

# Linkage disequilibrium in approximate models of recombination

Gil McVean  
Department of Statistics, Oxford

$$LD = \frac{1}{1 + 4N_e r}$$

$$LD \neq \frac{1}{1 + 4N_e r}$$

# Linkage disequilibrium in Wright-Fisher models

- Associations between alleles at linked loci (LD) are determined by a balance between genetic drift and recombination
- The compound parameter  $R = 4N_e r$  determines the rate of recombination in simple diploid populations
- For biallelic systems LD can be summarised by 2-locus statistics such as the squared correlation coefficient,  $r^2$

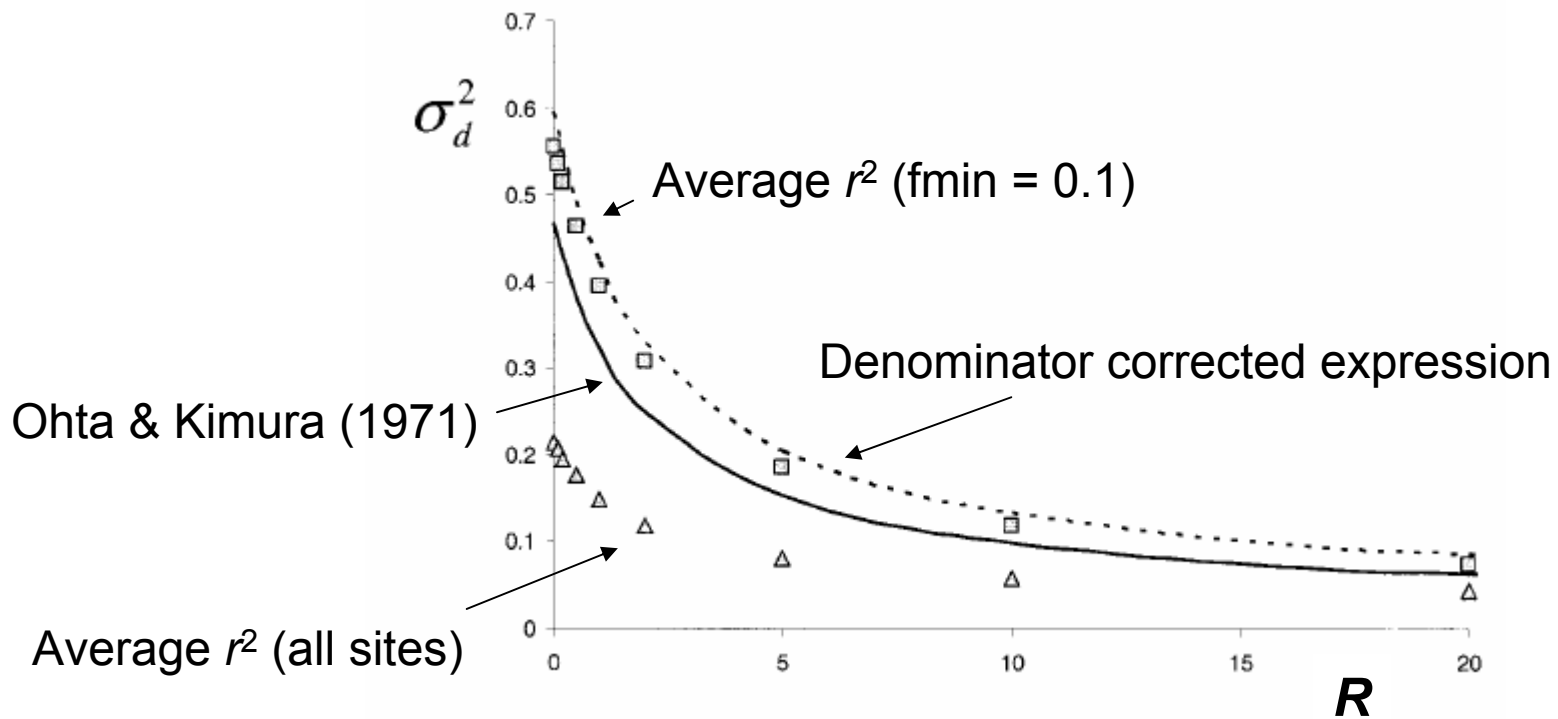
$$r^2 = \frac{D^2}{p_1 q_1 p_2 q_2}$$

- What is the relationship between 2-locus statistics and  $R$  ?

## The Ohta and Kimura result (1971)

$$\sigma_d^2 = \frac{E[D^2]}{E[p_1 q_1 p_2 q_2]} = \frac{10 + R}{22 + 13R + R^2}$$

# How good is the approximation?

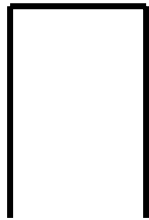


## The Sved result (1971)

$$E[r^2] = \frac{1}{1 + R}$$

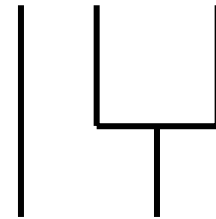
# How?

- Equate 2-locus IBD with 2-locus identity-in-state



1<sup>st</sup> event = coalescence

$$\Pr = \frac{1}{1 + R}$$



1<sup>st</sup> event = recombination

$$\Pr = \frac{R}{1 + R}$$

**Only true under very restricted conditions**



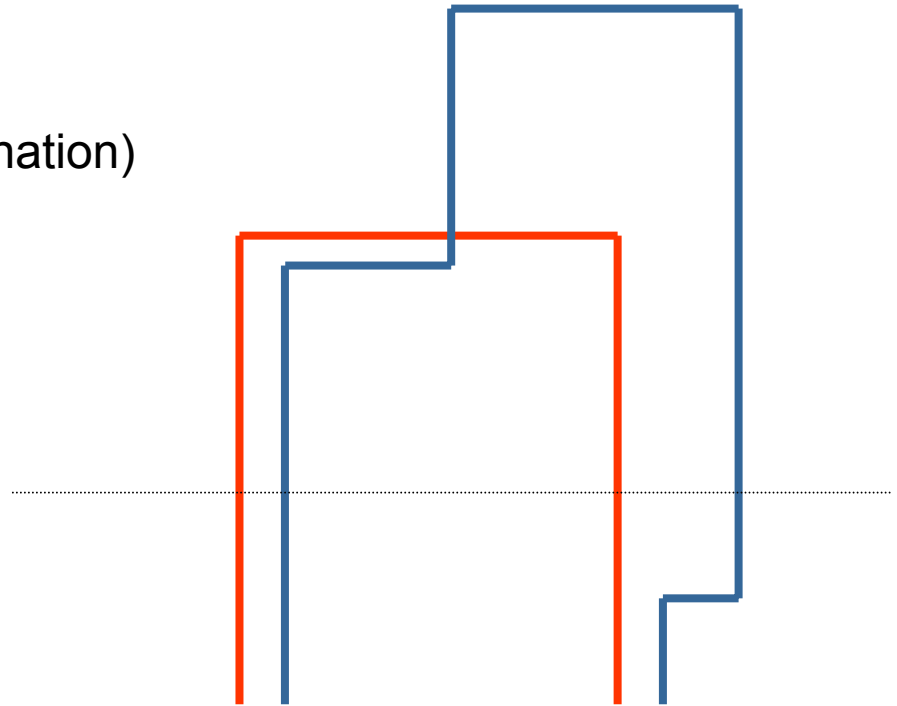
# Is the result ever valid? – a revised Sved model

## Coalescent phase

(Coalescence, mutation,  $\pm$ recombination)

## Recombination phase

(No mutation)



$$\Pr(\text{No rec}) = \frac{1}{1+R}$$

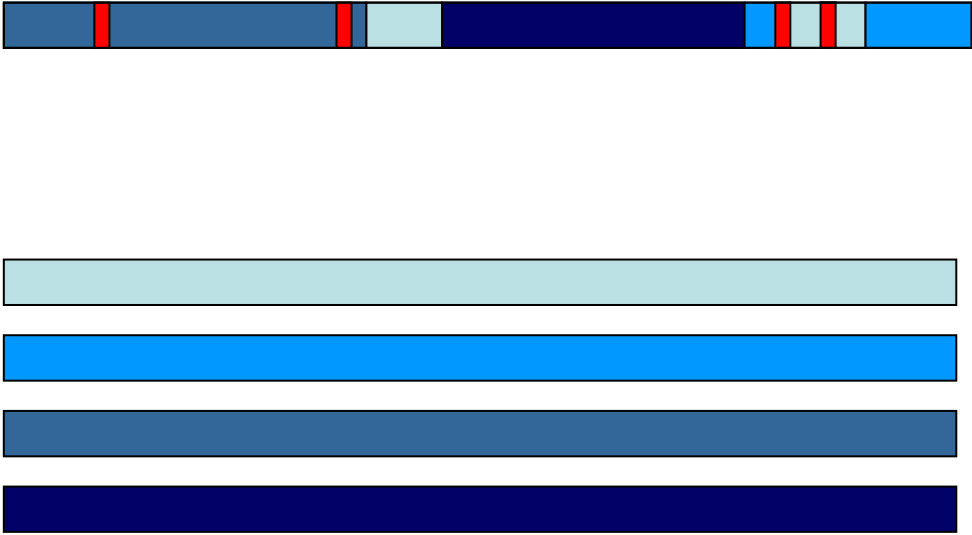
$$\Pr(1 \text{ rec}) = \frac{1}{1+R/2} - \frac{1}{1+R}$$

$$\Pr(2 \text{ rec}) = 1 - \frac{2}{1+R/2} - \frac{1}{1+R}$$

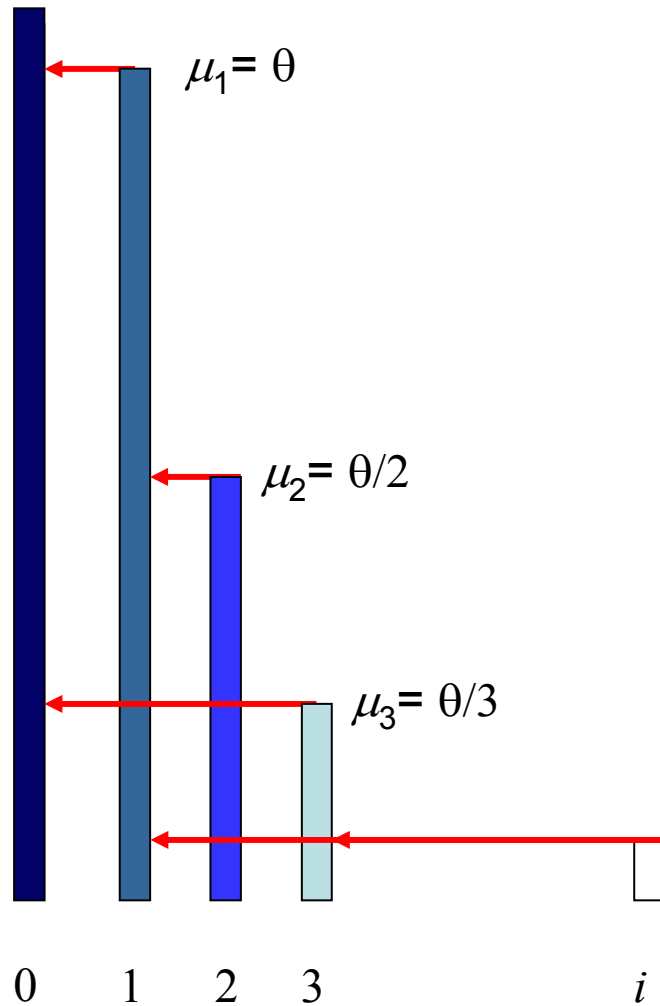
## LD in the revised Sved model

$$E[r^2] = \frac{1}{1+R} E[r_0^2]$$

# The Li and Stephens model



# A genealogical interpretation



$$M_{ij} = \sum_k I_k^{i,j} \mu_k$$

$$E[M_{ij}] = \theta$$

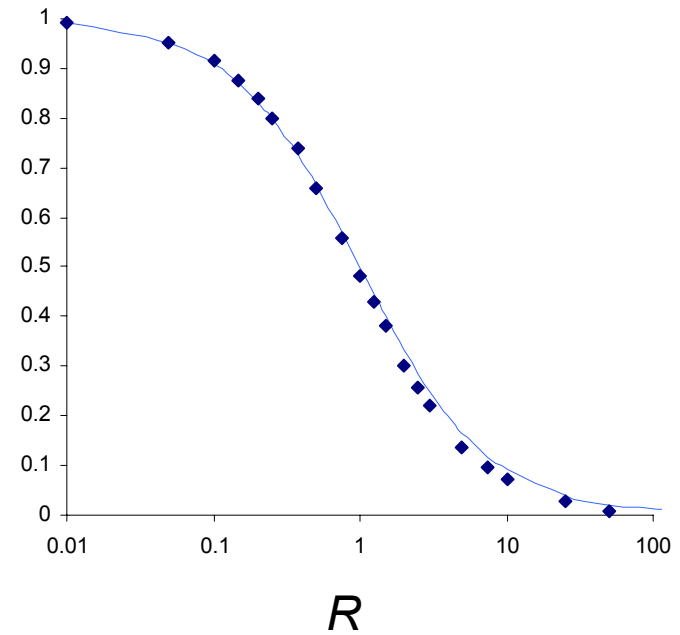
$$\mu_i = \theta/i$$

$$\Pr(\text{Rec}) = e^{-2R/i}$$

# Covariance of mutation weights

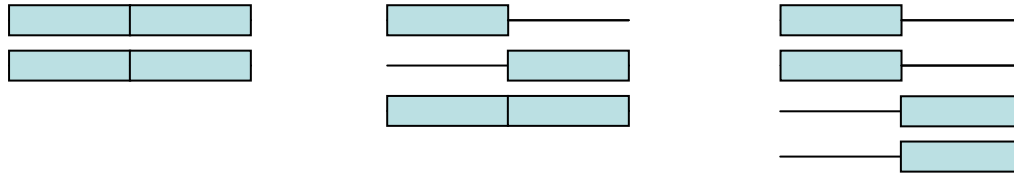
$$\text{Corr}(M_{ij(x)}, M_{ij(y)}) \approx \frac{1}{1+R}$$

$$\text{Var}(M_{ij}) \approx \theta^2 / 2$$



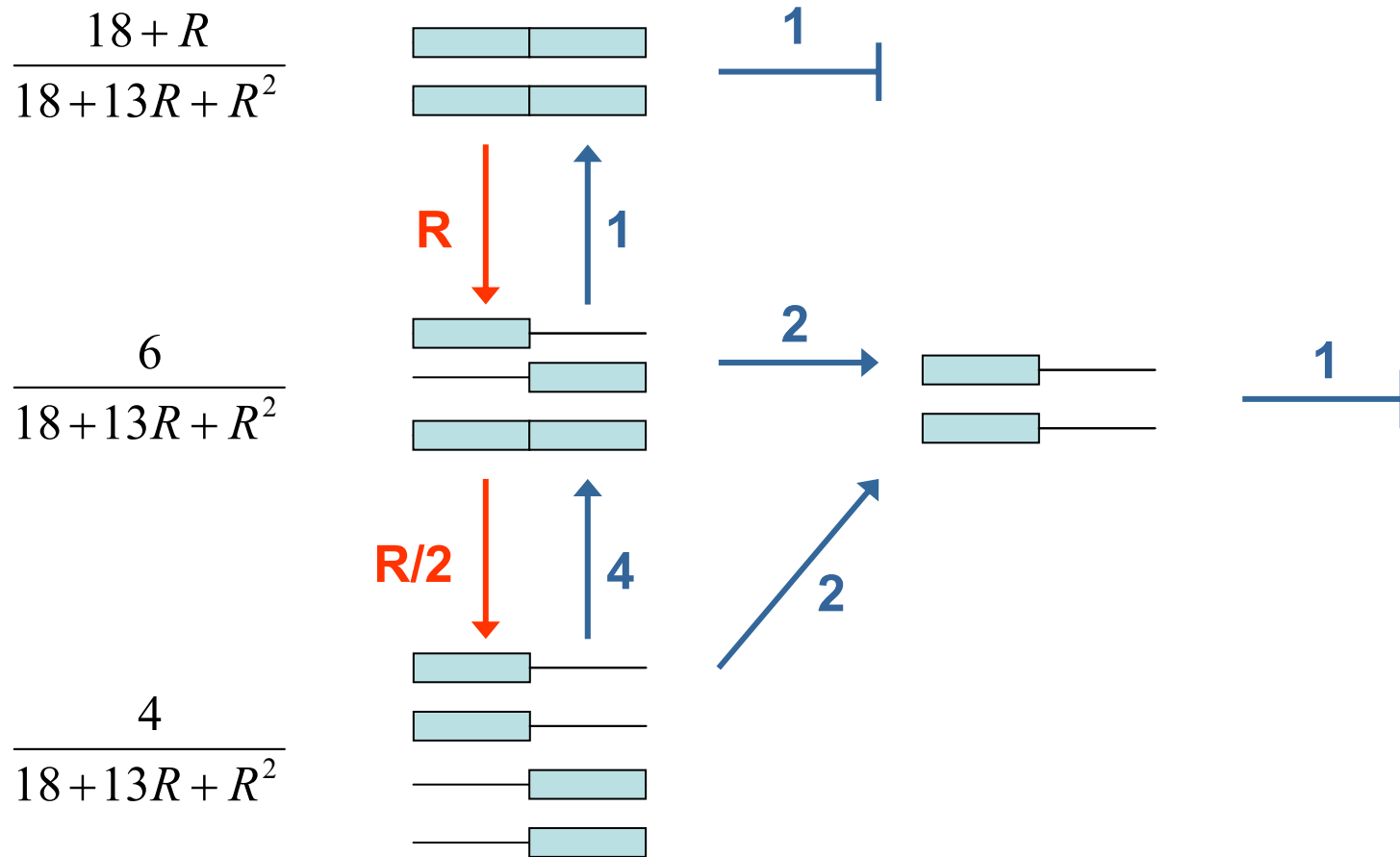
## Where does the Ohta and Kimura result come from?

$$D^2 = F_{ij(x)ij(y)} - 2F_{ij(x)ik(y)} + F_{ij(x)kl(y)}$$

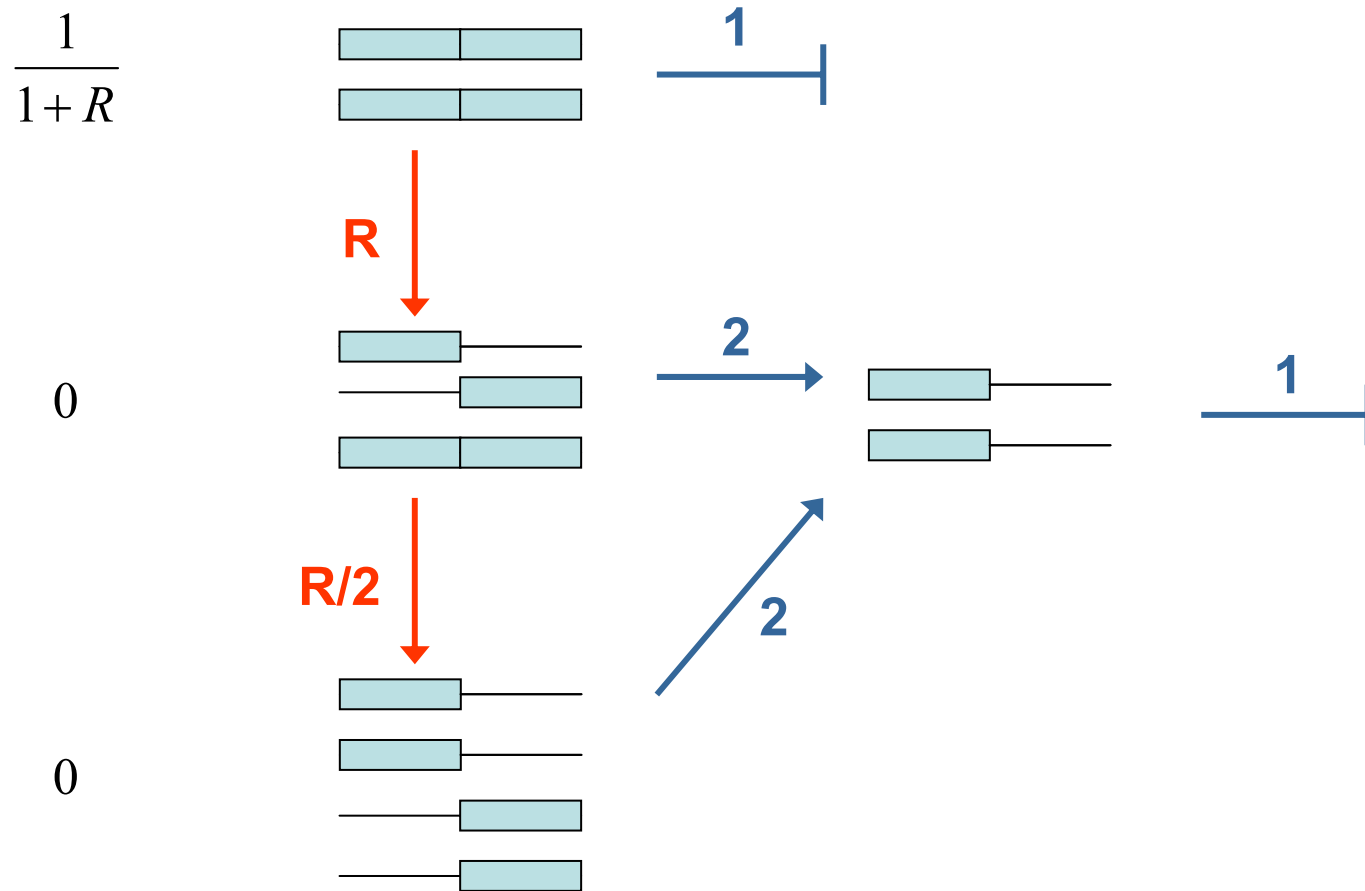


$$\sigma_d^2 = \frac{\text{Cov}(t_{ij(x)}, t_{ij(y)}) - 2\text{Cov}(t_{ij(x)}, t_{ik(y)}) + \text{Cov}(t_{ij(x)}, t_{kl(y)})}{E[t]^2 + \text{Cov}(t_{ij(x)}, t_{kl(y)})}$$

# The coalescent model structure



# A simplified coalescent model structure



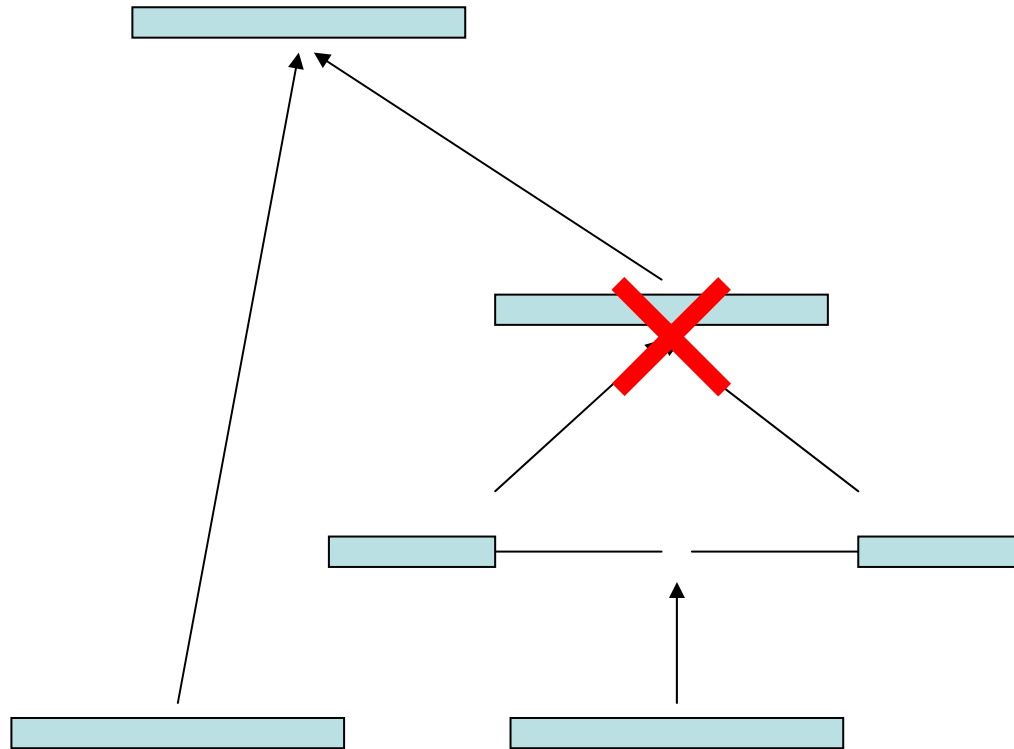


## LD under the simplified model

$$\sigma_d^2 = \frac{E[D^2]}{E[p_1 q_1 p_2 q_2]} = \frac{1}{1 + R}$$

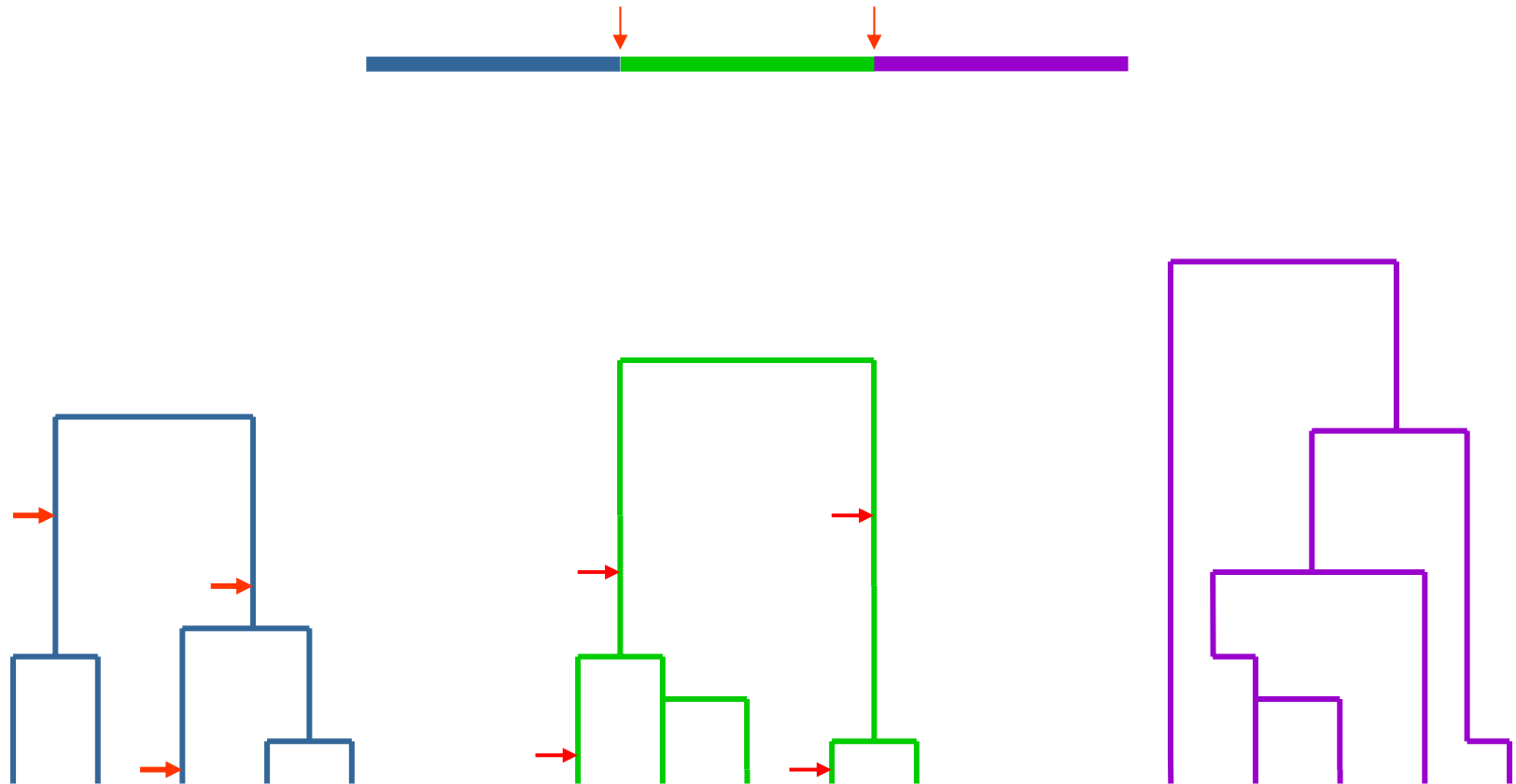
# What is the simplified model?

- Disallow coalescence between chromosomes that have no overlapping ancestral material



# Markov tree structure

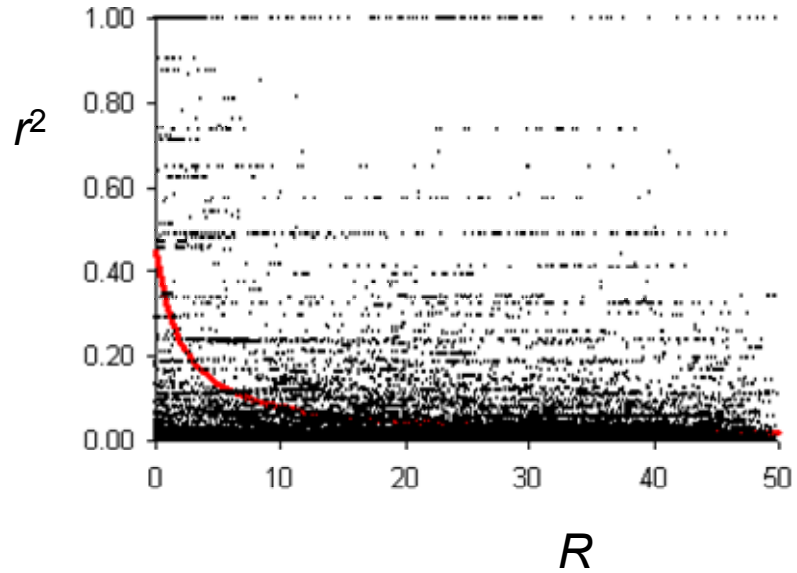
- This process generates a spatial coalescent process with a simple Markov structure on marginal genealogies



# Patterns of LD under alternative models

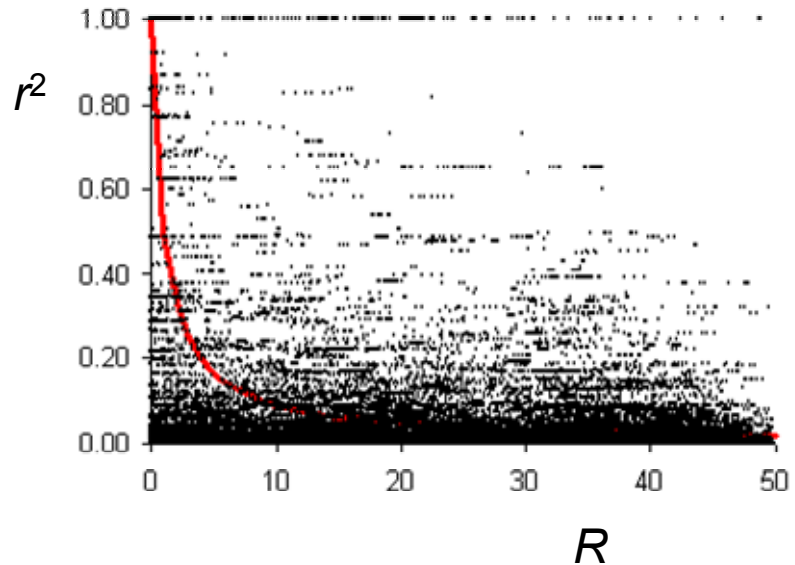
Full coalescent

$$\sigma_d^2 = \frac{10 + R}{22 + 13R + R^2}$$



Coalescent with banned events

$$\sigma_d^2 = \frac{1}{1 + R}$$



# Inference under the restricted model

# Conclusions

- $LD \neq 1/(1+R)$
- Results similar to Sved's can be derived using approximations to the coalescent that share the same conditional independence structure
- The coalescent with banned events is one such model that may provide a 'statistical' model for genetic variation under which the computational tractability of full-likelihood inference is greatly increased

**Many thanks to Niall Cardin and Chris Spencer**