Schemes for Efficient Realization of Effective Recombination Events to Enhance Breeding Progress

Seth C. Murray

Effective meiotic recombination events are recombination events that result in novel genetic combinations which can be directly observed; always less than the actual and expected number of recombination events. Modifying population designs can improve genome-wide effective recombination which is often a limiting factor in breeding and genetic linkage mapping. Using a simulation approach, this study sought to model and quantify the genome-wide effective recombination rate under various population designs. The number of markers needed to observe all effective recombination events and the distribution of the expected number of effective recombination events were then estimated. Three recombination models were used including one with recombination rates fit to the large Zea mays L. nested association mapping (NAM) dataset. Strong evidence was found in the empirical NAM dataset supporting a two- pathway modified Poisson model of recombination events with separate rate λ for each chromosome, reflecting the significant differences in effective recombination rates found between chromosomes. A positive linear relationship between the mean number of effective recombination events per generation and genome wide heterozygosity was observed. Primarily because of this phenomenon, dihybrid and doubled haploid populations increased the number of effective recombination events per generation when compared to traditional bi-parental recombinant inbred line populations. This study will be useful for quantitative geneticists and breeders in identifying efficient production of effective recombination events as well as researchers simulating recombination. Specific applications of these findings in both my applied breeding and quantitative genetics programs will be discussed.

Schemes for Efficient Realization of Effective Recombination Events to Enhance Breeding Progress

Seth C. Murray
Assistant Professor
Department of Soil and Crop Sciences
Texas A&M University

AgriLIFE RESEARCH
Texas A&M System

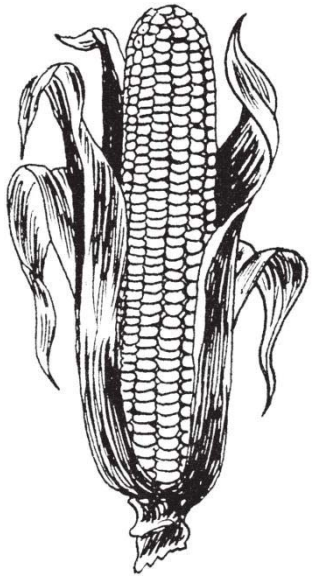
Program Overview

Two crops:

- **Maize**
- **Sorghum**

Multiple traits (genetics and/ or breeding) :

- **Yield**
- **Aflatoxin**
- **Drought**
- **Grain color (antioxidants)**
- **Perennialism**
- **Biomass (cellulosic bio-products)**
- **Composition (starch, oil, protein, phosphorus)**
- **Stem sugar**

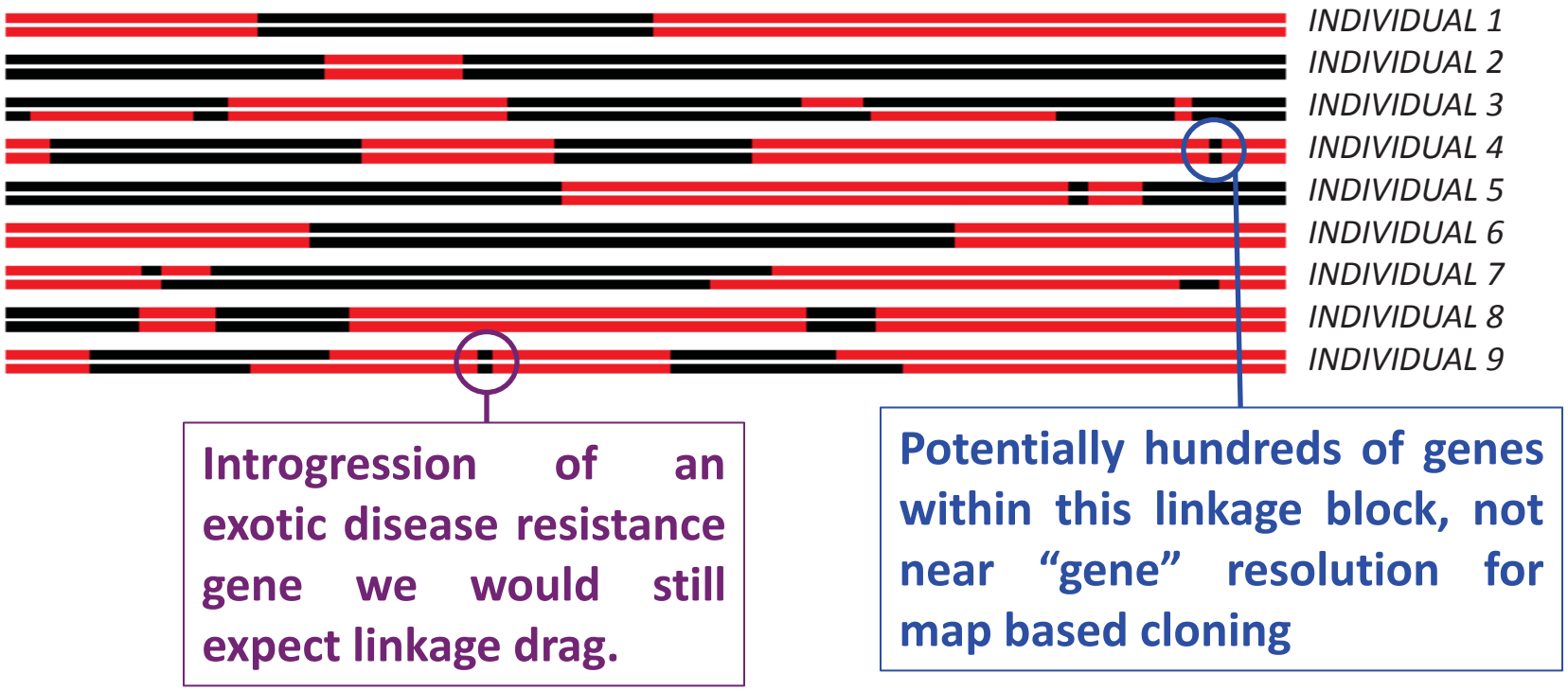


Multiple techniques for identification and improvement:

- **Bi-parental linkage QTL mapping**
- **Association QTL mapping**
- **Selection QTL mapping**
- **Breeding theory and simulation (recombination)**
- **Inheritance studies (diallel, design2)**
- **Near infrared spectroscopy (NIRS)**

In Plant Breeding and Genetic Linkage Mapping effective recombination is more often becoming the limiting factor

Effective recombination can directly be observed through rearrangement of polymorphic marker genotypes and always less than the actual and expected number of meiotic cross-over (recombination) events



With next generation sequencing technology we will soon be able to capture all effective recombination events

Breeding progress would be more efficient with more effective meiotic recombination events.

- Most especially for exotic introgression

Release of aflatoxin resistant lines
Mayfield et al. 2012. J. Plant Reg.

Breeding for perennial maize.

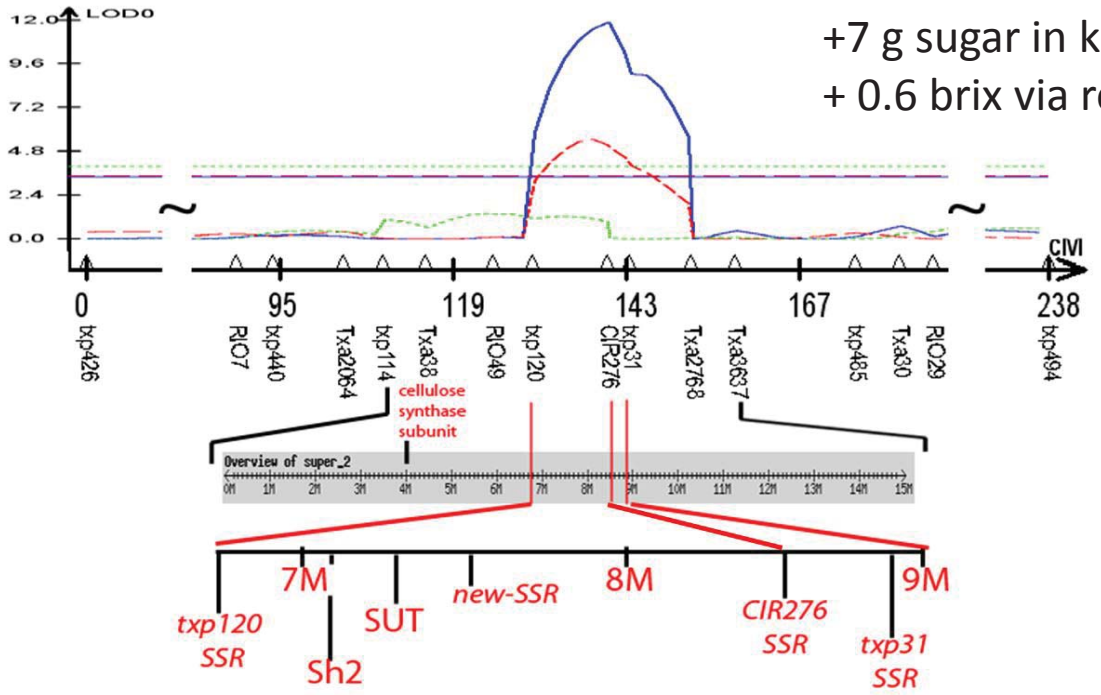
Germplasm enhancement of maize (GEM project)

- Transgenic event integration
  - Fewer backcrosses needed and/or smaller populations?
  - Targeted recombination sites would be preferred but is not reality yet
- GS models will be more accurate
- Even in elite x elite crosses the goal is to create new recombinants – we still select against anything deleterious.
- ***No perceived drawbacks from having more effective recombination***

Standard bi-parental QTL linkage mapping tends to identify large segments of chromosomes

- **Increased resolution to "gene-level" generally requires extensive follow up studies**
  - **Secondary and/ or much larger populations**
  - **Near isogenic lines (NILS)**
  - **Heterogeneous inbred families (HIFs)**
  - **Transcriptome, candidate genes, transformation, etc.**

- **Association mapping also has drawbacks and linkage mapping will always have a role**



+7 g sugar in kg of stem juice via HPLC
+ 0.6 brix via refractometer

Questions to be addressed in this study

Main question:

When creating a new linkage population - What distributions of effective number of recombination events per chromosome are expected to occur under common scenarios?

Background questions

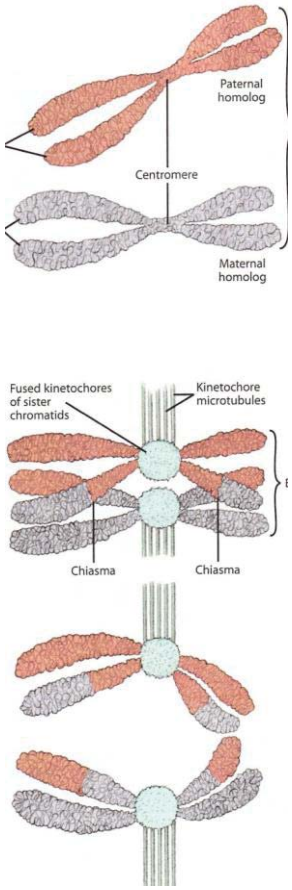
1) Using the model, how many loci need to be simulated?

2) Can a statistical distribution adequately be fit to model recombination in the high resolution *Zea mays* L. NAM dataset?

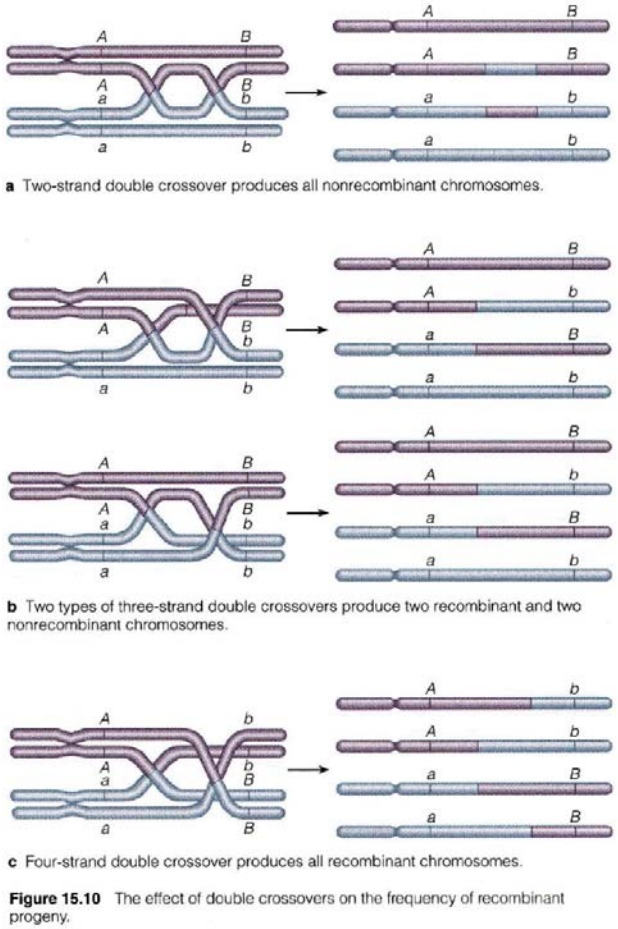
3) What proportion of effective recombination events in linkage populations might be missed due to insufficient and imperfect marker coverage?

4) What is the relationship between heterozygosity and effective recombination?

We tend to envision recombination as presented in introductory textbooks but have little idea how much is actually occurring



From: Raven, P.H., R.F. Evert and S.E. Eichhorn. 1999. *Biology of Plants; sixth ed.,* Freeman/Worth, New York



a Two-strand double crossover produces all nonrecombinant chromosomes.

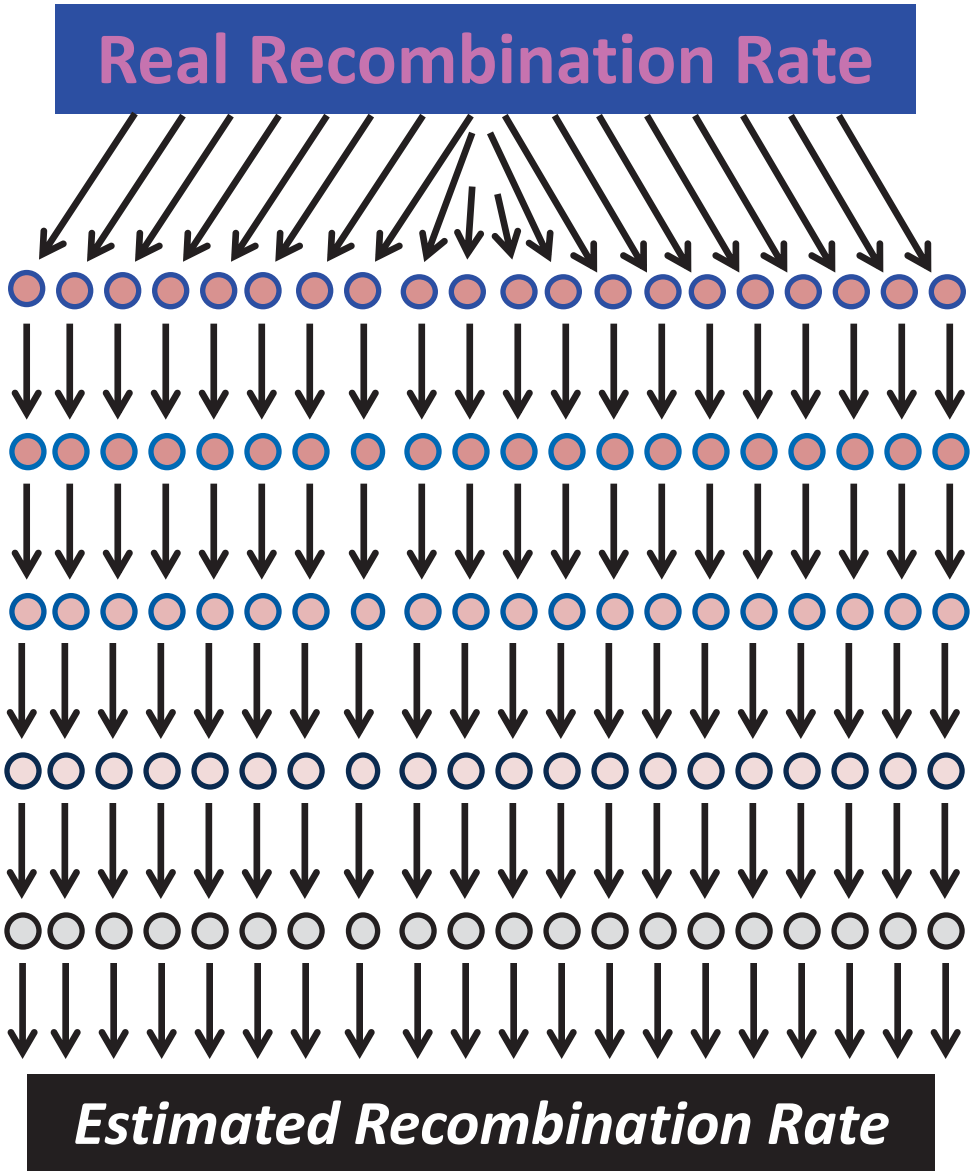
b Two types of three-strand double crossovers produce two recombinant and two nonrecombinant chromosomes.

c Four-strand double crossover produces all recombinant chromosomes.

Figure 15.10 The effect of double crossovers on the frequency of recombinant progeny.

From: Fairbanks, D. J., W. R. Andersen 1999 Genetics: the continuity of life. Brooks/Cole and Wadsworth, Pacific Grove, California, USA

Why do we need a simulation to measure recombination?



Masked by:

Recombination between homozygous regions (ineffective events)

Sampling only ¼ of chromatids per meotic tetrad

Each generation of selfing looses and masks more events

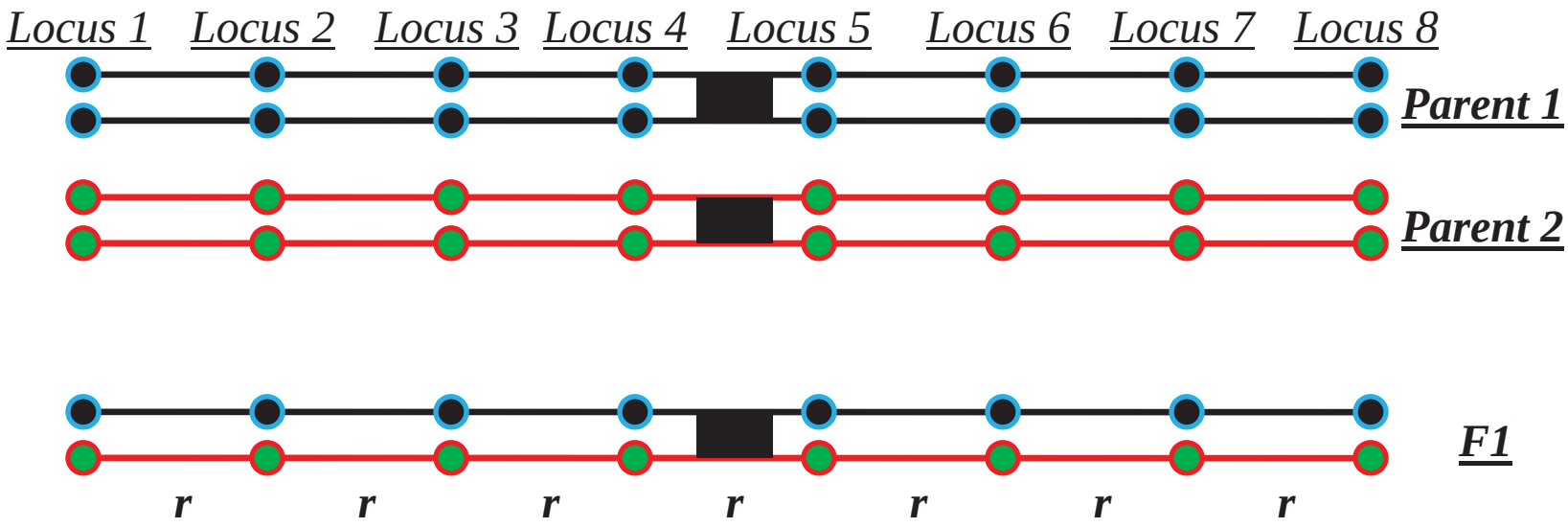
Sampling a limited number of individuals (less than ∞)

Sampling a limited number of markers (less than all base pairs in genome)

Mapping software assumptions may be imperfect

Method for simulating recombination and linkage

- **Using the software " *R* " using random number generators**
- **Incorporate all known information – mating design, number of markers, etc.**



Locus	1	2	3	4	C	5	6	7	8
Allele 1	*0*	*0*	*0*	*0*		*0*	*0*	*0*	*0*
Allele 2	*1*	*1*	*1*	*1*		*1*	*1*	*1*	*1*

LCrossover<- ceiling(runif(1,0,Loci/2))

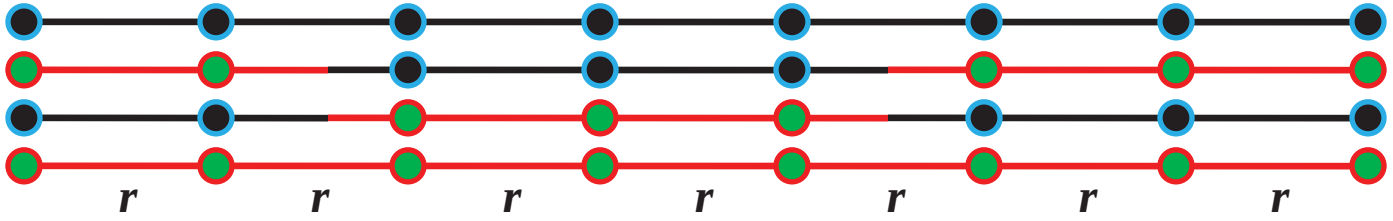
Example = 2

RCrossover<- ceiling(runif(1,Loci/2,Loci))

Example = 5

Method for simulating recombination and linkage

Where Loci = 8 and Loci/2 = centromere



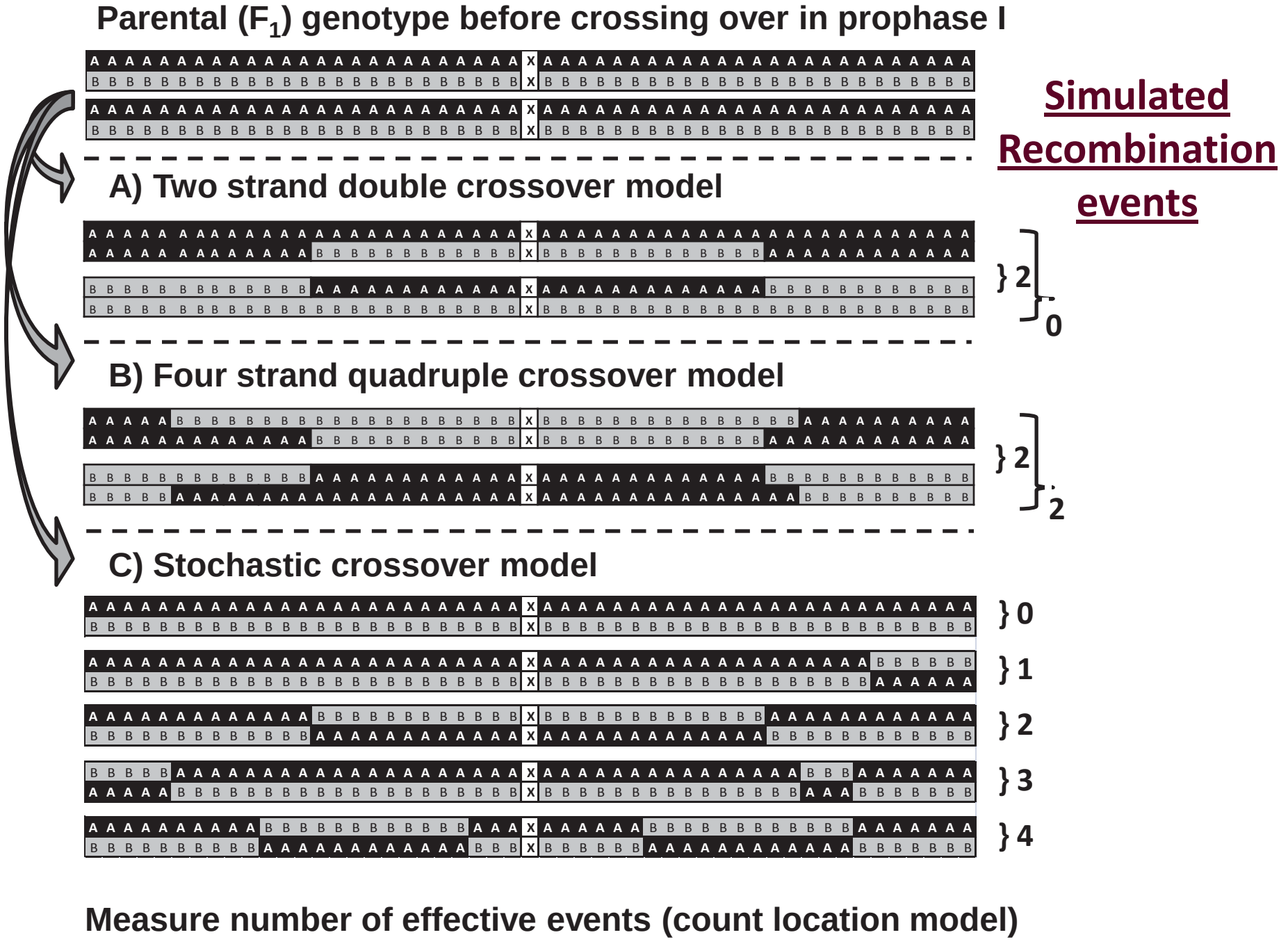
F1 gametes

Locus	1	2	3	C	5	6	7	8
Gamete 1	***0***	***0***	***0***	***0***	***0***	***0***	***0***	***0***
Gamete 2	***1***	***1***	***0***	***0***	***0***	***1***	***1***	***1***
Gamete 3	***0***	***0***	***1***	***1***	***1***	***0***	***0***	***0***
Gamete 4	***1***	***1***	***1***	***1***	***1***	***1***	***1***	***1***

Now within parent ***Ind.1*** we could choose which allele goes to the progeny using a uniform distribution.

Example: If gametes 1 & 2 were passed on:

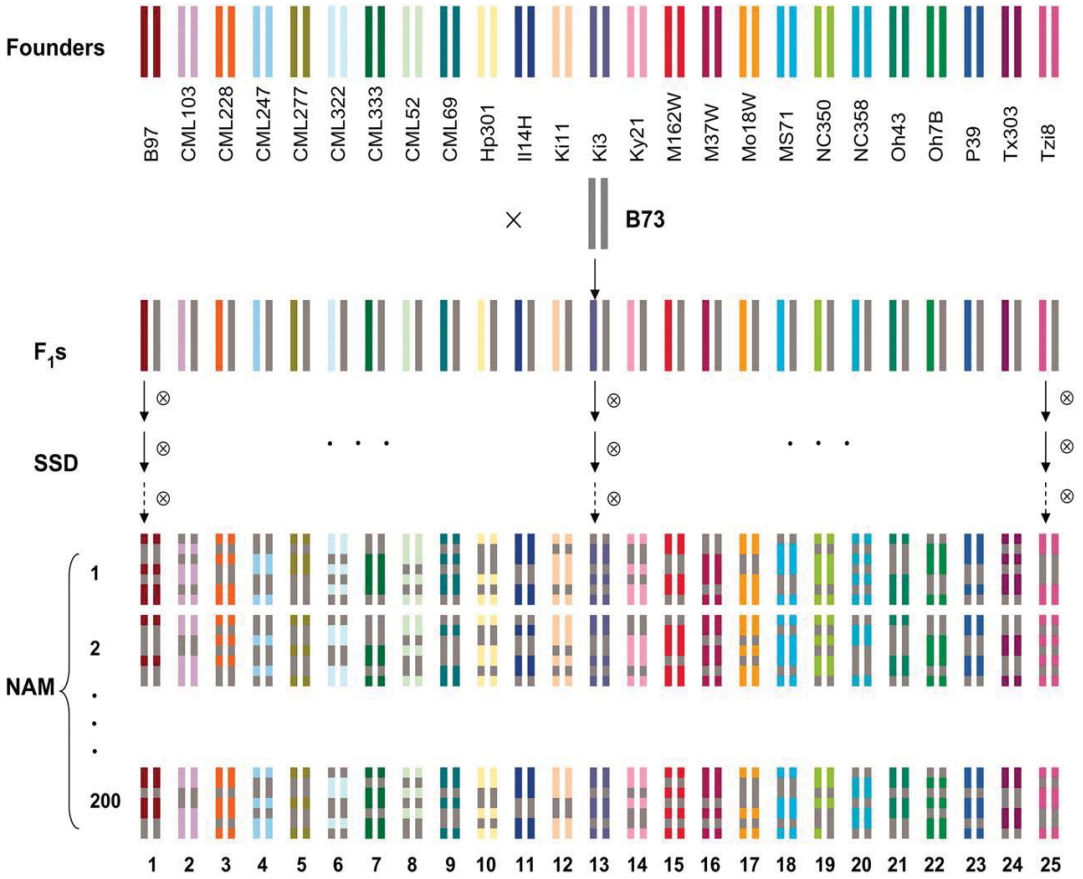
Locus	Ind.	1	2	3	C	5	6	7	8
Allele 1	1	***1***	***1***	***0***	***0***	***0***	***1***	***1***	***1***
Allele 2	1	***0***	***0***	***0***	***0***	***0***	***0***	***0***	***0***


Parental (F_1) genotype before crossing over in prophase I
A) Two strand double crossover model
B) Four strand quadruple crossover model
C) Stochastic crossover model
Simulated Recombination events
} 2
} 0
} 2
} 2
} 0
} 1
} 2
} 3
} 4
Measure number of effective events (count location model)

Genetic Properties of the Maize Nested Association Mapping Population

Michael D. McMullen,[1,2]|| Stephen Kresovich,[3]|| Hector Sanchez Villeda,[2]* Peter Bradbury,[1,2] Huihui Li,[4,5,3] Qi Sun,[6] Sherry Flint-Garcia,[1,2] Jeffry Thornsberry,[7] Charlotte Acharya,[3] Christopher Bottoms,[2] Patrick Brown,[3] Chris Browne,[1] Magen Eller,[1] Kate Guill,[1] Carlos Harjes,[3]† Dallas Kroon,[3] Nick Lepak,[1] Sharon E. Mitchell,[3] Brooke Peterson,[1] Gael Pressoir,[3]‡ Susan Romero,[1] Marco Oropeza Rosas,[8]¶ Stella Salvo,[1] Heather Yates,[3] Mark Hanson,[9] Elizabeth Jones,[10] Stephen Smith,[10] Jeffrey C. Glaubitz,[11] Major Goodman,[8] Doreen Ware,[1,12] James B. Holland,[1,8]|| Edward S. Buckler[1,3,13]||

McMullen et al., 2009 Science 325, 737-740

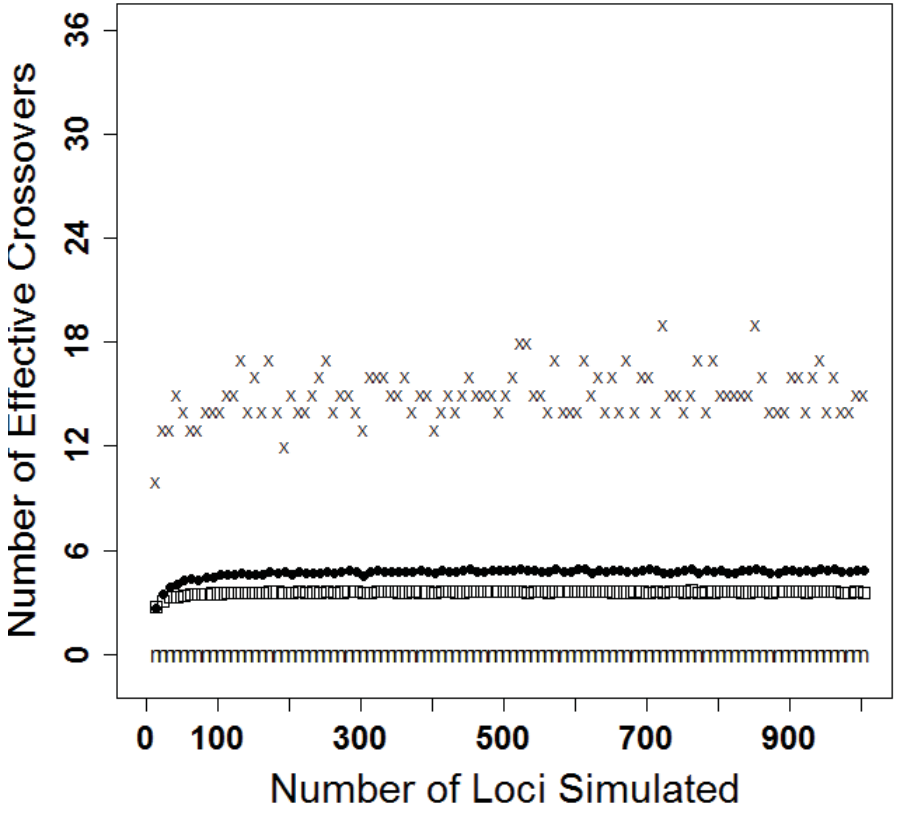


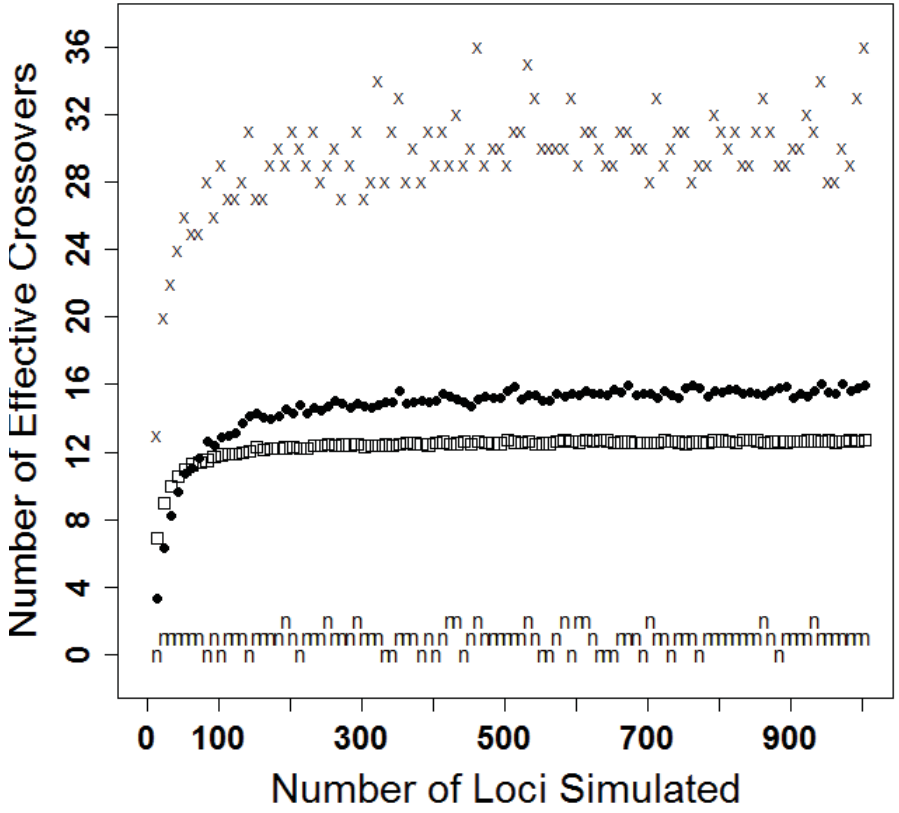
Yu J et al. Genetics 2008;178:539-551

- 25 F_5 Populations
- 4,699 RILS
- 1,144 SNPs
- 136,000 recombination events
- Effective recombination rates were statistically different for each chromosome (*but not necessarily each population)*

Minimum number of markers needed to simulate population recombination

- Counting recombination events NOT distance (cM)



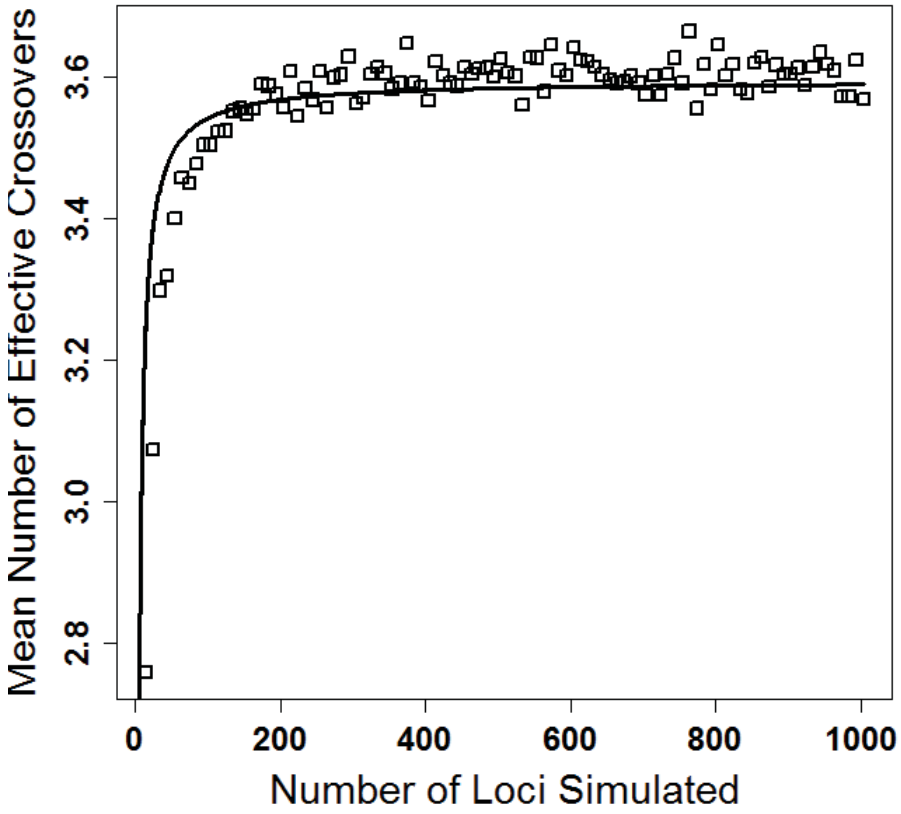


Biparental cross → selfed for 8 generations to RILs

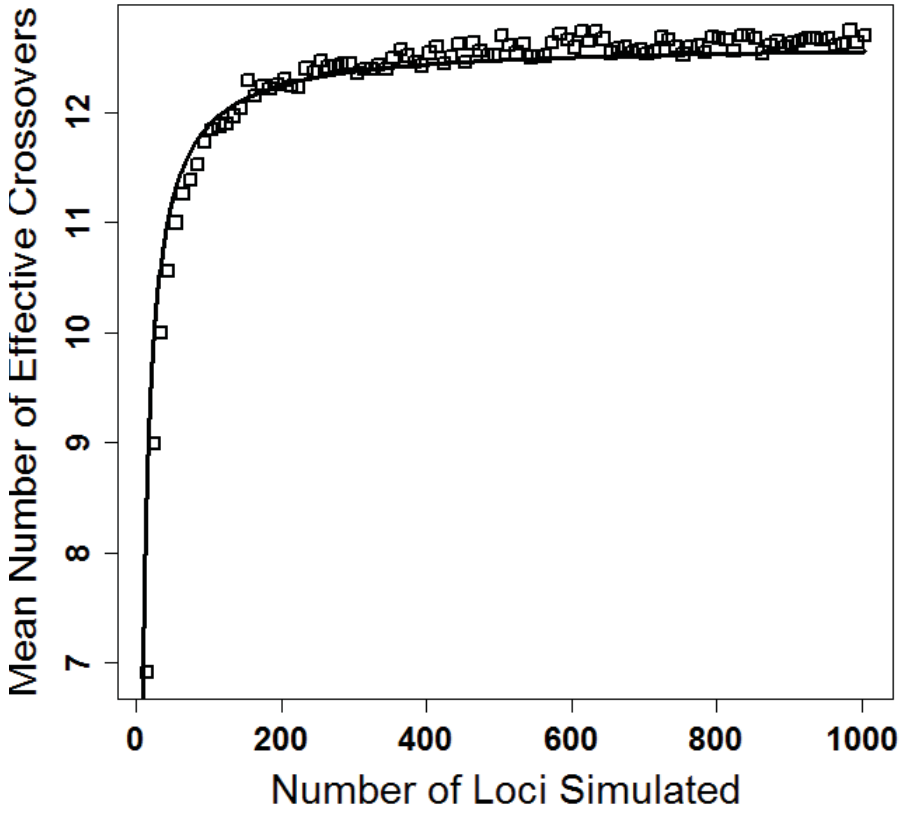
Dihybrid cross → 3 generations of sibing → 8 generations of selfing to RILs

Mean number of effective recombination events captured

$$= \beta - (((log\ (loci + \beta)) - (log\ (loci))) / \chi\ )$$



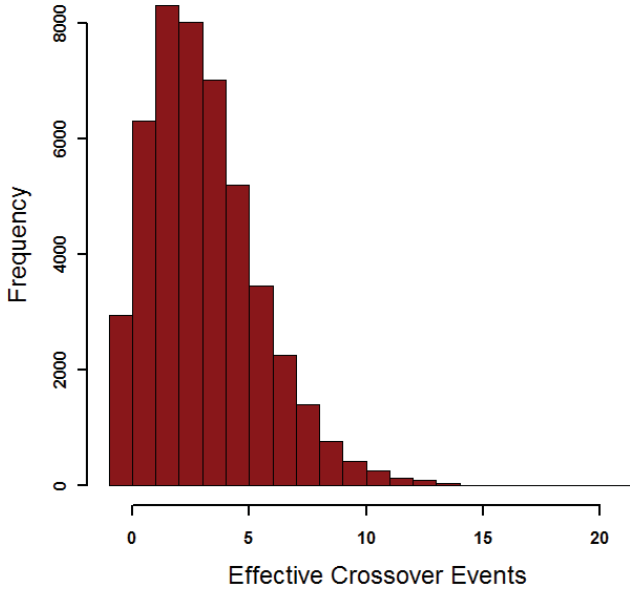
Biparental cross → selfed for 8 generations to RILs



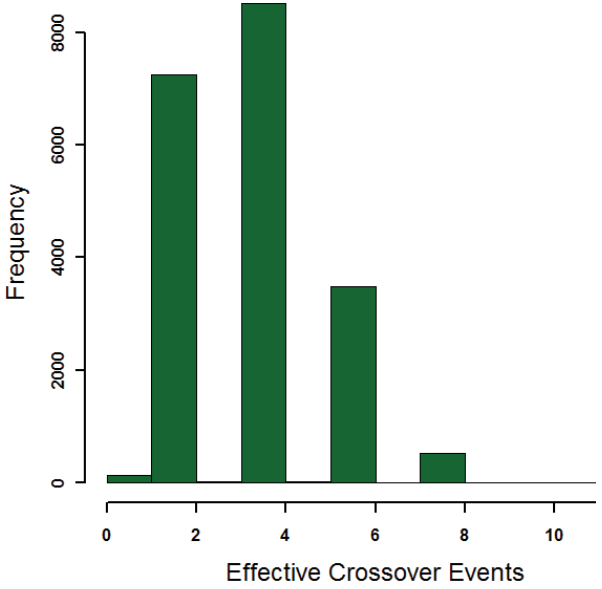
Dihybrid cross → 3 generations of sibing → 8 generations of selfing to RILs

Number of effective recombination events per chromosome in F_5

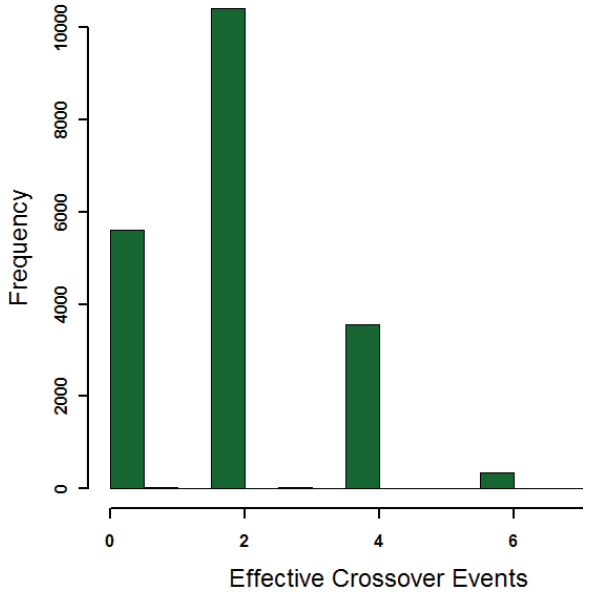
Actual NAM data



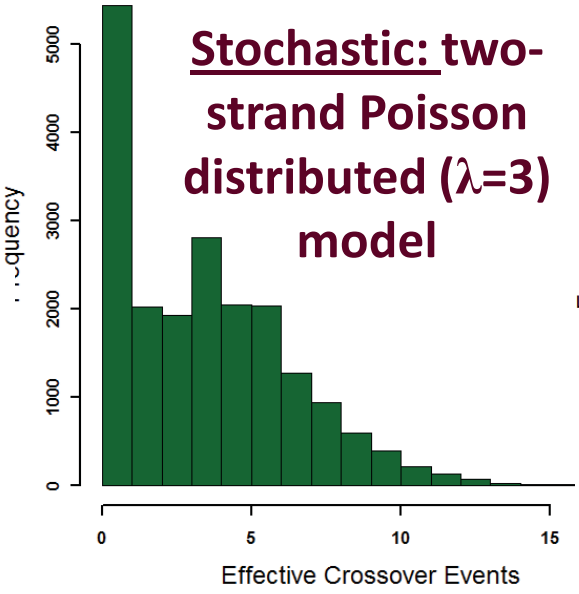
Fixed: four-strand quadruple crossover model



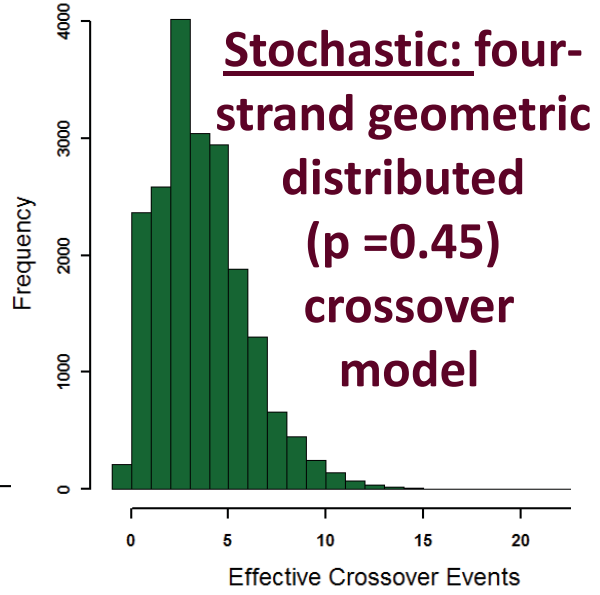
Fixed: two-strand double crossover model



Stochastic: two-strand Poisson distributed (λ=3) model



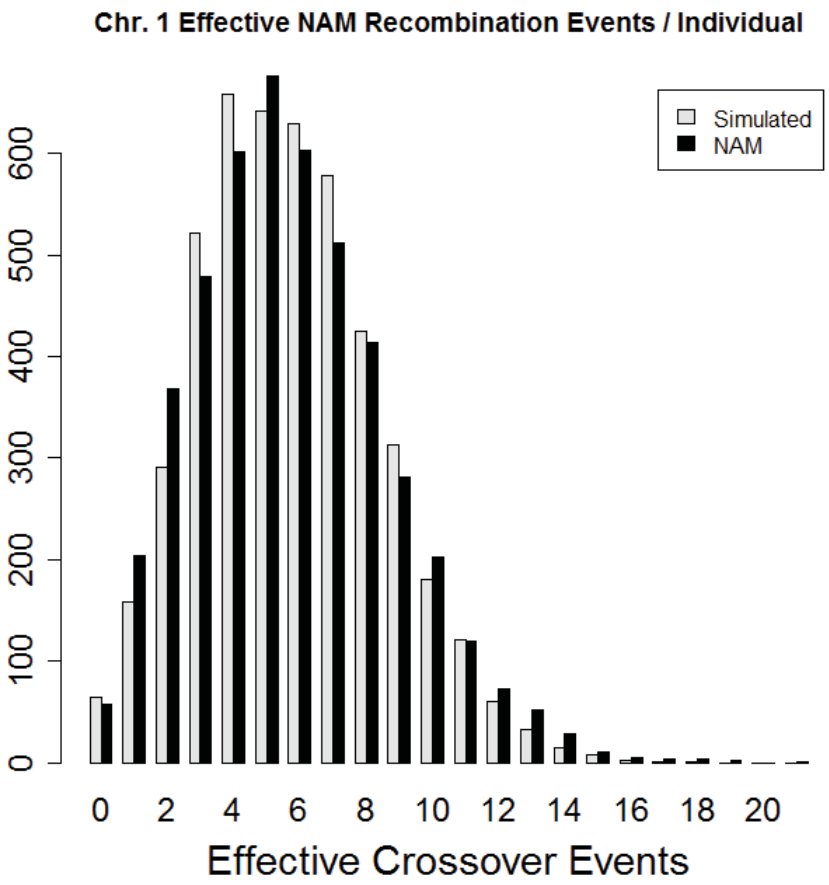
Stochastic: four-strand geometric distributed (p =0.45) crossover model

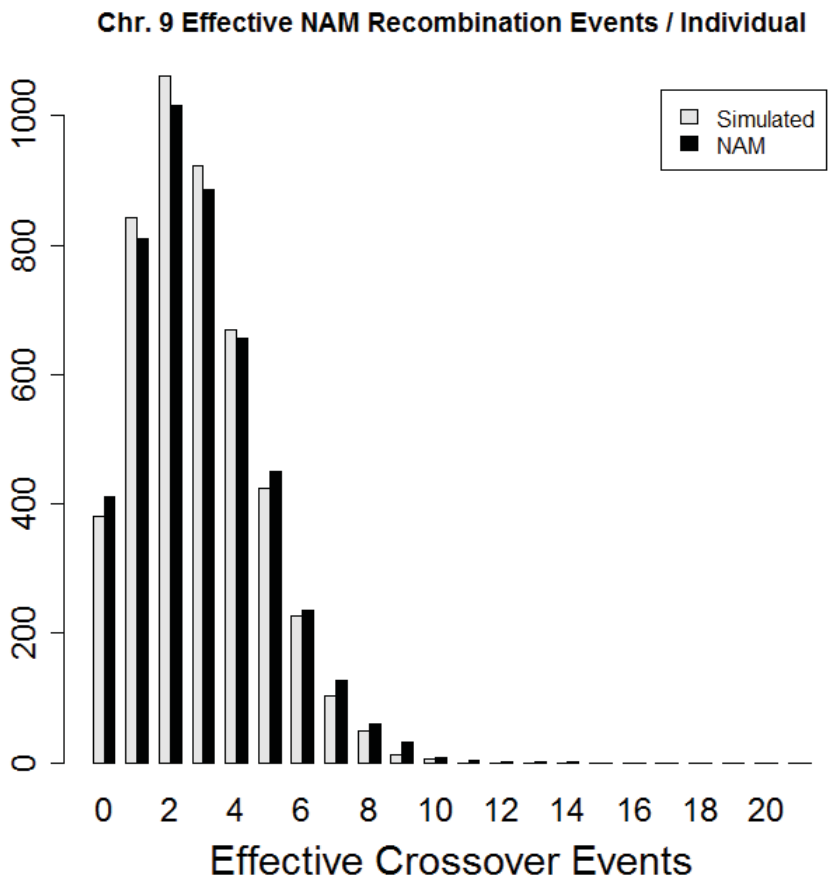


These models account for:

- Mating design
- Population size
- Markers
- Etc…

Stochastic Poisson + obligate crossover fit NAM data





Chromosome 1:

NAM sample μ = 5.73

NAM sample σ = 8.73

Sim. Poisson λ = 2.24

Fit Chi-sq. = p > 0.08

Chromosome 9:

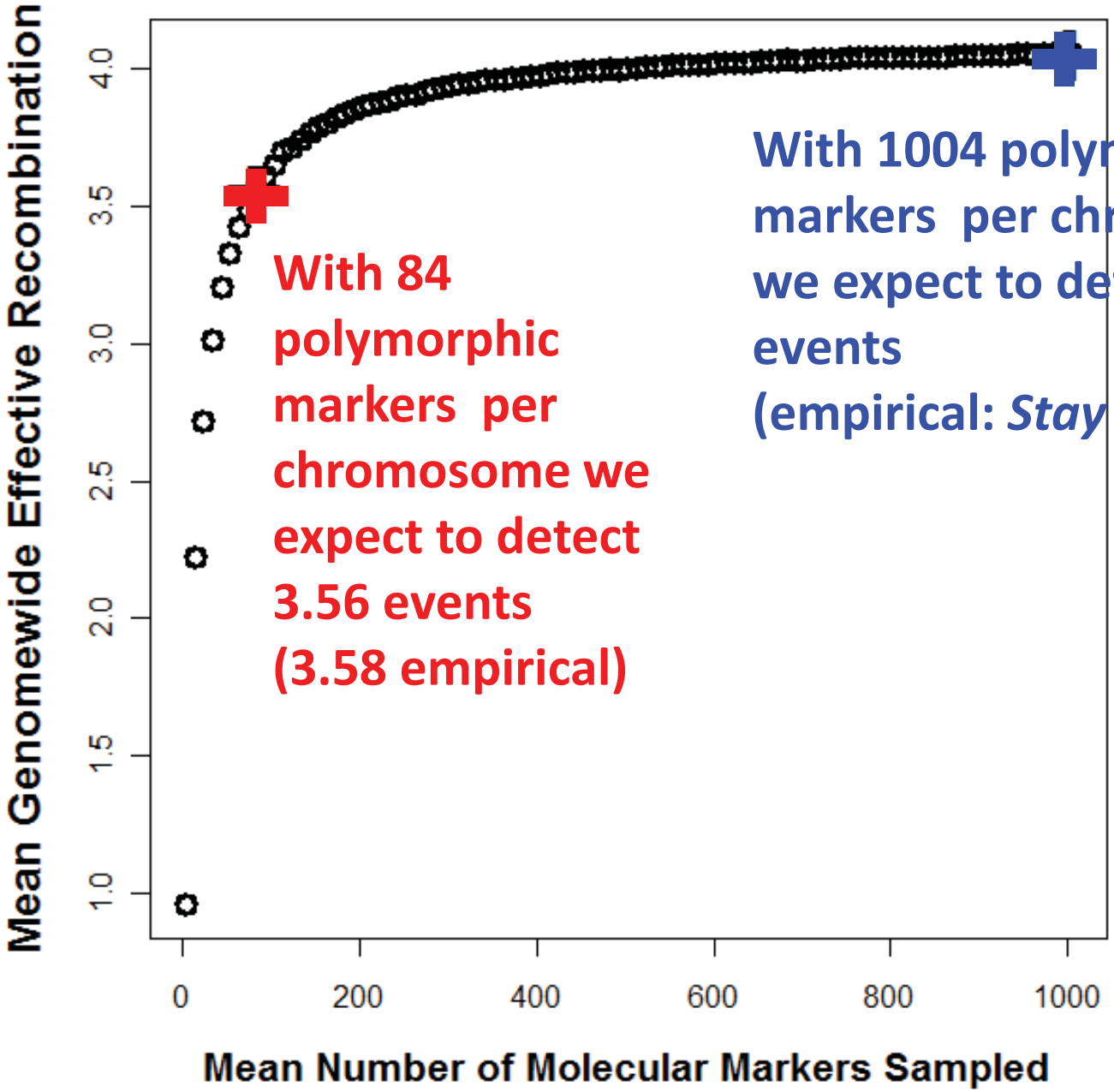
NAM sample μ = 2.90

NAM sample σ = 3.87

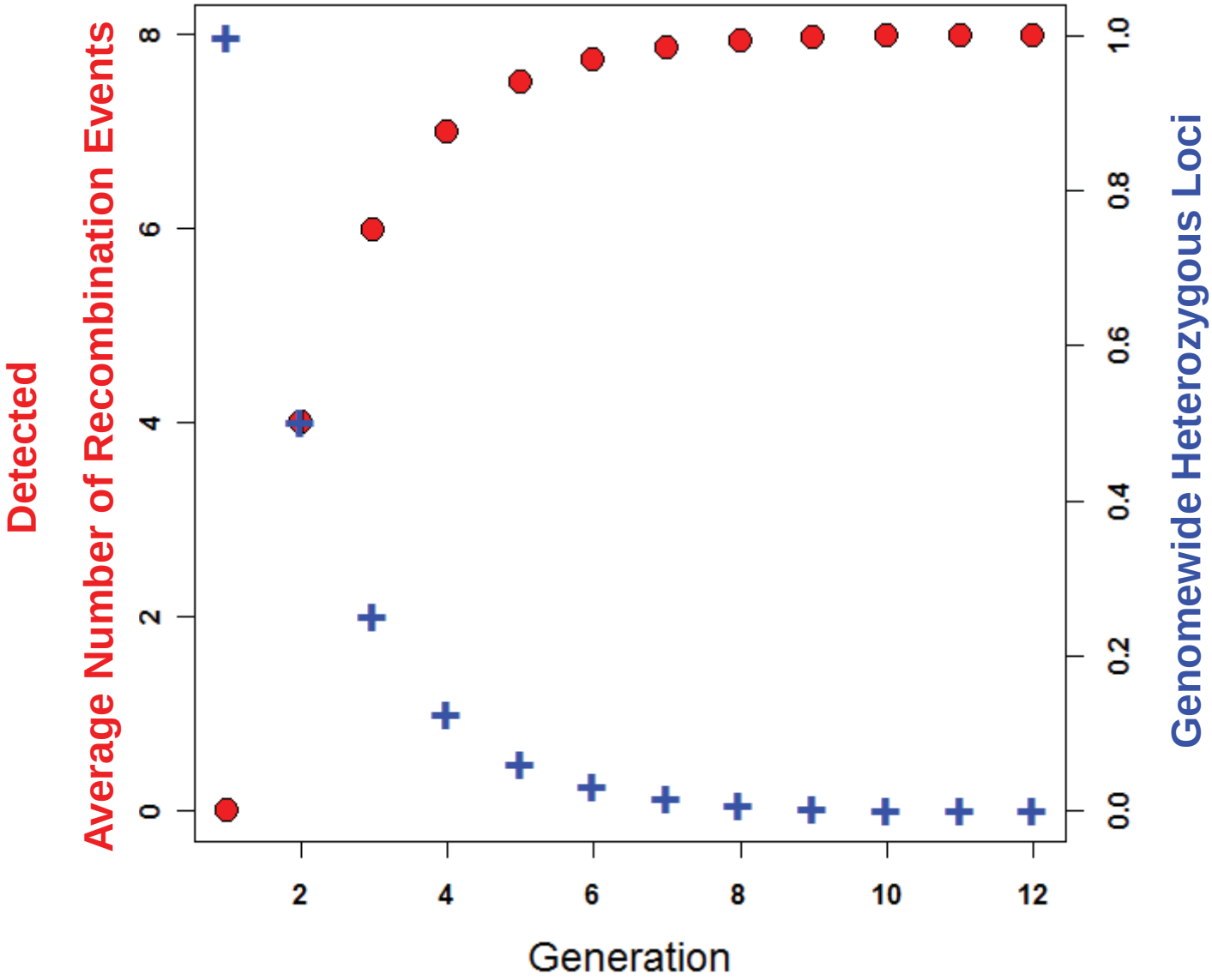
Sim. Poisson λ = 0.90

Fit Chi-sq. = p > 0.78

Predicted effect of incomplete marker data in NAM

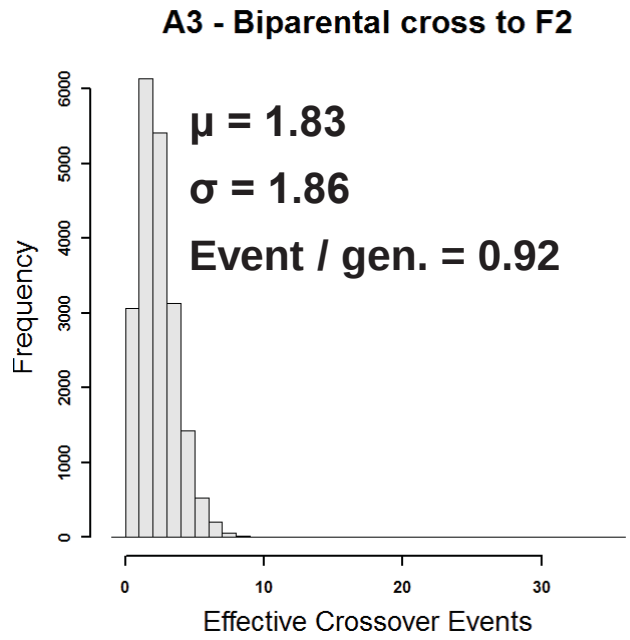


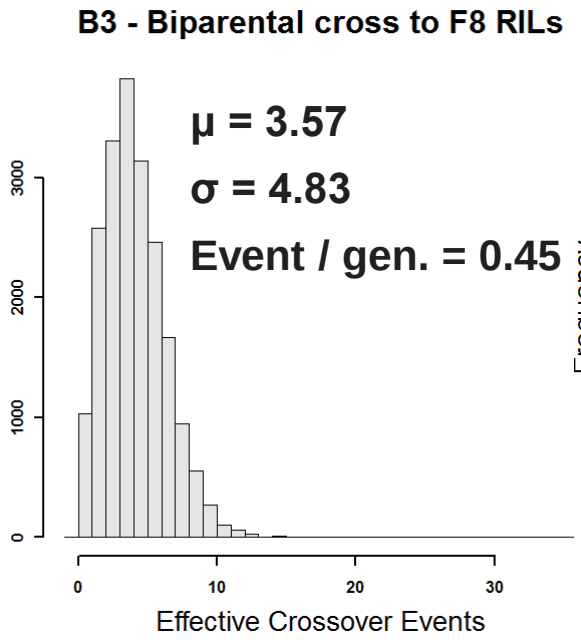
Direct relationship between heterozygosity and the effective number of recombination events (self pollination)

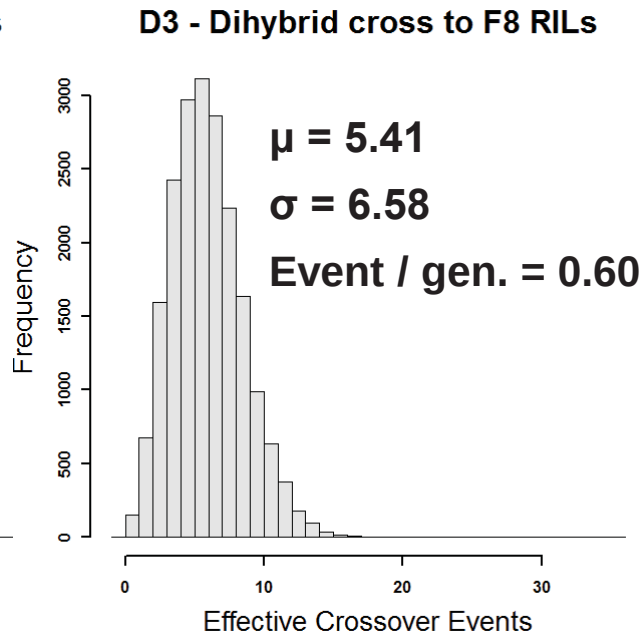


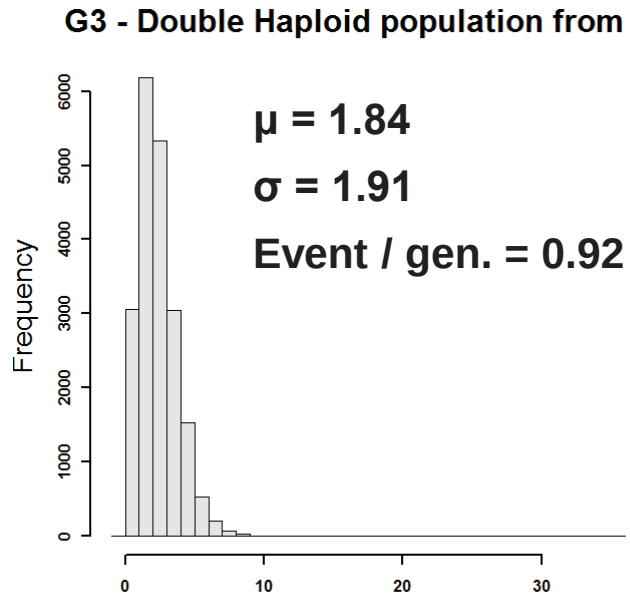
*EMRE = (crossovers * mean genomewide heterozygosity) + previous events*

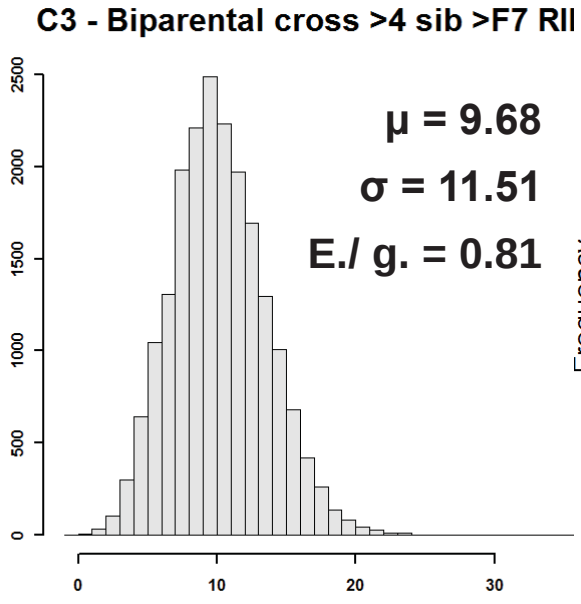
Predicted distributions of effective recombination events under different population development scenarios

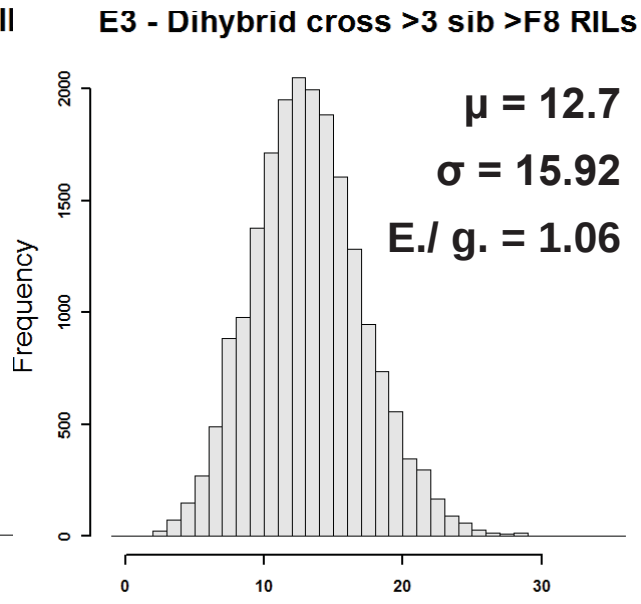












The highest number of effective recombination events per generation is expected to be found in dihybrid populations > F_2 ≥ doubled haploids F_1 > RILs. Intermating early always produces more recombination.

1) Depending on the model, a minimum of 300 – 1000 loci should be simulated.

2) Only a "stochastic Poisson + obligate" number of crossovers fit the high resolution NAM dataset.

3) Insufficient and imperfect marker coverage is expected to miss 0.5 effective recombination events / chromosome / individual in NAM.

5) Average effective recombination per generation is equal to the recombination rate * average genomewide heterozygosity.

Population designed to measure recombination

B73 Oleic

Lfy Blue

Tx772

Ethiopia B.

2 - Way F_1

2 - Way F_1

4 - Way F_1

Self

Self

Self

Self

Self

Self

INTERMATE

n = 158
2 way, 3 self

n = 133
4 way, 3 self

n = 253
4 way, 1 sib, 3 self

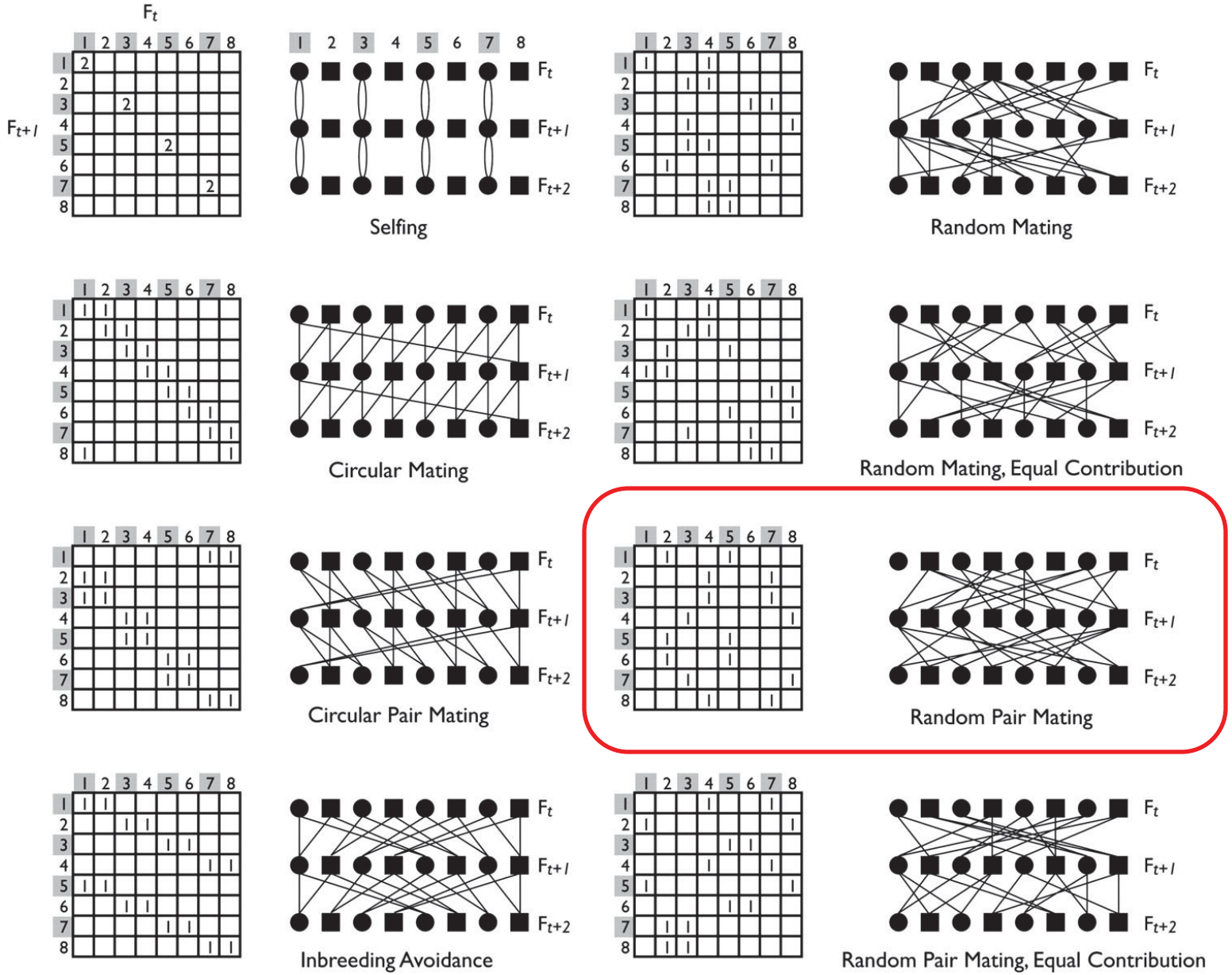
n = 140
2 way, 3 self

n = 575
4 way, 3 sib, 2 self

n = 173
4 way, 2 sib, 3 self

n = 1432 total
(Goal is 1000)

Mating Methods to Break Linkage Blocks



Rockman MV, Kruglyak L.2008. Breeding designs for recombinant inbred advanced intercross lines. Genetics 179(2):1069-78

Identifying global genetic diversity in genomewide recombination rates

- Simulations allow us to compare estimated recombination rate of different populations.
- It is known that genetic distance is affected by
  - Number of progeny screened
  - Number and distributions of genetic markers
  - Amount of sib or self pollination

- Gather 100's of populations from industry and the public sector and put through the simulation to estimate diversity in true genomewide-recombination rate
- Do you have populations that would be useful?

Use next generation sequencing

- Look at hotspots
- Detect most recombination events
- Might even detect non-crossovers (gene conversion) events

Is there really a genomewide recombination rate or is it an additive sum of local events?

Incorporating these findings into 'Selection Mapping' (detecting temporal selection)

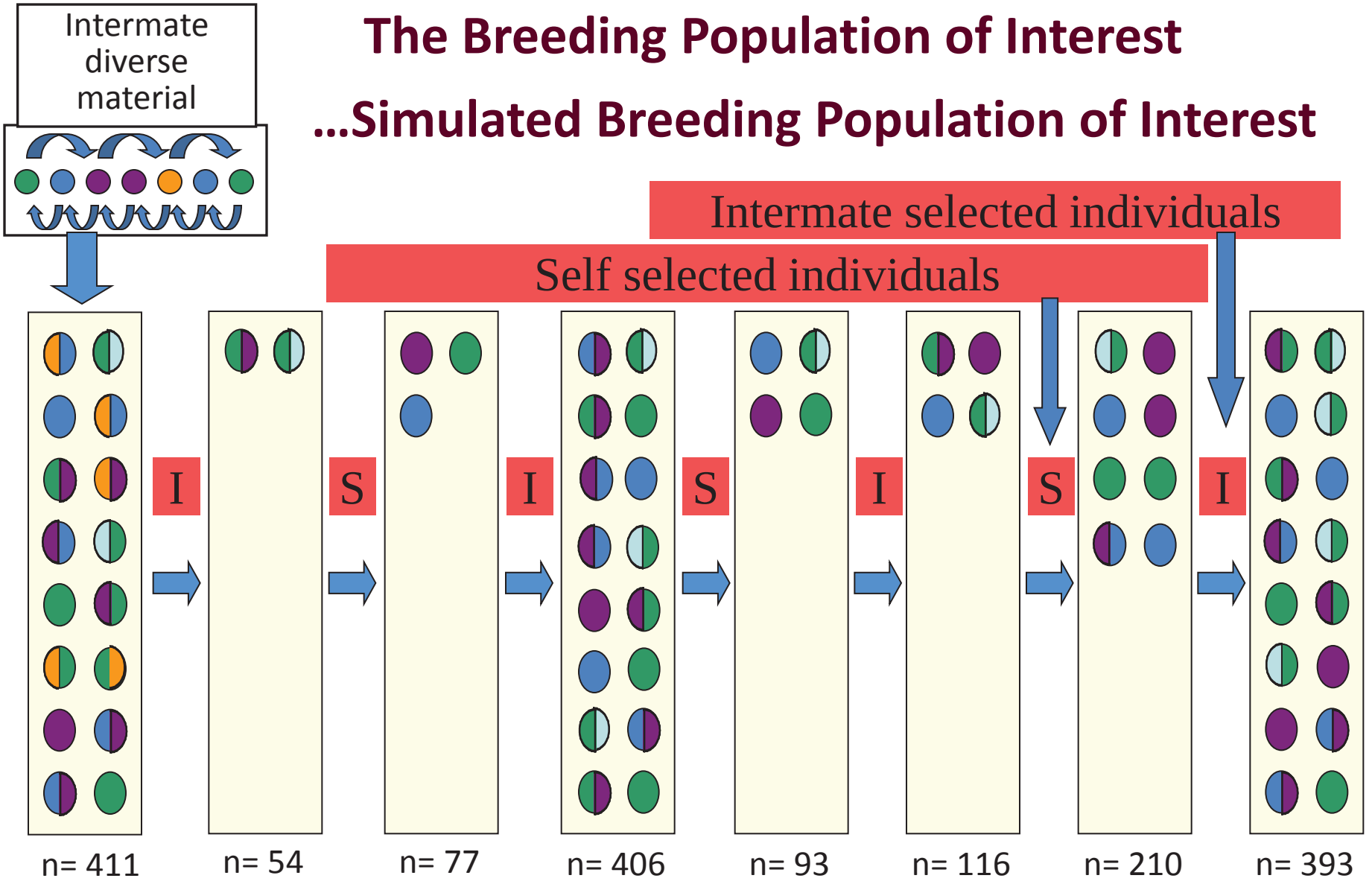
Wisser, R.J., S.C. Murray, J.M. Kolkman, H. Ceballos, and R.J. Nelson. 2008. Selection mapping of loci for quantitative disease resistance in a diverse maize population. Genetics.180:583 - 599.

Null hypothesis: Allele frequencies that differ between generations (before and after selection) can be explained by genetic drift.

Alternative hypothesis: The allele frequency difference between the generations can *NOT* be explained by genetic drift (mutation, migration, SELECTION).

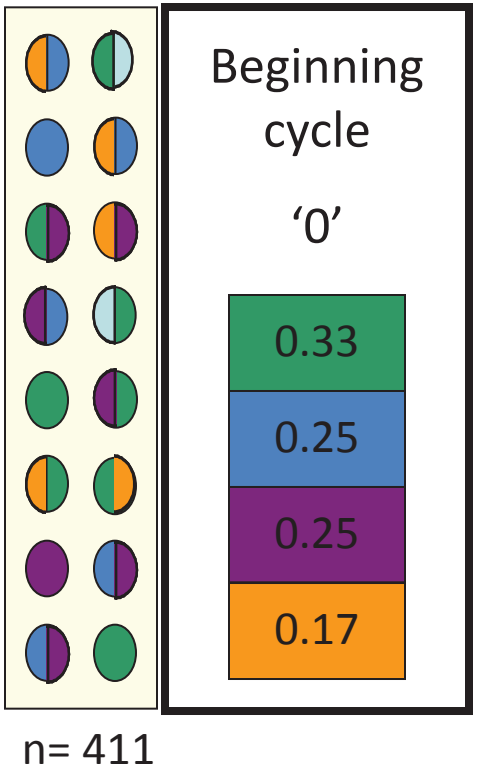
The Breeding Population of Interest

…Simulated Breeding Population of Interest

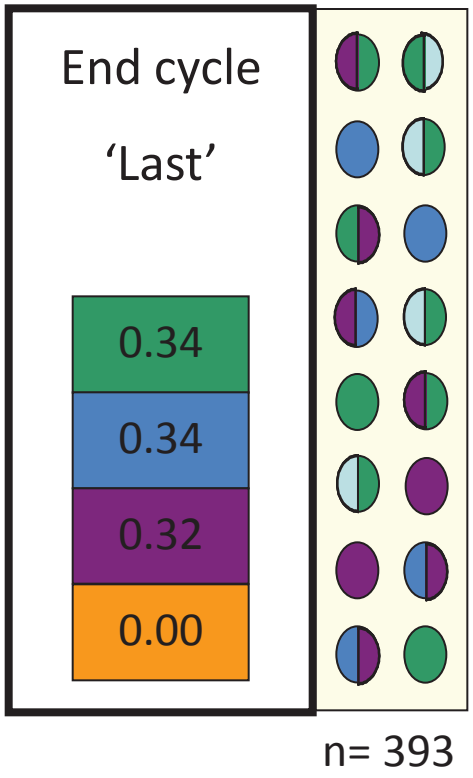


Example of a single locus in CIMMYT pool 30; a population selected for northern corn leaf blight. Ceballos *et al*. (1991) Crop Sci.

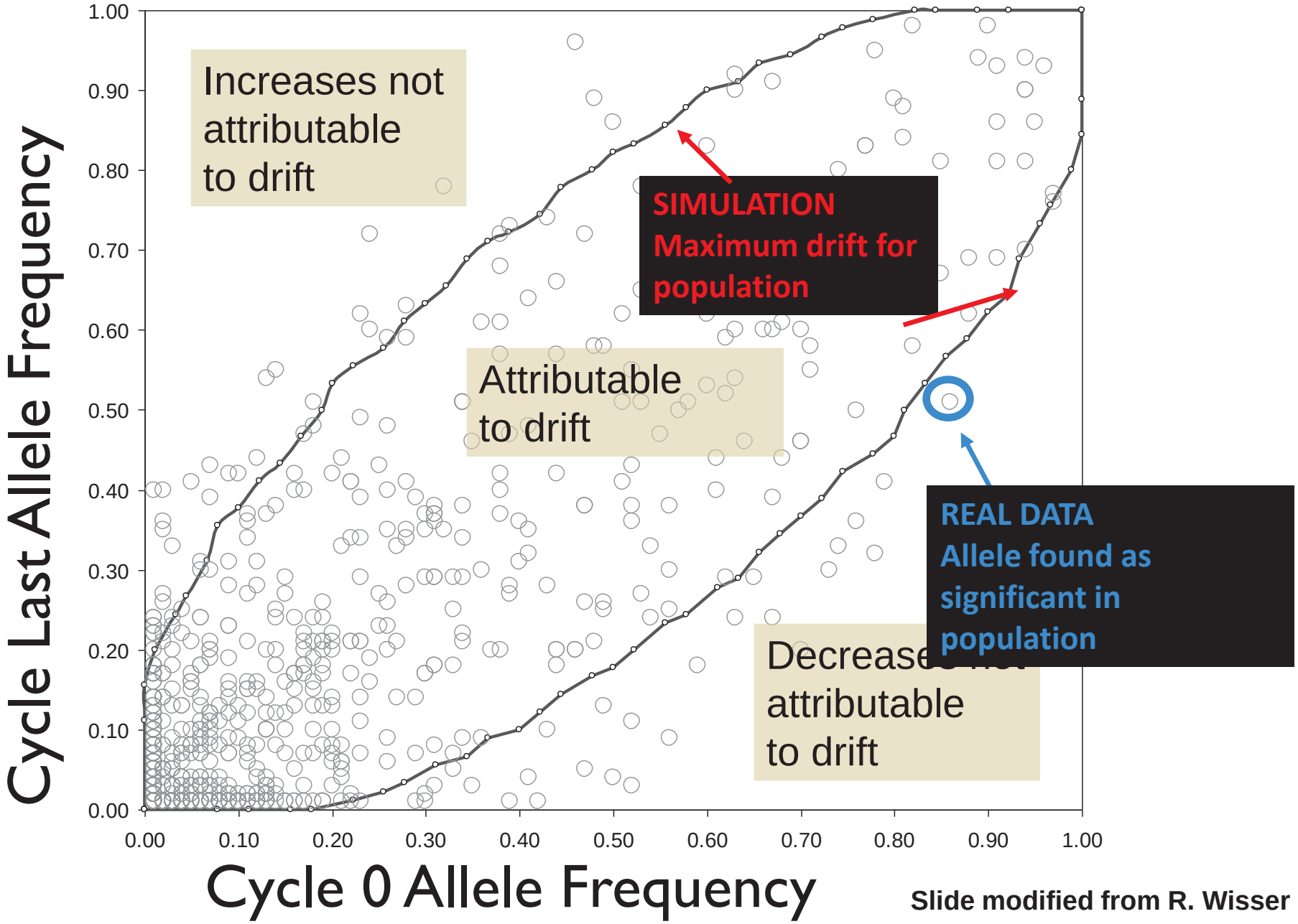
Use Molecular Markers Before and After Selection to Identify Allele Frequencies Shifts



…Compare These to a Sample Drawn From the Simulated Population to Test Significance.



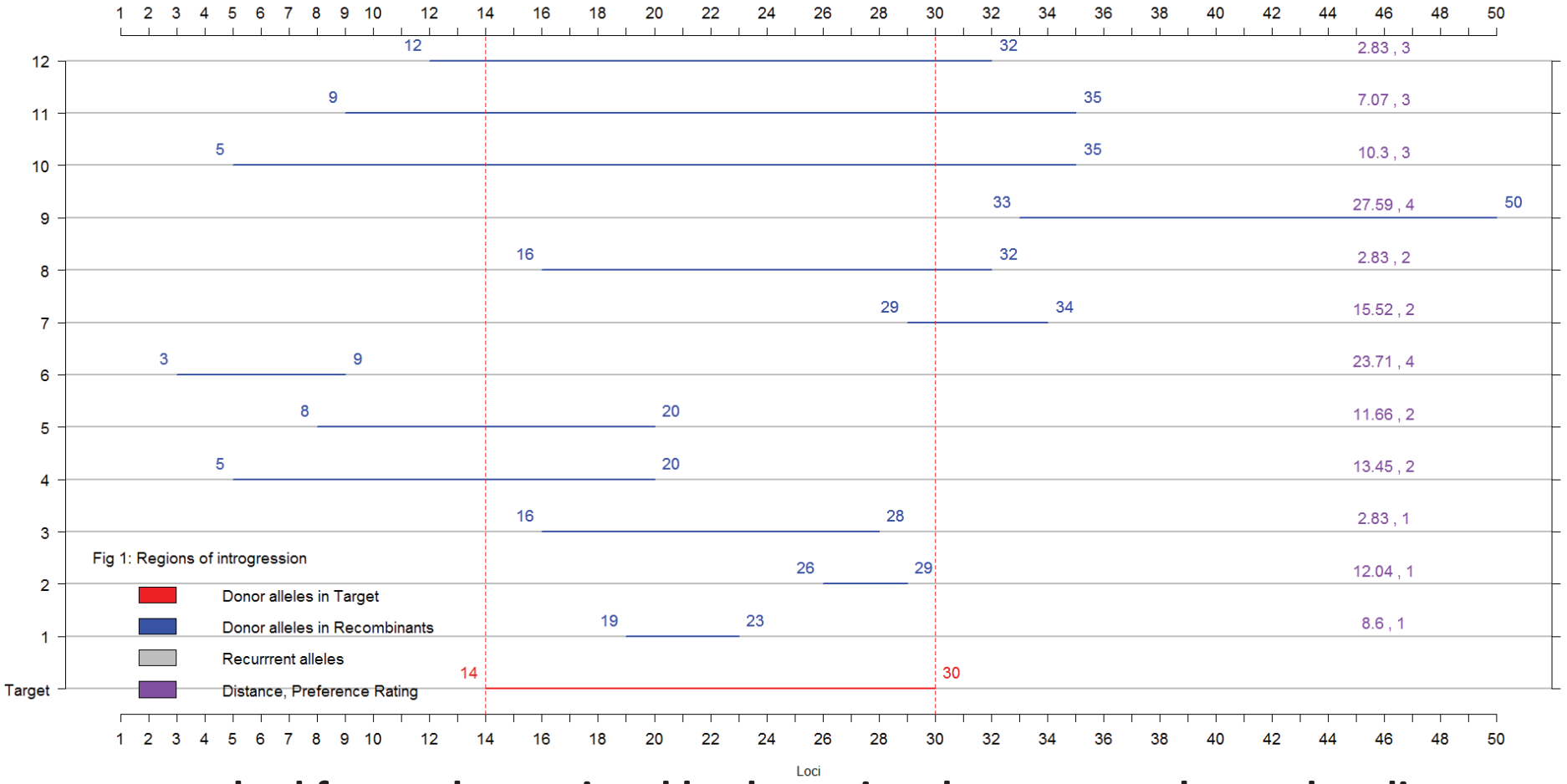
Selection Mapping



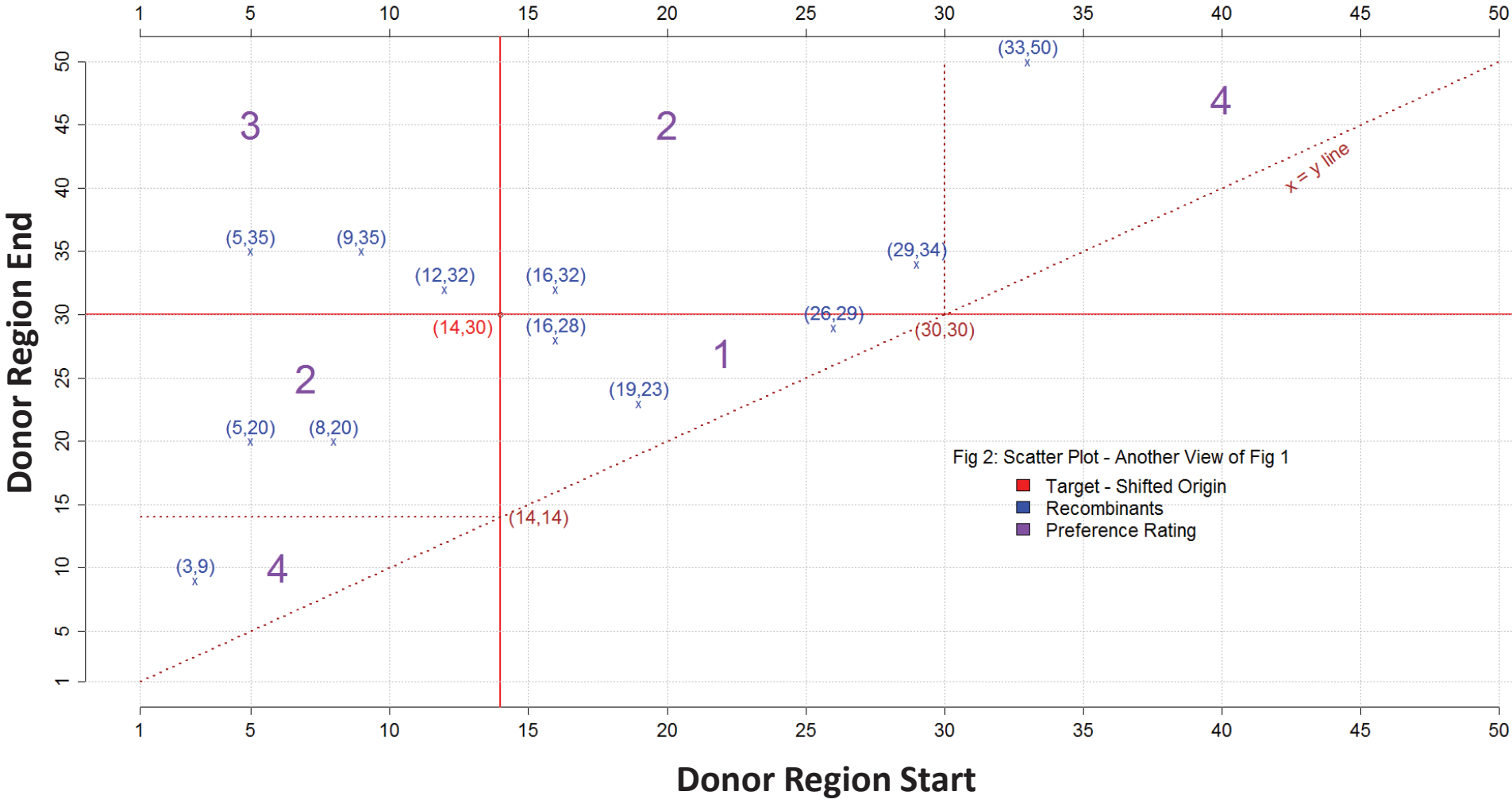
Slide modified from R. Wisser

Maize ATLAS

Adaptation Through Latitudinal Artificial Selection



…a method for marker-assisted backcrossing that uses marker and pedigree information to form tiling paths across chromosomal regions of interest



The method optimizes construction of tiling paths, throughout a series of successive backcrosses, by finding a *minimal spanning set* of individuals with crossovers in the region of interest

Simulation Study

- **Four distinct regions of the maize genome investigated, using the NAM linkage map and rates as a basis for the simulation parameters**

- **The number of *best* individuals obtained depends on**
  - **length of the chromosome**
  - **length of target region, and**
  - **a chromosome-specific Poisson parameter that determines the number of crossovers in each cross**

- **More efficient to "pool progeny" in each generation and to genotype more individuals towards the end of the backcross scheme**

NAM data
Dr. Jeffrey Glaubitz
Dr. Michael McMullen
The rest of the NAM project team

Collaborators and contributions on this work
Dr. Randall Wisser (University of Delaware)
Dr. Nick Lauter (USDA / Iowa State)
Dr. Patricia Klein (TAMU)

Helpful conversations
Dr. Mattieu Falque (INRA)
Dr. Wojtek Pawlowski (Cornell)
Dr. Michael Gore (USDA-ARS)
Dr. Jianming Yu (KSU)
Dr. Mark Wright (Cornell)
Dr. Sean Miles (NSAC)
Dr. George Hodnett (Texas A&M)
Dr. Martha Hamblin (Cornell)

Graduate Students
2010 SCSC643 class
Rupa Kanchi
Adam Mahan
Ivan Barrero
Gerald De La Fuente
Dr. Kerry Mayfield
Meghyn Meeks
Jeff Savage

United States Department of Agriculture
National Institute of Food and Agriculture

Part of this project was supported by a Agriculture and Food Research Initiative Competitive Grant from the USDA National Institute of Food and Agriculture.

Additional Funding
Texas Corn Producers Board
USDA-SCA
Texas AgriLife Research
Monsanto Fellowship
Pioneer Hybrid Fellowship

AgriLIFE RESEARCH
Texas A&M System