Coalescent Theory and its applications to Population Genetics

Based on:

Recent progress in coalescent theory – Nathanael Berestycki

Coalescent Theory – Magnus Nordborg

The Coalescent – John Wakeley

Combinatorial Stochastic Processes – Jim Pitman

November 10, 2011

Presented by:

Daniel Klein and Layla Oesper

Super Fast Biology Primer

- DNA can be thought of as a string containing only A,C,G,T.
- The letter present at a particular location is the string is often referred to as an **allele.**

Genome:
ACCTGGTACGGCGCGTTA

C allele at position 3

Super Fast Biology Primer

- Humans are **diploid** – meaning they have two copies of every chromosome.
- A **haploid** organism (e.g., bacteria) has a single copy of a chromosome.

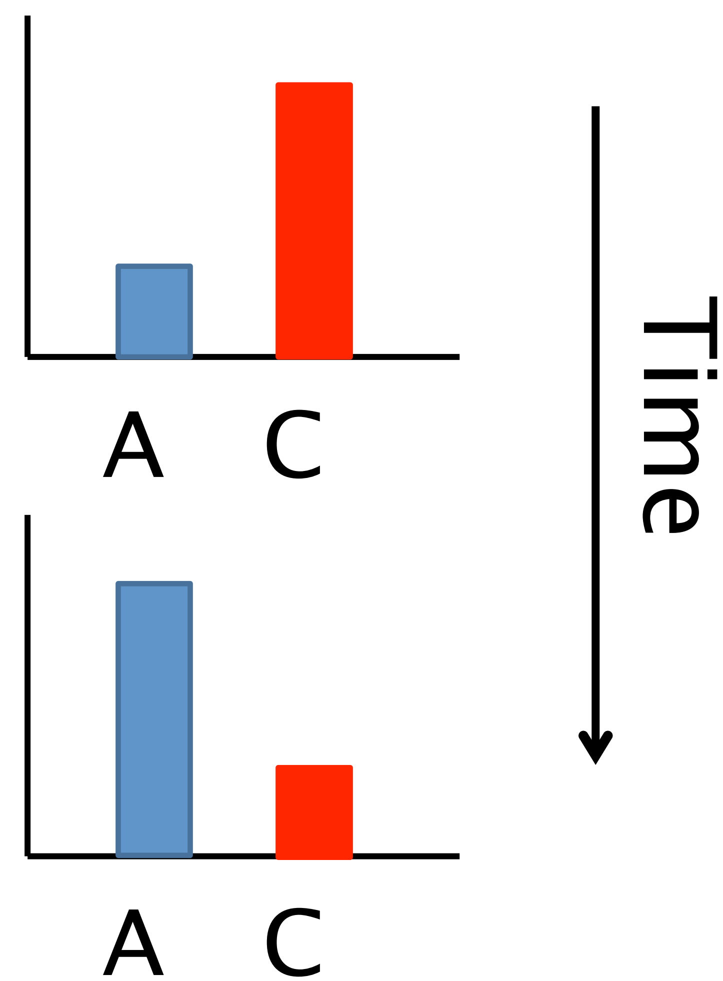
Diploid Genome:

ACCTGGTACGGCGCGTTA
ACGATGTAGGGCGCGTAA

CG genotype at position 3

Genetic Drift

A basic mechanism underlying evolution. Refers to the change in frequency of alleles in a population due to random sampling.



Various Theories

- Wright–Fisher Model
  - Generations do not overlap
- Cannings Model
  - Generations do not overlap
  - More control over number of offspring
- Moran Model
  - Assumes generations overlap

Various Theories

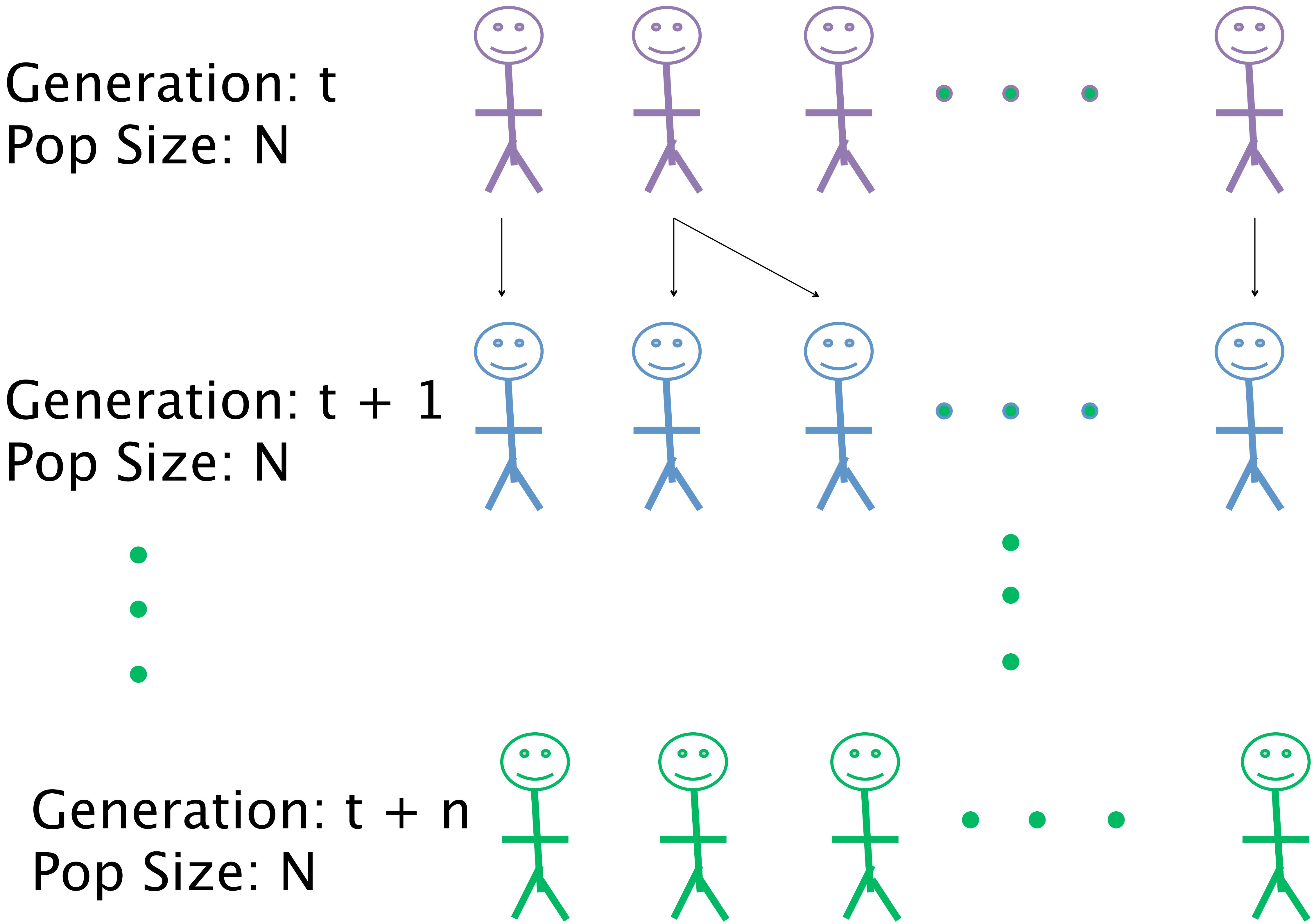
- Wright–Fisher Model
  - Generations do not overlap
- Cannings Model
  - Generations do not overlap
  - More control over number of offspring
- Moran Model
  - Assumes generations overlap

Various Theories

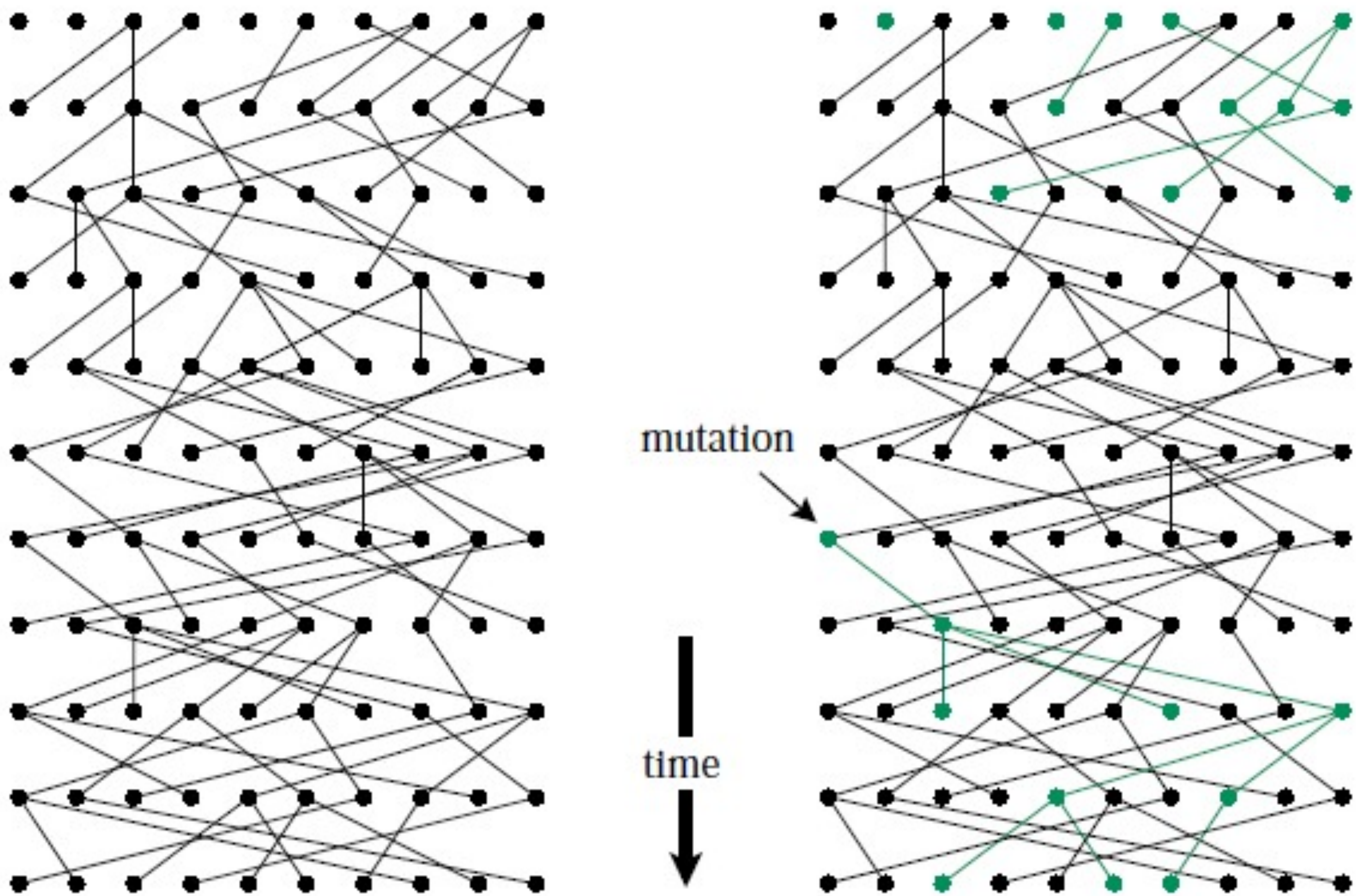
- Wright–Fisher Model
  - –Generations do not overlap
- Moran Model
  - –Assumes generations overlap

WARNING: For the purposes of this presentation, we will ignore the fact that people are diploid (have 2 copies of each chromosome).

Wright–Fisher Model



Wright-Fisher Model

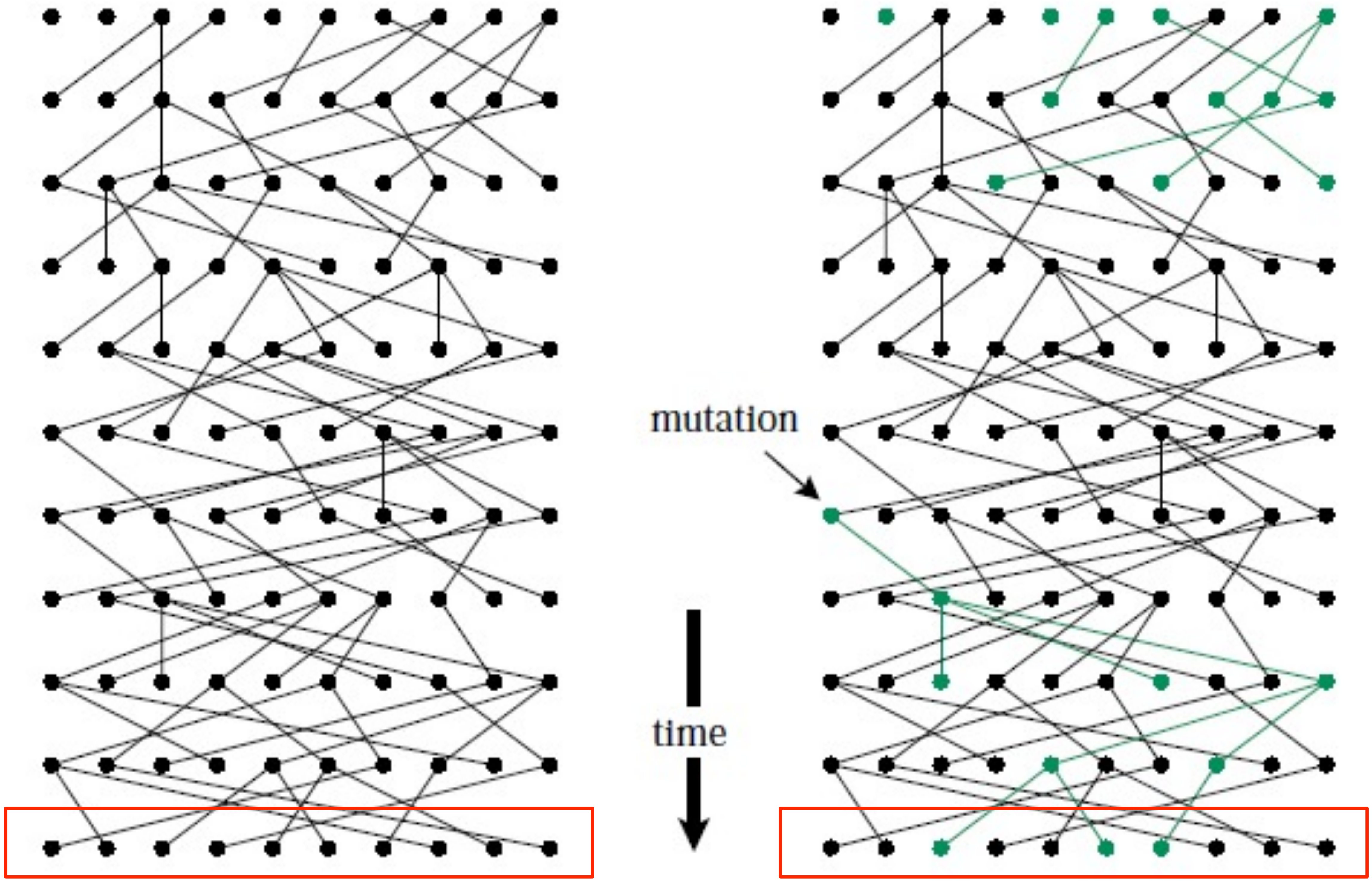


Wright–Fisher Alleles

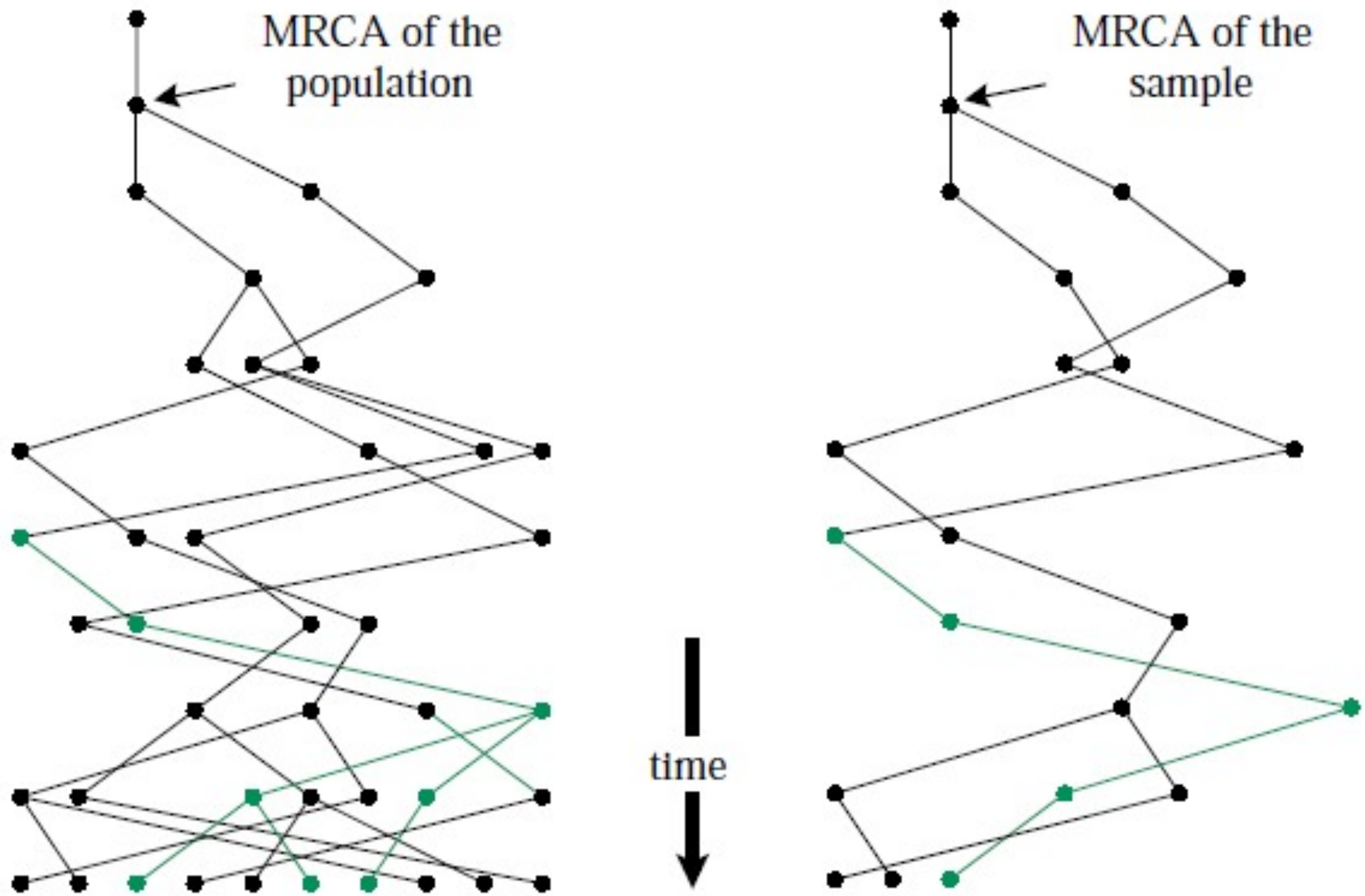
- Assume only two possible alleles (A or a) at any location in the genome.
  - –i copies of A in generation t, having frequency p = i/N
  - –N–i copies of a in generation t+1, having frequency 1–p
- Probability of j copies of A in generation t+1:

$$P_{ij} = \binom{N}{j} p^j (1-p)^{N-j} \qquad 0 \leq j \leq N,$$

Wright–Fisher Model



Coalescent Models



(N >> n)

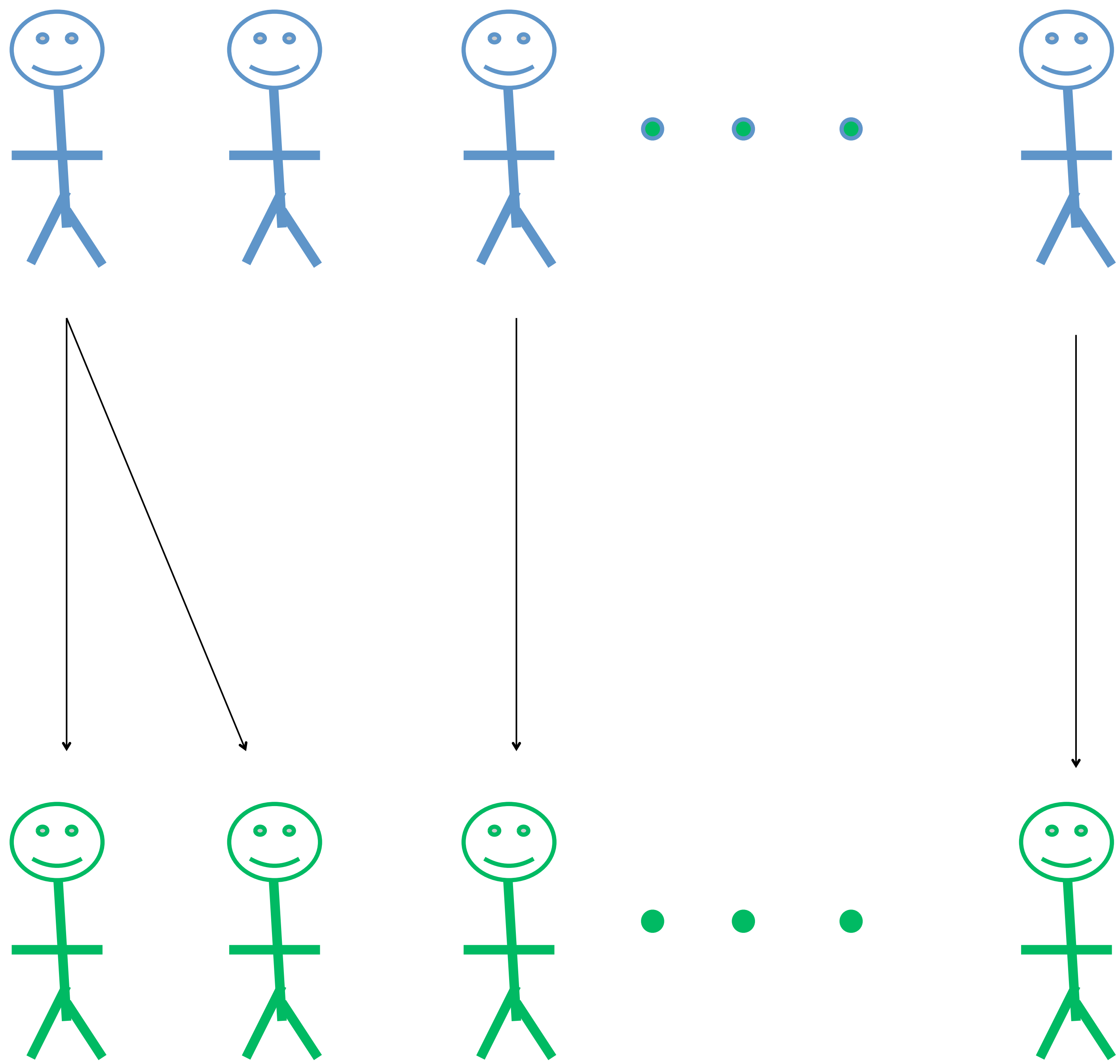
Ancestral Partitions

Let x_1, x_2,..., x_N be the current generation. The ancestral partition at generation t is just the partition where i~j if and only if x_i and x_j have a common ancestor in generation t.

Ancestral Partitions

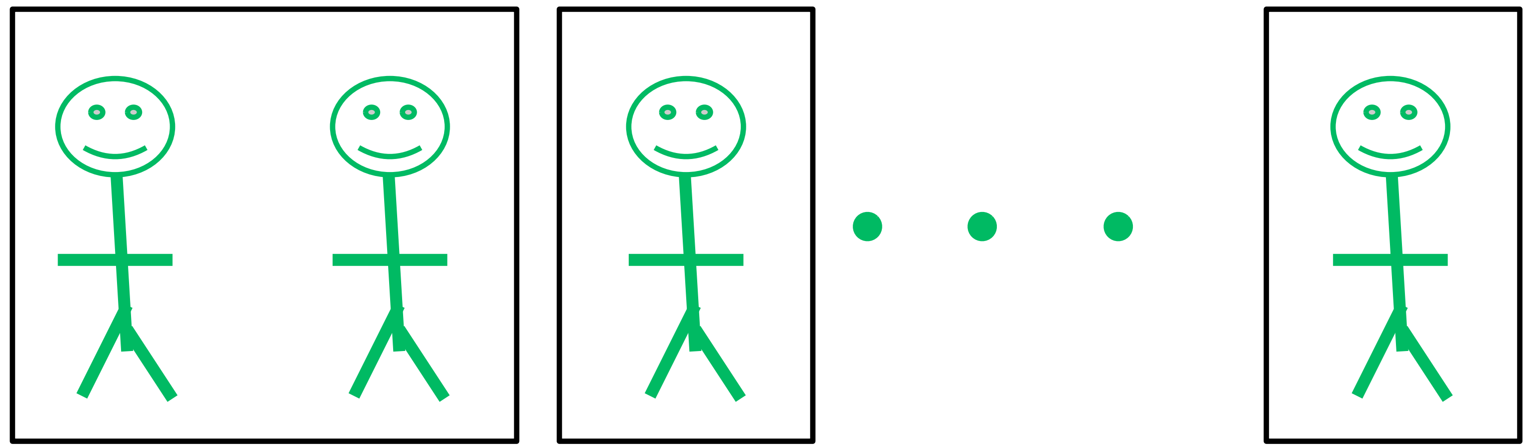
Generation: t
Pop Size: N

Generation: t + 1
Pop Size: N

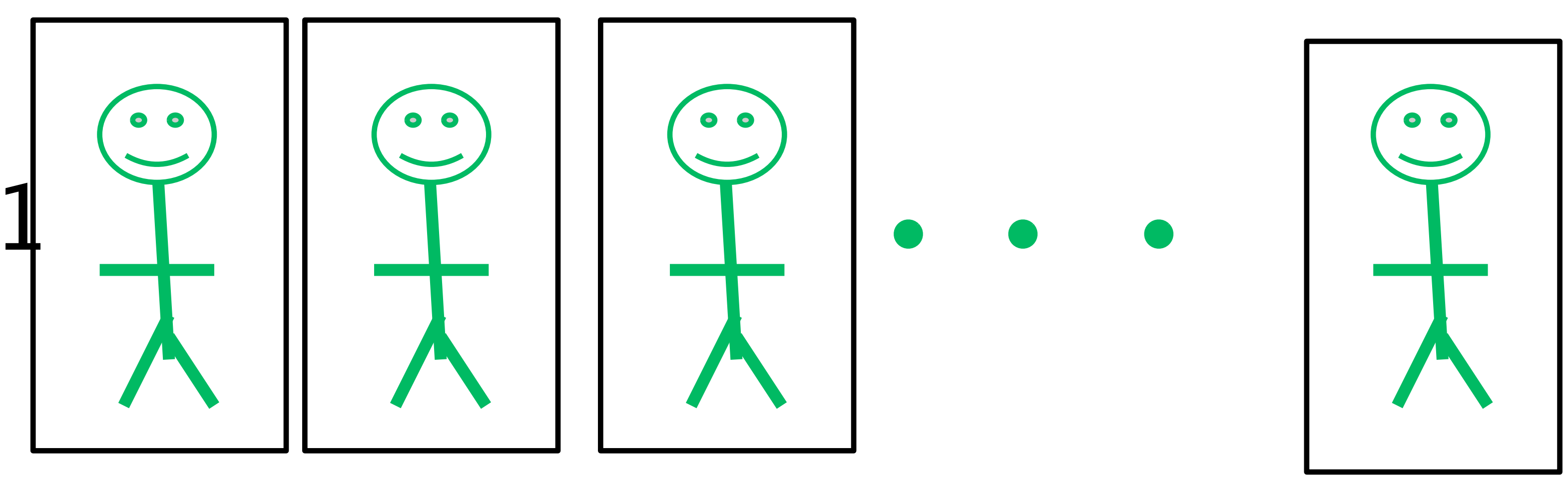


Ancestral Partitions

Generation: t
Pop Size: N



Generation: t + 1
Pop Size: N

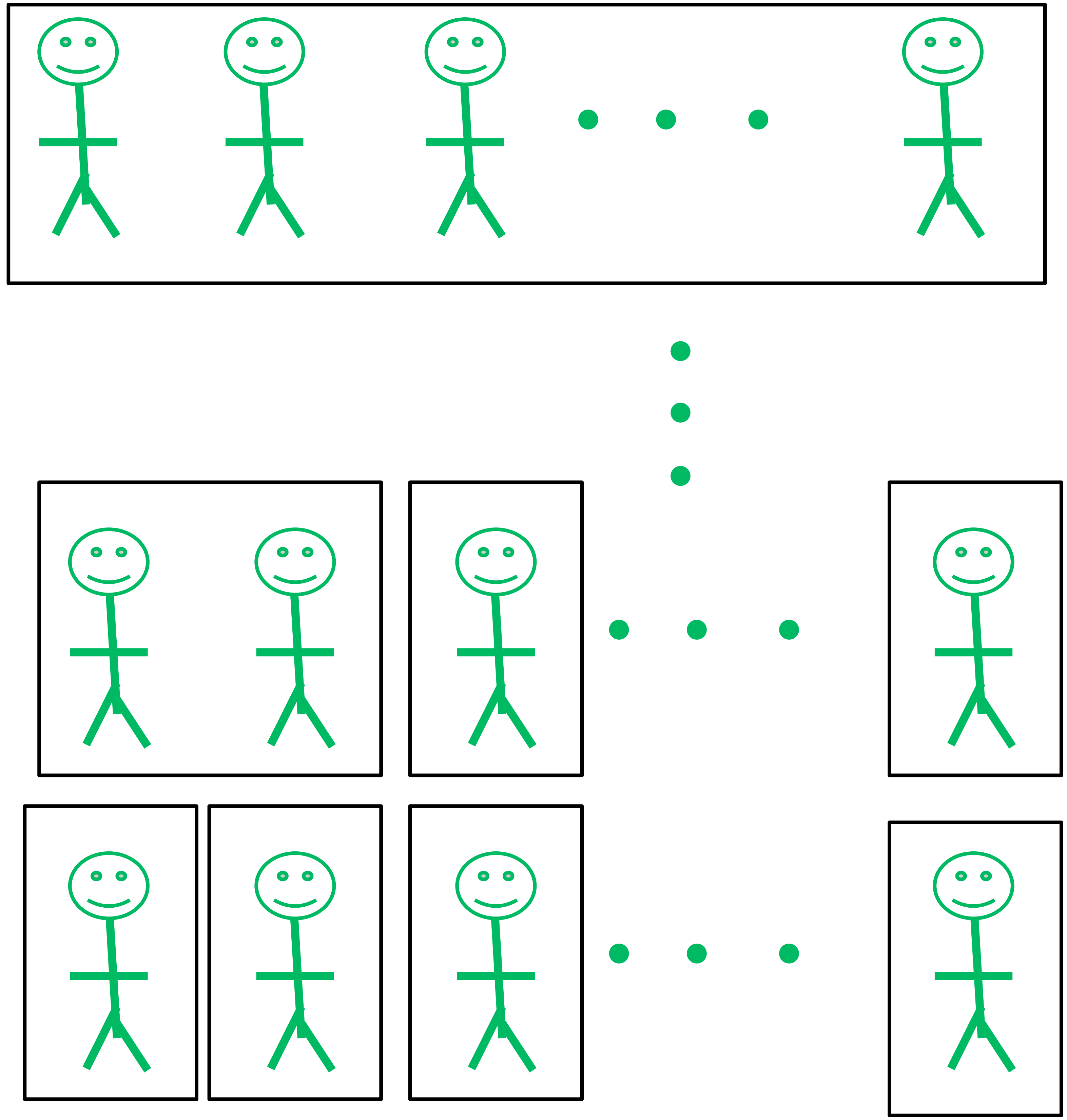


Wright-Fisher Model

Generation: t
Pop Size: N

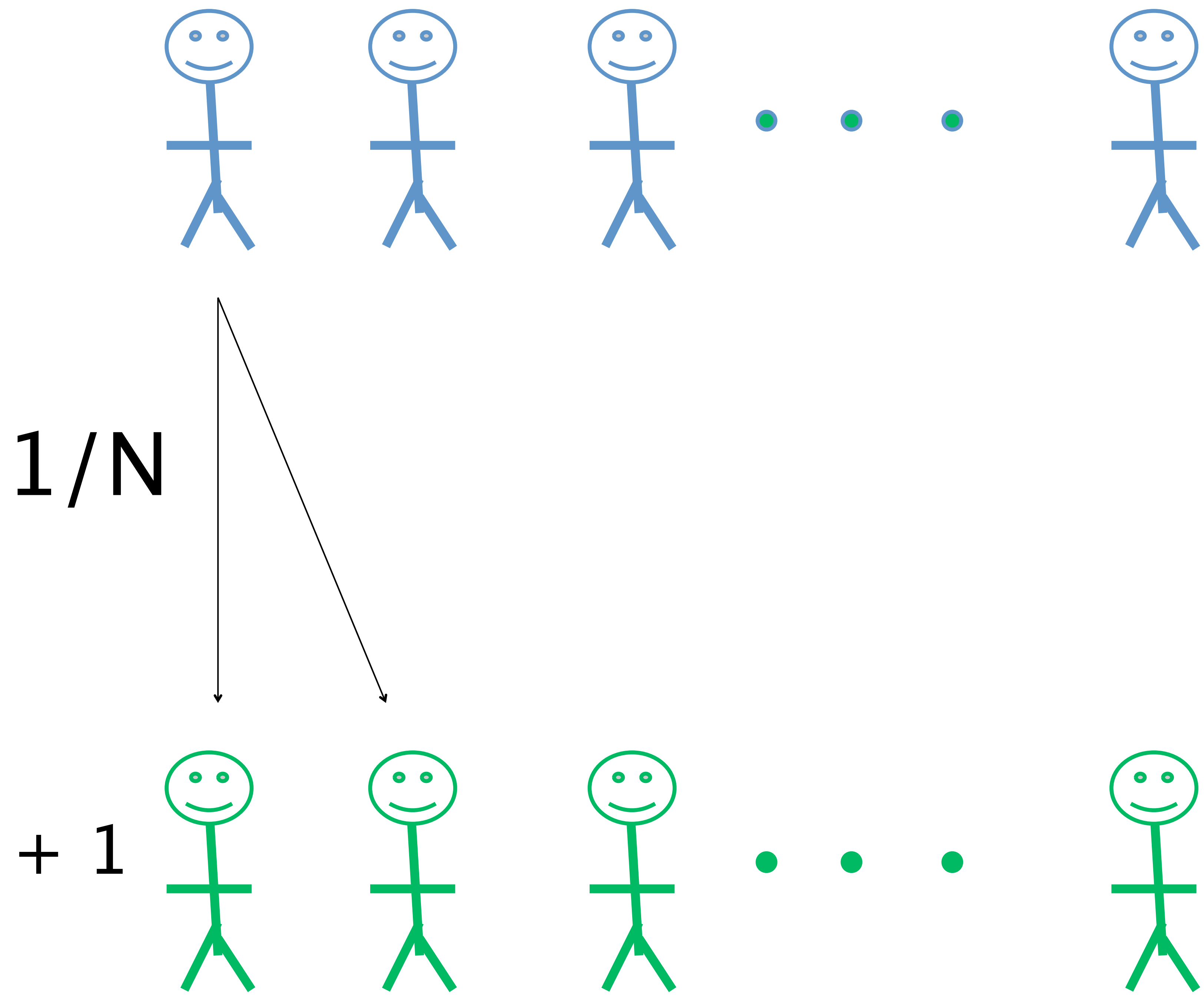
Generation: t + n – 1
Pop Size: N

Generation: t + n
Pop Size: N



Wright-Fisher Model

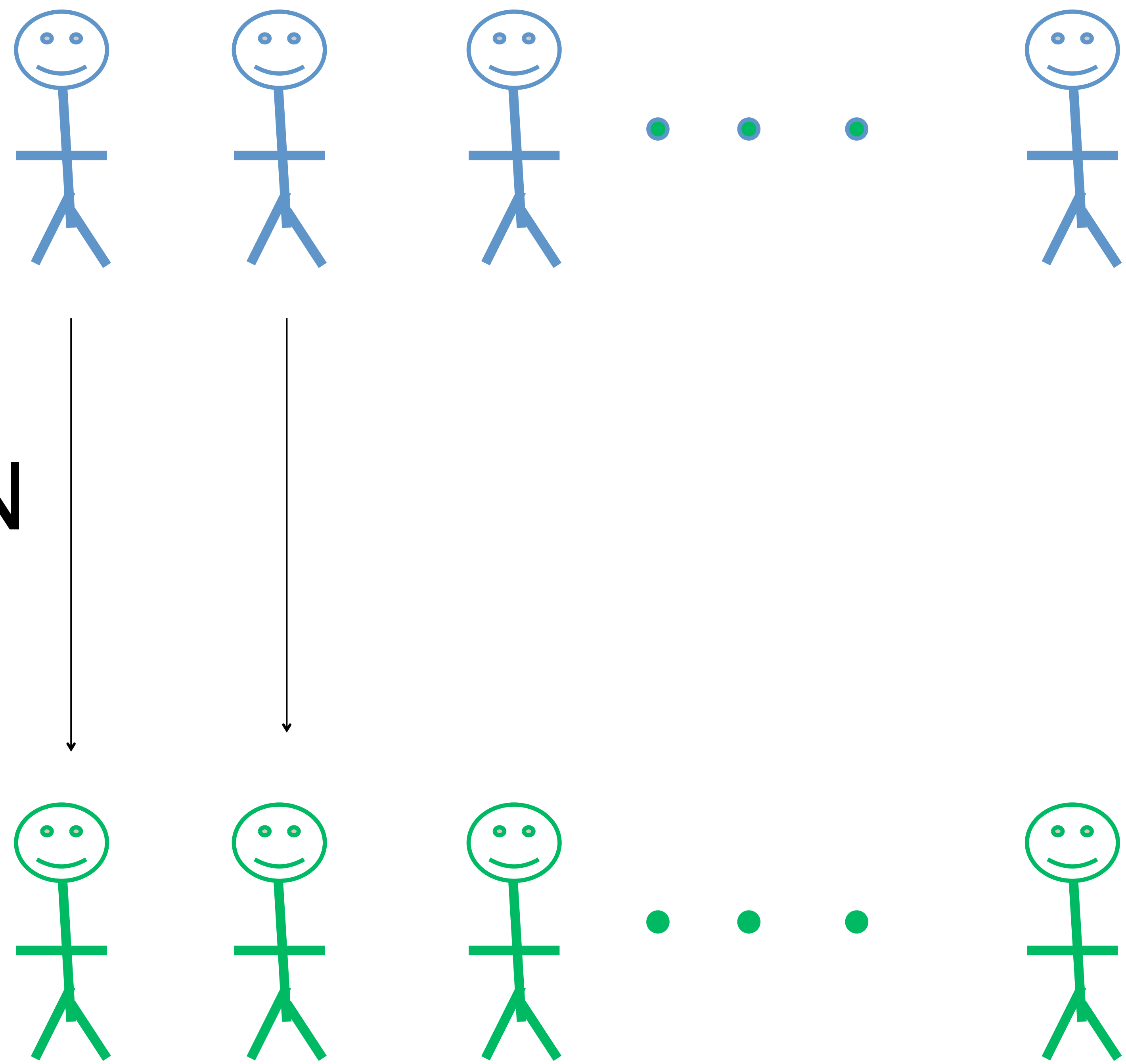
Generation: t
Pop Size: N



Generation: t + 1
Pop Size: N

Wright–Fisher Model

Generation: t
Pop Size: N

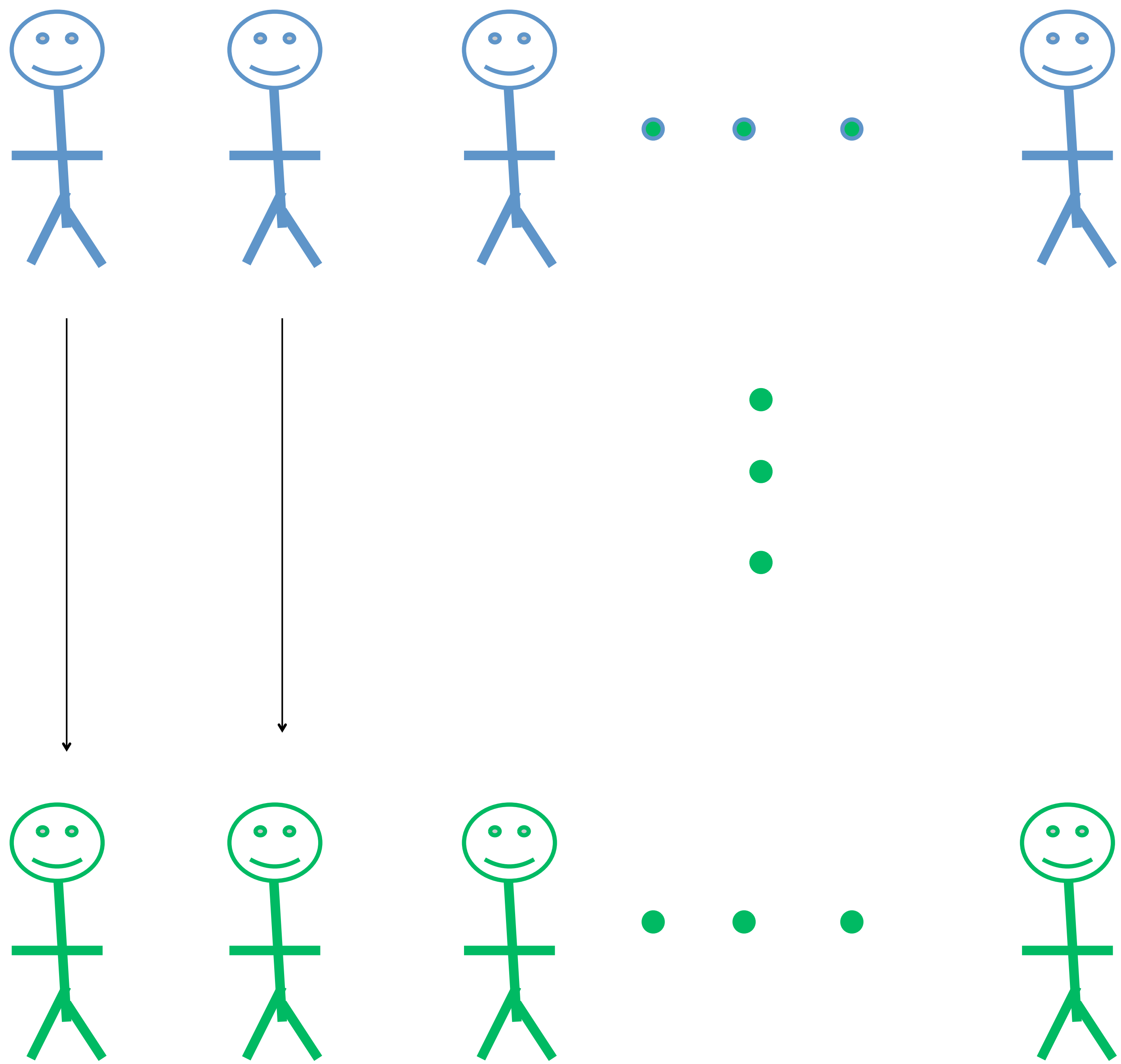


1 – 1/N

Generation: t + 1
Pop Size: N

Wright–Fisher Model

Generation: t
Pop Size: N



$(1 - 1/N)^n$

Generation: t + n
Pop Size: N

Wright-Fisher Model

- Expected amount of time for 2 lineages to join or coalesce is just N generations.

- Rescale time: 1 unit = N generations

Probability lineages stay distinct for x units of rescaled time:

$$(1 - 1/N)^{Nx} \to e^{-x}$$

(decay of heterozygosity interpretation)

Assumptions

1. Population of constant size, and individuals typically have few offspring.
2. Population is well-mixed. Everybody is liable to interact with anybody.
3. No selection acts on the population.

Assumptions

- We are assuming neutrality.
  - Different alleles do not have an affect upon survival.
  - This allows any generation to be viewed as an exchangeable partition.
- The biology is a lot more complicated than what we are presenting.
  - Recombination
  - Diploid genomes

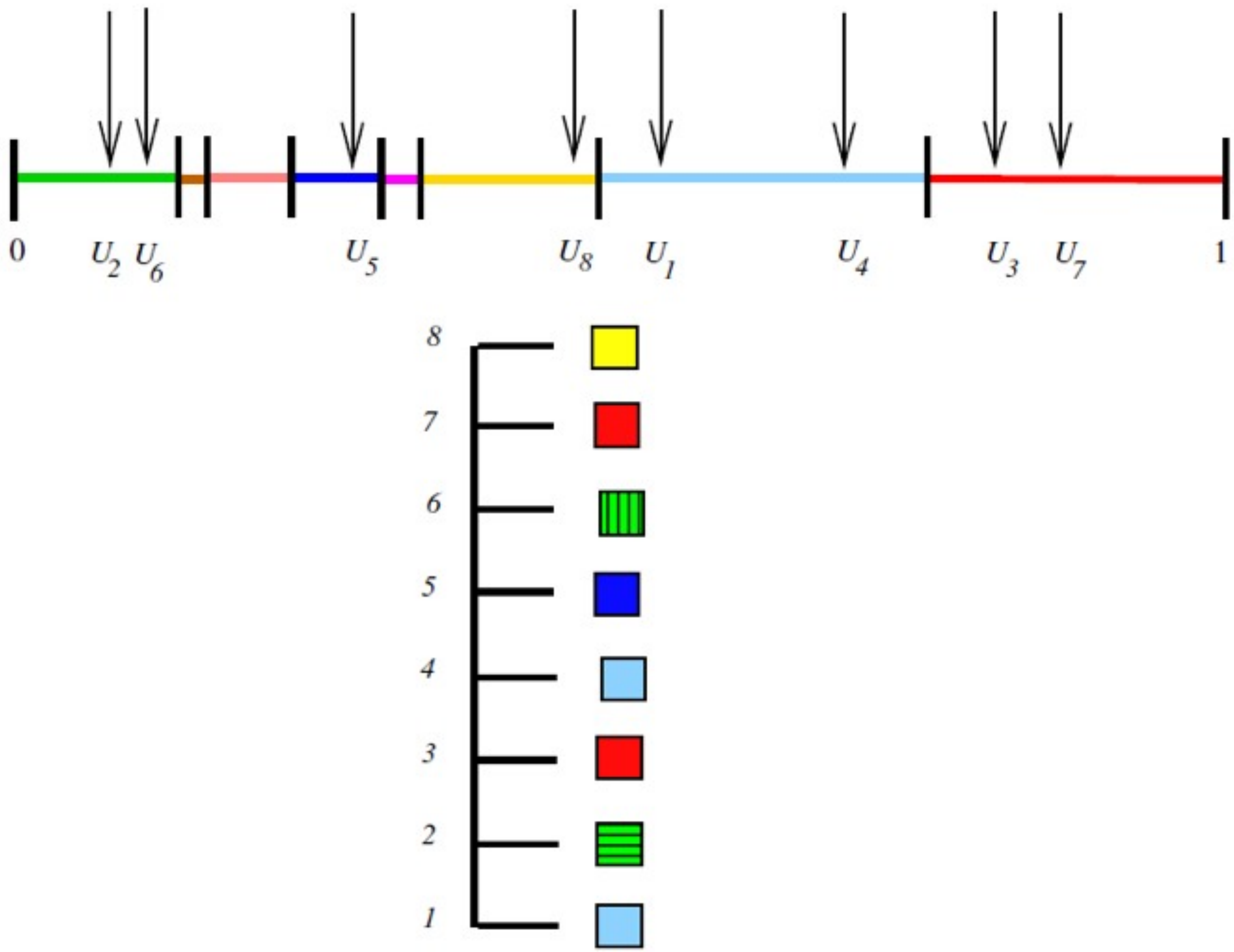
Preliminaries

- The coalescent is a stochastic process that takes values on exchangeable random partitions, so it is helpful to understand exchangeable random partitions.

Definition 1.1. *An exchangeable random partition* Π *is a random element of* $\mathcal{P}$ *whose law is invariant under the action of any permutation* σ *of* $\mathbb{N}$ *with finite support: that is,* Π *and* Π_σ *have the same distribution for all* σ.

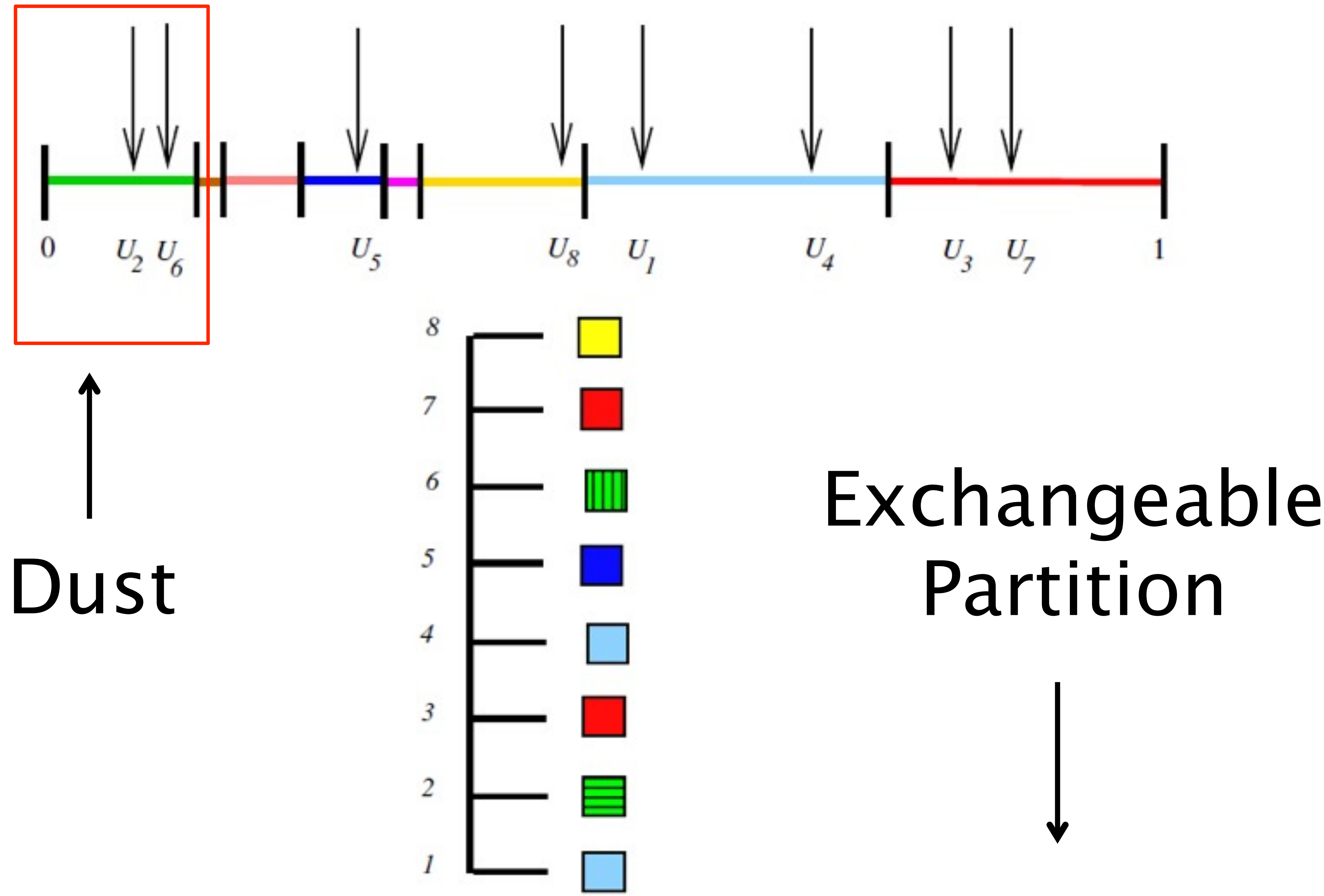
- **Observation**: Given a tiling of the unit interval, there is always a neat way to generate an exchangeable random partition associated with the tiling.
  - Stated formally as Kingman's correspondence

Tilings and Partitions



$$\Pi|_{[8]} = (\{1,4\},\{2\},\{3,7\},\{5\},\{6\},\{8\})$$

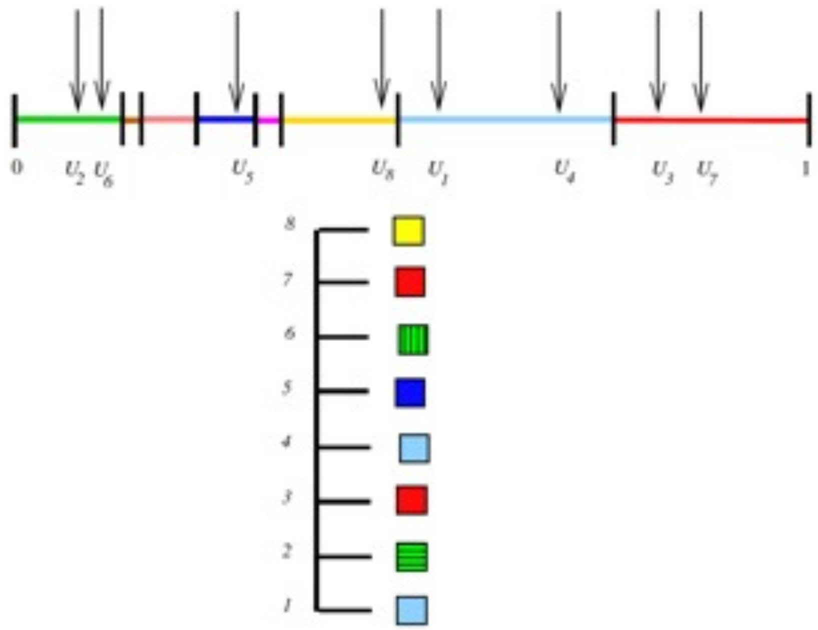
Tilings and Partitions



$$\Pi|_{[8]} = (\{1,4\},\{2\},\{3,7\},\{5\},\{6\},\{8\})$$

Formally: Paintbox process

$$\mathcal{S}_0 = \left\{ s = (s_0, s_1, \ldots) : s_1 \geq s_2 \geq \ldots, \sum_{i=0}^{\infty} s_i = 1 \right\}$$



Definition 1.2. Π *is the* paintbox *partition derived from* s.

Connection to De Finetti's Theorem:

Theorem 1.1. (Kingman [107]) *Let* Π *be any exchangeable random partition. Then there exists a probability distribution* $\mu(ds)$ *on* $\mathcal{S}_0$ *such that*

$$\mathbb{P}(\Pi \in \cdot) = \int_{s \in \mathcal{S}_0} \mu(ds) \rho_s(\cdot).$$

Kingman's correspondence

Measure–theoretic details to deal with dust: intuition is that dust cannot be characterized by a pdf, so judging convergence by pdf is not useful.

$$\Pi \in \mathcal{P} \longleftrightarrow s \in \mathcal{S}_0.$$

Corollary 1.1. *This correspondence is a 1-1 map between the law of exchangeable random partitions* Π *and distributions* μ *on* $\mathcal{S}_0$. *This map is* Kingman's correspondence.

Theorem 1.2. *Convergence in distribution of the random partitions* $(\Pi_\varepsilon)_{\varepsilon>0}$, *is equivalent to the convergence in distributions of their ranked frequencies* $(s_1^\varepsilon, s_2^\varepsilon, \ldots)_{\varepsilon>0}$.

Size–biased picking

- Mostly technical, but a few intuition–building results
  - Picking an arbitrary block is not well–defined
  - Introduce r.v. X, mass of block containing first individual
  - Exchangeability doesn't quite mean block containing first individual is typical, since larger blocks are more likely to contain any individual
- Results with the following flavor:

Theorem 1.4. *Let Π be a random exchangeable partition, and let N be the number of blocks of Π. Then we have the formula:*

$$\mathbb{E}(N) = \mathbb{E}(1/X).$$

Asymptotics

Definitions: K_n, which is the number of blocks of Π_n (the restriction of Π to $[n]$).

$K_{n,r}$, which is the number of blocks of size r, $1 \leq r \leq n$.

Theorem 1.11. *Let $0 < \alpha < 1$. There is equivalence between the following properties:*

(i) $P_j \sim Zj^{-\alpha}$ almost surely as $j \to \infty$, for some $Z > 0$.

(ii) $K_n \sim Dn^{\alpha}$ almost surely as $n \to \infty$, for some $D > 0$.

Furthermore, when this happens, Z and D are related through

$$Z = \left(\frac{D}{\Gamma(1-\alpha)}\right)^{1/\alpha},$$

and we have:
(iii) For any $r \geq 1$, $K_{n,r} \sim \frac{\alpha(1-\alpha)\ldots(r-1-\alpha)}{r!} Dn^{\alpha}$ as $n \to \infty$.

Asymptotics

The Pitman–Yor distribution verifies the assumptions of the theorem, hence:

Theorem 1.12. *Let Π be a $PD(\alpha, 0)$ random partition. Then there exists a random variable S such that*

$$\frac{K_n}{n^\alpha} \longrightarrow S$$

Power law for cluster sizes!

almost surely. Moreover S has the Mittag-Leffer distribution:

$$\mathbb{P}(S \in dx) = \frac{1}{\pi\alpha} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k!} \Gamma(\alpha k + 1) s^{k-1} \sin(\pi \alpha k).$$

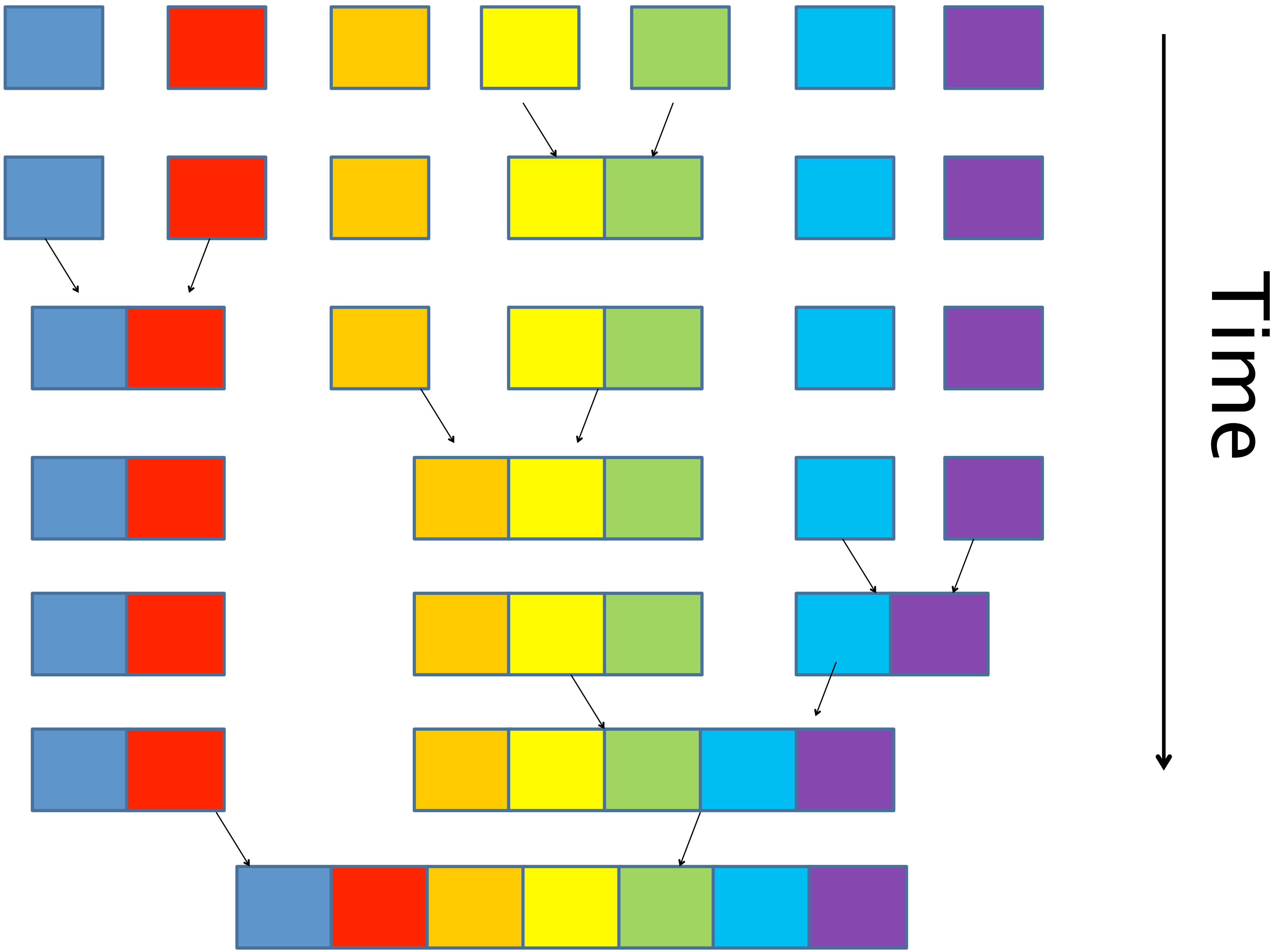
Kingman's n-Coalescent

A process $(\Pi_t^n, t \geq 0)$ with values in the space of partitions [n]={1, ..., n} defined by:

1. Initially Π_0^n is the trivial partition in singletons.

2. Π^n is a strong Markov process in continuous time, where the transition rates $q(\pi, \pi')$ are as follow: they are positive if and only if π' is obtained from merging two blocks of π, in which case $q(\pi, \pi') = 1$

Kingman's n-Coalescent



Kingman's n-Coalescent

<u>Consistency</u>: If we restrict Π^n to partitions of {1,…, m} where $m < n$, then $\Pi^{m,n}$ is an m-coalescent.

+

Kolmogorov's extension theorem

=

Proposition 2.1. *There exists a unique in law process* $(\Pi_t, t \geq 0)$ *with values in* $\mathcal{P}$, *such that the restriction of* Π *to* $\mathcal{P}_n$ *is an n-coalescent.* $(\Pi_t, t \geq 0)$ *is called Kingman's coalescent.*

Kingman's coalescent

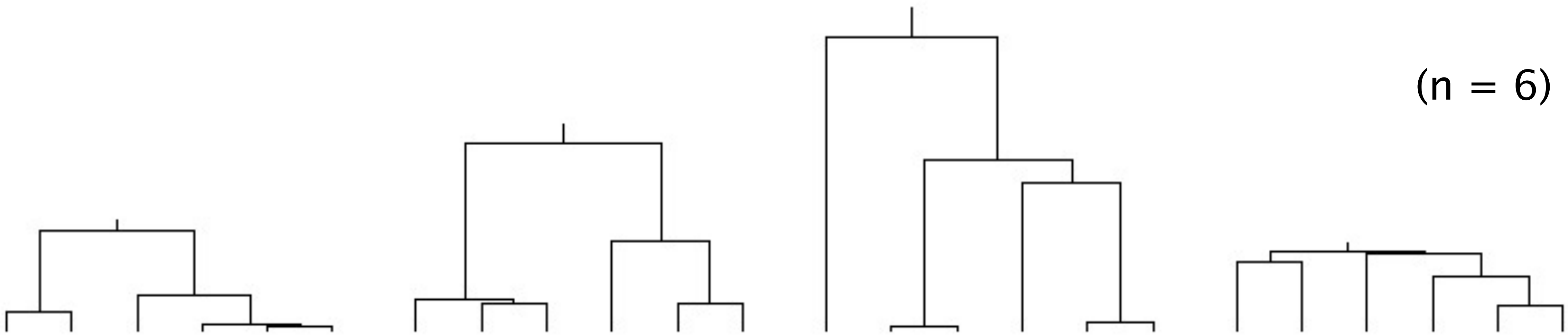
Theorem 5.1. Kingman [253] *There exists a uniquely distributed* $\mathcal{P}_{\mathbb{N}}$*-valued process* $(\Pi_\infty(t), t \geq 0)$, *called* Kingman's coalescent, *with the following properties:*

- $\Pi_\infty(0)$ *is the partition of* $\mathbb{N}$ *into singletons;*
- *for each* n *the restriction* $(\Pi_n(t), t \geq 0)$ *of* $(\Pi_\infty(t), t \geq 0)$ *to* $[n]$ *is a Markov chain with càdlàg paths with following transition rates: from state* $\Pi = \{A_1, \ldots, A_k\} \in \mathcal{P}_{[n]}$, *the only possible transitions are to one of the* $\binom{k}{2}$ *partitions* $\Pi_{i,j}$ *obtained by merging blocks* A_i *and* A_j *to form* $A_i \cup A_j$, *and leaving all other blocks unchanged, for some* $1 \leq i < j \leq k$, *with*

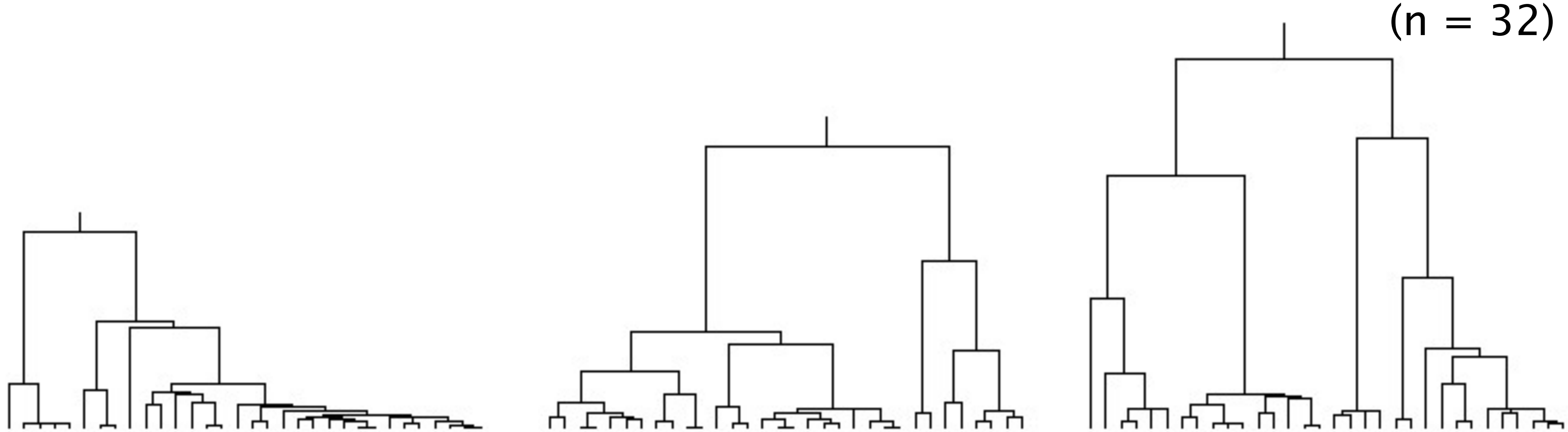
$$\Pi \to \Pi_{i,j} \text{ at rate } 1 \qquad (5.1)$$

Càdlàg (continuous from right, limits from left) paths suggests Skorokhod topology, can "wiggle space and time a bit"

What do the trees look like?



(n = 6)



(n = 32)

What do the trees look like?

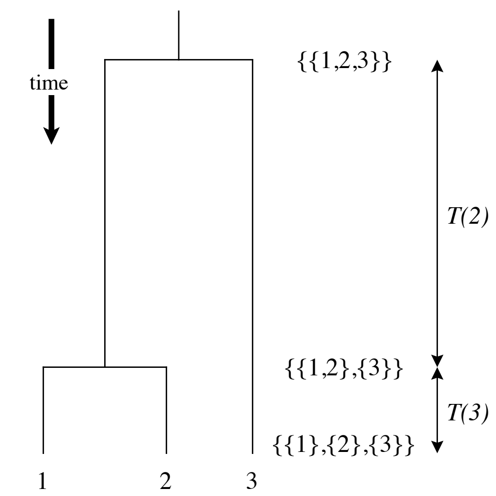
Let T(k) be scaled time until a coalescent event, when k lineages exist.

$$E\Big[\sum_{k=2}^{n} T(k)\Big] = \sum_{k=2}^{n} E[T(k)] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\Big(1 - \frac{1}{n}\Big),$$

Over half the expected time occurs for E[T(2)] = 1 (the last pair to coalesce)! The variance in total tree height is also dominated by T(2).

Properties

- Bifurcating tree
- Branch lengths are exponentially distributed



  - Exponential distribution is memoryless

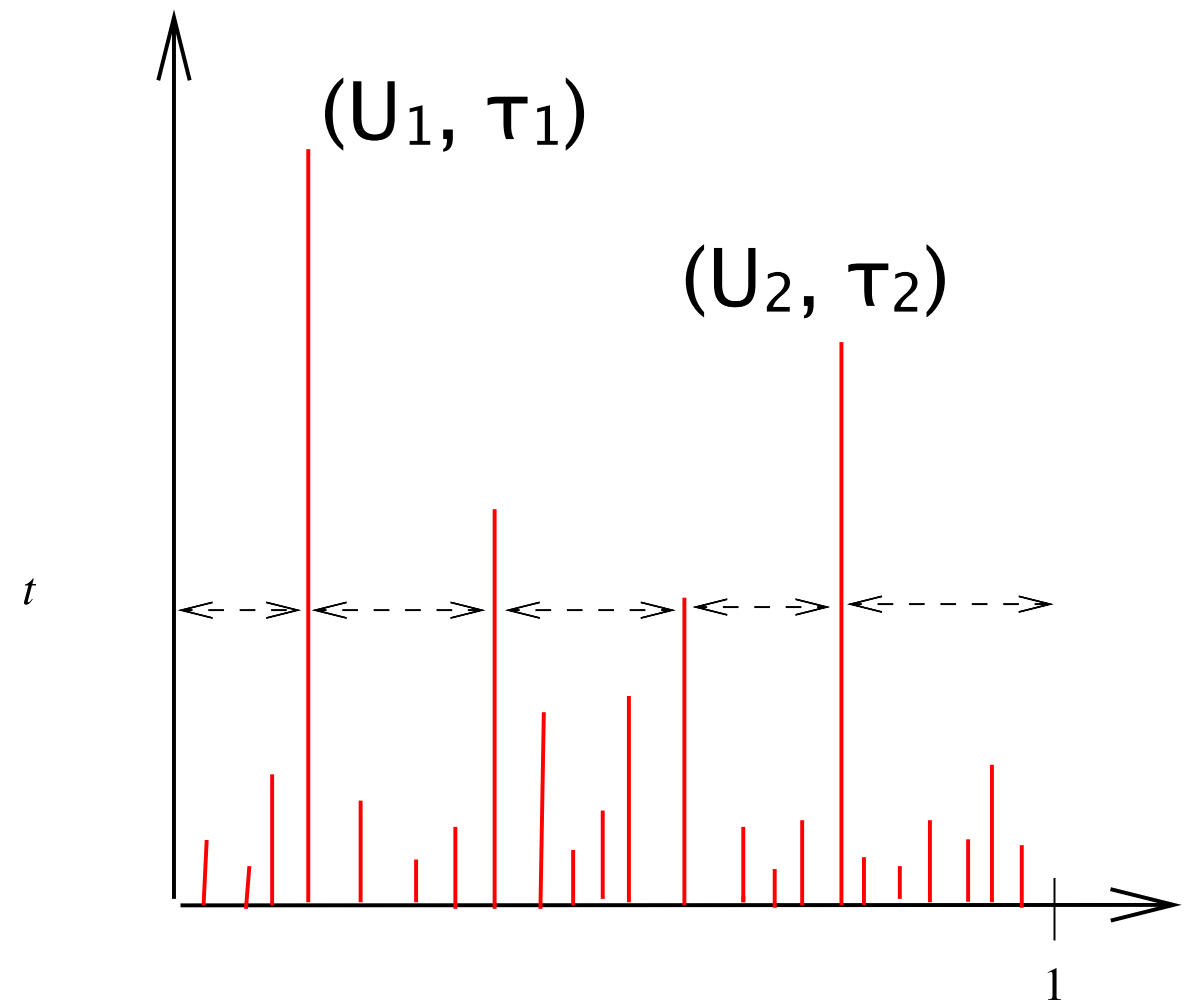
    $$\Pr(T > s + t \mid T > s) = \Pr(T > t) \text{ for all } s,t \geq 0$$

  - Allows all lineages to equal probability of coalescence at any time point.
  - Allows the partitions to remain exchangeable.

Constructions

- Kolmogorov's Extension Theorem
- Pure death process on partitions labeled by least element
- Large-population limits of biological models
  - Wright-Fisher
  - Moran
  - Cannings
- Aldous' construction
- Cutting a rooted random segment
- . . . diversity of constructions suggests universality

Aldous' construction



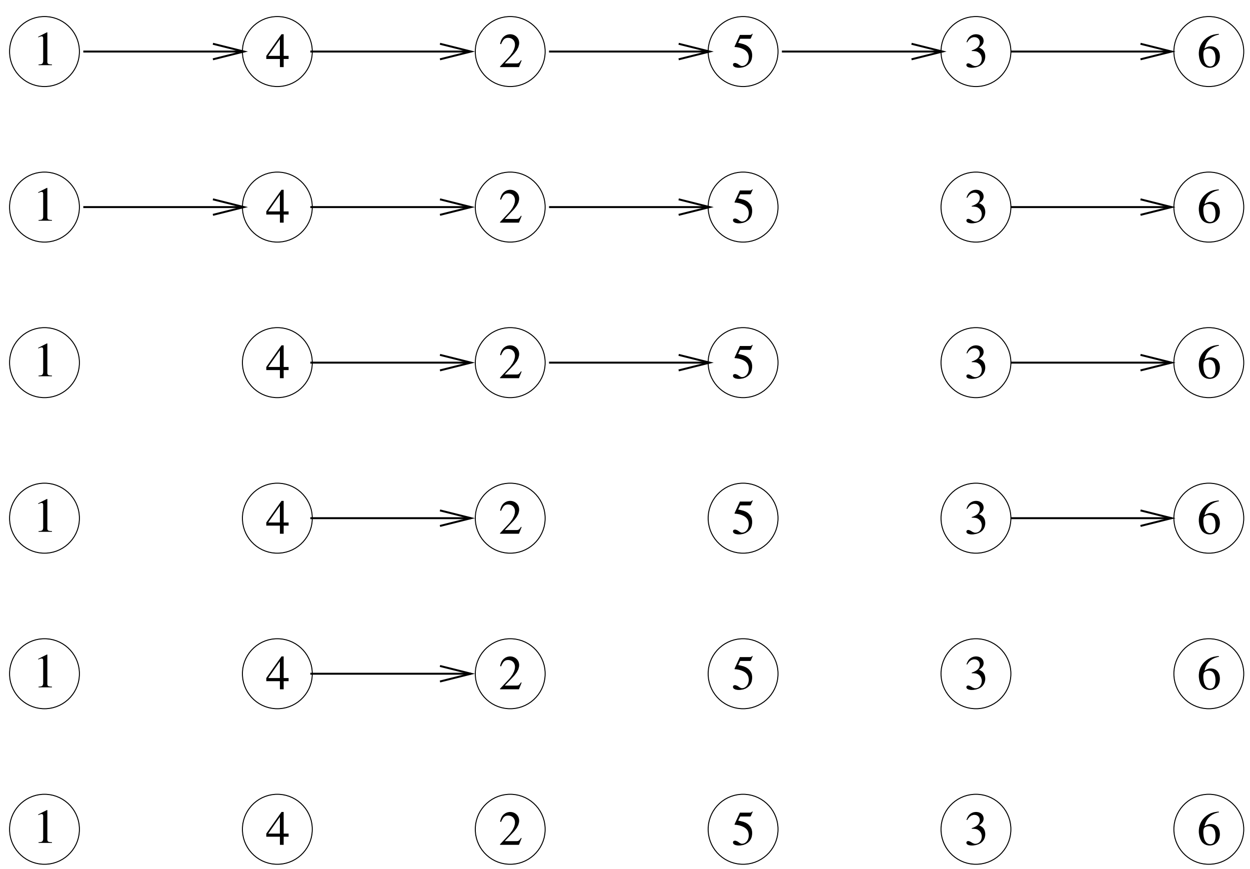
Stick locations uniform random

E_j exponential with rate $j(j-1)/2$

$$\tau_j = \sum_{k=j+1}^{\infty} E_k < \infty$$

Theorem 2.2. *$(S(t), t \geq 0)$ has the distribution of the asymptotic frequencies of Kingman's coalescent.*

Cutting a random rooted segment



$$\mathbb{P}(\Xi' = \xi') = \frac{n-k+1}{k(k-1)|\mathcal{R}_{n,k}|}$$

Lemma 2.1. *The random partition associated with a uniform element of $\mathcal{R}_{n,k}$ has the same distribution as Π_k^n, where $(\Pi_k^n)_{n\geq k\geq 1}$ is the set of successive states visited by Kingman's n-coalescent.*

Cutting a random rooted segment

Special payoff is a conditional version of
Ewens' Sampling Formula:

Corollary 2.3. *Let $1 \leq k \leq n$. Then for any partition of $[n]$ with exactly k blocks, say $\pi = (B_1, B_2, \ldots, B_k)$, we have:*

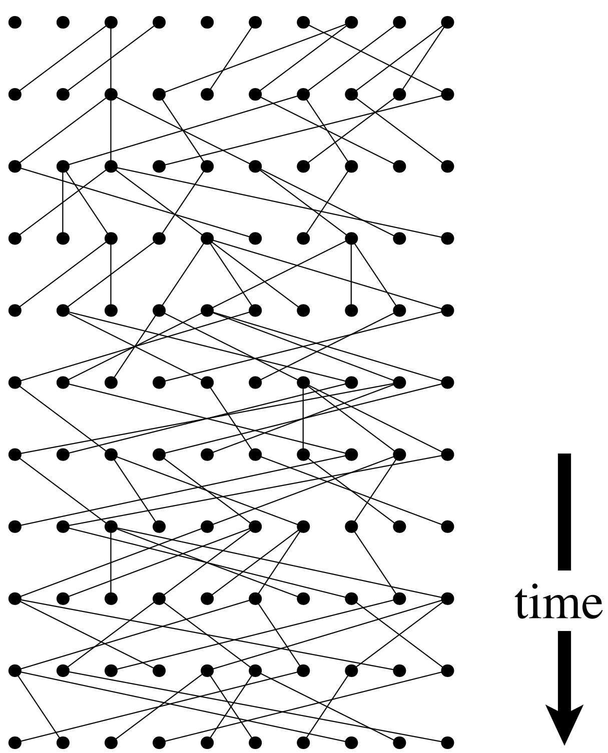
$$\mathbb{P}(\Pi_k^n = \pi) = \frac{(n-k)!k!(k-1)!}{n!(n-1)!} \prod_{i=1}^{k} |B_i|! \tag{2.6}$$

Wright–Fisher limit theorem

Theorem 2.4. *Fix* $n \geq 1$*, and let* $\Pi_t^{N,n}$ *denote the ancestral partition at time* t *of* n *randomly chosen individuals from the population at time* $t = 0$*. That is,* $i \sim j$ *if and only if* x_i *and* x_j *share the same ancestor at time* $-t$*. Then as* $N \to \infty$*, and keeping* n *fixed, speeding up time by a factor* N*:*

$$(\Pi_{Nt}^{N,n}, t \geq 0) \longrightarrow_d (\Pi_t^n, t \geq 0)$$

where $\longrightarrow_d$ *indicates convergence in distribution under the Skorokhod topology of* $\mathbb{D}([0,\infty), \mathcal{P}_n)$*, and* $(\Pi_t^n, t \geq 0)$ *is Kingman's* n*-coalescent.*



N → ∞

Moran limit theorem

Theorem 2.3. *Let $n \geq 1$ be fixed, and let $x_1, \ldots, x_n$ be n individuals sampled without replacement from the population at time $t = 0$. For every $N \geq n$, let $\Pi_t^{N,n}$ be the ancestral partition obtained by declaring $i \sim j$ if and only if x_i and x_j have a common ancestor at time $-t$. Then, speeding up time by $(N-1)/2$, we find:*

$$(\Pi^{N,n}_{(N-1)t/2}, t \geq 0) \text{ is an } n\text{-coalescent.}$$

Down from Infinity

Let N_t be the number of blocks of $\Pi(t)$.

Theorem 2.1. *Let E be the event that for all $t > 0$, $N_t < \infty$. Then* $\mathbb{P}(E) = 1$.

That is – all the “dust” has coagulated!

Proof relies on showing for all ε > 0, there exists M>0 such that:

$$\limsup_{n\to\infty} \mathbb{P}(N_t^n > M) \leq \varepsilon.$$

$$\begin{aligned}\mathbb{P}(N_t^n > M) &= \mathbb{P}\left(\sum_{k=M}^{n} E_k > t\right)\\ &\leq \frac{1}{t}\mathbb{E}\left(\sum_{k=M}^{n} E_k\right)\\ &\leq \frac{1}{t}\sum_{k=M}^{\infty}\frac{2}{k(k-1)}.\end{aligned}$$

Down from infinity

Intuitively, pass to continuum and model with differential equation:

$$\begin{cases} u'(t) & = -\dfrac{u(t)^2}{2} \\ u(0) & = +\infty. \end{cases}$$

Solving:

$$N_t \sim \frac{2}{t}, \quad t \to 0$$

How does it all fit together?

Wright–Fisher meets Kingsman

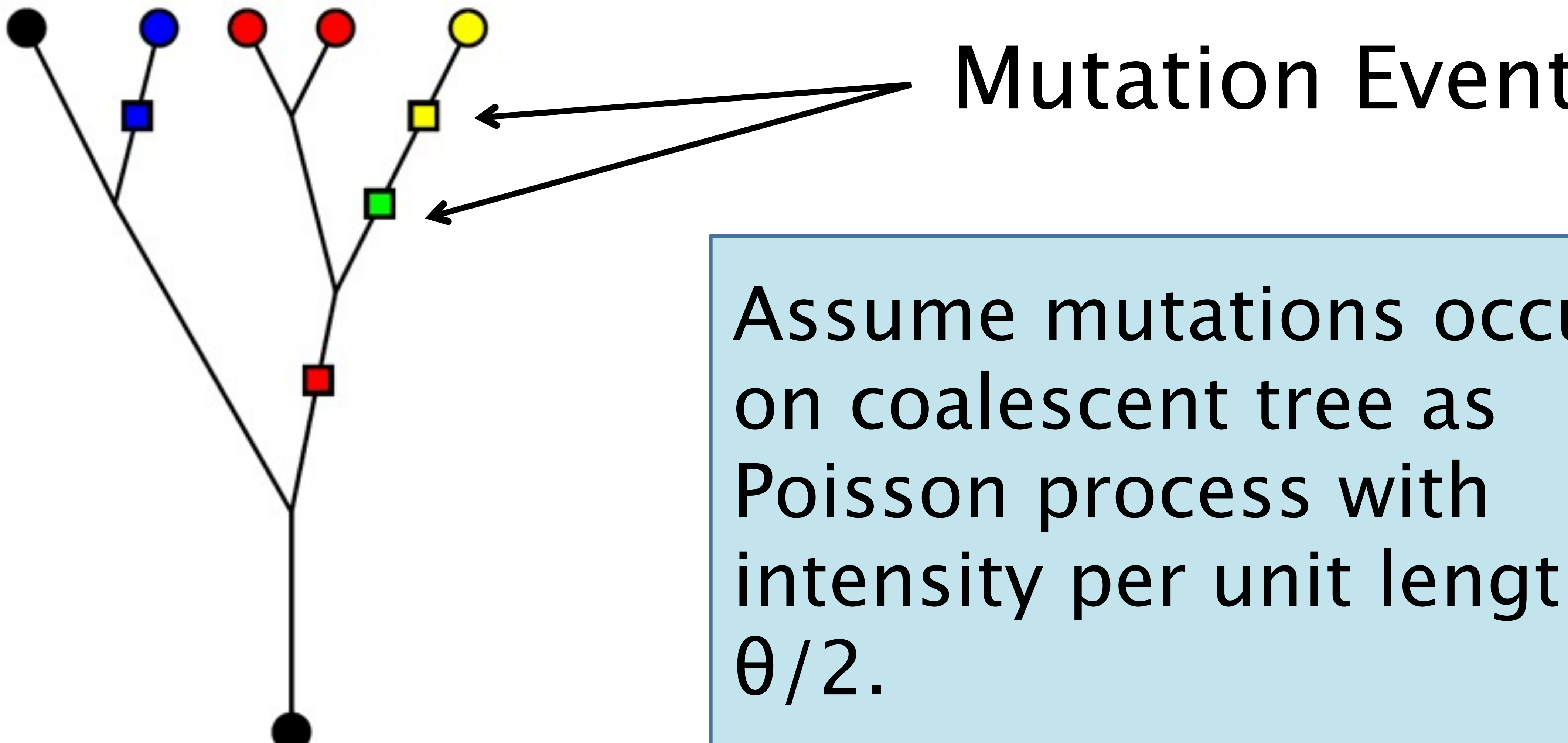
Theorem 2.7. *Let $\mathbb{E}^{\rightarrow}$ and $\mathbb{E}^{\leftarrow}$ denote respectively the laws of a Wright-Fisher diffusion and of Kingman's coalescent. Then, for all $0 < p < 1$, and for all $n \geq 1$, we have:*

$$\mathbb{E}_p^{\rightarrow}((X_t)^n) = \mathbb{E}_n^{\leftarrow}\left(p^{|\Pi_t|}\right) \tag{2.13}$$

where $|\Pi_t|$ denotes the number of blocks of the random partition Π_t.

Back to alleles

We want to analyze the allelic partition of Kingman's coalescent.



Assume mutations occur on coalescent tree as Poisson process with intensity per unit length $\theta/2$.

Ewens Sampling (original)

Theorem 1.6. *Let π be any given partition of $[n]$, whose block size are $n_1, \ldots, n_k$.*

$$\mathbb{P}(\Pi_n = \pi) = \frac{\theta^k}{(\theta)\ldots(\theta+n-1)} \prod_{i=1}^{k} (n_i - 1)!$$

Ewen Sampling (allelic)

Theorem 2.9. *Let Π be the allelic partition obtained from Kingman's coalescent and the infinite alleles model with mutation rate $\theta/2$. Then Π has the law of a Poisson-Dirichlet random partition with parameter θ. In particular, the probability that $A_1 = a_1,\ A_2 = a_2, \ldots, A_n = a_n$, is given by:*

$$p(a_1, \ldots, a_n) = \frac{n!}{\theta(\theta+1)\ldots(\theta+n-1)} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!}. \tag{2.19}$$

Inference?

Theorem 2.10. *let* Π *be a* $PD(\theta)$ *random partition, and let* Π_n *be its restriction to* $[n]$*, with* K_n *blocks. Then*

$$\frac{K_n}{\log n} \longrightarrow \theta, \quad a.s. \tag{2.23}$$

as $n \to \infty$*. Moreover,*

$$\frac{K_n - \theta \log n}{\sqrt{\theta \log n}} \longrightarrow_d \mathcal{N}(0,1). \tag{2.24}$$

Coalescent and coagulation?

Recall fragmentation/coagulation from Wood et al. (2009). Pitman (2006) claims coalescents are governed by coagulation operators, but the details are murky . . .

Theorem 5.7. [357, Theorem 6] *A coalescent process* Π_∞^π *starting at* π *with* $|\pi| = n$ *for some* $1 \leq n \leq \infty$ *is a* Λ*-coalescent if and only if* Π_n *defined by* (5.6) *is distributed as the restriction to* $[n]$ *of a* Λ*-coalescent. The semigroup of the* Λ*-coalescent on* $\mathcal{P}_\mathbb{N}$ *is thus given by*

$$\mathbb{P}^{\Lambda,\pi}(\Pi_\infty(t) \in \cdot) = p_t^\Lambda\text{-COAG}(\pi, \cdot) \tag{5.7}$$

where $p_t^\Lambda(\,\cdot\,) := \mathbb{P}^{\Lambda,1^\infty}(\Pi_\infty(t) \in \cdot)$ *is the distribution of an exchangeable random partition of* $\mathbb{N}$ *with the EPPF* $p_t^\Lambda(n_1, \ldots, n_k)$ *which is uniquely determined by Kolmogorov equations for the finite state chains* Π_n *for* $n = 2, 3, \ldots$.

Questions?