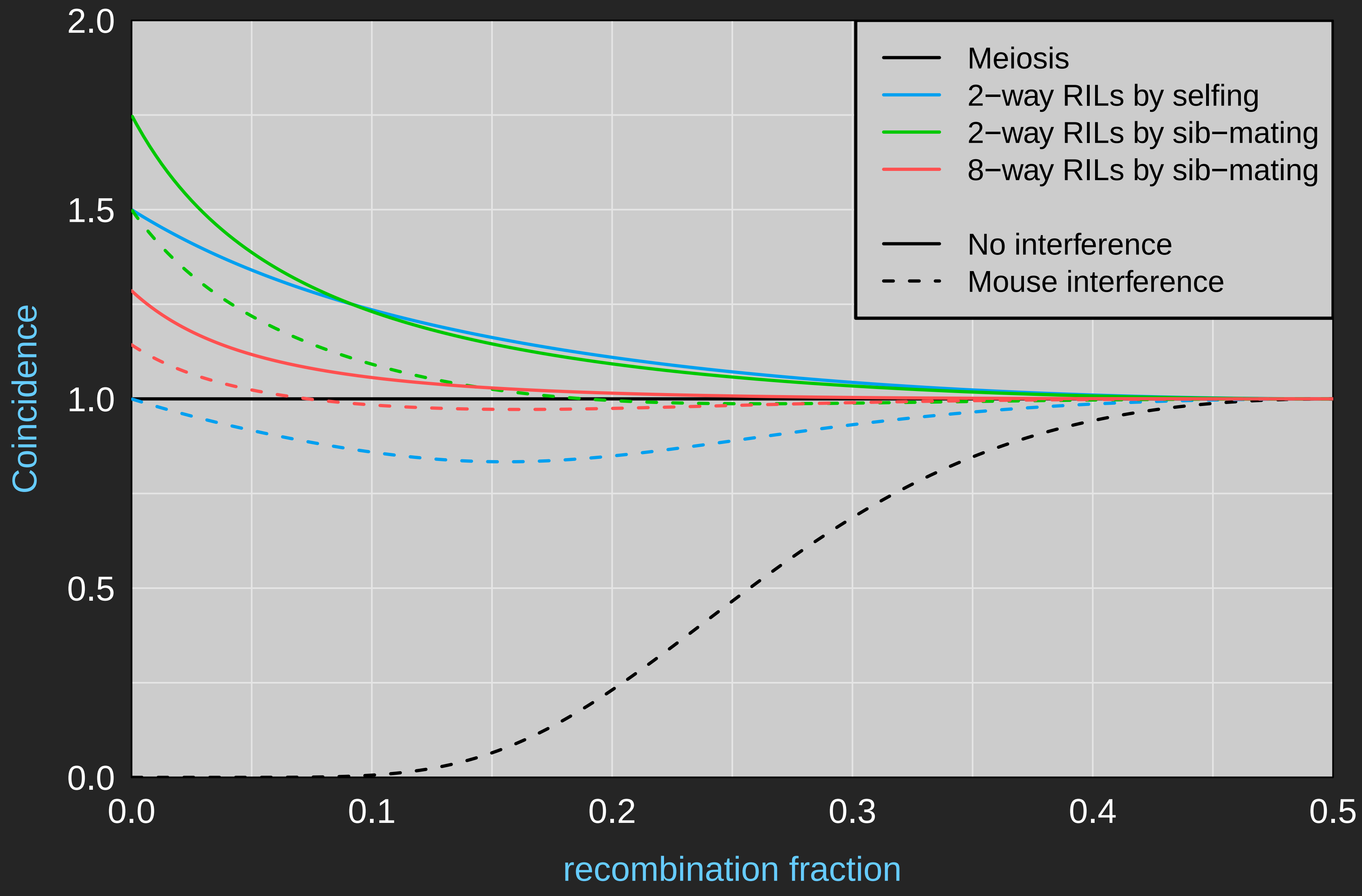
- No interference = 1

  Positive interference < 1

  Negative interference > 1

- Generally c is a function of r

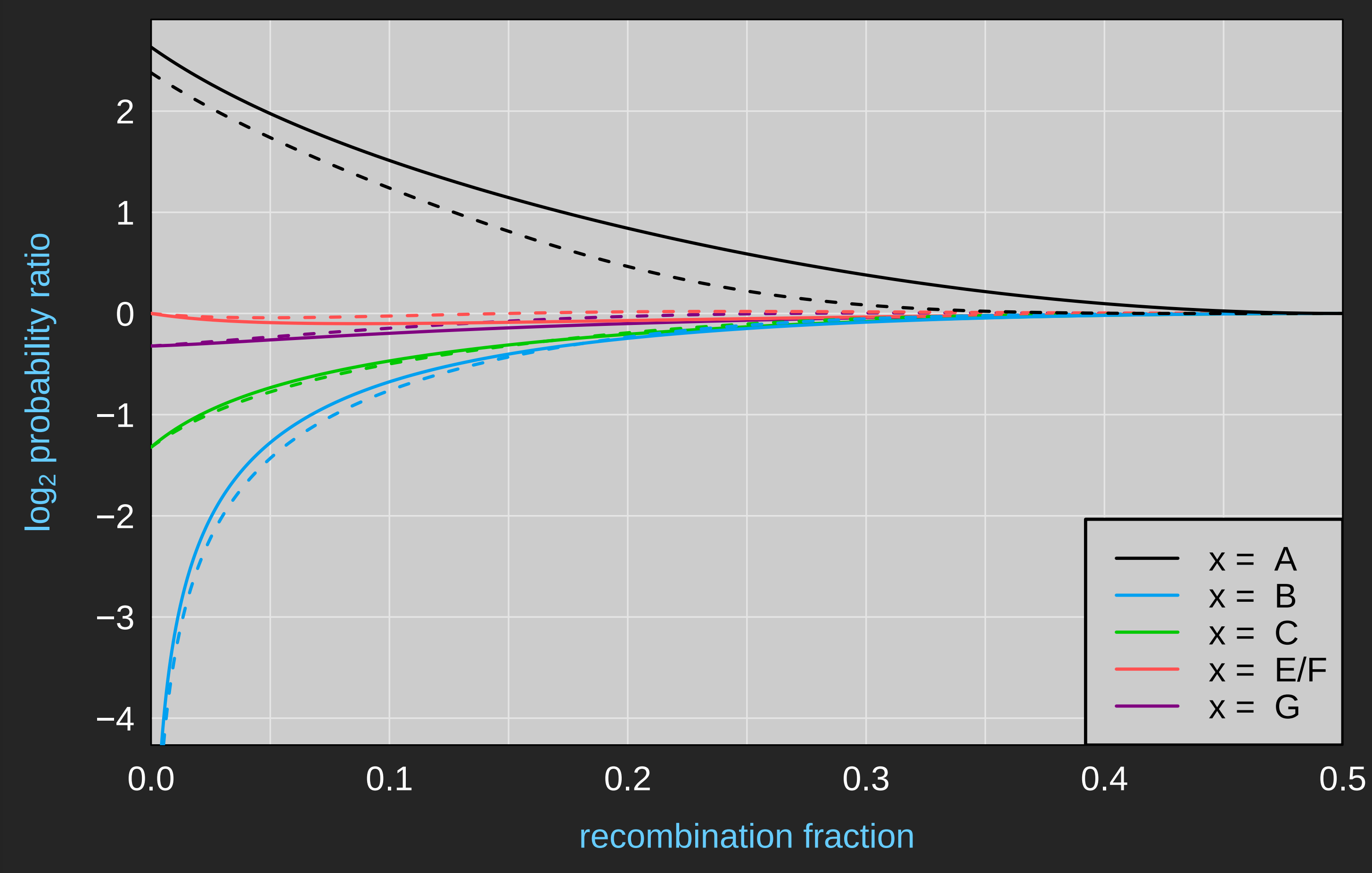
Coincidence function



Broman, Genetics 169:1133, 2005

non-Markov property

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



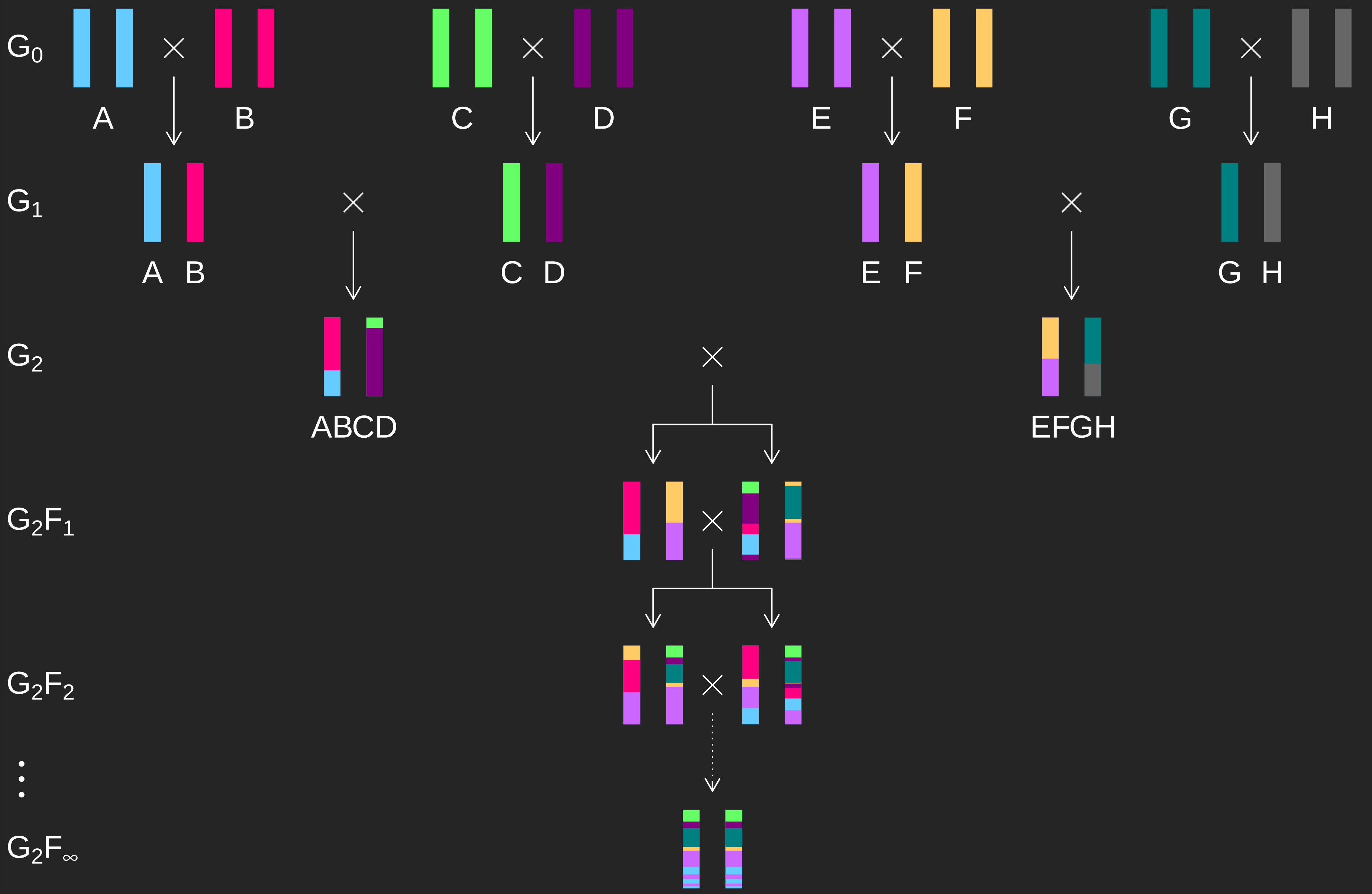
Broman, Genetics 169:1133, 2005

Coincidence formula

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

Teuscher & Broman, Genetics 175:1267, 2007

The CC again



Crazy table

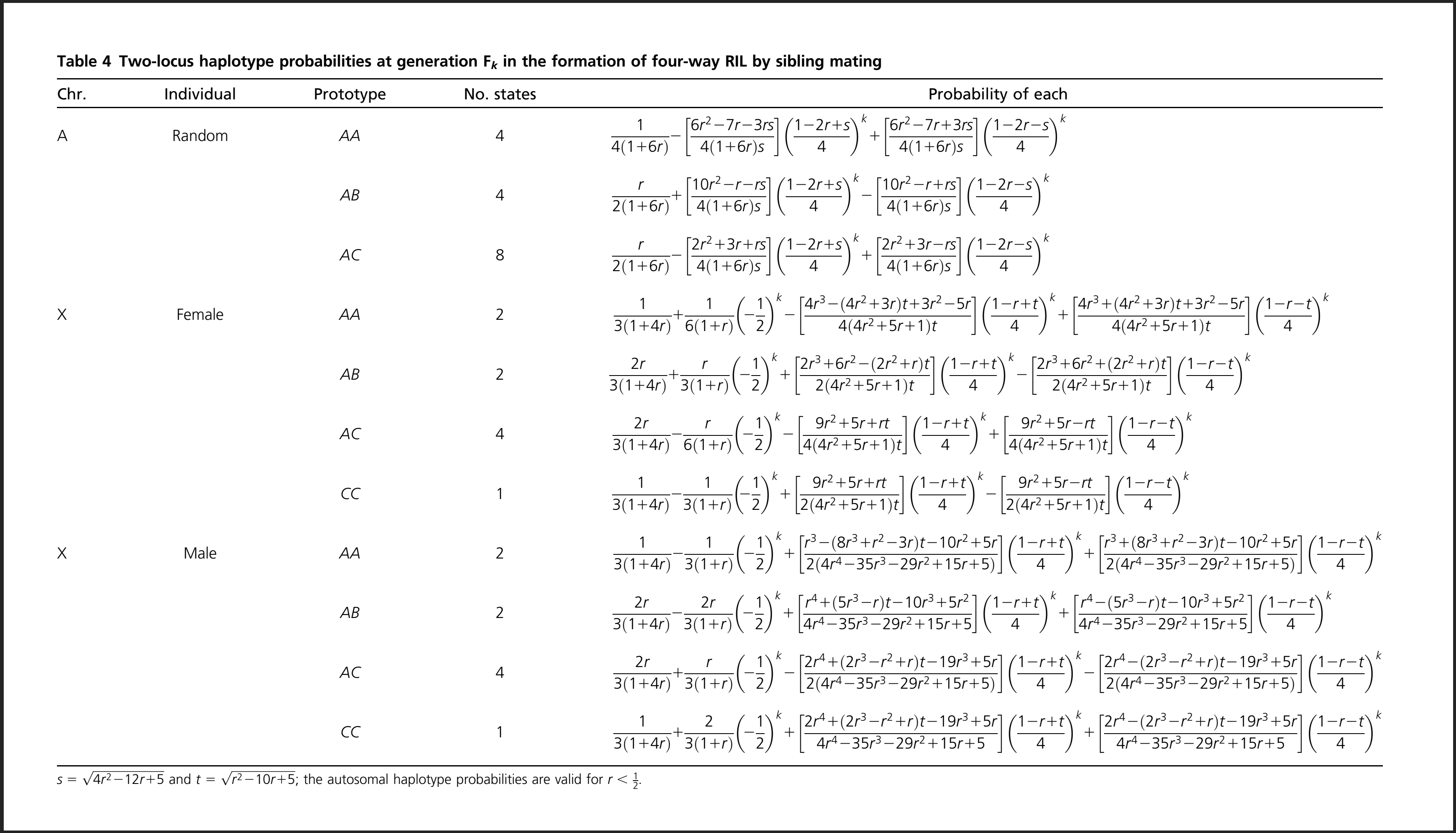


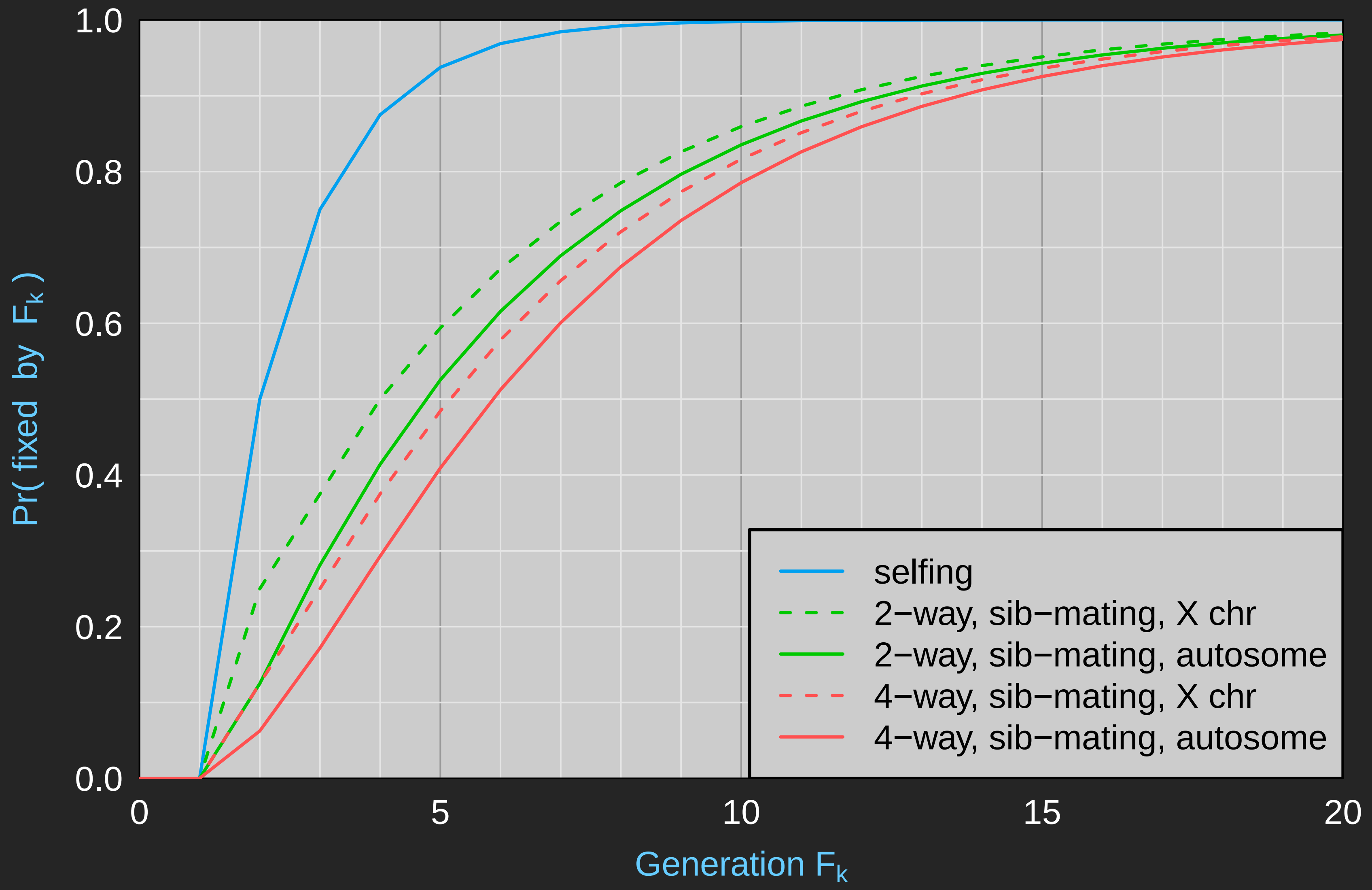
Table 4 Two-locus haplotype probabilities at generation F_k in the formation of four-way RIL by sibling mating

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 $r<\frac{1}{2}$.

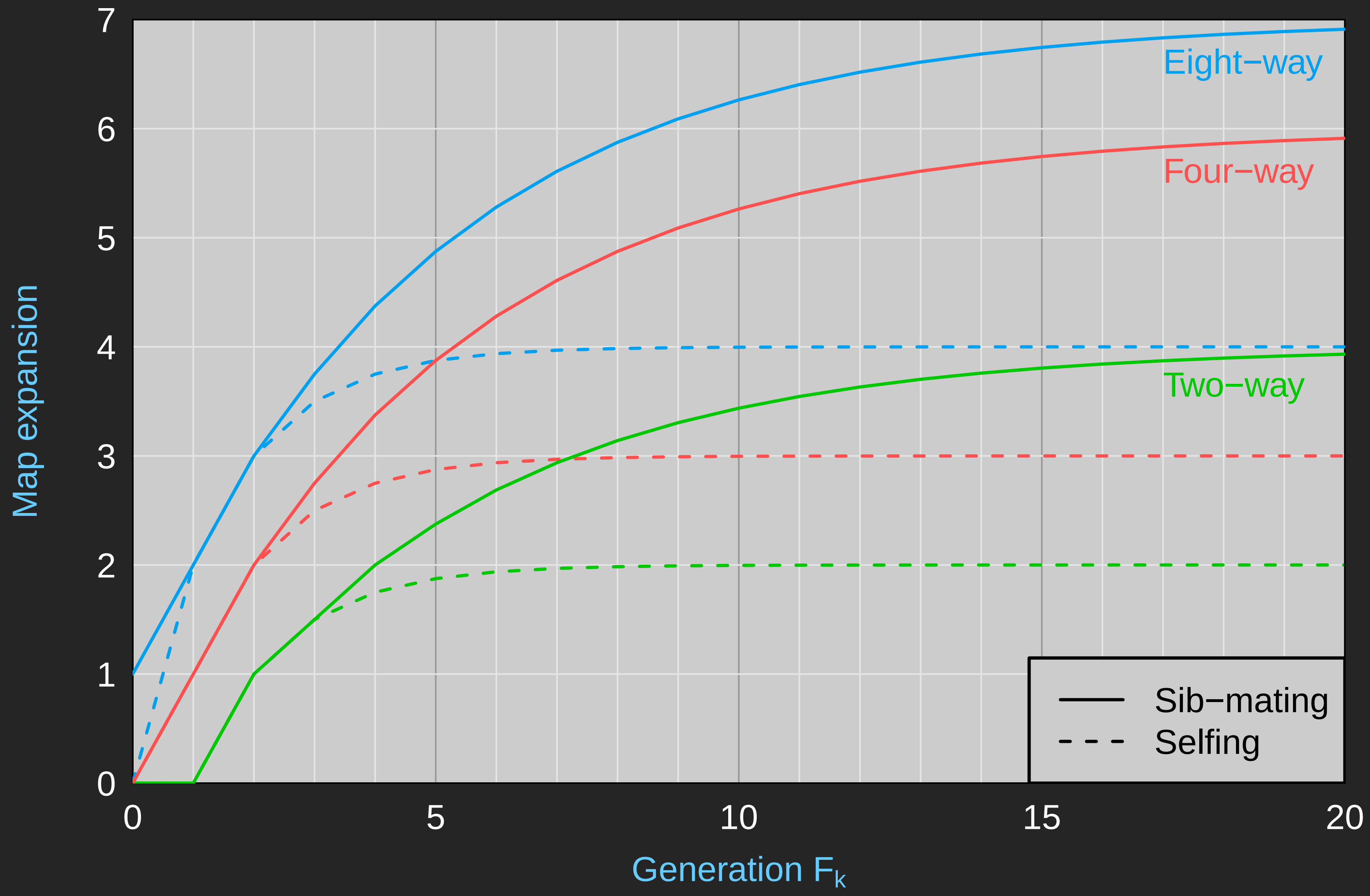
Broman, Genetics 190:403, 2012

Time to fixation



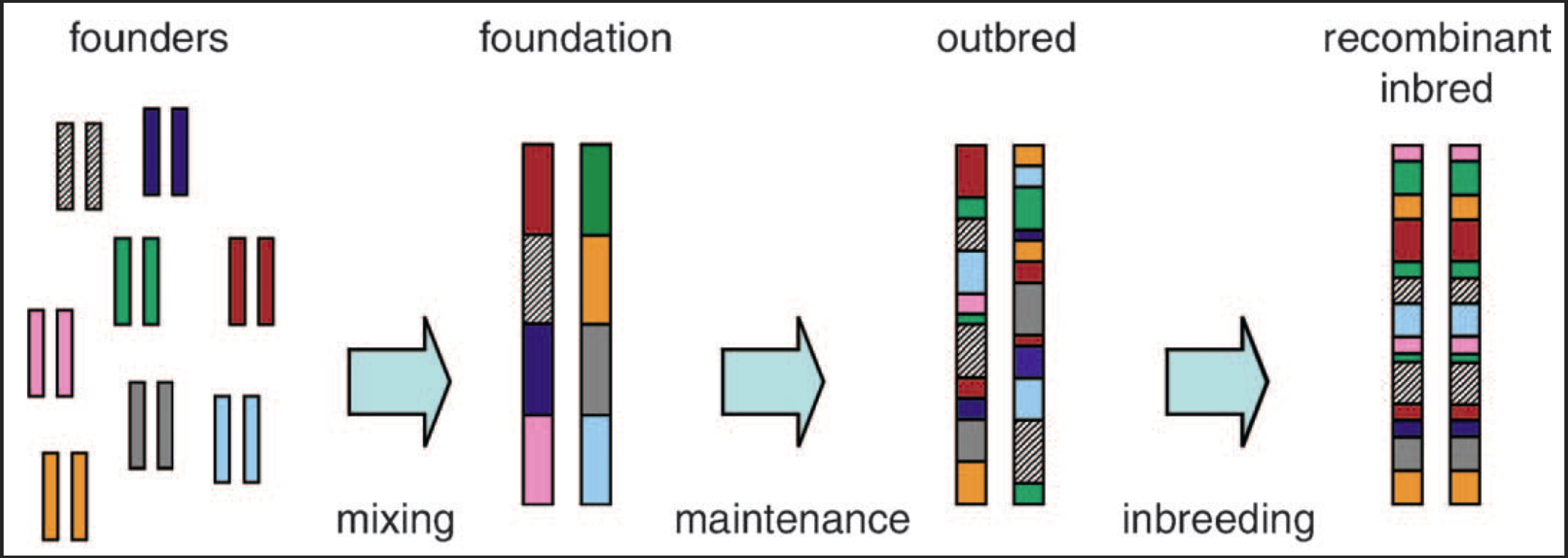
Broman, Genetics 190:403, 2012

Map expansion



Broman, Genetics 190:403, 2012

MAGIC lines



combine

mix

fix

How many?

How long?

How?

Which?

Valdar et al., Genetics 172:1783, 2006

MAGIC is magic

- Genetic diversity
- High-precision mapping
- Predictable linkage disequilibrium
- Phenotype replicates to reduce individual variation
- Pool phenotypes from multiple labs, environments, treatments
- Genotype once

MAGIC is magic

- Genetic diversity
- High-precision mapping
- Predictable linkage disequilibrium
- Phenotype replicates to reduce individual variation
- Pool phenotypes from multiple labs, environments, treatments
- Genotype once
- Cool name

The goal

Identify QTL

- Power
- Mapping precision

The goal

Identify QTG

- Power
- Mapping precision

The goal

Identify QTG

- Power
- Mapping precision
- Estimate QTL allele frequencies

Principles

- Avoid population structure
- Tradeoff between power for *de novo* discovery and mapping precision
- More QTL to find ⇒ more QTL getting in the way?
- More QTL alleles ⇒ less information about each
- Are QTL alleles common or rare?

How many founders?

More

- More general use
- More QTL
- Greater precision
- Estimate allele frequencies
- Haplotype analysis in founders

Fewer

- Lower residual variance
- Greater power for a particular QTL?
- Better power for epistasis
- Rare alleles are less rare

Which founders?

- Diverse
- Interesting
- No breeding problems
- Balanced: star phylogeny

How much mixing?

- More mixing ⇒ Greater mapping precision
- ...but lower power for *de novo* mapping
- Potential for population structure, missing alleles
- Random mating or curated mating?
- Start with many random cross directions?

Selfing or DH?

- Inbreeding gives added recombination
- But not so much as at the mixing stage
- If doubled haploids are feasible, use them

Key analysis issues

How to deal with the multiple alleles?

- Full model (an effect for each allele)
- Diallelic QTL model
- Random effects model (like BLUP)

How to account for multiple QTL?

- Stepwise selection
- Bayesian model averaging
- Random effect for polygenes

Sharing is also key

- The greatest power of MAGIC comes from sharing

  Pooling data, exploring multiple environments/treatments

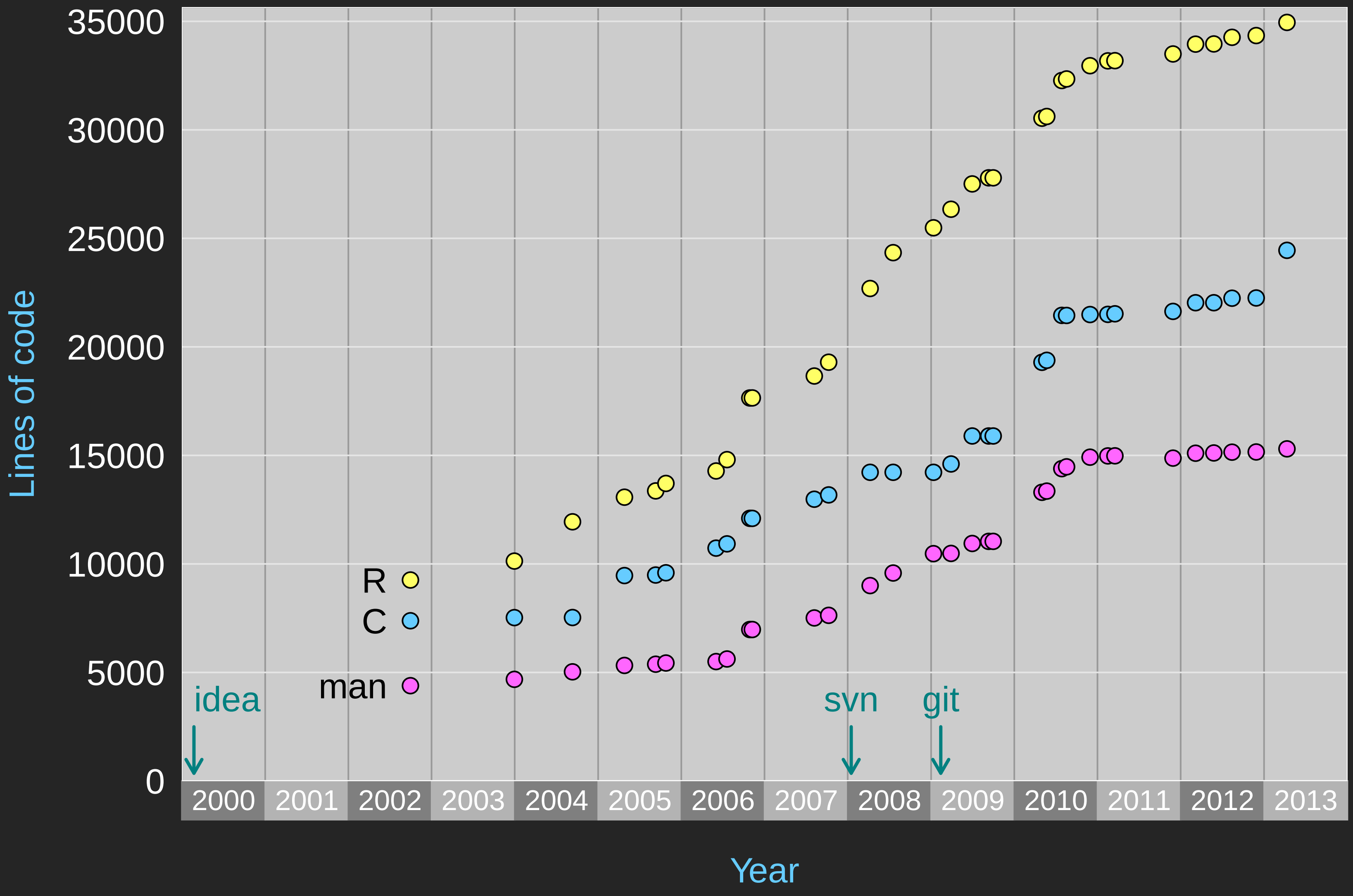
- Common software needs

  Analysis software, database infrastructure

- Our students need to learn the same stuff

  Joint training opportunities

R/qtl



Lines of code
35000
30000
25000
20000
15000
10000
5000
0
R
C
man
idea
svn
git
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Year

R/qtl: Good things

- Hidden Markov model code
- Many methods
- Extensible
- Open

R/qtl: Not-so-good

- Some really bad code
  - “Scantwo” is 4% of R code and 20% of C code, with a 1354-line R function
  - The stupidest R code ever:

```
for(i in 1:n) {
  temp[i] <- all(data[2,1:i]=="")
  if(!temp[i]) break
}
```

- The central data structure is too restrictive

  Can’t handle multiple individuals per genotype

- Memory mis-management
- Lack of connections to genome databases
- Largely one developer (who is also the support staff)

qtlHD

- Re-implementation of R/qtl

  Aimed at high-throughput computing

- High-dimensional data

  Dense markers, high-dim phenotypes, modern cross designs

- Separate from R

  But accessible from R (and ruby and python)

- Interactive graphics

- Connections to genome DBs/browsers

- `github.com/qtlHD/qtlHD`

  (Still at an exploratory stage)

Summary

- How many founders?

  Tradeoff between diversity and information about particular alleles

- Which founders?

  Diverse, interesting, no breeding problems, star phylogeny

- How long to mix?

  Tradeoff between power and precision

- How to fix?

  Doubled haploids are great if feasible

- Let's share!

  Lines, data, software, training

Summary

- How many founders?

  Tradeoff between diversity and information about particular alleles

- Which founders?

  Diverse, interesting, no breeding problems, star phylogeny

- How long to mix?

  Tradeoff between power and precision

- How to fix?

  Doubled haploids are great if feasible

- Let's collaborate!

  Lines, data, software, training

Acknowledgments

RIL calculations

James Crow
Bret Larget
David Levin
Dan Naiman
Timo Seppäläinen
Friedrich Teuscher

R/qtl

Danny Arends
Gary Churchill
Ritsert Jansen
Pjotr Prins
Śaunak Sen
Hao Wu
Brian Yandell

Funding: NIH R01 GM074244