Uncertainty and Imputation of Plant Genotypes

•••

Rachel Frantz (University of La Verne)
Nathan Ryder (Dordt College)
Jason Vander Woude (Dordt College)

Primary questions

- Can we quantify the uncertainty within a plant database due to genotyping errors?
- Can we adjust summary statistics to account for changes in Type I and Type II error?

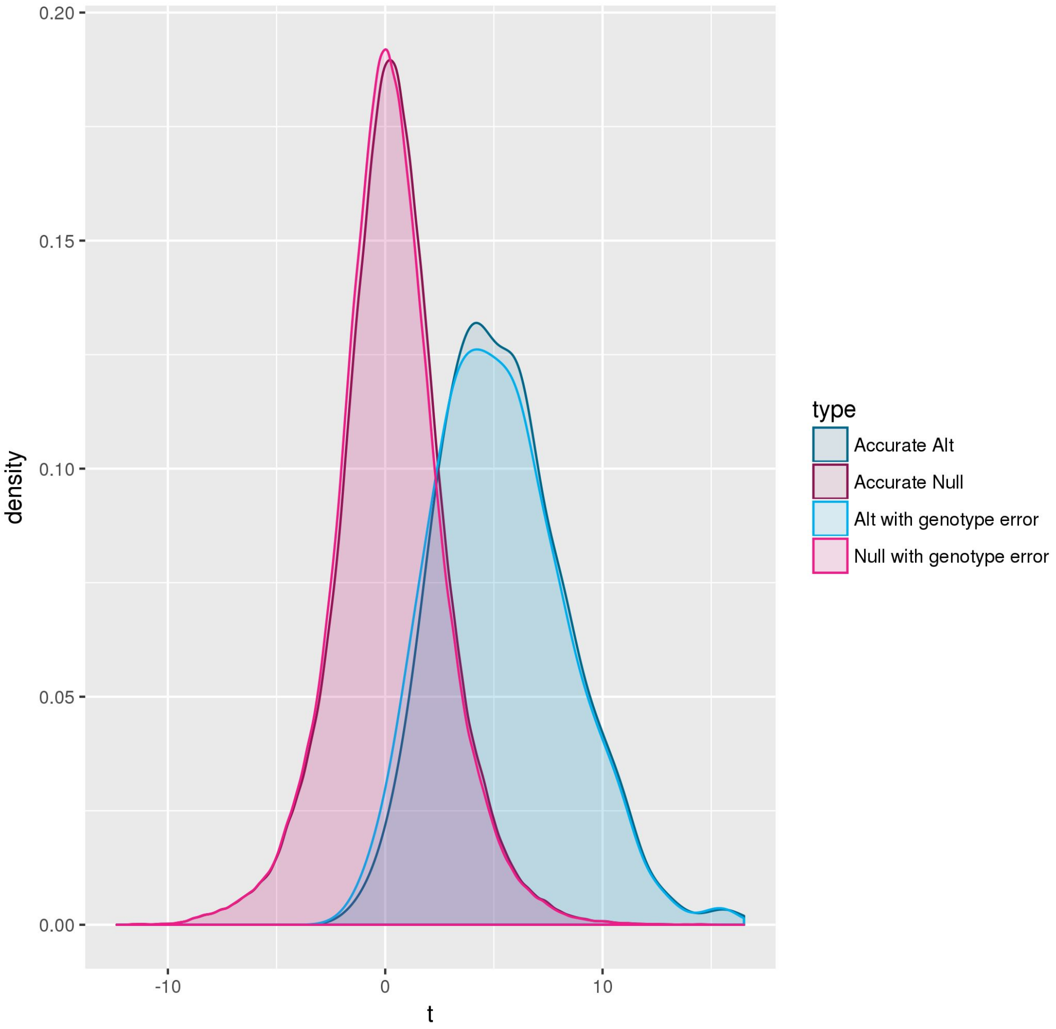
Background

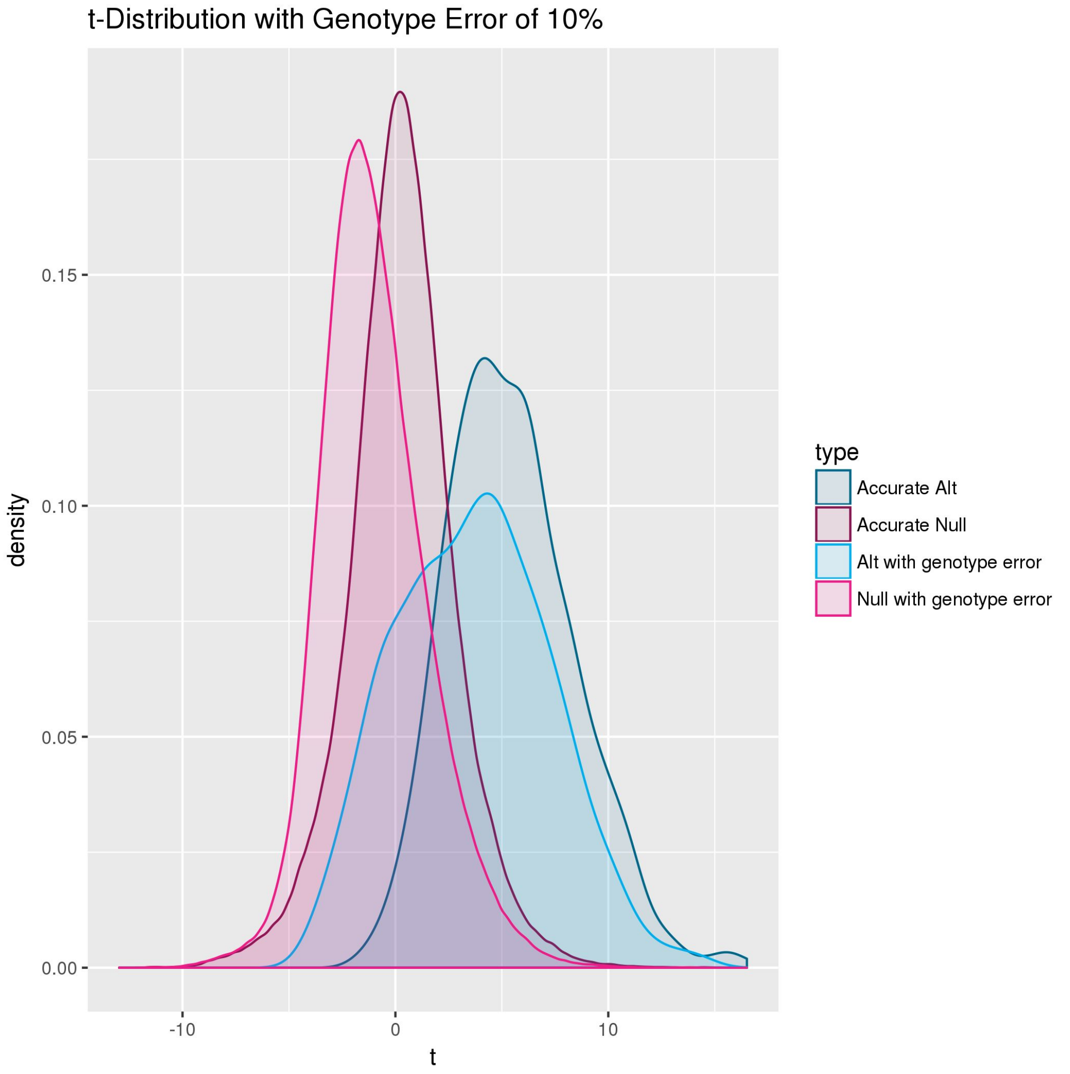
- Plant genetic data has characteristics distinct from human genetic data
  - Lower sequencing coverage
  - Allelic frequencies are dissimilar to humans
    - e.g. with breeding/selfing

Main methods

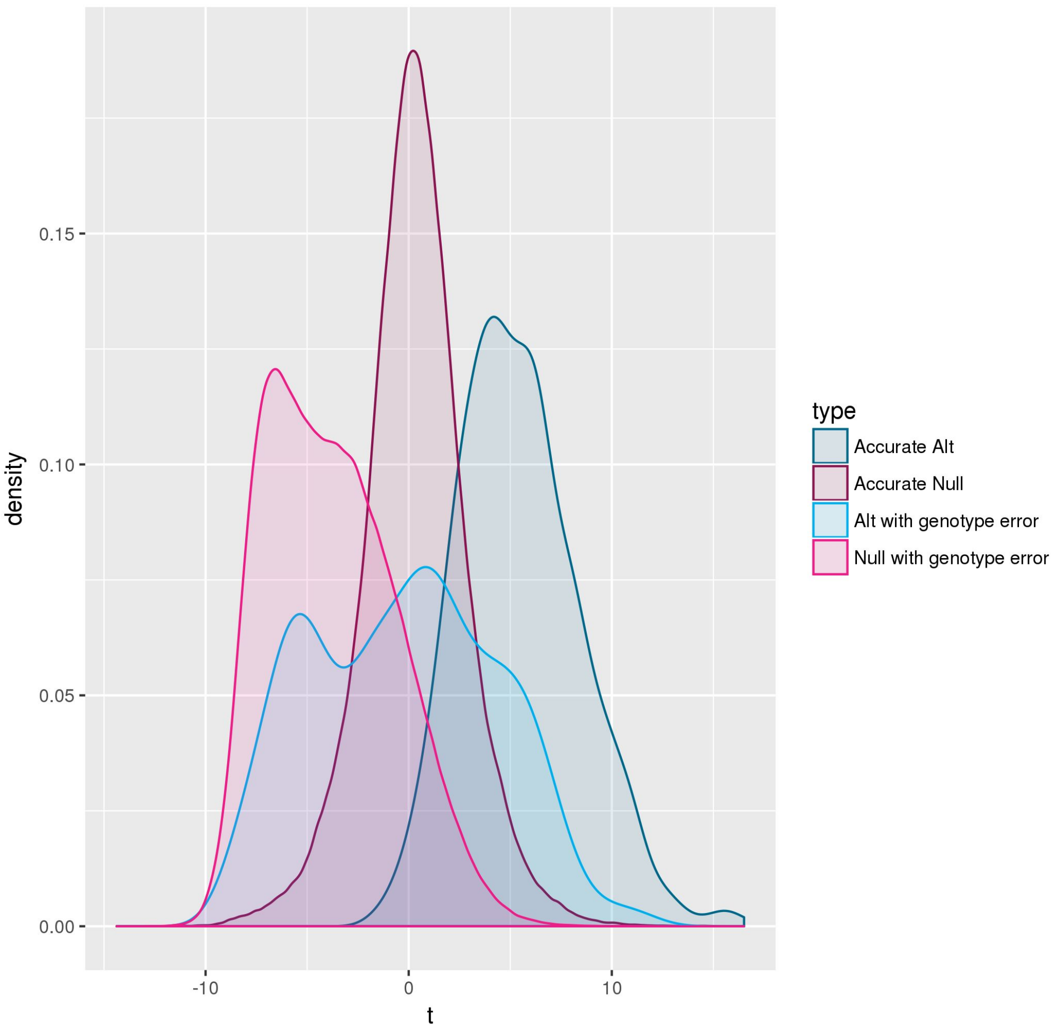
- Real wheat genetics data — simulated phenotypes
    - Minimal correlation between SNPs
- Error distribution
    - Haplotype level
    - Mapped geno → pheno → error rates → erroneous geno
- Parameters
    - Direction of the effect of causal SNPs (+/-)
    - Direction of correlation between phenotype and error (+/-)
    - Maximum error rates (1%, 10%, 30%)
- Test of association
    - Linear models

t-Distribution with Genotype Error of 1%




t-Distribution with Genotype Error of 10%
density
0.15
0.10
0.05
0.00
t
-10
0
10
type
Accurate Alt
Accurate Null
Alt with genotype error
Null with genotype error

t-Distribution with Genotype Error of 30%



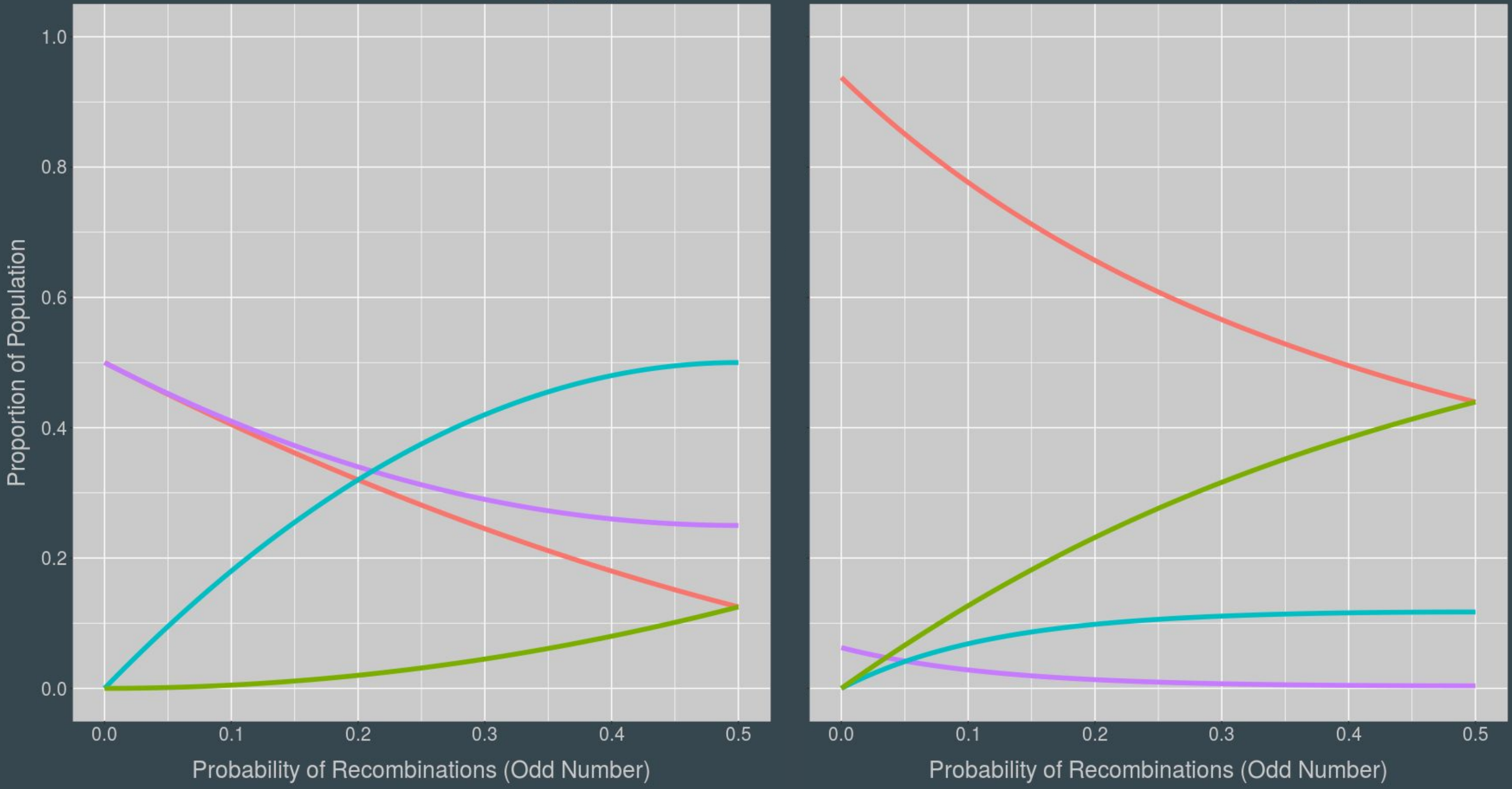
LaByRInth (Low-coverage Biallelic R Imputation)

- Accurate imputation can alleviate problems with differential errors
- LB-Impute is one software program to do this
  - Uses a hidden Markov model
- Last summer
  - Port LB-Impute from Java to R for accessibility
- This summer
  - Explore recombinant inbred line specific structure
  - Symbolic equations generated for recombinant lines
  - Transition frequencies
    - Estimate recombination probabilities instead of using physical distance mapping
    - Informed transition probabilities in hidden Markov model

Estimating probability of recombination

- For a pair of loci, look at frequencies of “transition types” across all members
- Dependent on generation of inbreeding and probability of an odd number of recombinations between the loci

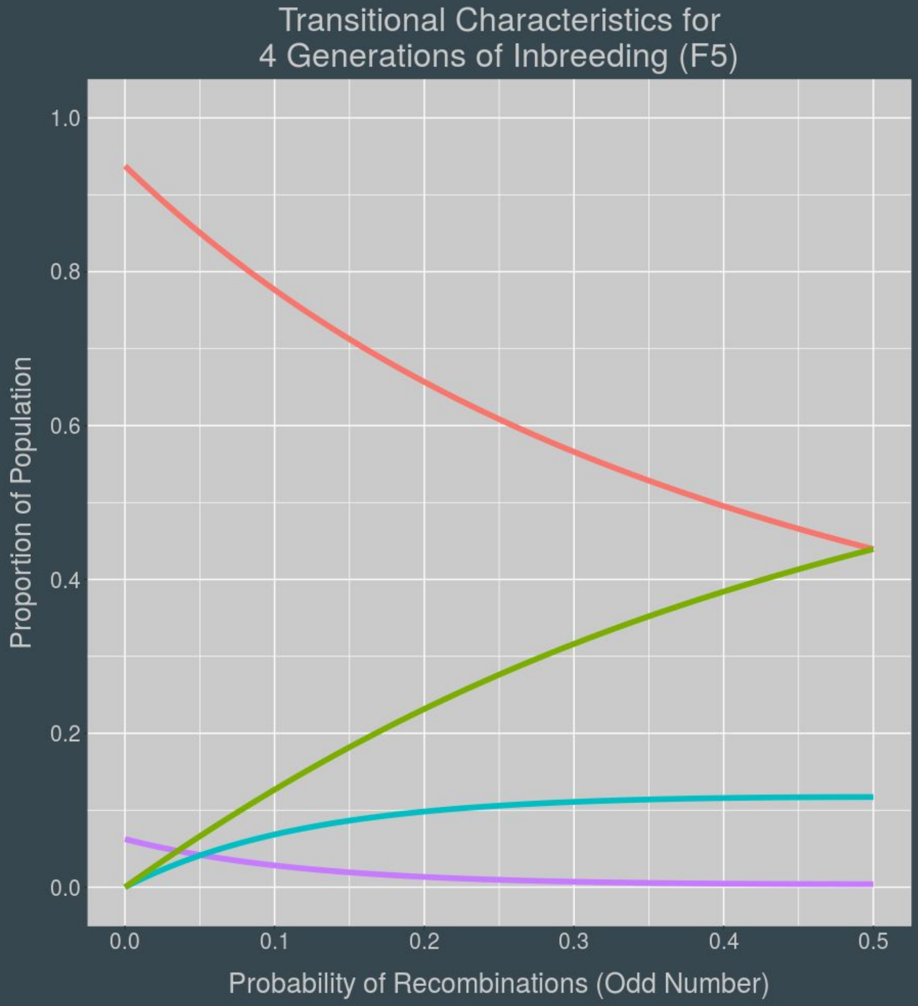
Marker *i*	**Marker *i+1***
Homozygous	Homozygous for same allele
Homozygous	Homozygous for other allele
Homozygous	Heterozygous
Heterozygous	Homozygous
Heterozygous	Heterozygous

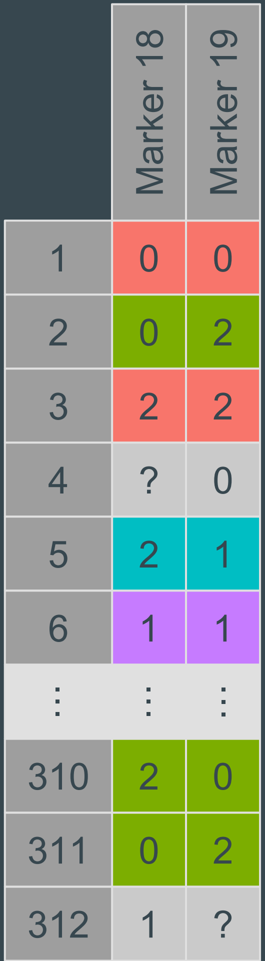


Transitional Characteristics for
1 Generation of Inbreeding (F2)

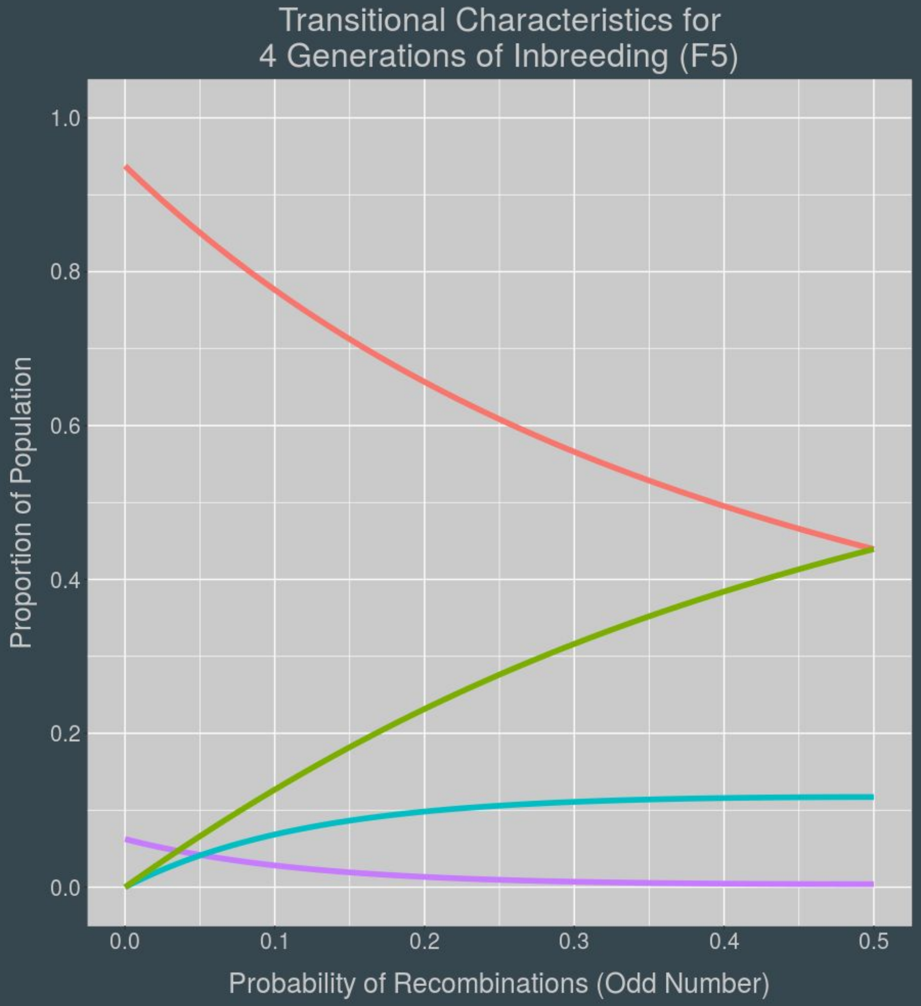
Transitional Characteristics for
4 Generations of Inbreeding (F5)

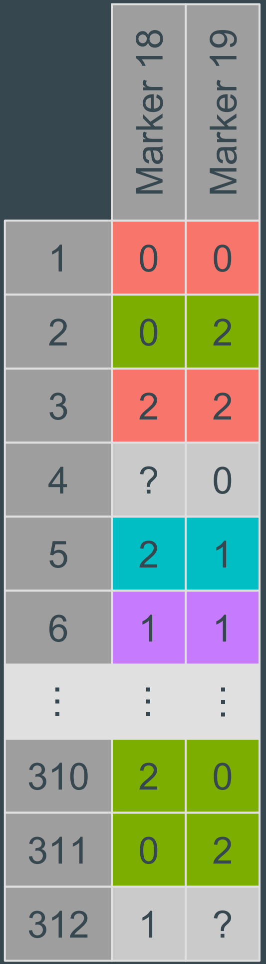
	Marker 18	Marker 19
1	0	0
2	0	2
3	2	2
4	?	0
5	2	1
6	1	1
⋮	⋮	⋮
310	2	0
311	0	2
312	1	?





	Marker 18	Marker 19
1	0	0
2	0	2
3	2	2
4	?	0
5	2	1
6	1	1
⋮	⋮	⋮
310	2	0
311	0	2
312	1	?

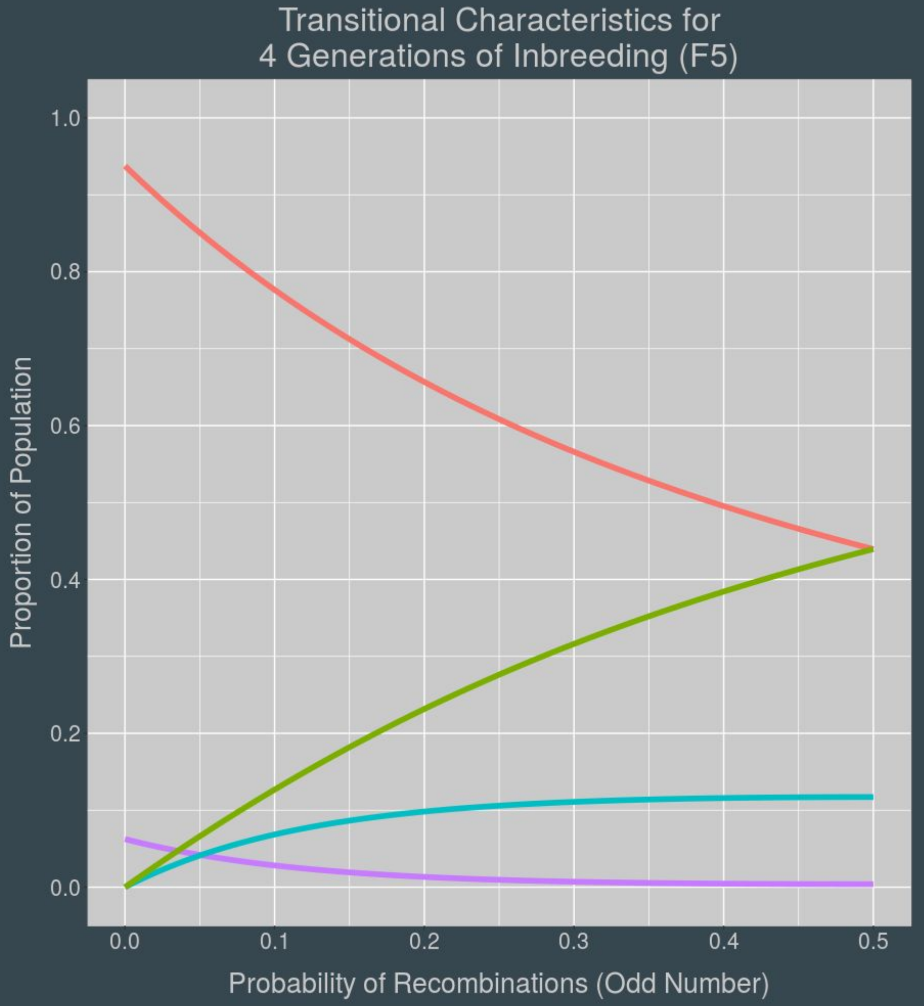


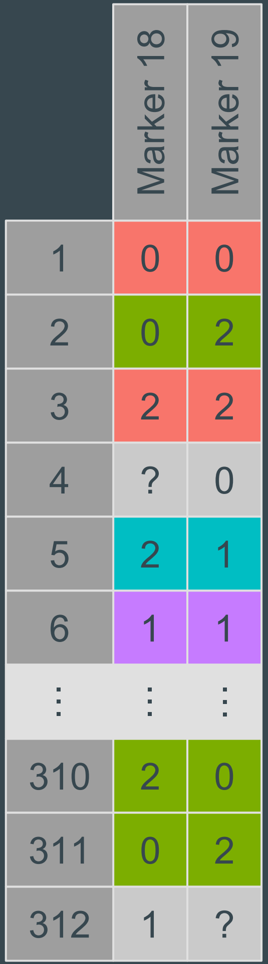


	Marker 18	Marker 19
1	0	0
2	0	2
3	2	2
4	?	0
5	2	1
6	1	1
⋮	⋮	⋮
310	2	0
311	0	2
312	1	?

Type	Count	Prop
RED	33	0.51
GREEN	24	0.37
BLUE	7	0.11
VIOLET	1	0.01
Total	65	1.00

Missing	247	N/A

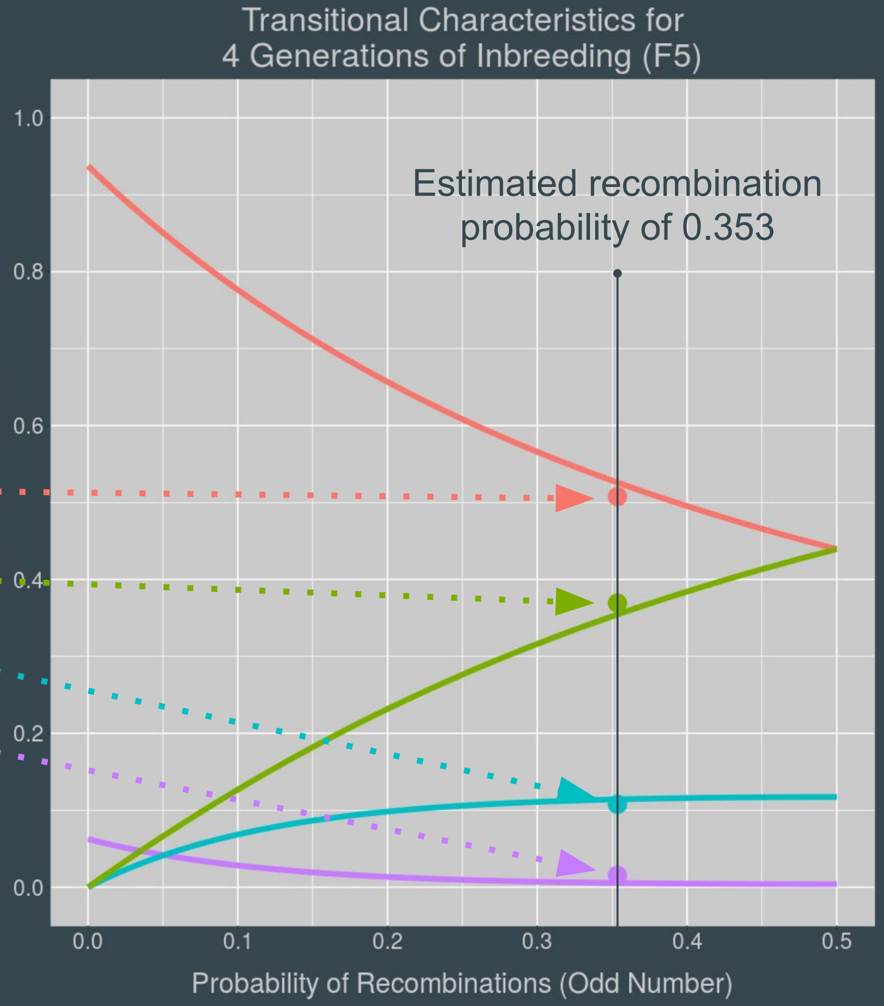


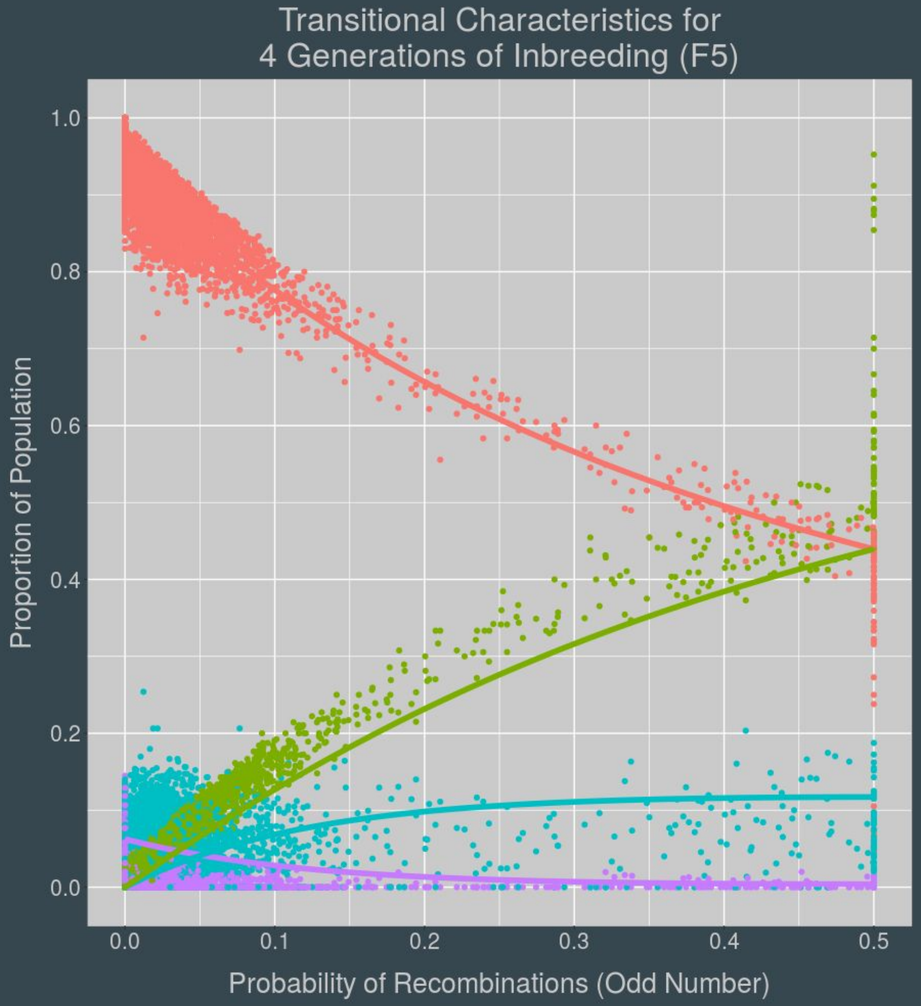


	Marker 18	Marker 19
1	0	0
2	0	2
3	2	2
4	?	0
5	2	1
6	1	1
⋮	⋮	⋮
310	2	0
311	0	2
312	1	?

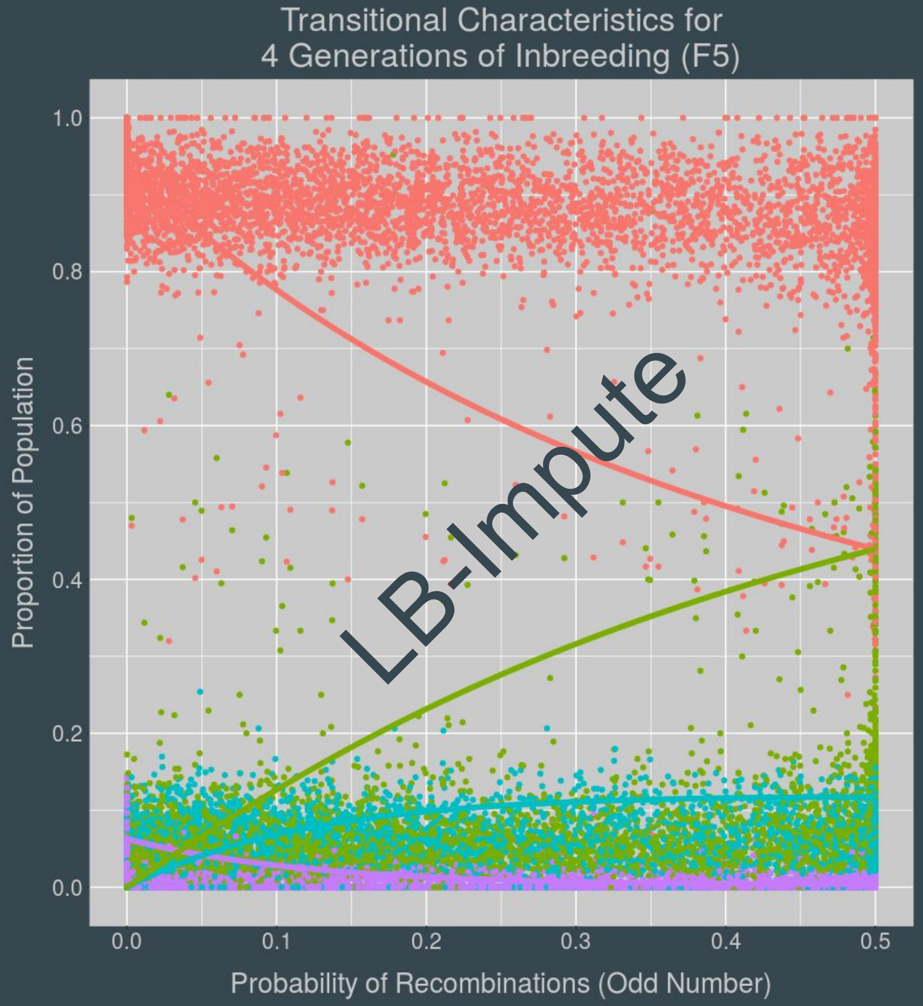
Type	Count	Prop
RED	33	0.51
GREEN	24	0.37
BLUE	7	0.11
VIOLET	1	0.01
Total	65	1.00

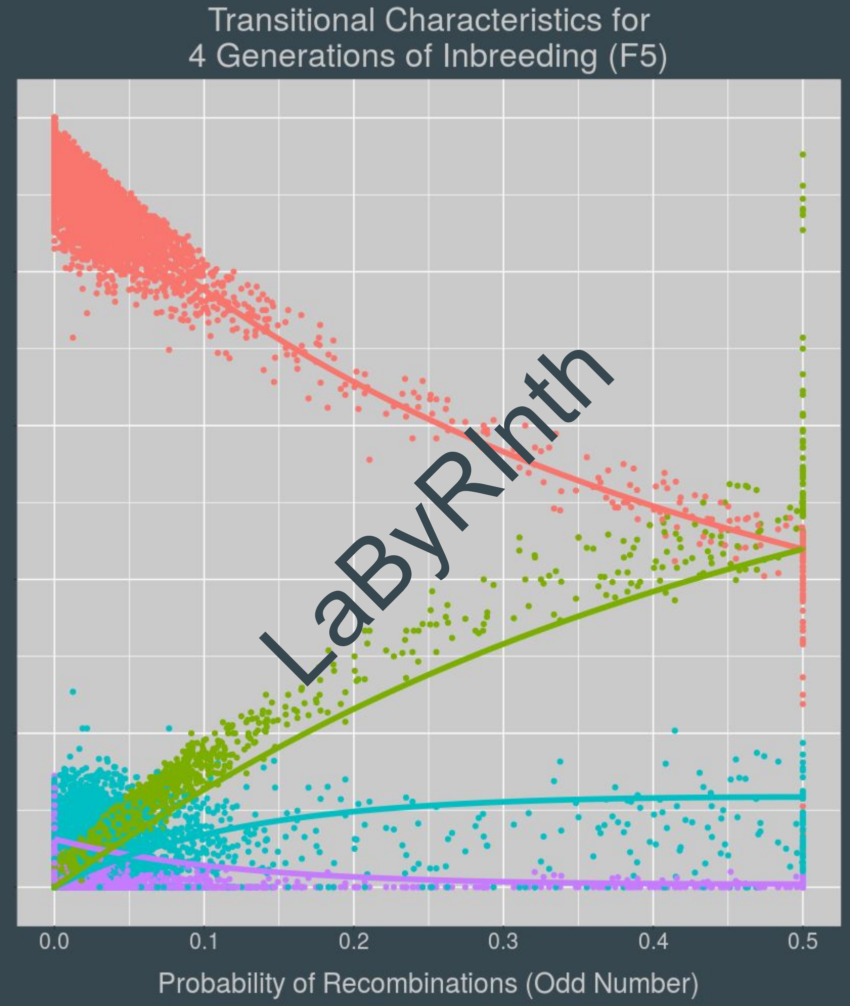
Missing	247	N/A


Transitional Characteristics for 4 Generations of Inbreeding (F5)



Transitional Characteristics for 4 Generations of Inbreeding (F5)


Transitional Characteristics for
4 Generations of Inbreeding (F5)
LB-Impute
Proportion of Population
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
Probability of Recombinations (Odd Number)


Transitional Characteristics for
4 Generations of Inbreeding (F5)
LaByRInth
0.0
0.1
0.2
0.3
0.4
0.5
Probability of Recombinations (Odd Number)

Future Work

- Uncertainty analysis
  - Develop methods
    - Estimate error rates
    - Adjust test statistics to compensate for inflated type I error
  - Extend results to phenotypic error
- Imputation
  - Update Markov transition probabilities using recombination probabilities
  - Release LaByRInth as an R package
  - Publish full LaByRInth paper (recently submitted software note)

References

- Edwards, B., Haynes, C., Levenstien, M., Finch, S., Gordon, D.: Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Gen. 6, 18. (2005). https://doi.org/10.1186/1471-2156-6-18
- Fragoso, C.A. , Heffelfinger, C., Zhao, H., Dellaporta, S.L.: Imputing genotypes in biallelic populations from low-coverage sequence data. Genetics, 202(2):487–495, (2016).
- Powers, S., Gopalakarishnan, S., Tintle, N.: Assessing the Impact of Non-Differential Genotyping Errors on Rare Variant Tests of Association. Hum. Hered. 152--159. (2011).

Acknowledgements

Dr. Nathan Tintle — Dr. Jesse Poland — Jason Westra — Dr. Jeff Ploegstra

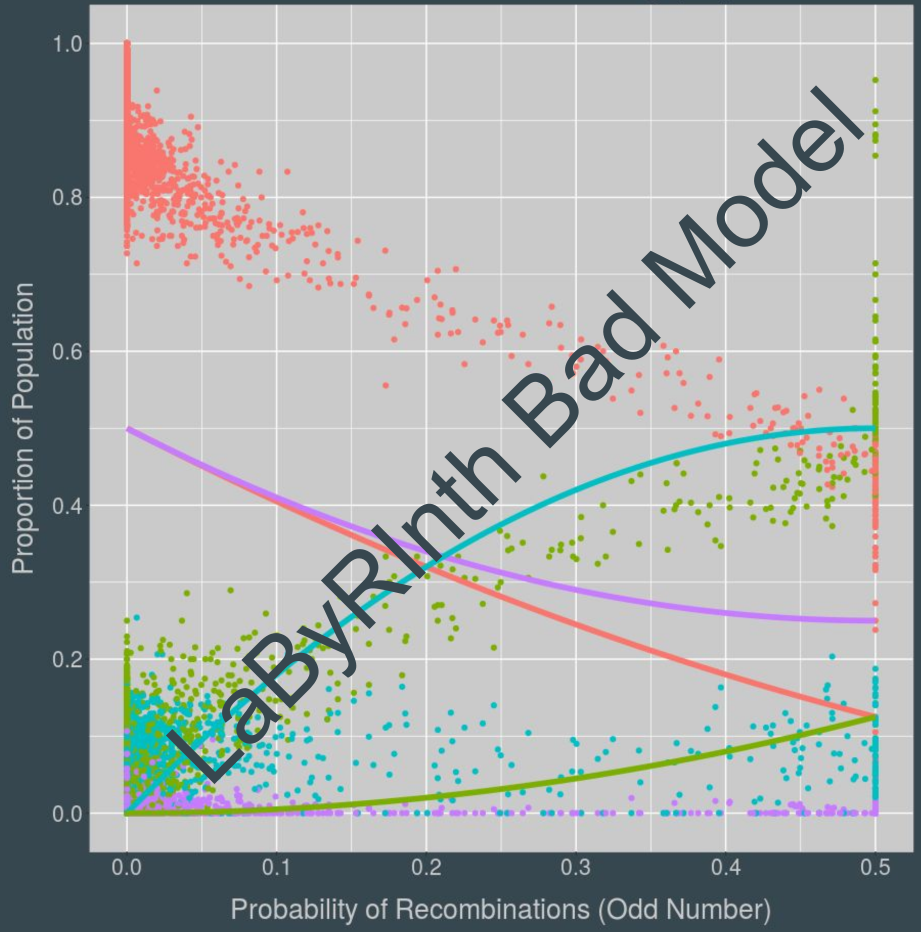
Funding

This work has been supported by the National Science Foundation grant IOS-1238187

Does the data fit too well in comparison?

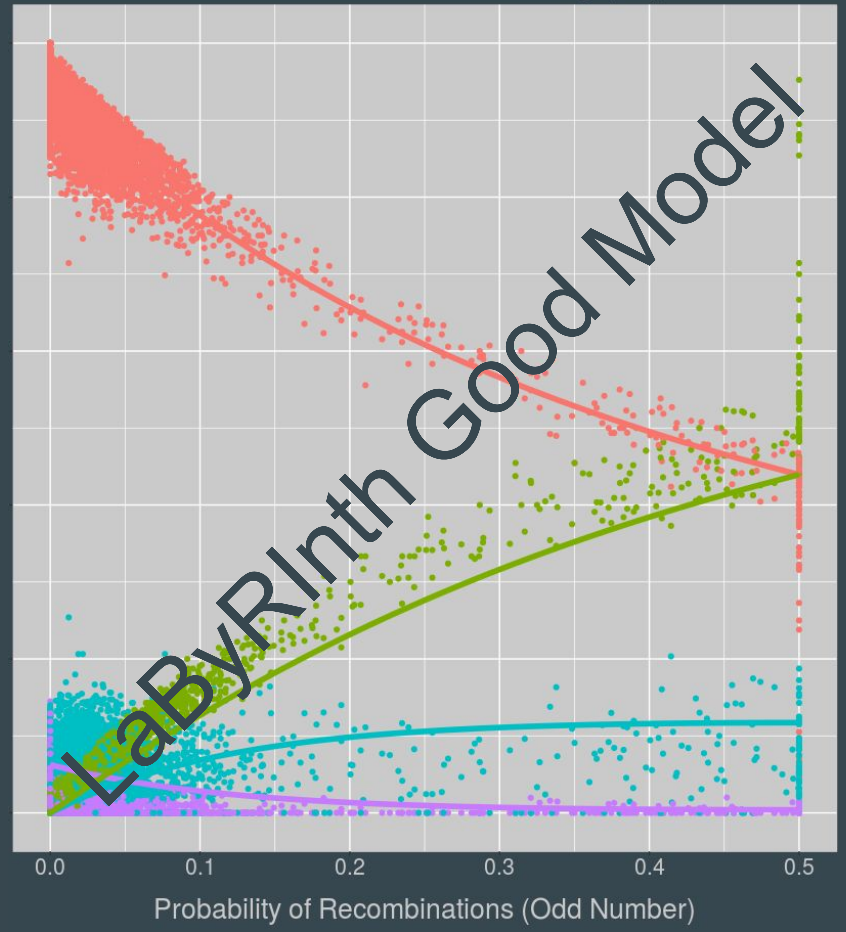
Are we fitting data well, but to the wrong model?

Transitional Characteristics for 1 Generation of Inbreeding (F2)



LaByRInth Bad Model

Transitional Characteristics for 4 Generations of Inbreeding (F5)



LaByRInth Good Model

Recombinant Inbred Line

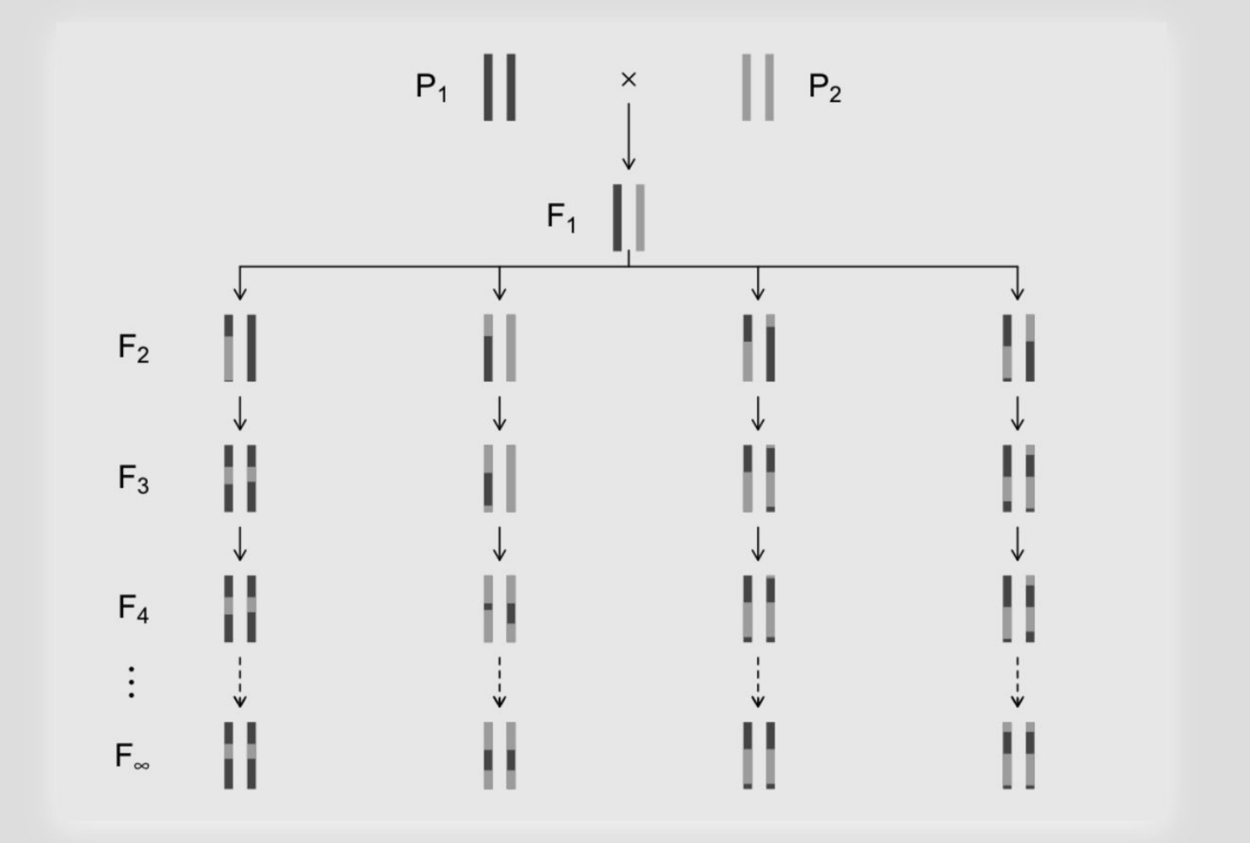


Image from: **https://www.biostat.wisc.edu/~kbroman/talks/FunQTL/**