Approximate Genealogies under Genetic Hitchhiking

Peter Pfaffelhuber

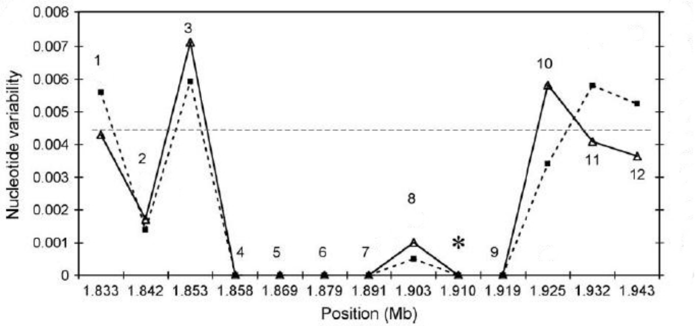
Oxford, November 2008

Selective Sweeps

- Goal: detect selection in a genome
- Use variation data to find candidate genes
- Maynard Smith, Haigh (1974):
  Variation around a strongly beneficial allele reduced
- This talk: approximations of genealogies
  - at (one or more) neutral loci linked to the selected site
  - with or without recurrent occurrence of the beneficial allele
- Useful for
  - analytical predictions
  - simulation of selective sweeps

Example

- The *wapl*-region in a European sample of *Drosophila melanogaster* (Beisswanger et al. 2006)



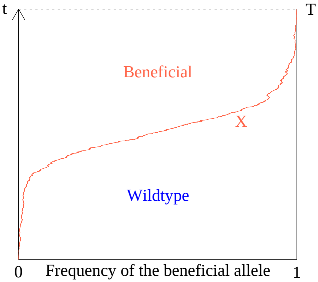
The Wright-Fisher diffusion

- Condition on fixation; frequency path of beneficial allele is

$$dX = \alpha X(1-X)\coth(\alpha X)dt + \sqrt{X(1-X)}dW, \qquad X_0 = 0$$

s	selective advantage
N	population size
α	$:= sN \gg 1$
dt	$\equiv Ndt$ generations

$$\mathbb{E}[T] \approx \frac{2\log\alpha}{\alpha}$$



The structured coalescent

- **Goal:** describe common ancestry of sample at time T

- **Idea:** Kaplan, Hudson, Langley (1989), using deterministic frequency path
- refined by Barton, Etheridge, Sturm (2004)

- Genealogy given by **structured coalescent**

The structured coalescent

- Discrete model: given $X_t = x$, birth events of beneficial alleles:

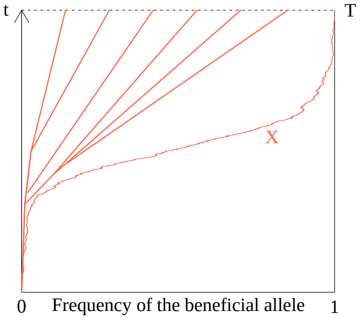
$$\text{rate } \frac{Nx}{2}$$

common ancestry of a given pair

$$\text{probability } \frac{1}{\binom{Nx}{2}}$$

$\Rightarrow$ scaled coalescence rate

$$\frac{1}{x}$$



The structured coalescent

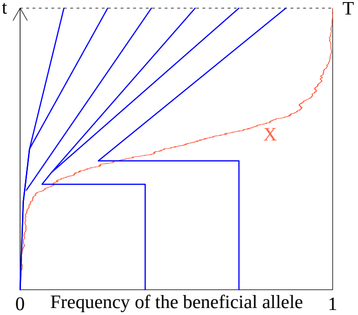
- Discrete model: given $X_t = x$, Frequency of recombinants of beneficial allele with wild-type is

  $$rx(1-x)$$

  Probability of picking a beneficial allele is x.

  $\Rightarrow$ scaled recombination rate

  $$\rho := rN, \quad \rho(1-x)$$



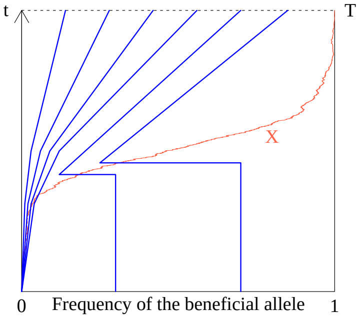
The star-like approximation

- MRCA: founder of the sweep
- Recombination events on each line with probability

$$p = 1 - \exp\left(\rho \cdot \mathbb{E}[T]/2\right)$$

- H_t: Expected heterozygosity:

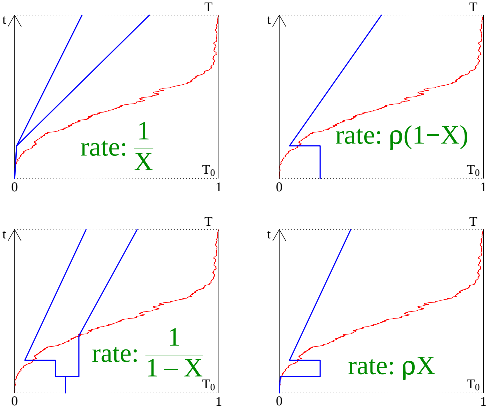
$$H_T = \left(1 - (1-p)^2\right) H_0$$



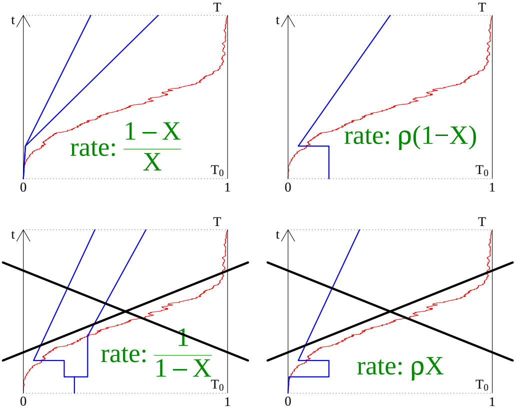
The structured coalescent

- **Goal:** finer description of the genealogy
- **Goal:** error bounds for the approximation

The structured coalescent



The structured coalescent



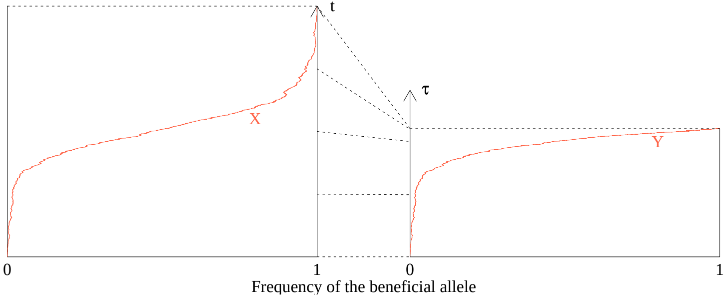
Probability
$\mathcal{O}\left(\frac{\rho^2}{\alpha^2}\right)$

Yule approximation

- Time transform $d\tau = (1 - X)dt$ gives

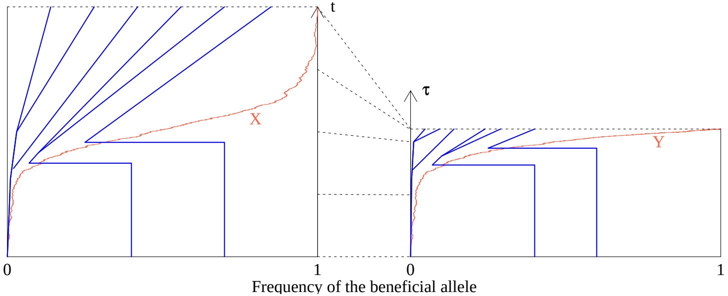
$$dY = \alpha Y \coth(\alpha Y) d\tau + \sqrt{Y} dW$$

- Y: supercritical branching process



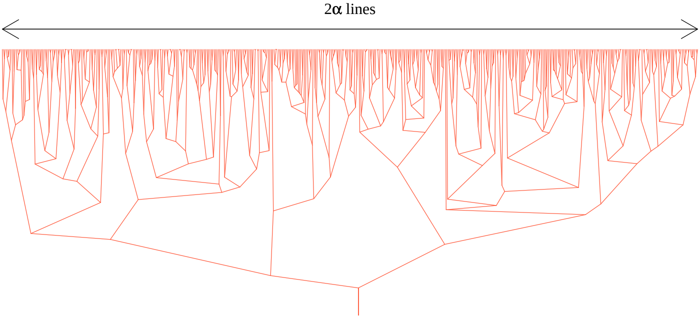
Yule approximation

- Coalescence rate $1/Y$
- Recombination rate ρ



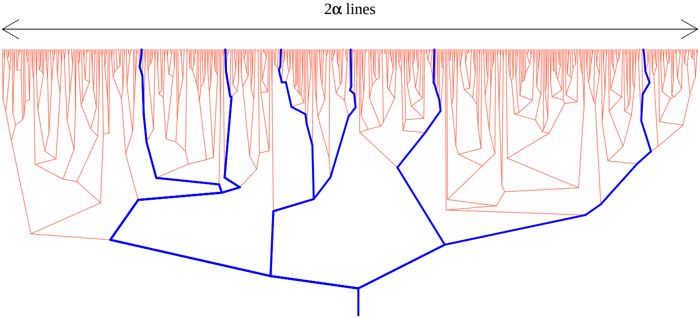
Yule approximation

- Durrett, Schweinsberg; Etheridge, Haubold, P, Wakolbinger
- Genealogy of Y: Yule process, stopped at 2α lines
- recombinations: rate ρ along Yule tree



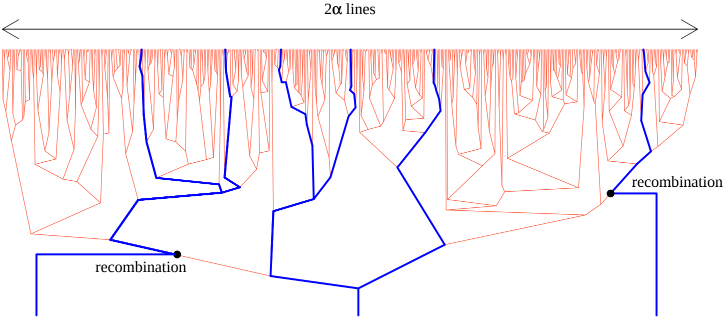
Yule approximation

- Durrett, Schweinsberg; Etheridge, Haubold, P, Wakolbinger
- Genealogy of Y: Yule process, stopped at 2α lines
- recombinations: rate ρ along Yule tree



Yule approximation

- Durrett, Schweinsberg; Etheridge, Haubold, P, Wakolbinger
- Genealogy of Y: Yule process, stopped at 2α lines
- recombinations: rate ρ along Yule tree



Accuracy

- **Star-like** approximation error: $\mathcal{O}(\rho/\alpha)$
- **Yule** approximation error: $\mathcal{O}(\rho^2/\alpha^2)$

- **Difference:** with probability $\mathcal{O}(\rho/\alpha)$, coalesced lines recombine
- Both approximations best for large α

- Durrett, Schweinsberg: Yule approximation holds for finite Moran models

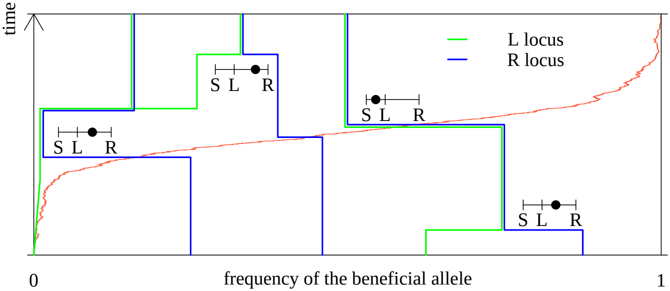
Association between neutral loci

- Simplest case: Three-locus model:

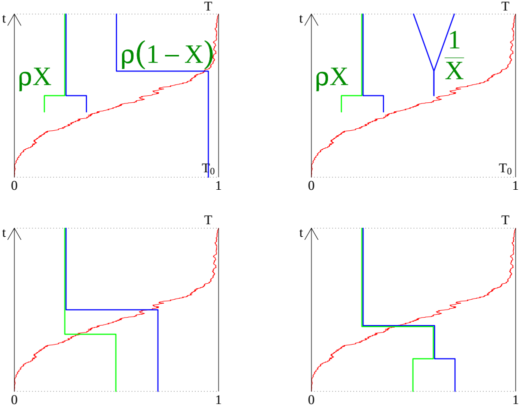
- What is the joint distribution of families for all loci?
- Stephan, Song, Langley (2006) and McVean (2007) describe approximate DNA pattern during the sweep

Ancestral recombination graph

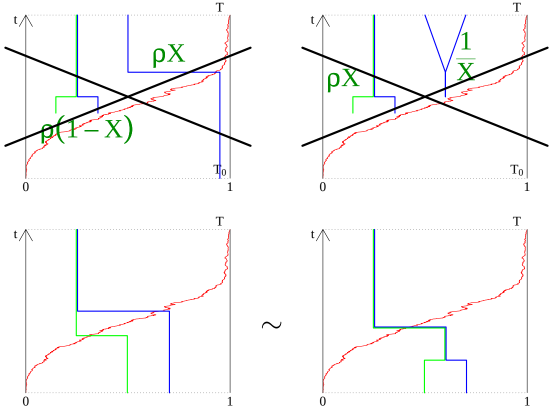
- Lines may split in beneficial and wild-type background



The structured ancestral recombination graph

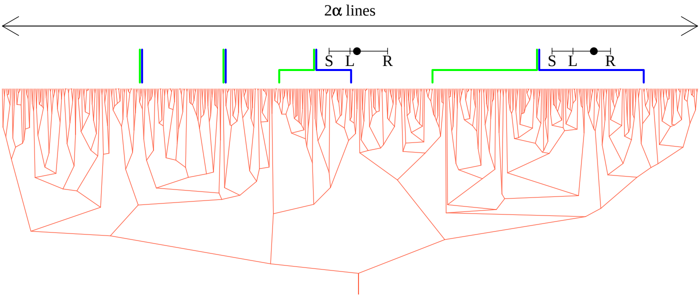


The structured ancestral recombination graph



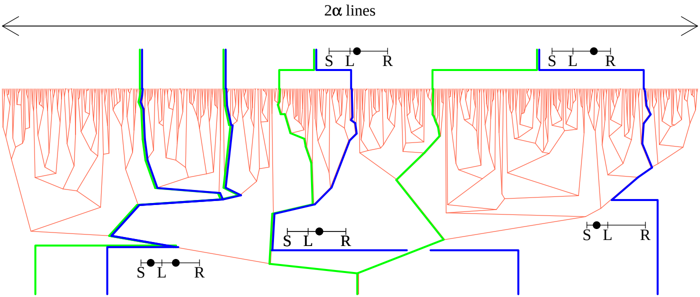
Yule approximation

- P, Studeny (2007)
- Splits in beneficial background generated first
- Other recombination events at constant rate



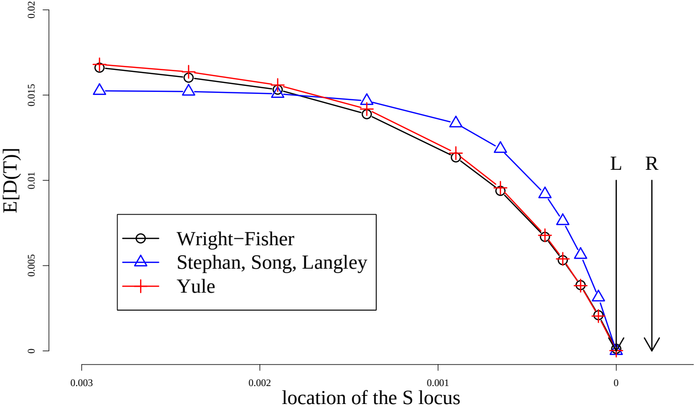
Yule approximation

- P, Studeny (2007)
- Splits in beneficial background generated first
- Other recombination events at constant rate



Application: Linkage Disequilibrium

For $\mathbb{E}[D(0)] = 0.0242, N = 10^5, S = 0.01$ we find a good fit to simulations.



Soft Sweep Patterns

- Classical sweep: neutral variation dragged to high frequency together with beneficial allele
- Soft sweeps: Multiple mutants introduce different patterns of neutral variation
- Consequence: Different haplotype blocks around the selected site

Lactose gene (from Tishkoff et al (2007))

- Not all adults can digest milk (→ lactase persistence LP)
- Probably connection to cattle domestication
- Europe: Swedes 90% LP, Spanish 50% LP;
  SNP C/T-13910 associated with LP
- Asia: Chinese 1% LP
- Africa: West-African agriculturalists 5-20% LP;
  G/C-14010 most significantly associated SNP with LP
- **⇒: Different origins of LP**

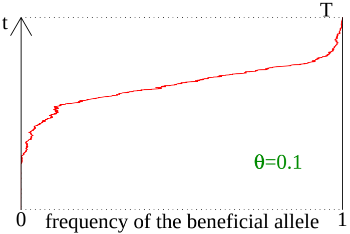
The Wright-Fisher diffusion

- Frequency path of beneficial allele is

$$dX = \big(\tfrac{\theta}{2}(1-X)+\alpha X(1-X)\big)dt+\sqrt{X(1-X)}dW, \qquad X_0 = 0$$

u mutation rate

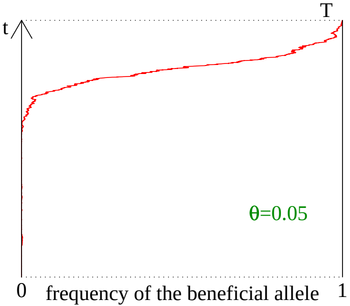
θ $:= 2uN$



The Wright-Fisher diffusion

- Frequency path of beneficial allele is

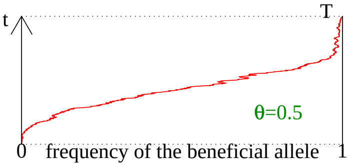
$$dX = \big(\tfrac{\theta}{2}(1-X)+\alpha X(1-X)\big)\,dt+\sqrt{X(1-X)}\,dW, \qquad X_0 = 0$$



The Wright-Fisher diffusion

- Frequency path of beneficial allele is

$$dX = \Big(\frac{\theta}{2}(1-X)+\alpha X(1-X)\Big)dt+\sqrt{X(1-X)}dW, \qquad X_0 = 0$$



Fixation times

- Let

$$T_0 := \sup\{t \geq 0 : X_t = 0\}, \qquad T^* := T - T_0.$$

- Fixation times for $\theta > 0$,

$$\mathbb{E}[T] \approx \frac{1}{\alpha\theta} + \frac{2\log\alpha}{\alpha}$$

$$\mathbb{E}[T^*] \approx \frac{2\log\alpha}{\alpha}$$

The structured coalescent

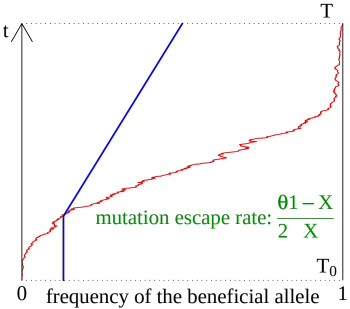
- Discrete model: given $X_t = x$, probability of following a mutant is

$$u(1-x).$$

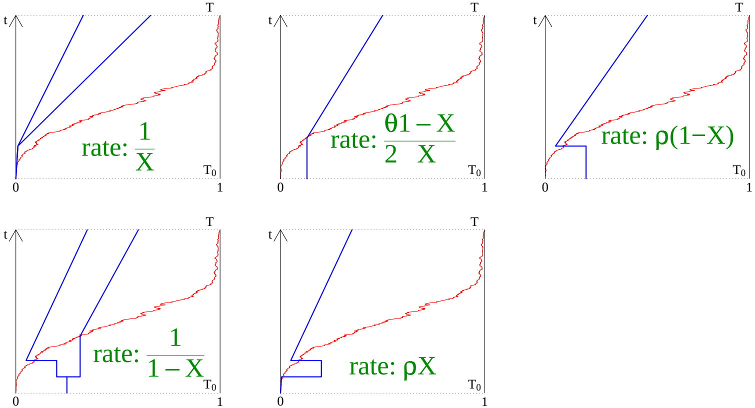
Probability of picking a beneficial allele is x.

⇒ unscaled mutation rate

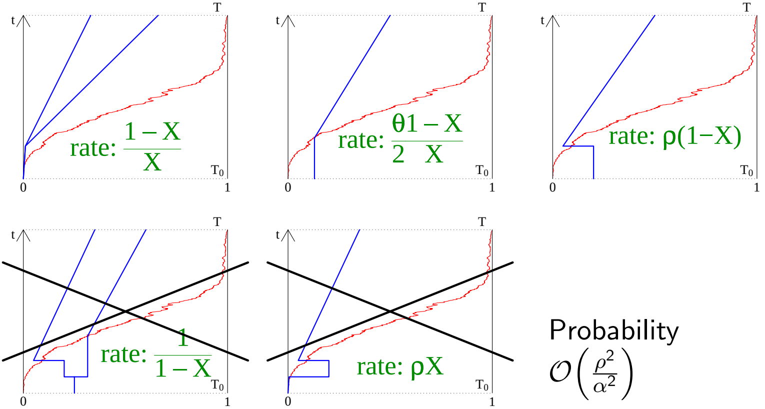
$$\frac{u(1-x)}{x}$$



The structured coalescent



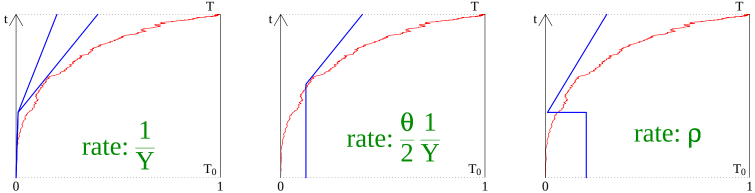
The structured coalescent



Probability
$\mathcal{O}\left(\frac{\rho^2}{\alpha^2}\right)$

The structured coalescent

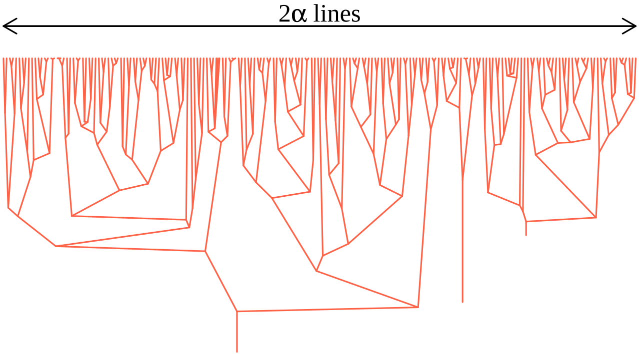
- After time scaling $d\tau = (1 - X)dt$



- $\Longrightarrow$ marked (rate ρ) genealogy of supercritical branching process with immigration (rate θ)

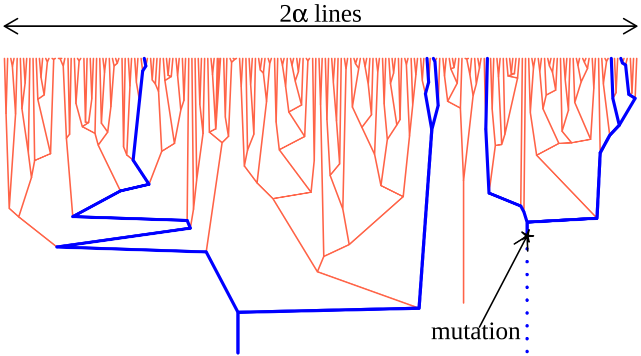
Yule approximation

- Hermisson, P (2008)
- mutation: immigration rate θ
- recombinations: rate ρ



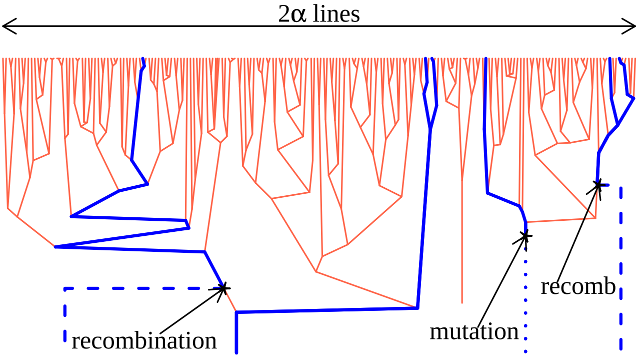
Yule approximation

- Hermisson, P (2008)
- mutation: immigration rate θ
- recombinations: rate ρ

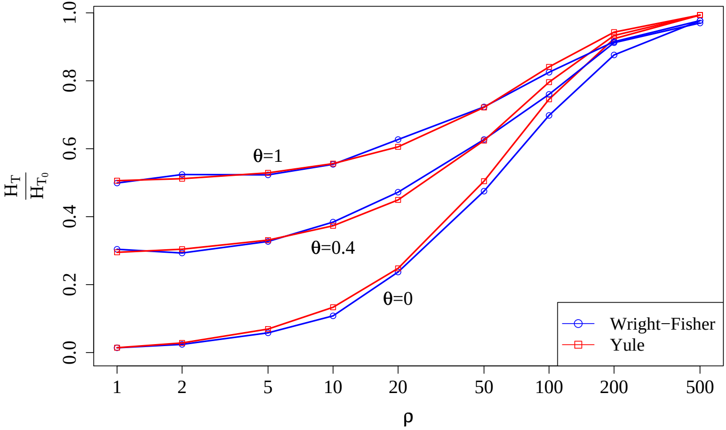


Yule approximation

- Hermisson, P (2008)
- mutation: immigration rate θ
- recombinations: rate ρ



Application: expected heterozygosity



Conclusion

- Yule approximation refines star-like approximation
- Yule approximation analytically tractable
- Soft selective sweeps show a pattern different from classical sweeps
- Future application of Yule approximation: sweeps in structured populations