

Combinatorics & the Coalescent (26.2.02)

Tree Counting & Tree Properties .

Basic Combinatorics .

Allele distribution.

Polya Urns + Stirling Numbers.

Number of ancestral lineages after time t .

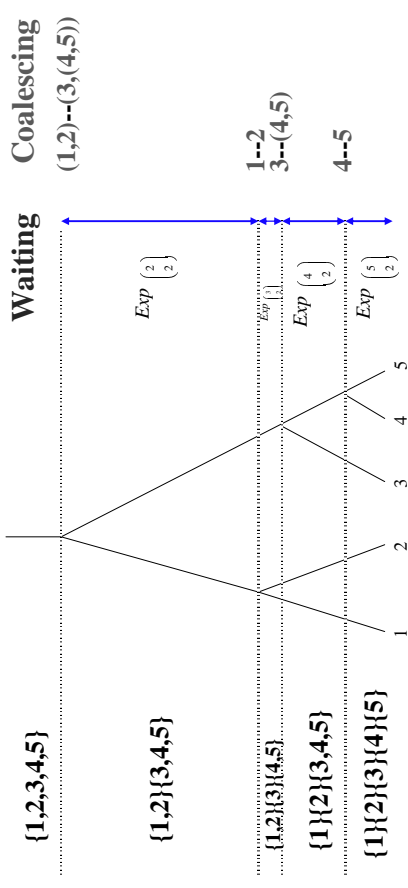
Inclusion-Exclusion Principle .

The Standard Coalescent

Two independent processes

Continuous: Exponential Waiting Times

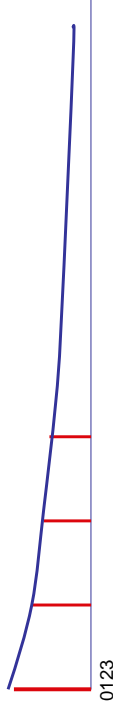
Discrete: Choosing Pairs to Coalesce.



The Exponential Distribution.

The Exponential Distribution: R^+ Expo(a)

Density: $f(t) = ae^{-at}, P(X > t) = e^{-at}$



Properties: $X \sim \text{Exp}(a), Y \sim \text{Exp}(b)$ independent

i. $P(X > t_2 | X > t_1) = P(X > t_2 - t_1) | (t_2 > t_1)$

ii. $E(X) = 1/a$.

iii. $P(X < Y) = a/(a+b)$.

iv. $\min(X, Y) \sim \text{Exp}(a+b)$.

v. Sums of k iid X_i is $\Gamma(k, a)$ distributed

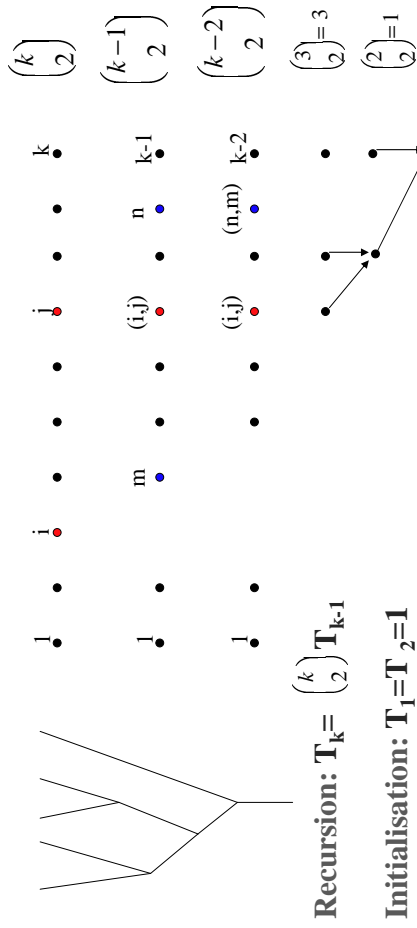
$$\frac{a^k x^{k-1} e^{-ax}}{\Gamma(k)}$$

As set of realisations

(from Felsenstein)



Trees: Rooted, bifurcating & node time -ranked.



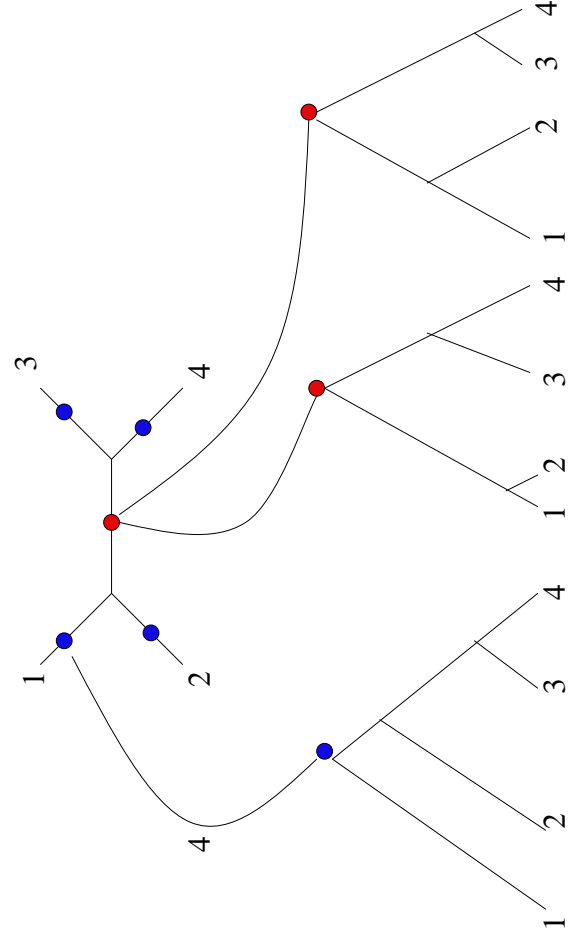
Recursion: $T_k = \binom{k}{2} T_{k-1}$
 Initialisation: $T_1 = T_2 = 1$

$$\prod_{j=2}^k \binom{j}{2} = \prod_{j=2}^k \frac{j!}{2^{j-1}} = \frac{j!(j-1)!}{2^{j-1}}$$

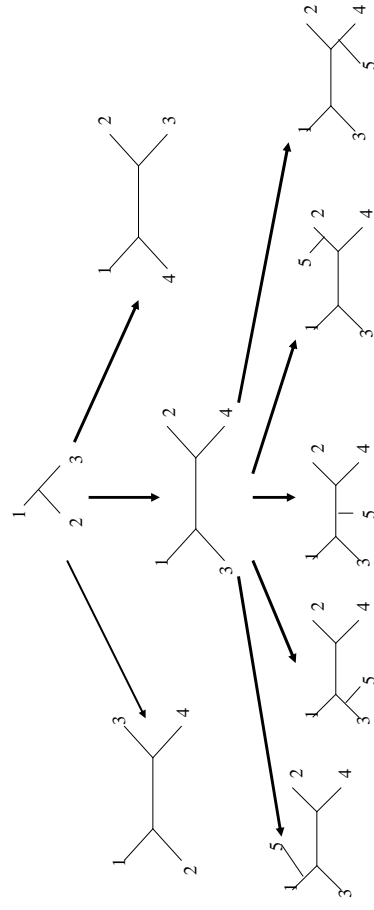
3	4	5	6	7	8	9	10	15	20
3	18	180	2700	5.710 ⁴	1.510 ⁶	5.710 ⁷	2.510 ⁹	6.910 ¹⁸	5.610 ²⁹

Coalescent versus unrooted tree topologies

4 leaves: 3 unrooted trees & 18 coalescent topologies.
 1 unrooted tree topology contains 6 coalescent topologies.



Trees: Unrooted & valency 3



Recursion: $T_n = (2n-5) T_{n-1}$ Initialisation: $T_1 = T_2 = T_3 = 1$

$\prod_{j=3}^n (2j-3) = \frac{(2n-5)!}{(n-2)! 2^{n-2}}$	4	5	6	7	8	9	10	15	20
	3	15	105	945	10345	1.410 ⁵	2.010 ⁶	7.910 ¹²	2.210 ²⁰

Inner & outer branches

Fu & Li (1993)

External (ϵ) versus Internal (i) Branches.

$E(\epsilon) = 2$ $E(i) = \left(\sum_{i=1}^{n-1} \frac{1}{i} \right) - 2$



Let $l_{i,n}$ be length of i 'th external branch in a n -tree. Obviously

$E(\epsilon) = nE(l_{n,i})$

$l_{n-1} + t_n$ $Pr = 1 - 2/n$

$l_n = t_n$ $Pr = 2/n$

Red - external. Others internal.

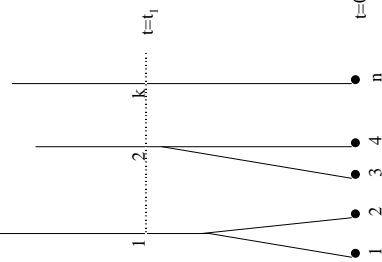
Except for green branch, internal-external corresponds to single/non-single segregating sites if only one mutation can happen per position.

- ACTTGATCGA
- ACTTGTACGA
- ACTTGTACGA
- TCTTATACGA
- ACTTATACGA
- sn

Probability of hanging Sub-trees.

Kingman (1982b)

For a coalescent with n leaves at time 0, with k ancestors at time t_1 , let ξ be the group of leaves of the k subtrees hanging from time t_1 . Let $\lambda_1, \lambda_2, \dots, \lambda_k$ be the number of leaves of these sub-trees.



$$P\{R_k = \xi\} = \frac{(n-k)!k!(k-1)!}{n!(n-1)!} \lambda_1! \lambda_2! \dots \lambda_k!$$

Example: $n=8, k=3$. Classes observed: 4, 3, 1

$$\frac{5!3!2!}{8!7!} 4!3!1! = 0.0012$$

The basal division splits the leaves into $(k, n-k)$ sets with probability: $1/(n-1)$.

Nested subsamples

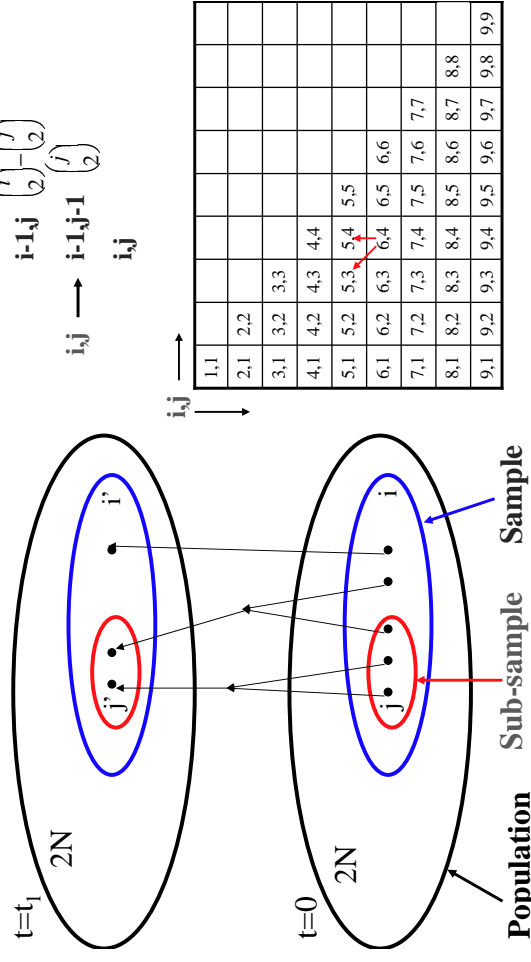
(Summers et al. (1986) Adv. Appl. Prob. 16:471-91.)

Transitions

$$i_j \rightarrow i-1, j-1 \binom{i}{2} \binom{j}{2}$$

$$i_j \rightarrow i-1, j \binom{i-1}{2} \binom{j}{2}$$

$$i_j \rightarrow i, j-1 \binom{i}{2} \binom{j-1}{2}$$



Nested subsamples

(Summers et al. (1986) Adv. Appl. Prob. 16:471-91.)

$$\frac{(i+1)(j-1)}{(i-1)(j+1)}$$

$\Pr\{\text{MRCA}(\text{sub-sample}) = \text{MRCA}(\text{sample})\} =$

$$\frac{(j-1)}{(j+1)}$$

$\Pr\{\text{MRCA}(\text{sub-sample}) = \text{MRCA}(\text{population})\} =$

Age of a Mutation

Wu & Donnelly (1999) Wu & Donnelly (2000), Mathews (2000)

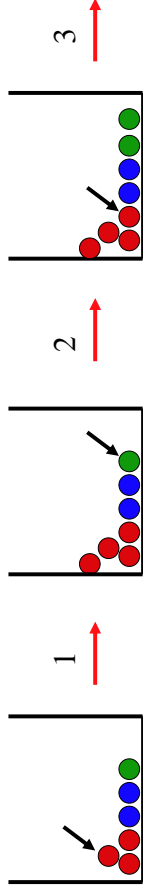
The probability that there are k differences between two sequences. Going back in time 2 kinds of events can occur (mutations Θ) - or a coalescent (1). This gives a geometric distribution:

$$\frac{1}{1+\theta} \left(\frac{\theta}{1+\theta}\right)^k$$



Classical Polya Urns

Feller I.



Let X_0 be the initial configuration of the initial Urn.

Astep: take a random ball from the urn and put it back together with an extra of the same colour.

X_k be the content after the k 'th step. Let Y_k be the colour of the k 'th picked ball.

$$i. P\{Y_k = j\} = P\{Y_1 = j\}.$$

ii. Sequences $Y_1 \dots Y_k$ resulting in the same X_k - has the same probability.

Labelling, Polya Urns & Age of Alleles

(Donnelly, 1986+Hoppe, 1984+87)

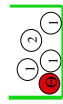
- ● ● As they come
- ● ● By size
- ● ● By age



A ball is picked proportionally to its weight. Ordinary balls have weight 1.

If the initial Θ -size ball is picked, it is replaced together with a completely new type.

If an ordinary ball is picked, it is replaced together with a copy of itself.

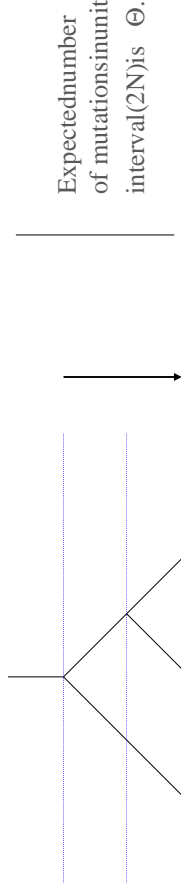


There is a simple relationship between the distribution of "the alleles labeled with age ranking" is the same as "the alleles labeled with size ranking".

Polya Urns & Infinite Allele Model

(Donnelly, 1986+Hoppe, 1984+87)

The only observation made in the infinite allele model is identity/non-identity among all pairs of alleles. i.e. The central observation is a series of classes and their sizes.

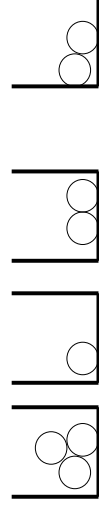
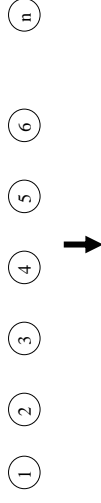


This model will give rise to distributions on partitions of $\{1, 2, \dots, n\}$ like $\{1, 4, 7\}, \{2, 3\}, \{5\}, \{6\}$. Since the labelling is arbitrary, only the information about the size of these groups is essential for instance represented as $1^2 2^1 3^1$.

What is the next event - a duplication of an existing type or a introduction of a "new" allele.

Stirling Numbers

Partitioning into k sets - Stirling Numbers (of second kind) - $S_{n,k}$



1 2 3 k
k unlabelled bins - all non-empty.

k	1	2	3	4	5	6	7
n	1	1	1	1	1	1	1
1	1						
2	1	1					
3	1	3	1				
4	1	7	6	1			
5	1	15	25	10	1		
6	1	31	90	65	15	1	
7	1	63	301	350	140	21	1

Bell Numbers - B_n - Partitioning into any number of sets.

Obviously: $B_n = \sum_{k=1}^n S_{n,k}$

Stirling Numbers

n-1 items - k classes: $(n-1, k-1): \{ \dots, \{ \dots, \{ \dots, \dots \} \}$

$\{ \dots, \{ \dots, \{ \dots \} \}$ $\xrightarrow{+''n''}$ $\{ \dots, \{ \dots, \{ \dots, \dots \} \}$ $\xrightarrow{+''n''}$

$(n, k): \{ \dots, \{ \dots, \{ \dots, \dots \} \}$

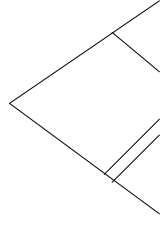
Basic Recursion: $S_{n,k} = kS_{n-1,k} + S_{n-1,k-1}$

Initialisation: $S_{n,1} = S_{n,n} = 1$.

Ewens' formula - example.

(1972TPB3.87 -112)

Assume ● ● ● ● has been observed and that 0.5 mutations is expected per unit (2N) time.



$$P_5(2,0,1,0,0) = \frac{n!}{\Theta(\Theta+1) \dots (\Theta+n-1)} \prod_{j=1}^n \binom{\Theta a_j}{j} = \frac{5!}{0.5 * 1.5 * 2.5 * 3.5 * 4.5} * \frac{0.5^3}{3}$$

$$E_5(\text{types}) = \sum_{j=1}^n \frac{\theta}{\theta + j - 1} = 0.5 \left(\frac{1}{0.5} + \frac{1}{1.5} + \frac{1}{2.5} + \frac{1}{3.5} + \frac{1}{4.5} \right)$$

$$P_5(2,0,1,0,0;3) = \frac{n!}{S_k^{\Theta} 1^{a_1} 2^{a_2} \dots k^{a_k}} = \frac{5!}{25 * 3 * 2!}$$

Ewens' formula.

(1972TPB3.87 -112)

$P_5(2,0,1,0,0)$ is the probability of seeing 2 single and one allele in 3 copies in a sample of 5.

Obviously, $a_1 + 2a_2 + \dots + ia_i + na_n = n$

$$P_n(a_1, a_2, \dots, a_n) = \frac{n!}{\Theta(\Theta+1) \dots (\Theta+n-1)} \prod_{j=1}^n \binom{\Theta a_j}{j}$$

$$E_n(\text{types}) = \sum_{j=1}^n \frac{\theta}{\theta + j - 1}$$

$$P_n(a_1, a_2, \dots, a_n; k) = \frac{n!}{S_k^{\Theta} 1^{a_1} 2^{a_2} \dots k^{a_k} a_1! a_2! \dots a_k!}$$

k is minimal sufficient statistic for Θ . The probability of the data conditioned on k is Θ -less and there is no simple sufficient statistic.

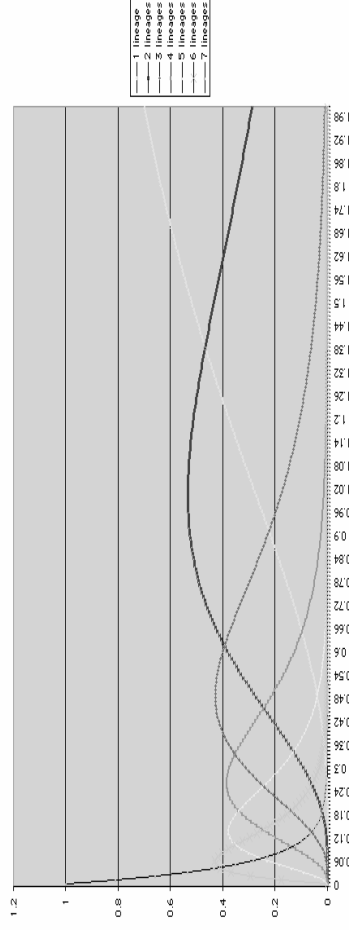
Ancestors to Ancestors

Griffiths (1980), Tavaré (1984)

$h_{i,j}$ = probability that i individuals have j ancestors after time t .

$$h_{i,j} = \sum_{k=j}^i e^{-\theta} \binom{i}{k} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} j_{(k)}}{j!(k-j)! i_{(k)}} \quad i_{[k]} = i(i-1) \dots (i-k+1) \quad i_{(k)} = i(i+1) \dots (i+k-1)$$

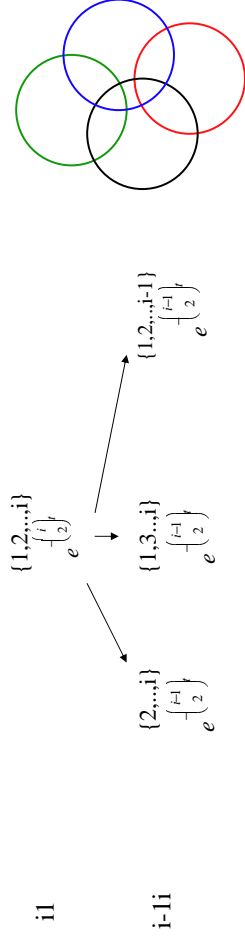
Example: Disappearance of 7 lineages.



Surviving Lineages

Which probability statements can be made? Let \mathbf{s} be a subset of $\{1, 2, \dots, i\}$ and \mathbf{s}' be the event that no coalescence has happened to \mathbf{s} . Additionally, if \mathbf{s}' is a subset of \mathbf{s} , then $S(\mathbf{s})$ implies $S(\mathbf{s}')$.

Size number

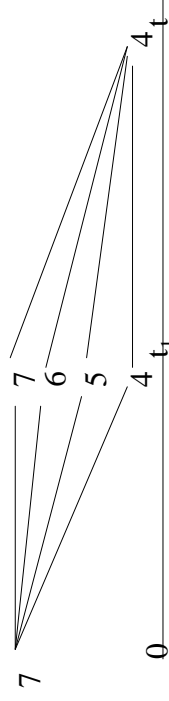


$$j \binom{i}{j}$$

$$2 \binom{i}{2} e^{-t}$$

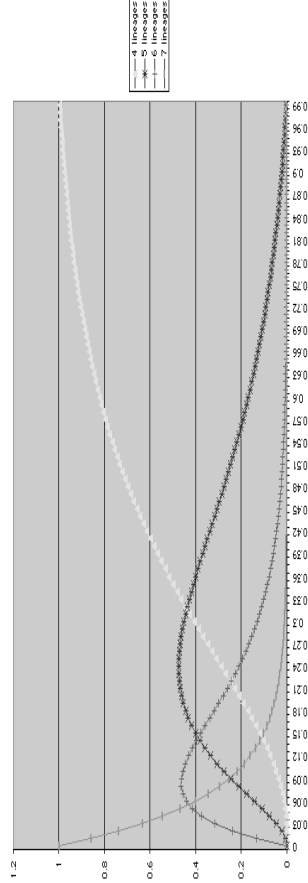
$$(i-1, i) e^{-t}$$

Conditional Ancestral Lineages



$$P_k(t_1) = h_{i,k}(t_1) * h_{k,j}(t-t_1) / h_{i,j}(t)$$

Example: 7 --> 4 lineages.



Surviving Lineages

There are $r = \binom{i}{j}$ sets. We want the number of members of these sets.

$$\sum_{k=1}^r (-1)^{k-j} \binom{k}{j} A_k \quad \text{where} \quad A_k = \sum_i S_i$$

Summation is over all k -subsets of $\{1, \dots, r\}$ and intersection is between the sets chosen.

Summary

Tree Counting & Tree Properties .

Basic Combinatorics .

Allele distribution.

Polya Urns + Stirling Numbers.

Number of ancestral lineages after time t .

Inclusion-Exclusion Principle .

Recommended Literature

- Bender (1974) Asymptotic Methods in Enumeration Siam Review vol 164-485.
- Donnelly (1986) Theor. Pop. Biol.
- Ewens (1989)
- Feller (1968-71) Probability Theory and its Applications I-II Wiley
- Fu & Li (1993) Statistical Tests of Neutrality of Mutations" Genetics 133:693-709.
- Griffiths (1980)
- Griffiths & Tavaré (1988) "The Age of amutation on a general coalescent tree."
- Griffiths & Tavaré (1999) "The ages of mutations in gene trees"
- Griffiths & Tavaré (2001) "The genealogy of a neutral mutation"
- Hoppe (1984) J. Math. Biol.
- Kingman (1982) "On the Genealogy of Large Populations" 27-43.
- Kingman (1982) "The Coalescent" Stochastic Processes and their Applications 13:235-248.
- Kingman (1982)
- Matthews, S. (1999) "Times on Trees and the Age of an Allele" Theor. Pop. Biol. 58:61-75.
- Möhle
- Pitman
- Schweinsberg
- Simonsen & Churchill (1997)
- Saunders et al. (1986) "On the genealogy of the sub-samples from a haploid population" Adv. Appl. Prob. 16:471-511.
- Tajima (1983) Evolutionary Relationships of DNA Sequences in Finite Populations Genetics 105:437-60.
- Tavaré (1984) Line-of-Descent and Genealogical Processes and Their Application in Population Genetics Models. Theor. Pop. Biol. 26:119-164.
- Thompson, R. (1998) "Ages of mutations on a coalescent tree" Math. Bios. 153:41-61.
- van Lint & Wilson (1991) A Course in Combinatorics - Cambridge
- Wiuf (2000) On the Genealogy of a Sample of Neutral Rare Alleles. Theor. Pop. Biol. 58:61-75.
- Wiuf & Donnelly (1999) Conditional Genealogies and the Age of a Mutant. Theor. Pop. Biol. 56:183-201.