UNIVERSITY OF HOHENHEIM

Coalescent theory

3502-470 Plant Genetic Resources

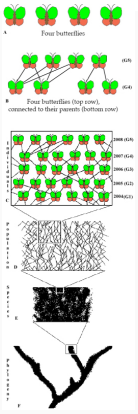
Prof. Karl Schmid

SS 2025

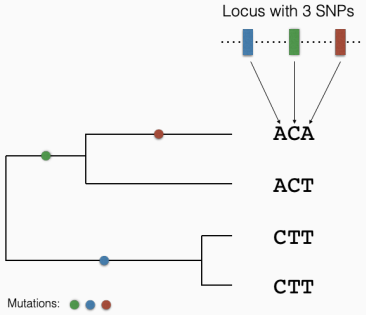
Institute of Plant Breeding, Seed Science and Population Genetics
University of Hohenheim

Key question:

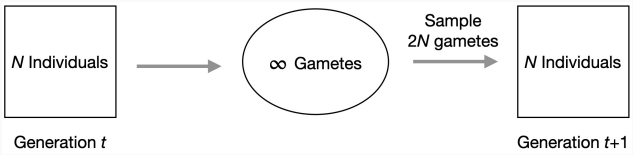
How are current patterns of genetic variation linked to evolutionary processes in the past?



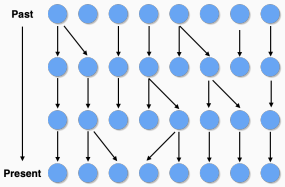
Modeling of SNP data with a genealogy



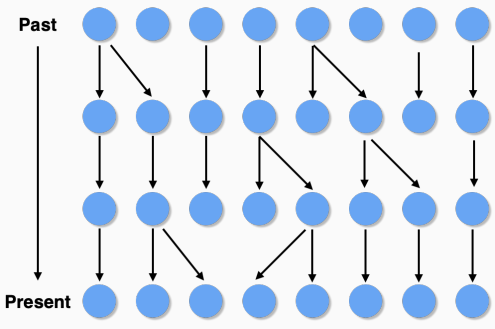
The Wright-Fisher model

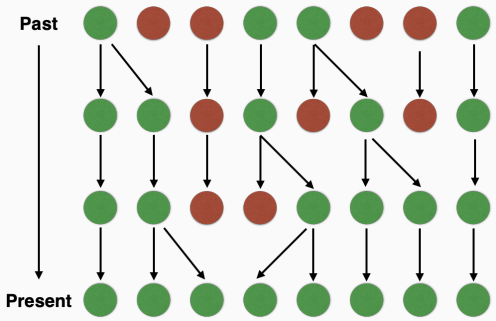


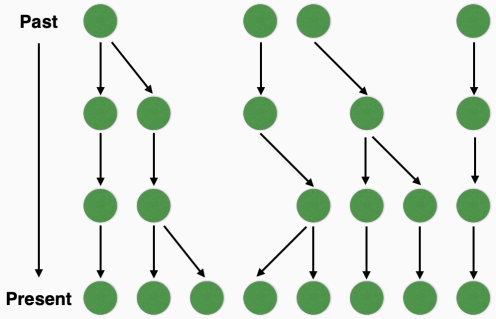
The Wright-Fisher model

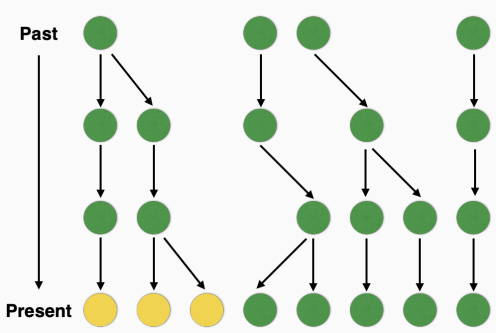


- **Population** model for N individuals with $2N$ alleles; N remains constant
- Alleles in the next generation are randomly sampled from the previous generation
- Probability that two alleles have the same parent allele: $\frac{1}{2N}$
- What is the history of a sample of n alleles backwards in time?
- Samples are small: $n \ll N$

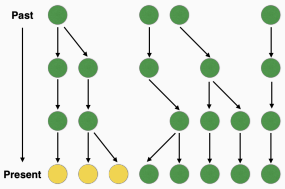



Past
Present


Past
Present


Past
Present

Discrete coalescent



- Sample size: $n = 3$, Population size: $N = 8$
- What is the genealogy/ancestral tree of the sample of alleles!
- From now on, we measure time backwards

Ancestral tree should resemble tree of Wright-Fisher model: random tree with random mutations

What is coalescence?

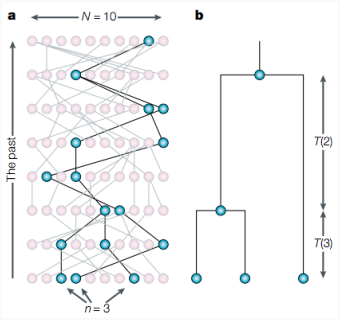
- Coalescence event: Two (or more) alleles have the same parent allele in the previous generation
- Probability of coalescent event for two specific alleles in the previous generation: $1/2N$
- Probability of coalescent event for three or more alleles: much smaller (ignore it)

Coalescence theory

Key idea: Combine genealogical process with process of mutation

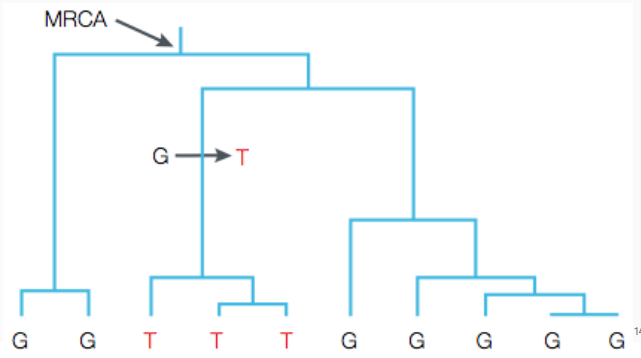
- All chromosomes in a sample have a evolutionary relationship that can be expressed as gene genealogy.
- Genetic variation originates by mutation on a branch on genealogy whose descendants inherit the mutation

Genealogy of a sample



Rosenbert et al. Nat Rev Genet. (2002)

Mutations in a sample



Rosenberg et al. Nat Rev Genet. (2002)

What is coalescence?

Probability of a coalescent event of two specific alleles:

$$\frac{1}{2N}$$

Sample of n alleles: How many different allele pairs are possible?

$$\frac{n(n-1)}{2}$$

Probability of a coalescence event of any pair of alleles:

$$p \approx \frac{1}{2N}\frac{n(n-1)}{2} = \frac{n(n-1)}{4N}$$

When was the first coalescent event?

Probability of no coalescence event in the first generation:

$$1 - p \approx 1 - \frac{n(n-1)}{4N}$$

Probability that the first coalescent event happened t generations ago:

$$P(T_n = t) \approx (1-p)^{t-1} p$$

→ geometric distribution.

Expected time T_n to first coalescence event (sample size n):

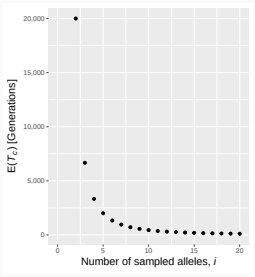
$$E(T_n) = \frac{1}{p} \approx \frac{1}{\left(\frac{n(n-1)}{4N}\right)} = \frac{4N}{n(n-1)}.$$

Expected time to coalescence with i alleles to $i-1$:

$$E(T_i) \approx \frac{4N}{i(i-1)}$$

Coalescence time of two randomly chosen alleles: $2N$ generations

Expected time to first coalescent event



Population size: $N = 10,000$

Waiting times between coalescence events

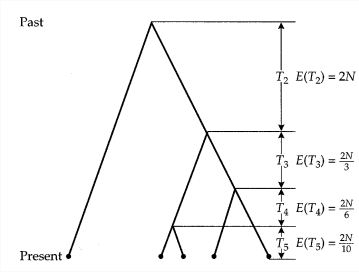
- After one coalescence event in the sample, there are $n-1$ individuals left (with high probability, ignore multiple coalescences)
- Time between first and second coalescence event: T_{n-1}

⋮

Expected (= average) time to coalescence events ($N = 10,000$, $n = 5$)

Coalescent event	Generations	per N
1^{st} coalescence event	2,000	2N/10
2^{nd} coalescence event	3,333	2N/6
3^{rd} coalescence event	6,666	2N/3
4^{th} coalescence event	20,000	2N

Coalescences in a gene genealogy



MRCA of a sample

- MRCA: Most recent common ancestor (of the sample)
- Total time to the coalescence of the whole sample:

$$E(T_{MRCA}) = \sum_{i=2}^{n} E(T_i) \approx 4N \sum_{i=2}^{n} \frac{1}{i(i-1)} = 4N \left(1 - \frac{1}{n}\right)$$

- Limit for looking back in a neutral genealogy: $4N$ Generations!

Total length of a genealogy

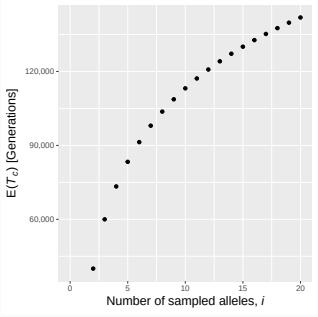
- Corresponds to total time (generations) of all branches:

$$T_{total} = \sum_{i=2}^{n} iT_i,$$

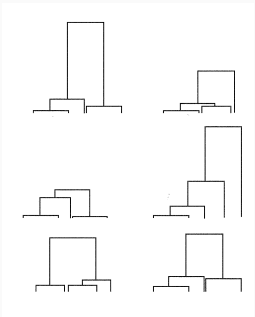
- Expected sum of random variables equals sum of expectations of variables:

$$E(T_{total}) = \sum_{i=2}^{n} iE(T_i) \approx \sum_{i=2}^{n} i\frac{4N}{i(i-1)} = 4N \sum_{i=2}^{n} \frac{1}{i-1}$$

Total number number of generations in a tree

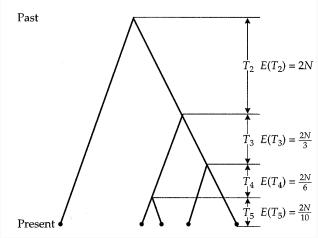


Random genealogies of 5 individuals



Modeling neutral mutations with the coalescent

Separate the genealogical from the mutational process (neutral mutation!)



Coalescence time	Generations (approx.)
T_5	$5 \times 2,000 = 10,000$
T_4	$4 \times 3,333 = 13,333$
T_3	$3 \times 6,666 = 20,000$
T_2	$2 \times 20,000 = 40,000$
Total	$83,333$

- Mutation rate μ per generation (locus-wide, per allele): Each allele inherits a new mutation w. probability μ
- Total number $E(S)$ of expected mutations if $\mu = 10^{-4}$:

$$E(S) = \mu E(T_{total}) \approx 10^{-4} \times 83,333 = 8.33$$

Problem: Where do we get μ and N?

- Discrete genealogy does not depend on the genealogical relations of unsampled alleles, BUT on the total population size N. To add mutations, we need to know μ
- N can be either the census or the effective population size
- N nearly always unknown!
- μ could be known, e.g. through mutation accumulation experiments, but also usually unknown

Idea 1: Express all coalescent times in units of $2N$
Idea 2: Estimate $\theta = 4N\mu$ from data!

Approximating discrete genealogies: (Kingman's) coalescent

- Rescale time: Branch length of $t2N$ generations correspond to branch length t in the rescaled genealogical tree
- J.F.C Kingman (1982): The discrete genealogy, for N large ($N > 1000$) is very well approximated by a bifurcating tree
  - Waiting time T_i for next coalescence of i ancestral lineages: $E(T_i) = \frac{2}{i(i-1)}$ (exponentially distributed)
  - Always 2 lineages coalesce
- Mutations: Put them randomly on the genealogy with scaled mutation rate $\frac{\theta}{2}$
- Probability for j mutations on a branch of length t is $e^{-\frac{\theta}{2}t}\frac{(t\frac{\theta}{2})^j}{j(j-1)\ldots 1}$

Estimating θ from data

- Remember: $\theta = 4N\mu$, the scaled mutation rate
- Infinite sites model: each mutation hits another site
  $\Rightarrow$ each mutation causes a polymorphism in the sample
- S_n: number of polymorphisms in the sample

Discrete WFM: $E(S_n) = \mu E(T_{total}) \approx 4N\mu \sum_{i=2}^{n} \frac{1}{i-1} = \theta \sum_{i=2}^{n} \frac{1}{i-1}$

Continuous coalescent: $E(S_n) = \frac{\theta}{2} \sum_{i=2}^{n} i \frac{2}{i(i-1)} = \theta \sum_{i=2}^{n} \frac{1}{i-1}$

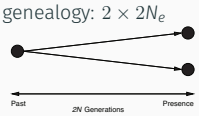
- Watterson's estimator: Estimate θ so that using this estimate causes, on average, S_n mutations

$$\hat{\theta}_W = \frac{S_n}{1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}}$$

Coalescent with two alleles

What is the difference (in polymorphisms) of two randomly chosen alleles?

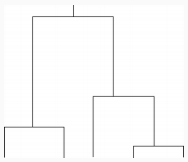
- Infinite sites model
- Two random alleles are expected to have a common ancestor $2N_e$ generations ago.
- Total expected time in genealogy: $2 \times 2N_e$



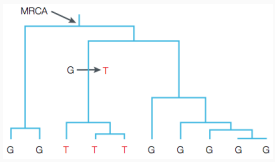
- Add mutation rate μ_s per site and length of gene m:

$$k = 4N_e\mu_s m$$

Coalescent with and without mutation



Coalescent with sample size $n = 5$

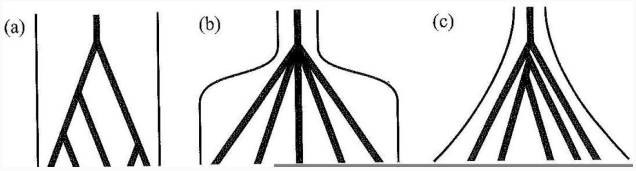


Rosenberg and Nordborg, Nat. Rev. Gen. 2002

Coalescent of 10 alleles/1 SNP with mutation

Changes in population size

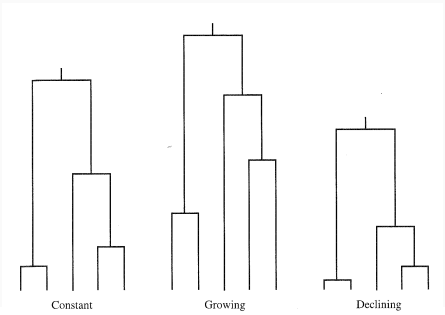
Random gene genealogies



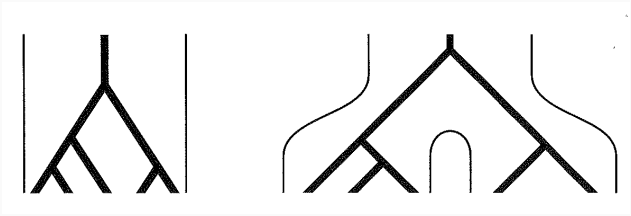
a Constant population size
b Stepwise growth
c Exponential growth

Changes in population size

Random gene genealogies



Genealogies in the presence of a population structure



Left No population structure
Right Two populations

Why is coalescent theory useful?

- Explicit model of the genealogical tree of closely related alleles
- Different demographic histories (e.g. exponential growth) → different tree likelihoods → different distributions of genetic diversity
- Statistical inference which demographic history shaped observed diversity: Decide which demographic model is most consistent with data and estimate model parameters.
- Selection tests: Scan genome for regions whose genetic diversity is significantly different from the best-fitting neutral model
- Model recombination by changing genealogical tree across genome

Outline of coalescent simulations

The genetic diversity produced by a coalescent model can be assessed via simulation

1. Define demographic models for a locus
2. Simulate a random gene genealogy under models
3. Distribute random mutations on the genealogy with θ estimated from data (in general: $\hat{\theta} = \frac{2S}{E(T_{total}^{(model)})}$)
4. Create sample of alleles
5. Calculate diversity statistic(s) of sample
6. Repeat steps 1-4 many times
7. Compare distribution under models with observed diversity of locus using appropriate statistical test (e.g. via Monte Carlo test, Approximate Bayesian Computation,...)

Applications in the context of plant genetic resources

- What is the genetic relationship of modern crop species?
- At which time and where in the genome did artificial selection occur?
- How are the heterotic pools defined and which exotic germplasm can we use to expand them?
- Allele mining: How many alleles of a gene are contained in a gene bank collection?
- When was a crop domesticated and how did it occur?
- In which geographic regions can be find potentially useful novel genetic variation in land races and wild ancestors?

Literature

Cited:

- Cavanagh, Colin R., et al. "Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars." Proceedings of the national academy of sciences 110.20 (2013): 8057-8062.

Some supporting literature

- Rosenberg and Nordborg (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms.

Review questions

1. Describe the conceptual basis of coalescent theory. Which two processes are investigated separately?
2. What is the probability of a coalescence event in a random sample of two alleles in the previous generation?
3. Is the time back to the most recent common ancestor of all alleles in a sample dependent on the size of the total population?
4. How can you estimate the mutation rate if you know the number of polymorphisms in a sample?
5. How do demographic processes such as an increase or a decrease of the population size affect the shape of genealogical trees?
6. What are potential applications of coalescence theory in the field of plant genetic resources?
7. How many polymorphisms would you expect between two randomly chosen alleles? Assume a population size of $20,000$ alleles.

References i

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3(5):380–390, DOI $10.1038/\mathrm{nrg795}$