

Fast “Coalescent” Simulation

Paul Marjoram, Keck School of
Medicine, University of Southern
California, Los Angeles.

Goal

- n A faster way of producing coalescent data for chromosomal-length regions (cf. existing methods such as Hudson's *ms*)

Why?

Why? – natural progression

slow

quick

Why? – natural progression

slow

quick



Why? – natural progression

slow



quick



Why? – natural progression

slow

quick



Why? – natural progression

slow



quick



Why? – natural progression

slow

quick



Why? – natural progression

slow



quick



Why? - Growth of genome-wide data

- n e.g. SNP-chips
- n New analysis methodologies being developed
- n Need to test them somehow.
 - Usual strategy: simulate test data
 - Problem: traditional (coalescent) models too slow.
- n Simulation-based analysis methods (Rejection algorithms, Importance Sampling)

Generating test data

n Real data + perturbation
– e.g. bootstrap resampling

n Model + simulation
– e.g. coalescent

Real data + perturbation

- n Advantage – ‘model’ is correct.
 - Don’t know how the data got there, but it used the correct model.
- n Disadvantage – subsequent perturbation adds noise (over-dispersion). What do we end up with?

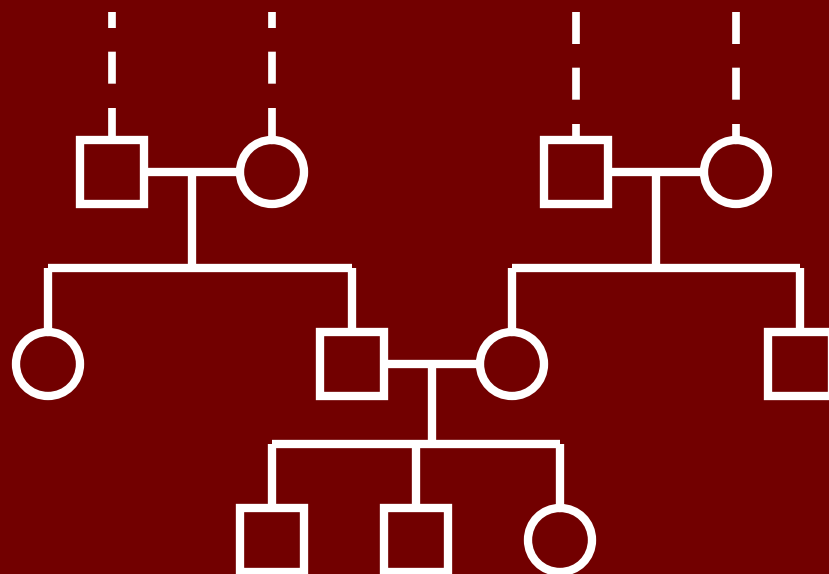
Model + simulation

- n Advantage – Know what you are getting
- n Disadvantage – May take a long while to get it + how accurate is the model?

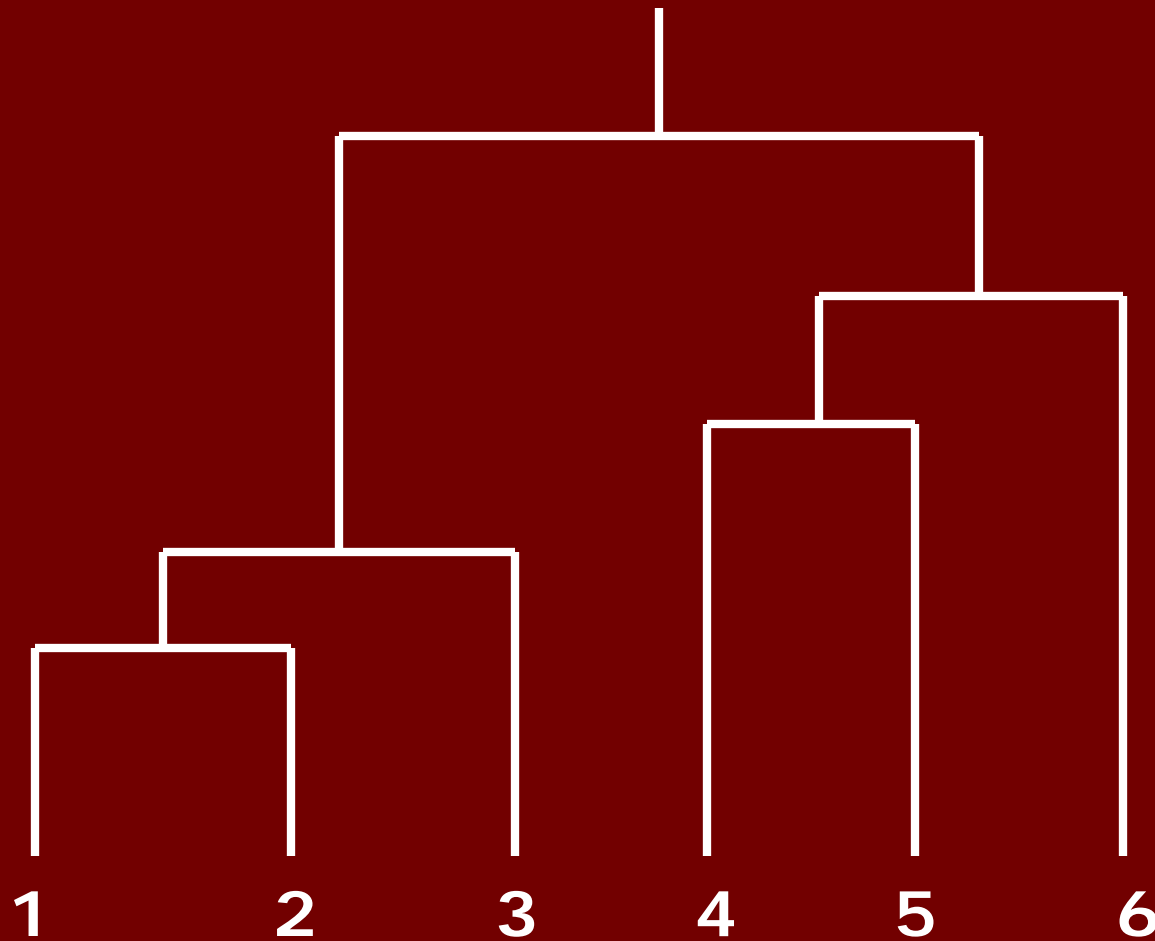
Model-based approach

- n Traditionally, many groups have used coalescent models
- n Such models are slow for chromosomal-length regions

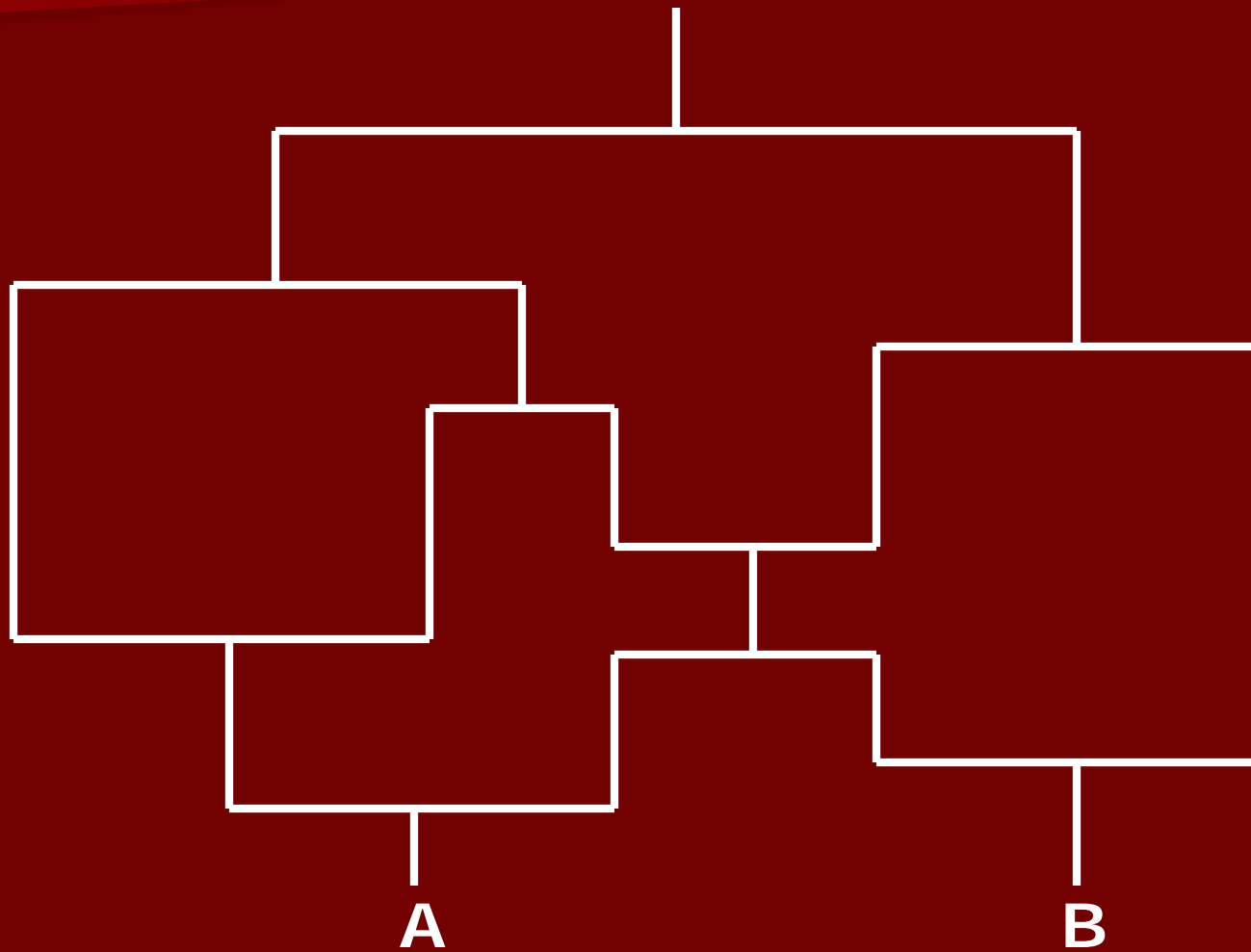
Ancestral history



Genes/SNPs have ancestors too



Coalescent – Longer region



Full coalescent models are slow for chromosomal-length regions

Run-times (secs) for ms (3 GB RAM)

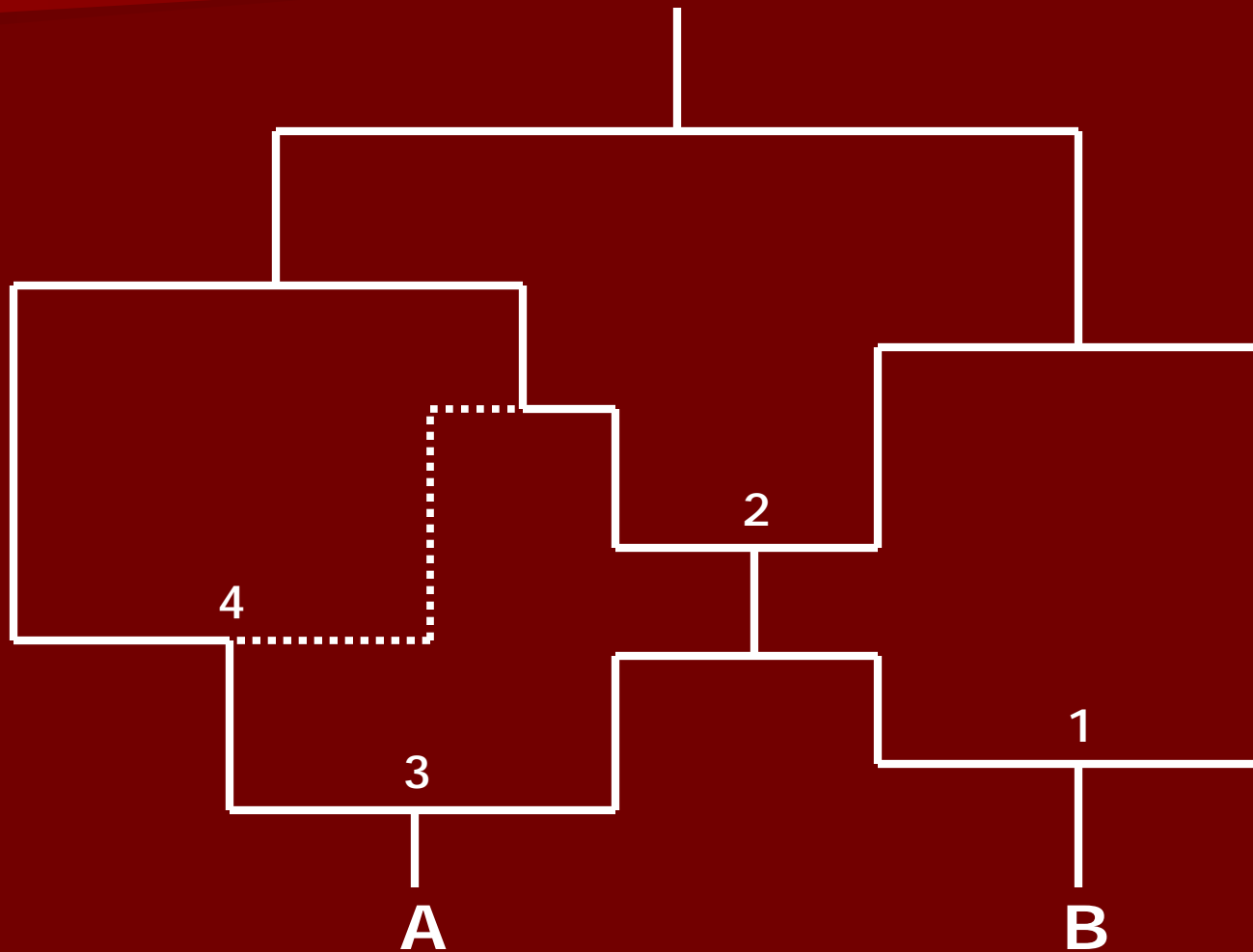
| Sample size | Length (Mb) | ms |
|-------------|-------------|--------|
| 1000 | 2 | 7.2 |
| | 5 | 62.6 |
| | 10 | 473.6 |
| | 20 | 6459.6 |
| | 50 | - |
| | 100 | - |
| | 200 | - |

Human chromosomes range from 50-200 Mbs

Run-times (secs) for ms (3 GB RAM)

| Sample size | Length (Mb) | ms |
|-------------|-------------|------|
| 4000 | 2 | 10.6 |
| | 5 | - |
| | 10 | - |
| | 20 | - |
| | 50 | - |
| | 100 | - |
| | 200 | - |

Why so slow? – Longer region.
Memory + Parts of ARG are not
used



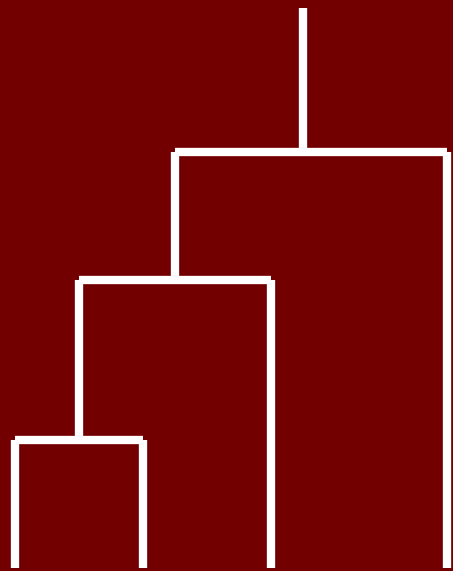
Find a faster way....How?

- n Use an approximation to the coalescent
- n Advantage - it will be faster
- n Disadvantage – it's an approximation (to an approximation)

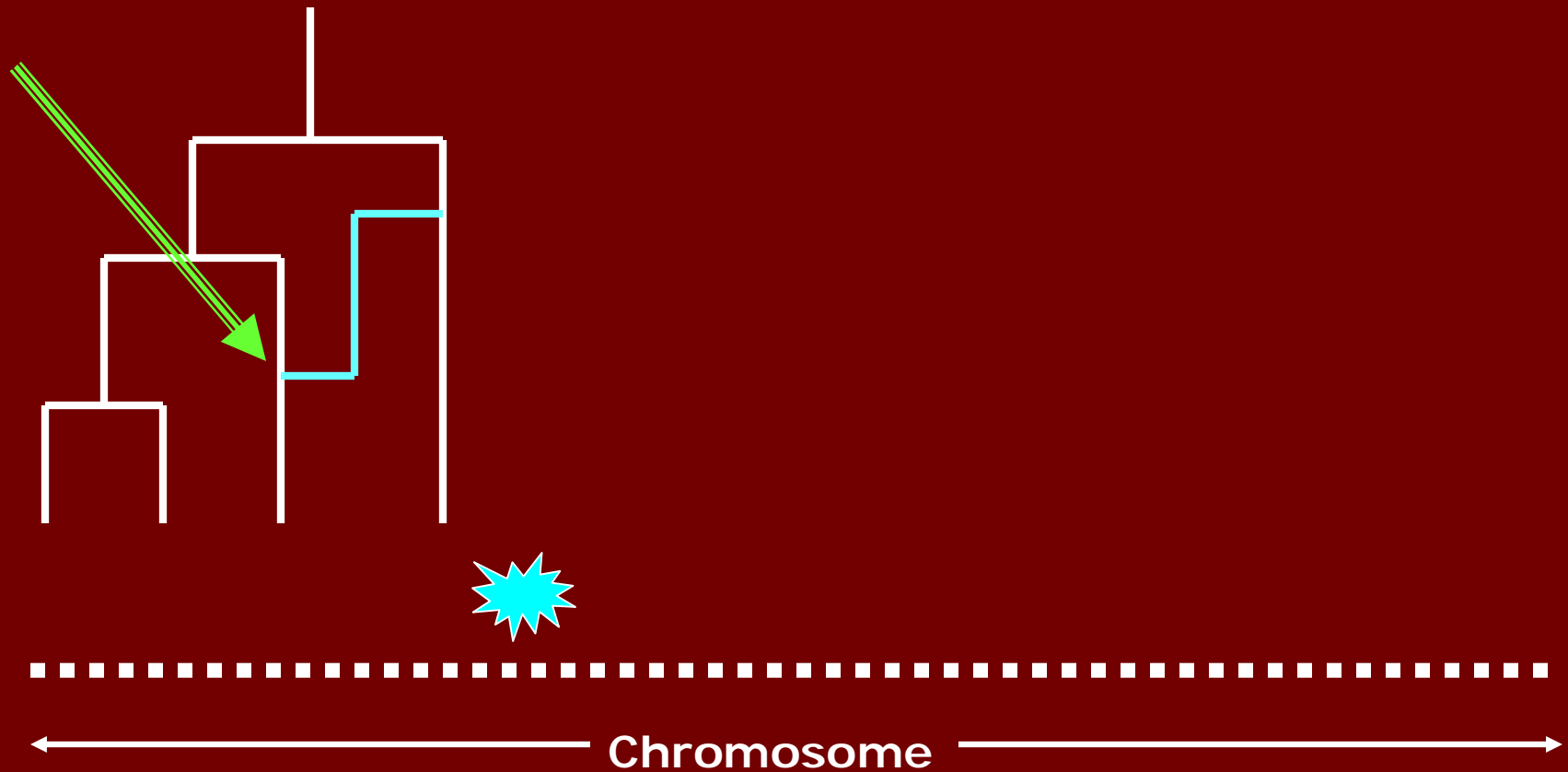
Something to worry about?

- n Is the coalescent model appropriate for chromosomal-length regions?
- n It's an asymptotic model that assumes events are rare (i.e. prob. of order $1/N$).
- n e.g. Recombination is not rare in this context.

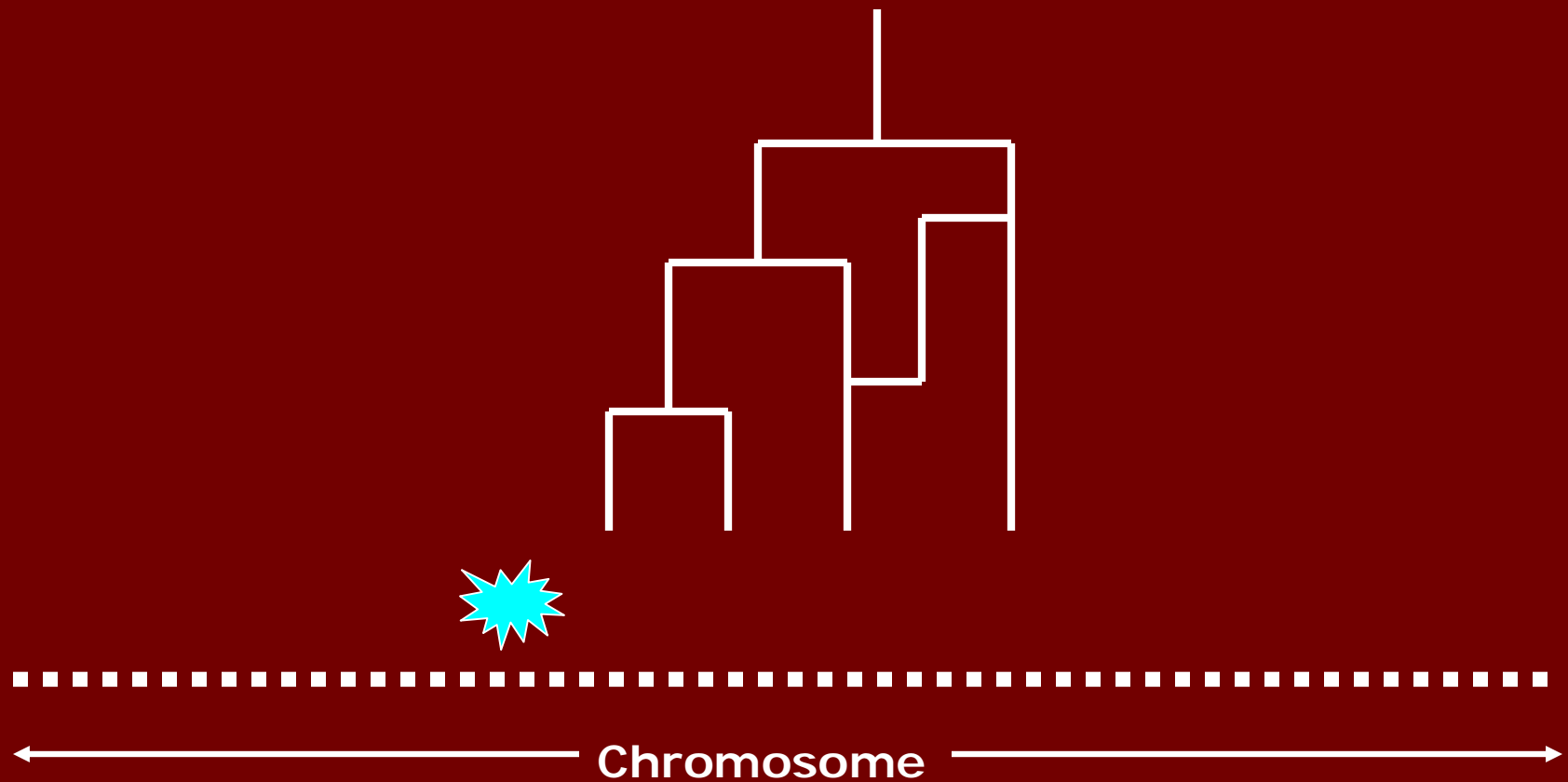
Wiuf and Hein "along the chromosome" algorithm



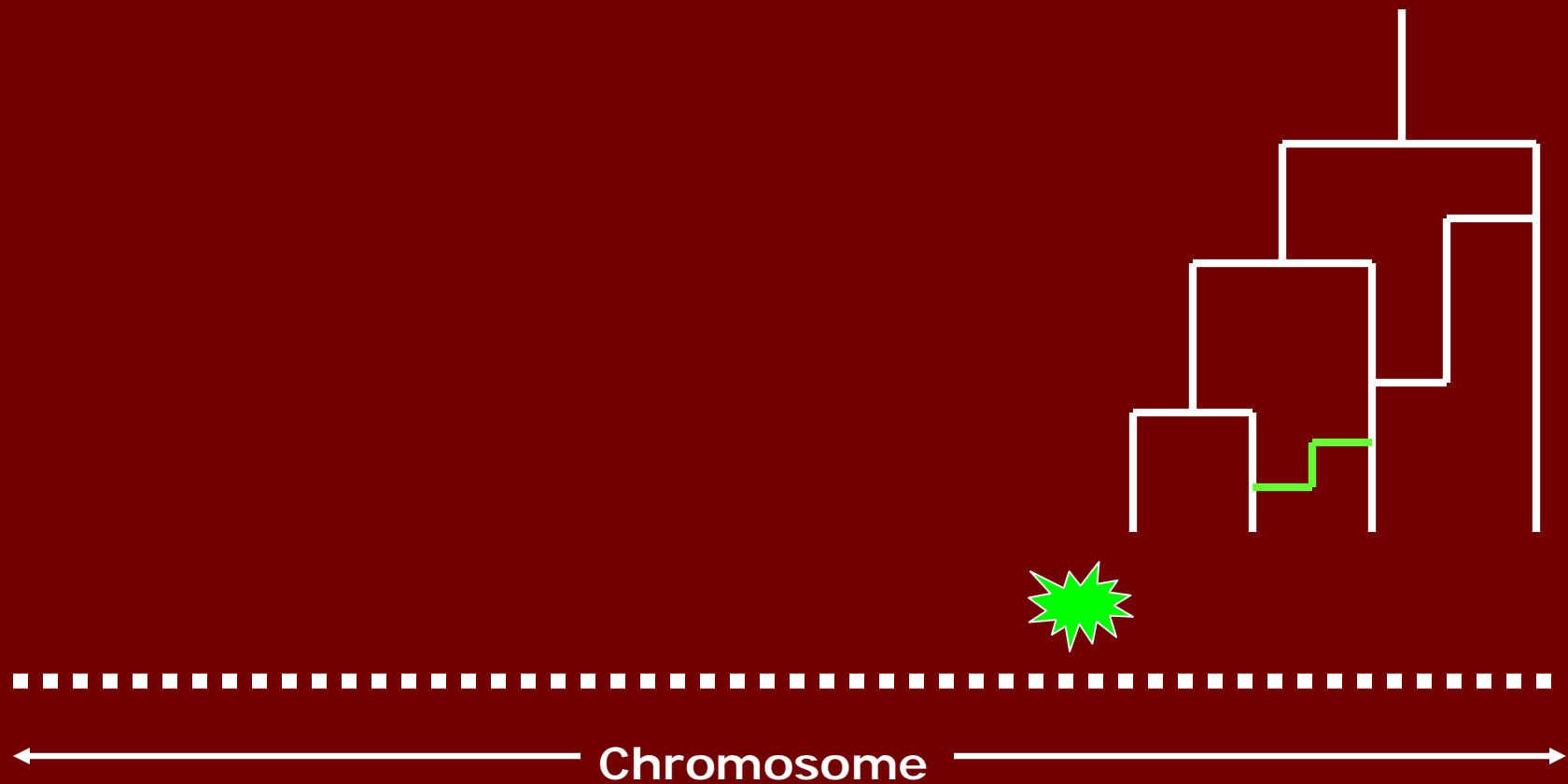
Wiuf and Hein "Along the Chromosome" algorithm



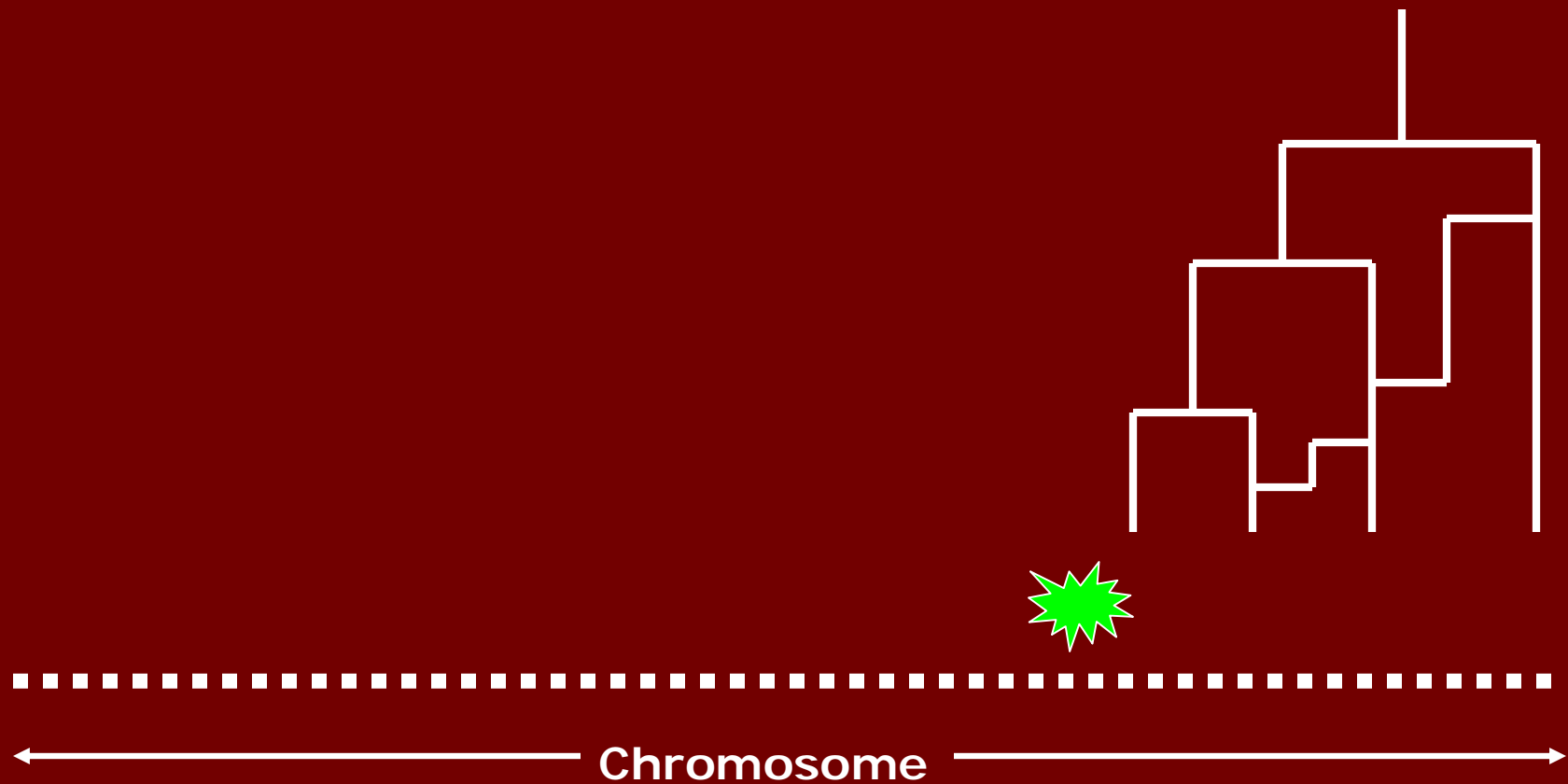
Wiuf and Hein "Along the Chromosome" algorithm



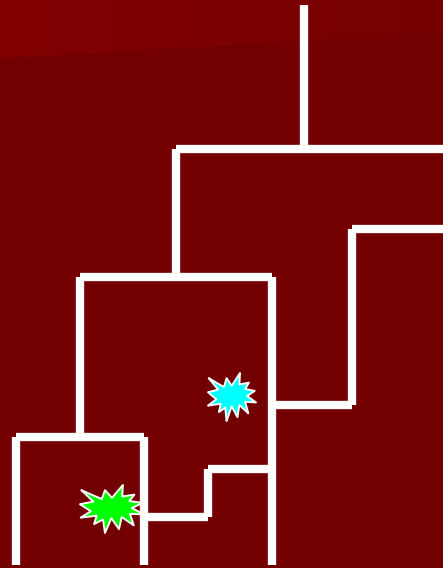
Wiuf and Hein "Along the Chromosome" algorithm



Wiuf and Hein "Along the Chromosome" algorithm



Builds subset of ARG

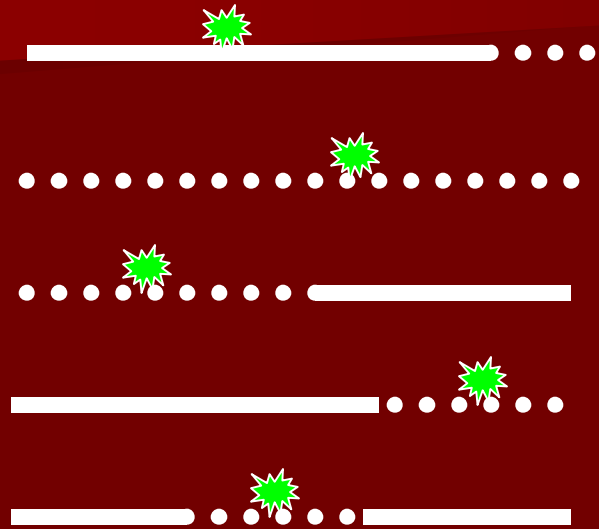


← Chromosome →

Comments

- n Slower than *ms* (larger subset)
 - Includes many recombinations in non-ancestral material
- n Suggests a simplification

Types of recombination



1. Ancestral material
2. Non-ancestral material
3. Non-ancestral material
4. Non-ancestral material
5. Non-ancestral material

—— ancestral material

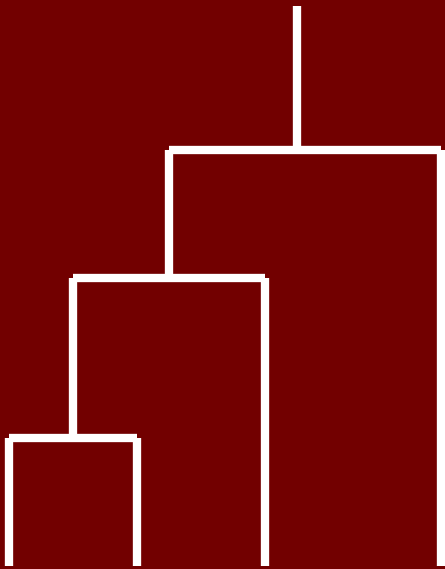
..... non-ancestral material

Types of recombination

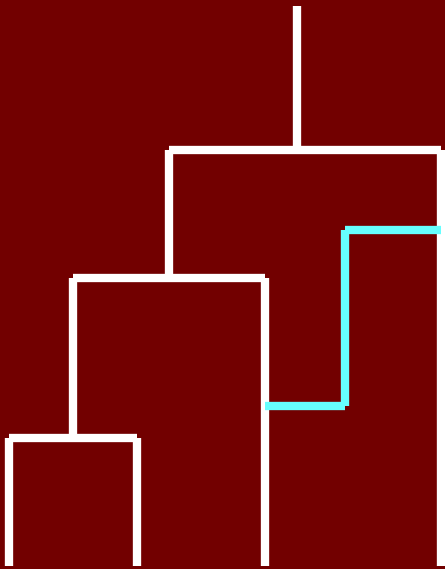


SMC algorithm (McVean and
Cardin 2005)

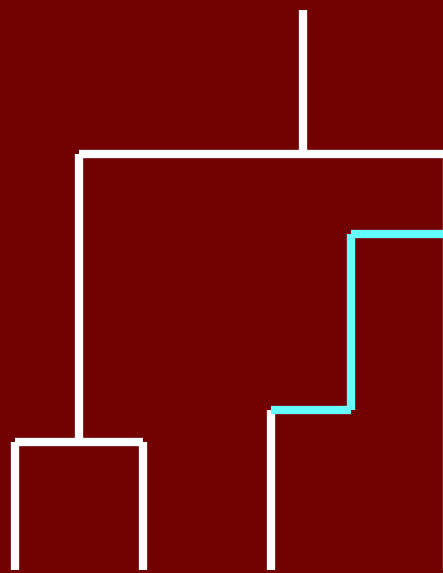
SMC' algorithm (Marjoram and Wall
2006)



Chromosome



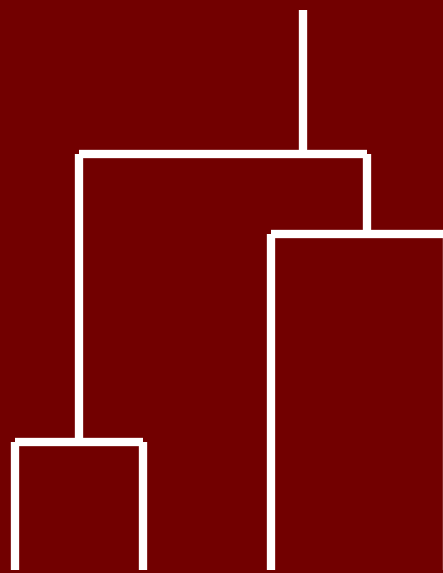
Chromosome



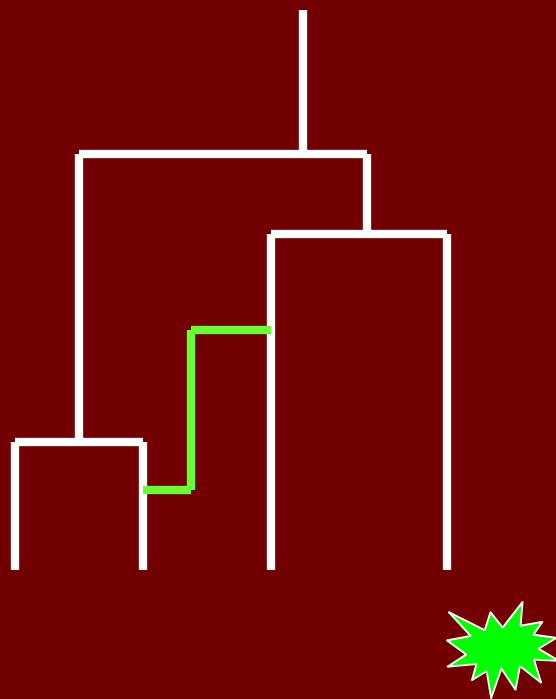
Chromosome

n SMC - delete old line, then add new;

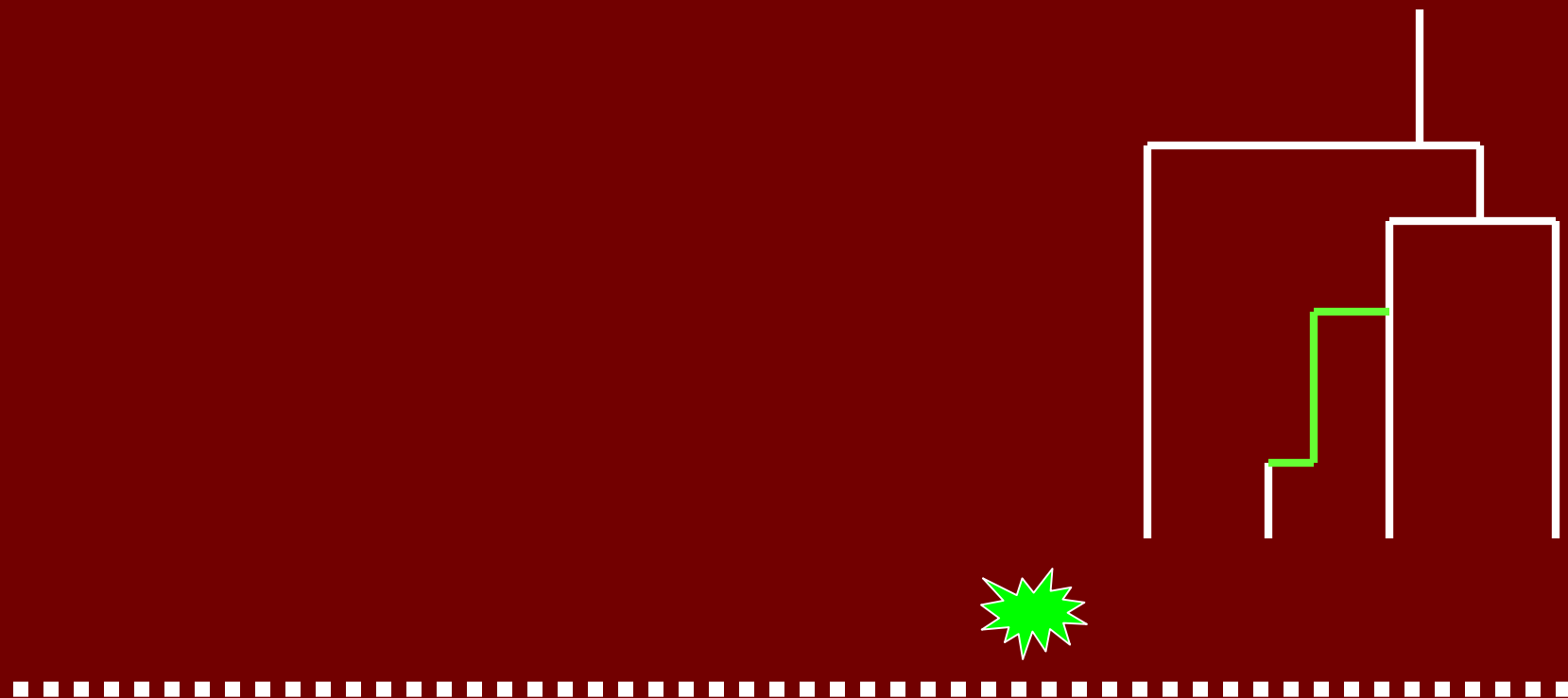
n SMC' – add new line, then delete old.



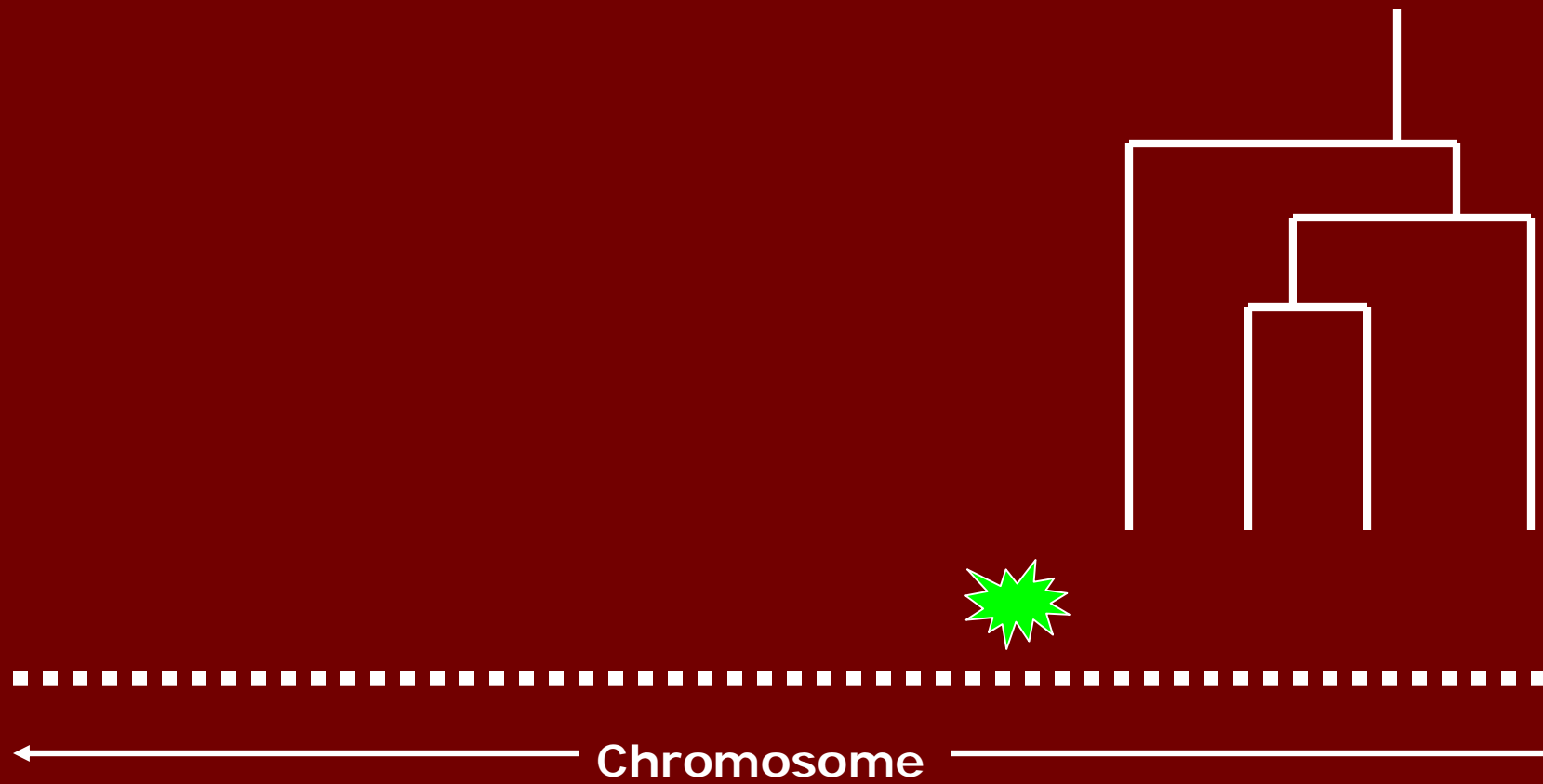
Chromosome

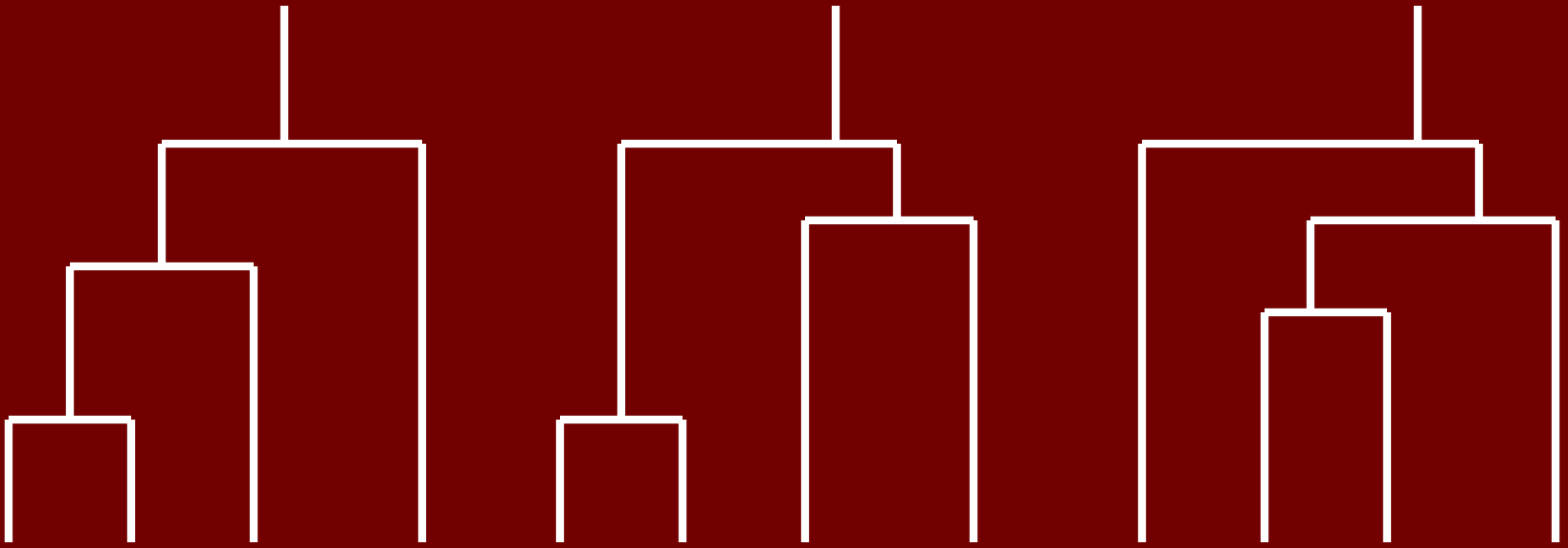


Chromosome



← Chromosome →





Chromosome

Outline of formal statement

- n $L(x)$: length of tree at $x \in [0,1]$
- n Simulate $y \sim \text{Exp}(L(x)\rho/2)$
- n If $x+y < 1$
 - Start next tree at $x+y$ by adding a recombination at a point chosen uniformly over the current tree
 - Add new line using usual coalescent prior
 - Delete old line
- n Else
 - Stop

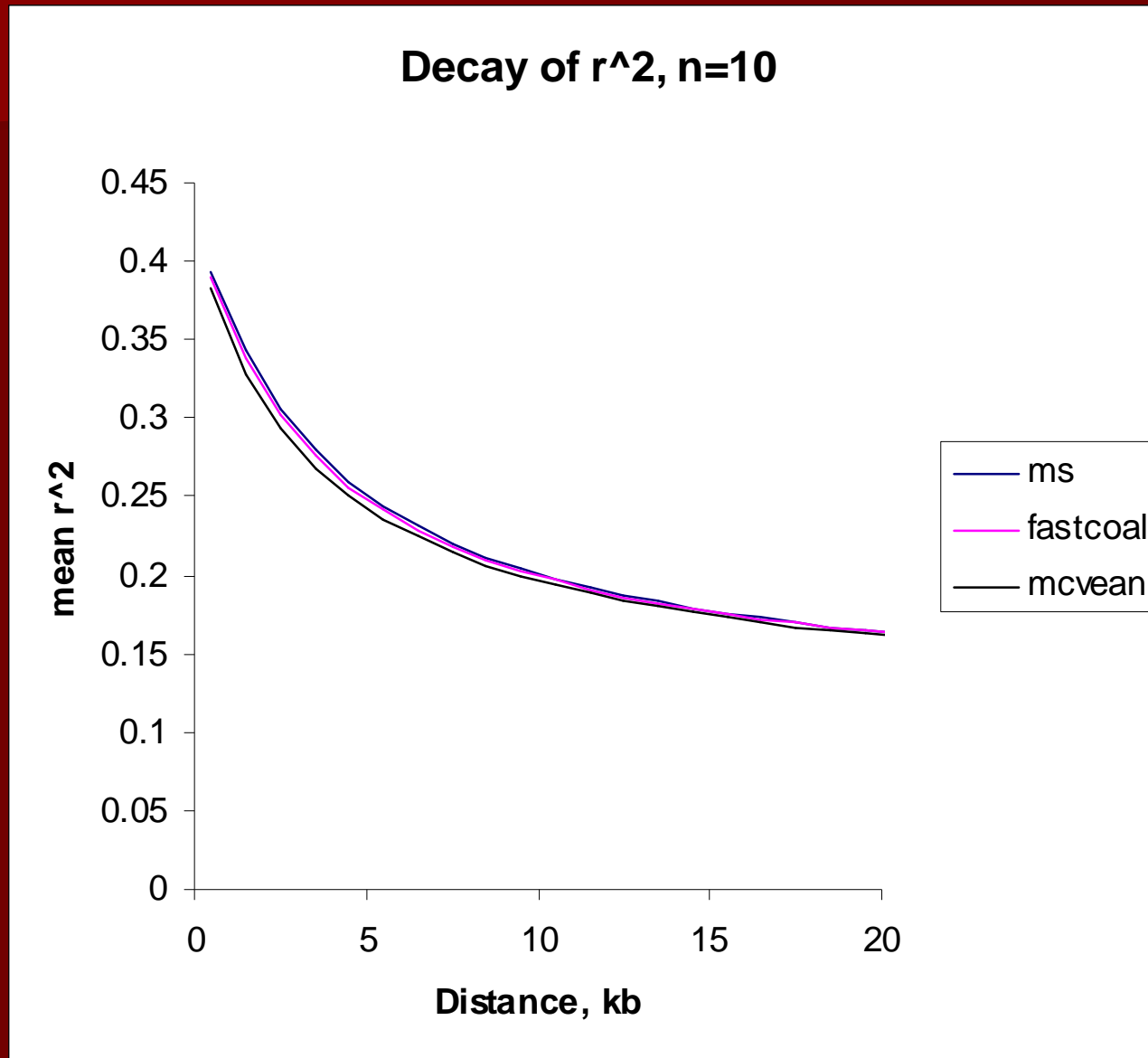
Run-times (secs) for ms (3 GB RAM)

| Sample size | Length (Mb) | ms | FastCoal |
|-------------|-------------|--------|----------|
| 1000 | 2 | 7.2 | 0.9 |
| | 5 | 62.6 | 2.1 |
| | 10 | 473.6 | 4.3 |
| | 20 | 6459.6 | 8.3 |
| | 50 | - | 20.9 |
| | 100 | - | 41.6 |
| | 200 | - | 83.9 |

Run-times (secs) for ms (3 GB RAM)

| Sample size | Length (Mb) | ms | FastCoal |
|-------------|-------------|------|----------|
| 4000 | 2 | 10.6 | 4.0 |
| | 5 | - | 10.4 |
| | 10 | - | 22.2 |
| | 20 | - | 40.7 |
| | 50 | - | 105.8 |
| | 100 | - | 201.5 |
| | 200 | - | 406.1 |

Behavior of LD



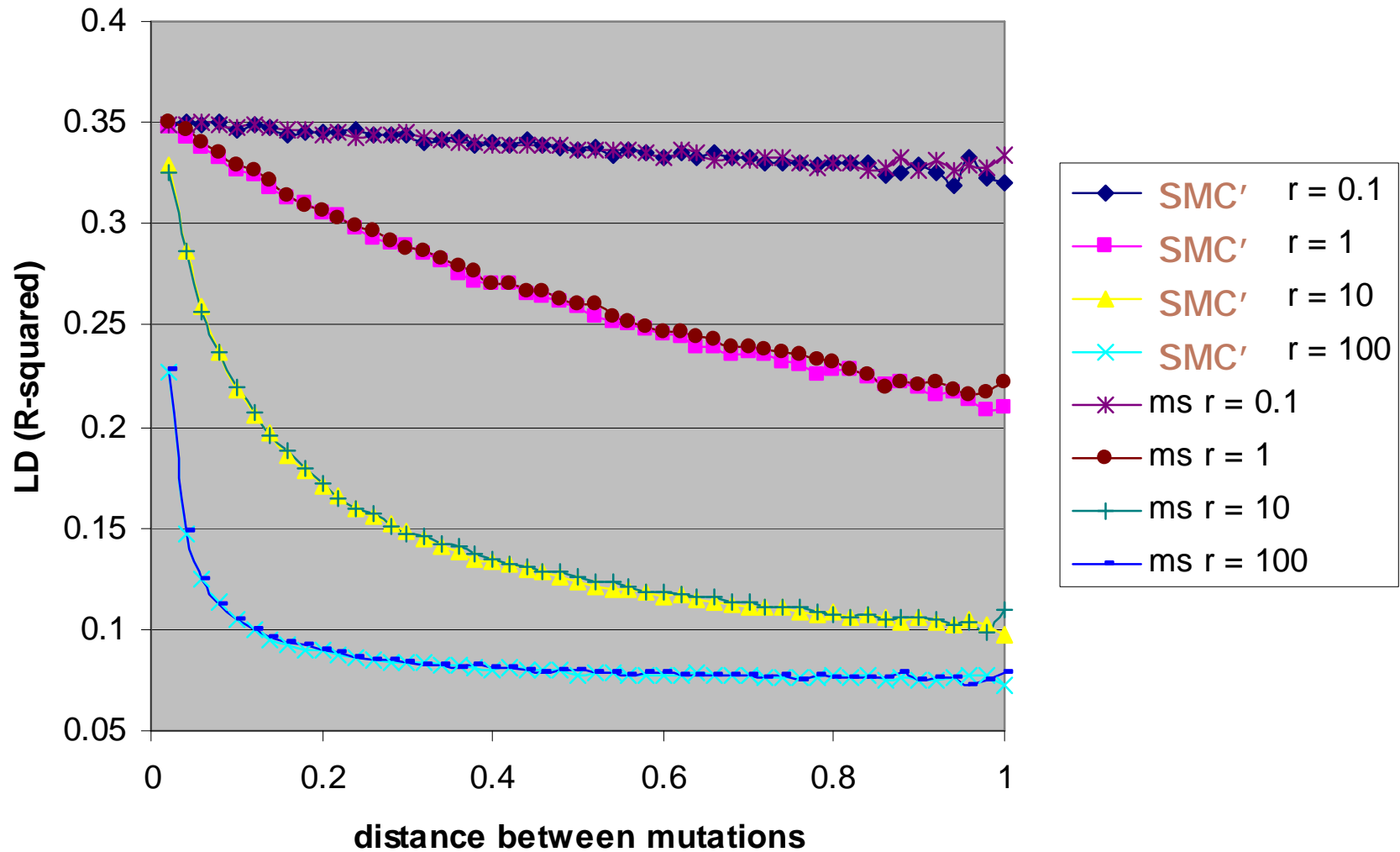
Ongoing work: Generalizations

Recombination/mutation rate variation

- n Appeal to Poisson process theory
- n Post-process the data.
- n To simulate a region of width D , with a rec. param. that is R times above norm:
 - Simulate region of width RD .
 - Contract to become width D .
 - Remove each mutation with prob. $(R-1)/R$

Population structure – mig. rate=1

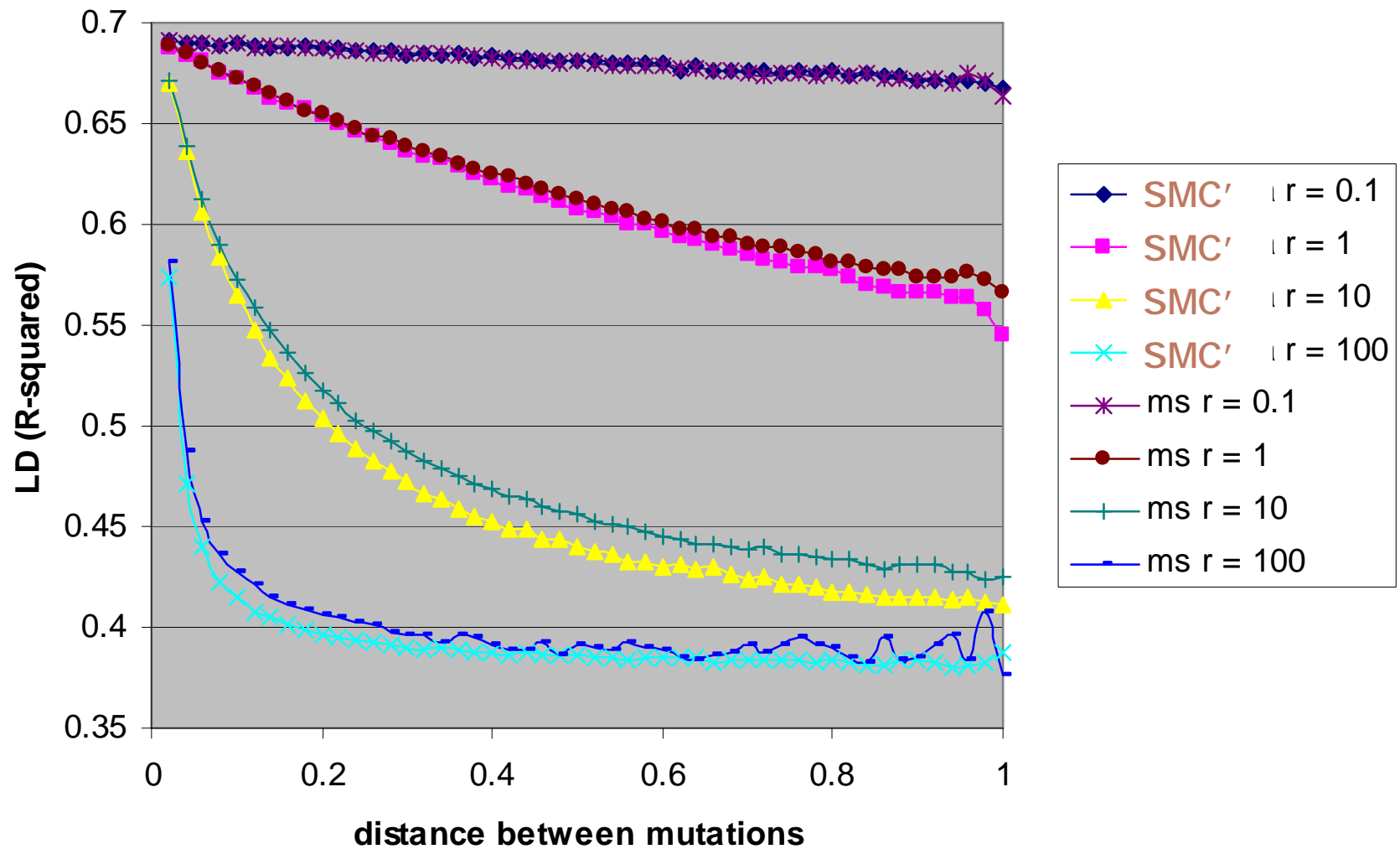
(n=20, theta=1, m=1, b=0, repeat=100000)



Population structure – mig.

rate 0.1

($n=20$, $\theta=1$, $m=0.1$, $b=0$, repeat=100000 (last of $m_s=10000$))

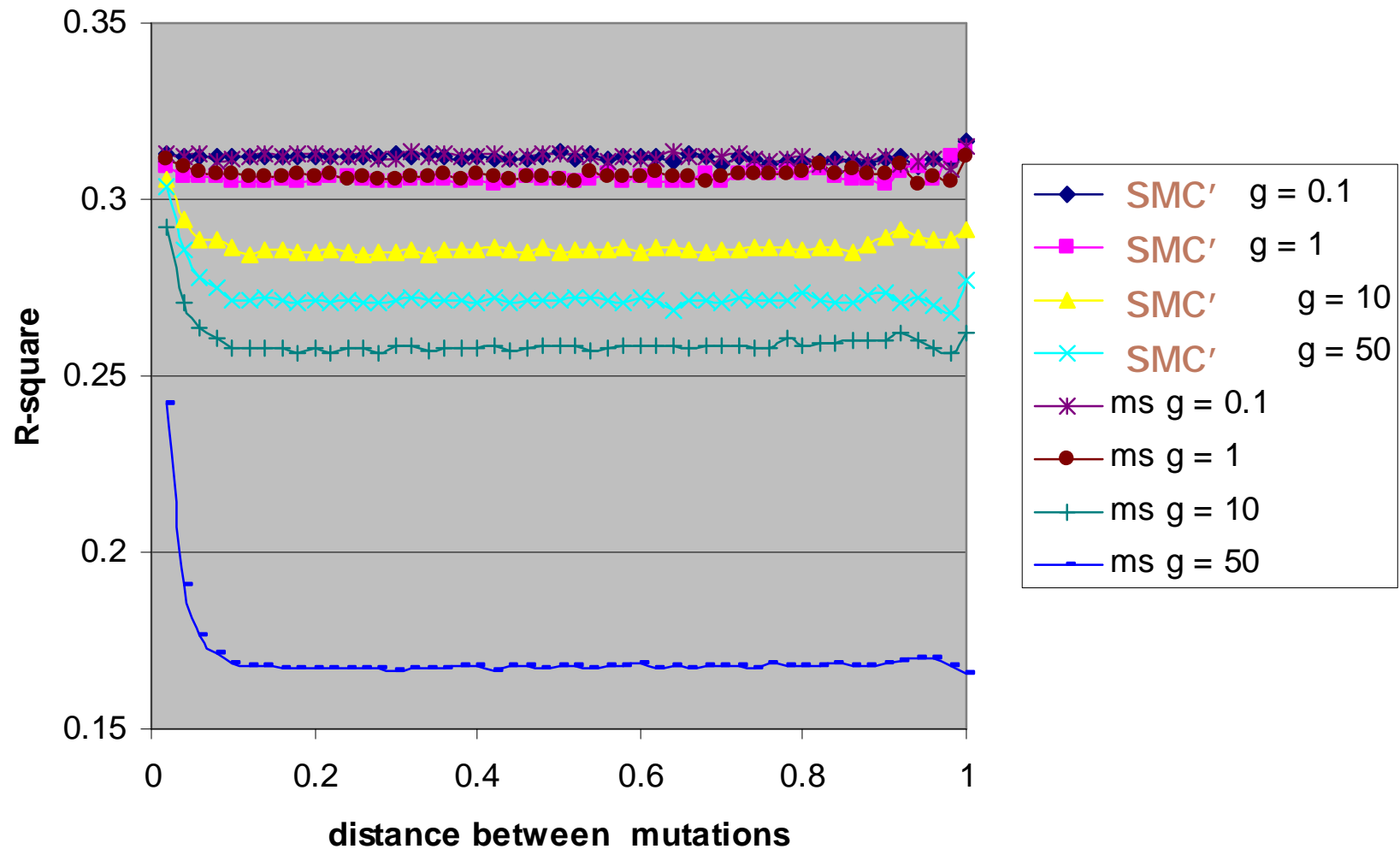


Gene conversion

- n Add gene conversion event in same way as a recombination
- n Return to previous tree at end of tract
- n Disallow other recombination/gene conversion within length of tract *on any line.*

Gene conversion

($n=20$, $\theta=5$, $\beta=0$, repeat=100000)



Remedy?

- n Allow other events to occur within the distance covered by a gene conversion tract.
- n Must keep more than one, last tree.

Conclusions

- n When you can use ms, you should do so.
- n For long regions, SMC provides a very close approximation to an exact answer that is otherwise unobtainable
- n For gene conversion, low migration rates,?

Acknowledgments

n Jeff Wall

n Christina Jung

Refs:

- Recombination as a point process along sequences, Wiuf and Hein, *Theor. Pop. Biol.* 55:28-259, 1999.
- Approximating the coalescent with recombination, McVean and Cardin, *Phil. Trans. R. Soc. B* 360:1387–1393, (2005).
- Fast “Coalescent” Simulation. P. Marjoram and J. Wall. *BMC Genetics*, 7:16, 2006.
- Algorithm available via email: pmarjora@usc.edu or at <http://chp220mac.hsc.usc.edu/Marjoram/Software.html>

The End

Coalescent (Kingman)

- n Stochastic process
- n Parameters: θ (mutⁿ), ρ (recombⁿ)
- n Given there are K lines:
 1. Coalescence w.p. $(K-1)/(\theta+\rho+K-1)$
 2. Mutation w.p. $\theta /(\theta+\rho+K-1)$
 3. Recombination w.p. $\rho /(\theta+\rho+K-1)$
- n Times between events
~ $\text{Exp}(K(\theta+\rho+K-1)/2)$